



UNIVERSITAT DE
BARCELONA

Mapping protein aggregation by deep mutational scanning

Mireia Seuma Areñas

ADVERTIMENT. La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX (www.tdx.cat) i a través del Dipòsit Digital de la UB (diposit.ub.edu) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX ni al Dipòsit Digital de la UB. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX o al Dipòsit Digital de la UB (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

ADVERTENCIA. La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR (www.tdx.cat) y a través del Repositorio Digital de la UB (diposit.ub.edu) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR o al Repositorio Digital de la UB. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR o al Repositorio Digital de la UB (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

WARNING. On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX (www.tdx.cat) service and by the UB Digital Repository (diposit.ub.edu) has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized nor its spreading and availability from a site foreign to the TDX service or to the UB Digital Repository. Introducing its content in a window or frame foreign to the TDX service or to the UB Digital Repository is not authorized (framing). Those rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA I CIÈNCIES DE L'ALIMENTACIÓ

MAPPING PROTEIN AGGREGATION BY DEEP MUTATIONAL SCANNING

Mireia Seuma Areñas
2022/2023

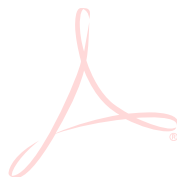
UNIVERSITAT DE BARCELONA
FACULTAT DE FARMÀCIA I CIÈNCIES DE L'ALIMENTACIÓ
DOCTORAT EN BIOTECNOLOGIA

MAPPING PROTEIN AGGREGATION BY DEEP MUTATIONAL SCANNING

Memòria presentada per **Mireia Seuma Areñas**
per optar al títol de doctor per la universitat de Barcelona

Doctoranda: Mireia Seuma Areñas

Directora: Dra. Benedetta Bolognesi



Tutora: Dra. Laura Baldomà Llavines

2022/2023

A la tieta Imma,

Acknowledgments

My biggest thank you goes to Benni for engaging me in this incredible adventure. It's been a bunch of fun working in the lab, from the good old times driving across Barcelona with the car full of yeast and boxes, to the million dumplings we can't stop eating while I share my dramas, or the Sunday nights going to the lab and: oh! you are here too, what a life! Thank you for the tons of great science you taught me, the opportunities and guidance during all these years, and for constantly reminding me how scientists are first of all humans, with a million feelings, experiences and ambitions - fears included!

I am also grateful to Ben, for being such a clever mind, always willing to share ideas. Thank you for giving crucial feedback and pushing my work forward, but also for being supportive especially when I needed it the most.

Thanks to Marta! Everyone needs someone like you to share the PhD journey with. Your enthusiasm for the things you do is just contagious and I am grateful for having learnt so much from you. Thank you for all the music, the long chats and wine late at night and for understanding all my ups and downs. I will really miss you and your hugs. Thanks to Trini, for the million stories I will never forget (who would!?) but especially for your energy and for always being ready to help us ¡Qué bonita la ciencia eh! Thanks to Mariano, for your kindness and for being hungry all the time and making terrible excuses to have an early lunch. Thank you to everyone in the Lehner lab at the CRG, the new but also the old ones, with whom I share my first PhD memories. A special thank you to Andre, for patiently answering my very naive questions about coding and for the cool projects we share. Thank you to people in Salvatella's lab at the IRB for great discussions, especially Xavier and Carla for the support and all the hard work with the impossible peptides! I also want to thank everyone who has been part of our lab during these years and the people I met around IBEC, the PCB and the PRBB – you all made my days better.

Thanks to all my friends, the ones in my hometown, the ones from university, the ones in Barcelona and abroad. I can't wait to see what life brings to us, but surely there will be many wild trips, hikes, dinners, drinks and laughs. I am deeply happy to have you all in my life.

Gràcies mama, papa, nanu, per ser el meu exemple de treball, esforç i il·lusió, i perquè sense entendre massa el món del doctorat m'heu acompanyat incondicionalment. Us estic infinitament agraïda! Gràcies també a tota la meva família i en especial al meu cosí Pol per regalar-me aquesta il·lustració tan bonica.

To all of you, and to all I am not naming but meant to, I couldn't be more grateful, thank you so much for the journey !

Abstract

Proteins can adopt multiple conformations and material states, from soluble states to self-assemblies, such as liquid condensates or amyloids. Protein aggregation has been associated with many human diseases but how mutations impact protein conformation and cell toxicity is still not fully understood, partially due to the low-throughput connotation of both *in vitro* and *in vivo* mutational approaches to date.

To address this shortcoming, I developed a Deep Mutational Scanning (DMS) method to report on the aggregation of thousands of sequences in parallel. I applied this systematic approach to the study of the amyloid beta (A β) peptide, as a model of classical amyloids. Self-assembly of A β into amyloid fibrils is a hallmark of Alzheimer's disease (AD) and dominant mutations in A β also cause rare familial AD (fAD). By quantifying the effect of >16,000 A β variants, we generated the first comprehensive atlas of how mutations alter the nucleation of amyloids by any protein *in vivo*. The atlas also represents the first comparison of the effects of substitutions, insertions, truncations and deletions in a human disease gene. Variants that increase nucleation from all mutation types are highly enriched in the polar N-terminus of A β , revealing a modular organization of mutational effects along the sequence. Strikingly, the *in vivo* nucleation scores, unlike computational predictors and previous measurements, accurately discriminate all fAD mutations, suggesting that accelerated nucleation is the fundamental molecular mechanism by which mutations cause fAD. Moreover, the atlas prioritizes many variants beyond substitutions that accelerate aggregation and are likely to be pathogenic, providing a resource for future clinical interpretation.

In parallel, I have also pioneered the use of DMS to report on cellular toxicity induced by >50,000 mutations in a disordered protein domain, namely the prion-like domain of TDP-43, a protein associated with amyotrophic lateral sclerosis. While identifying increased hydrophobicity as the one feature able to reduce toxicity in yeast cells, this study also revealed that this putatively disordered domain actually adopts secondary structure inside the cell.

Overall, my thesis provides a global picture of how sequence changes modulate protein self-assembly or toxicity. More generally, it illustrates the power of DMS in illuminating sequence-to-activity relationships suggesting that this approach should be employed to systematically target other self-assembling protein sequences.

Resum

Les proteïnes poden adoptar diferents conformacions estructurals i materials, des d'estats solubles a auto-assemblatges, com condensats líquids o amiloides. L'agregació de proteïnes s'ha associat a diverses malalties humanes, però es desconeix l'impacte de les mutacions en la conformació estructural i la toxicitat a la cèl·lula, degut a que els mètodes de mutagènesi actuals són a petita escala, tant *in vitro* com *in vivo*.

Aquí, he desenvolupat un mètode d'escaneig profund de mutacions (conegut com DMS) per mapar l'agregació de milers de seqüències en paral·lel. He utilitzat aquest mètode sistemàtic en el pèptid amiloide beta (A β), com a model clàssic d'amiloide. L'auto-assemblatge d'A β en fibres amiloides és característic de la malaltia d'Alzheimer (AD), i mutacions en A β amb herència dominant també causen formes minoritàries d'AD familiar (fAD). Quantificant >16,000 variants d'A β he generat el primer atlas exhaustiu de com les mutacions alteren la nucleació d'amiloides en qualsevol proteïna *in vivo*. L'atles també representa la primera comparació dels efectes de substitucions, insercions, truncaments i delecions en qualsevol gen humà associat a malaltia. Les variants de qualsevol tipus de mutació que acceleren la nucleació es troben majoritàriament a l'N-terminal d'A β , mostrant una organització modular dels efectes de les mutacions al llarg de la seqüència. Sorprenentment i, a diferència de predictors computacionals i altres estudis realitzats prèviament, la quantificació de nucleació *in vivo* en aquest estudi discrimina acuradament totes les mutacions associades a fAD, suggerint que l'increment de la nucleació d'amiloide és el mecanisme molecular fonamental pel qual les mutacions en A β causen fAD. Més enllà de les substitucions, l'atles prioritza moltes variants que incrementen l'agregació i que són candidates a ser patogèniques, proporcionant un recurs per a la futura interpretació clínica de l'impacte de les mutacions.

En paral·lel, també he utilitzat un estudi DMS pioner per reportar en l'efecte tòxic de >50,000 mutacions en el domini desordenat de TDP-43, una proteïna associada a l'esclerosi lateral amiotròfica. A més d'identificar que un increment en la hidrofobicitat redueix la toxicitat en cèl·lules de llevat, l'estudi també indica que aquest domini desordenat adopta una estructura secundària dins la cèl·lula.

En resum, aquesta tesi proporciona una imatge global de com els canvis en la seqüència modulen l'auto-assemblatge o la toxicitat en les proteïnes. De manera més general, també il·lustra com la tècnica DMS pot il·luminar la relació seqüència-activitat en una proteïna, suggerint que aquest mètode hauria de ser utilitzat sistemàticament en altres seqüències proteïques amb capacitat d'auto-assemblatge.

Abbreviations

A β	amyloid β
AD	Alzheimer's disease
AFM	atomic force microscopy
α -syn	α -synuclein
ALS	amyotrophic lateral sclerosis
apo A-I	apolipoprotein A-I
APP	amyloid precursor protein
APR	aggregation-prone regions
ARTAG	aging-related tau astrogliopathy
CBP	CREB-binding protein
CNT	classical nucleation theory
cryoEM	cryo-electron microscopy
CSF	cerebrospinal fluid
DHFR	dihydrofolate reductase
DIM	deep indel mutagenesis
DLB	dementia with Lewy bodies
DMC	double mutant cycle
DMS	deep mutational scanning
FACS	fluorescence-activated cell sorting
fAD	familial Alzheimer's disease
FTD	frontotemporal dementia
FUS	fused in sarcoma
GWAS	genome-wide association studies
HET-s	heterokaryon incompatibility s
Htt	Huntingtin
IDP	intrinsically disordered protein
IDR	intrinsically disordered region
indel	insertion and/or deletion
MAVE	multiplexed assay of variant effects
MCI	mild cognitive impairment
MPRA	massively parallel reporter assays
NMR	nuclear magnetic resonance
PA	pathological aging
PCA	protein-fragment complementation assay
PD	Parkinson's disease
polyQ	polyglutamine
PRD	prion-like domains
PrP	Prion protein
PS1 or PS2	presenilin-1 or presenilin-2
sAD	sporadic Alzheimer's disease
SGE	saturation genome editing
SNP	single nucleotide polymorphism

Sup35N	nucleation domain of Sup35
$t_{1/2}$	half-time
T2D	type II diabetes
TDP-43	TAR DNA-binding protein 43
ThT	thioflavin T
VUS	variant of uncertain (clinical) significance
WT	wild type
Y2H	yeast two-hybrid

Table of contents

Acknowledgements	1
Abstract	3
Resum	5
Abbreviations	7
Table of contents	9
Introduction	11
1. Protein folding	13
2. Amyloids and disease	14
3. Principles of amyloid structures and their stabilizing interactions	15
3.1. Thermodynamics of amyloid formation	17
3.2. Monomeric subunit	17
3.3. Stacked monomeric subunits constitute protofilaments	18
3.4. Arrangement of protofilaments to form fibrils	19
3.5. Flanking regions	20
3.6. Polymorphism	21
4. Macroscopic and microscopic mechanisms in amyloid aggregation	22
4.1. Primary nucleation	23
4.2. Secondary nucleation mechanisms	24
4.2.1. Surface-catalyzed secondary nucleation	24
4.2.2. Fibril fragmentation	25
4.3. Growth mechanisms: elongation and dissociation	25
4.4. Global kinetic analysis	26
4.5. Oligomeric species	27
5. Amyloid β	29
6. Intrinsically disordered proteins and other types of phase transitions	33
7. Variants of uncertain significance: the new challenge in human genetics	35
7.1. Genetic variation beyond substitutions	36
8. Multiplexed assays of variant effects and deep mutational scanning	36
8.1. Different approaches for library construction	38
8.2. Engineering selection	40
8.3. Using DMS to track amyloid nucleation	41

8.4. Massively parallel sequencing	42
8.5. Inferring protein structure using DMS	42
Objectives	45
Results	47
Summary	49
Thesis director report	51
Chapter I. <i>The genetic landscape for amyloid beta fibril nucleation accurately discriminates familial Alzheimer's disease mutations</i>	53
Chapter II. <i>An atlas of amyloid aggregation: the impact of substitutions, insertions, deletions and truncations on amyloid beta fibril nucleation</i>	75
Chapter III. <i>The mutational landscape of a prion-like domain</i>	107
Discussion	121
Predicting mutational effects in amyloid forming sequences	123
Discriminating disease variants and prioritizing disease candidates	125
Structural insights from deep mutagenesis	127
Aggregation kinetics and the transition state	130
Genetic interactions to infer structural conformations	131
One assay, multiple proteins; one protein multiple assays	132
Conclusions	135
References	137
Annex I	149
Annex II. <i>Understanding and evolving prions by yeast multiplexed assays</i>	153

Introduction

1. Protein folding

The structure of globular proteins is encoded in their amino acid sequence. During the process of folding, protein sequences sample the conformational space in search for a free energy minimum - the native state. The shape of the free energy surface is defined by a large number of intermediate states, local minimums, energy barriers and protein interactions that control the kinetics and thermodynamics of the folding reaction (**Figure 1**) (Rumbley et al. 2001; Jahn and Radford 2005).

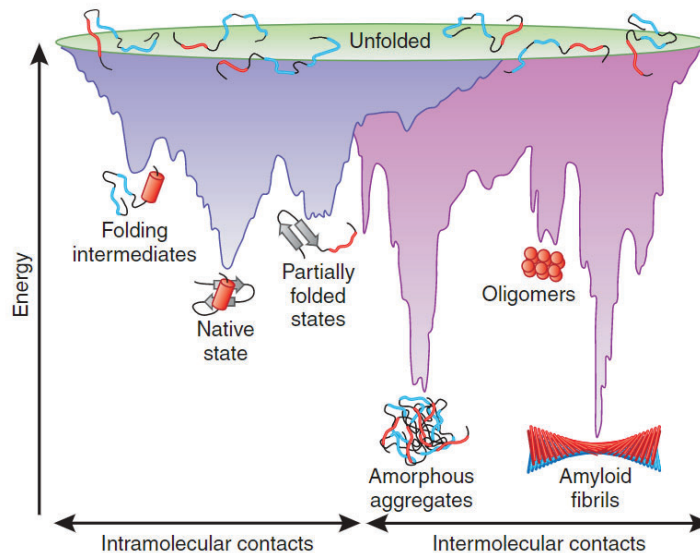


Figure 1. Free energy surface of protein folding. Adapted from (Hartl and Hayer-Hartl 2009).

In the intermediate states, proteins can be partially unfolded, for example, while they are being synthesized in the ribosome, while they are transported through membranes or because they temporarily unfold upon environmental fluctuations (Meinema et al. 2011). These intermediates may represent on-pathway metastable states that require additional reorganization to reach the native state. In these partially unfolded states, proteins are more vulnerable to misfold and aggregate. Therefore, there is a complex and delicate equilibrium between states regulated by molecular chaperones, quality control machinery and kinetic barriers (Dobson 2003).

Protein aggregates can be amorphous and disordered, but also adopt highly ordered structures such as an amyloid conformation. In fact, amyloids represent the most thermodynamically stable states in the free energy surface, while the native functional state only represents a local minimum (Dobson 2003; Chiti and Dobson 2017). Generally, the formation of amyloids is not favored due to a high kinetic barrier, especially for large peptide chains where topological constraints may impede packing into amyloid folds. However, it is thought that for shorter peptides with <150 residues, the amyloid state is somewhat more accessible. Interestingly, all the amyloidogenic proteins associated with human diseases have <700 residues with half of them of <100 residues (Chiti and Dobson 2017). In contrast, the average number of residues for 30,000 human proteins was calculated to be 500 residues, suggesting that evolution has constrained protein length against self-assembly (Monsellier et al. 2008; Baldwin et al. 2011; Tiessen, Pérez-Rodríguez, and Delaye-Arredondo 2012). In addition, protein sequences have incorporated charged and polar regions that act as gatekeepers preventing aggregation (Houben, Rousseau, and Schymkowitz 2022).

In physiological conditions, the native state of proteins is not always globular: in some cases only certain regions of the protein are structured, and in others, the full protein sequence adopts a completely disordered structure. These are known as intrinsically disordered proteins (IDPs) and they constitute more than 30% of the human proteome (Alberti and Hyman 2016). Their intrinsic disorder and the presence of multiple interaction motifs make them well suited to interact with several partners, facilitating the formation of dynamic assemblies. Thus, many IDPs function as central hubs of signaling and regulation pathways (Wright and Dyson 2015), for example, the CREB-binding protein (CBP) and p300 acetylate histones and transcription factors and also act as scaffolds to recruit and assemble the transcriptional machinery (Dyson and Wright 2005).

IDPs have also been largely associated with disease (Uversky 2015). They populate multiple conformations that depend on interactions with specific proteins, post-translational modifications or perturbations, such as mutations or environmental changes, and importantly, some of these conformations may have pathological consequences. For example, TAR DNA-binding protein 43 (TDP-43) is a nucleic acid binding protein that contains an intrinsically disordered region (IDR) at the C-terminus. TDP-43 generally localizes at the nucleus and is involved in alternative splicing and mRNA stability. However, the self-assembly of TDP-43 into insoluble aggregates, driven by its IDR, has been associated with amyotrophic lateral sclerosis (ALS) and other neurodegenerative conditions (J. P. Taylor, Brown, and Cleveland 2016).

2. Amyloids and disease

Aberrant protein folding and aggregation can result in a loss of function of the corresponding protein and/or in the generation of toxic species. Currently, there are around 50 different diseases which are associated with misfolding and protein aggregation (Chiti and Dobson 2017; Iadanza et al. 2018). Many of these disorders are associated with aging, such as Alzheimer's disease (AD) or lifestyle, such as type II diabetes (T2D), and have a great social and economic impact. For example, AD is the most common cause of dementia worldwide and the fifth cause of death in adults older than 65 years (Wong 2020). T2D, which affects >6% of the population, causes >1 million deaths per year (M. A. B. Khan et al. 2020).

Many of the pathogenic aggregating proteins assemble into amyloids, such as the amyloid β (A β) peptide causing AD, α -synuclein (α -syn) causing Parkinson's disease (PD), Prion protein (PrP) causing Creutzfeldt-Jakob disease, Tau causing Pick's disease, Huntingtin exon 1 (Htt) causing Huntington's disease or β 2-microglobulin causing dialysis-related amyloidosis, to name a few. Other types of assemblies have also been observed, the precise structural nature of which is still debated, such as intracellular inclusions of TDP-43 or fused in sarcoma (FUS), both involved in Frontotemporal dementia (FTD) and ALS, or p53 associated to cancer (Chiti and Dobson 2017; Iadanza et al. 2018; Zbinden et al. 2020). For sake of simplicity, only the most common disease for each protein is mentioned here, but many of these proteins are involved in various disorders.

A total of 37 proteins are found in amyloid deposits in disease, most of them in the extracellular space. Importantly, these proteins have little in common between them, since they differ in sequence, native structure and function. Many of the proteins involved in disease are intrinsically

disordered in their native state, such as A β 42 and α -syn (Knowles, Vendruscolo, and Dobson 2014), although there are also examples of globular proteins such as the β 2-microglobulin or the transthyretin (Colon and Kelly 1992; White et al. 2009; Knowles, Vendruscolo, and Dobson 2014). A total of 7 proteins form deposits in the central nervous system resulting in neurodegenerative diseases, like in AD and PD, 15 of them aggregate in many tissues resulting in systemic amyloidosis, while the other 15 aggregate in specific tissues, for example in the case of T2D (Chiti and Dobson 2017).

One third of the amyloid diseases are hereditary, generally autosomal dominant and with an early age of onset. Half of the diseases are sporadic (note that some conditions can be both familial or sporadic) and in this case they have a later age of onset (Chiti and Dobson 2017). Familial forms of amyloid diseases have been extensively studied, and it has been shown that mutations drive pathogenicity through several different mechanisms. For example, 100 mutations described in transthyretin, which has a globular fold in the native state, are known to destabilize the tetrameric form of the protein, enhancing the population of amyloidogenic monomers (Sekijima et al. 2005). Mutations in the apolipoprotein A-I (apo A-I) induce proteolysis, resulting in N-terminal unstructured peptides with increased amyloid propensity (Raimondi et al. 2011). For other proteins like Tau, the pathogenicity of more than 50 known mutations varies: some of them cause a loss of binding to microtubules, resulting in unfolded and aggregation-prone state (Spillantini and Goedert 2013), while others affect alternative splicing and generate a more amyloidogenic isoform (Niblock and Gallo 2012). In the case of A β 42, out of 32 known pathogenic mutations in the amyloid precursor protein (APP), 18 fall inside the A β 1-42 region and increase aggregation rates (Hatami et al. 2017; Seuma et al. 2021). The rest alter the cleavage sites in APP, generating different A β isoforms (De Jonghe et al. 2001). In some disorders, aggregation is accelerated due to aberrant extensions of the protein, such as the expansion of the polyglutamine (polyQ) repeats in the exon 1 of the Htt (Aronin et al. 1995).

Finally, it is worth mentioning that not all amyloids are necessarily associated with disease, and although many functional amyloids are found in bacteria and fungi, some examples in higher organisms such as humans also exist (Otzen and Riek 2019). Amyloids have some unique structural properties that nature has exploited with functional purposes. For example, the amyloid structure provides a dense and stable packing that is used as a storage system for peptides in pituitary secretory granules (Maji et al. 2009); the ability to oligomerize is used for the formation of the RIP1/RIP3, involved in programmed cell necrosis in mammals (J. Li et al. 2012); or the repetitive structural pattern of amyloids has an avidity effect in the surface and can act as a large binding pocket, a feature exploited for the formation of bacterial biofilms (Chapman et al. 2002). Overall, why the same type of structures are in some cases exploited to be functional and in others strongly associated with disease, is still a very open question in the amyloid field.

3. Principles of amyloid structures and their stabilizing interactions

Many, if not all, amino acid sequences can form amyloids under certain conditions. This implies that amino acid sequences and native structures of proteins capable of forming amyloids are

highly diverse. However, when adopting the amyloid structure, all proteins share a generic architecture named the cross- β fold and based on β -sheet structures (Knowles, Vendruscolo, and Dobson 2014). The X-ray diffraction pattern characteristic of cross- β has 4.7 Å meridional reflection and 10 Å equatorial reflection, arising from the regular spacing in all fibrils (Sunde et al. 1997).

X-ray diffraction, atomic force microscopy (AFM), solid-state nuclear magnetic resonance spectroscopy (ss NMR) and cryo-electron microscopy (cryoEM) have provided very detailed information about the molecular structures of amyloids. Amyloids have a hierarchical organization, meaning that symmetrical and identical structural units associate at different length scales (**Figure 2**). The highest level of hierarchy are amyloid plaques, with the assembly of mature fibrils. Plaques are observed for example, as A β 42 deposits in the brain of AD patients. Amyloid fibrils are elongated, unbranched, with diameters of 7-13 nm and up to 10 μ m in length (Xue, Homans, and Radford 2009). In turn, fibrils are formed by protofilaments wrapping around each other, and each protofilament is formed by monomeric subunits stacked one on top of the other and oriented perpendicularly to the protofilament axis. Monomeric subunits are composed of β -strands and upon stacking, they form β -sheets with hydrogen bonds running along the length of the fibril. Hydrogen bonds between carbonyl and amide groups of stacked β -strands impose the spacing of 4.7 Å seen by X-ray diffraction experiments (Sunde et al. 1997).

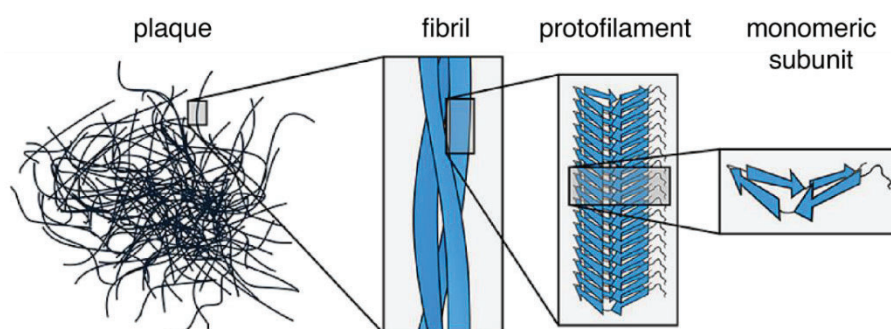


Figure 2. Hierarchical organization of amyloids. Monomeric subunits with a β fold conformation stack one on top of the other and form protofilaments. Two or more protofilaments twist around each other forming fibrils, which in turn assemble into amyloid plaques. Adapted from (A. I. P. Taylor and Staniforth 2022).

There is a wide range of intra and inter-monomer and fibril interactions that contribute to the overall high stability of amyloids (Sawaya et al. 2021; A. I. P. Taylor and Staniforth 2022). These interactions are similar to those in folded proteins, although the balance may differ. For example, while the hydrophobic effect determines most of the stability of native proteins, hydrogen bonding is the main source of stabilizing interactions in amyloids (A. W. Fitzpatrick et al. 2011).

Despite all amyloids sharing the cross- β fold and being stabilized mainly by hydrogen bonds, structures formed by different peptide sequences differ in how subunit side chains arrange, how subunits stack one on top of the other and how and how many protofilaments form a fibril (Knowles, Vendruscolo, and Dobson 2014; Chiti and Dobson 2017; Sawaya et al. 2021; A. I. P. Taylor and Staniforth 2022).

3.1. Thermodynamics of amyloid formation

As mentioned previously, the amyloid state is thermodynamically more stable than the native state of the protein (Baldwin et al. 2011). Amyloids are highly ordered structures with many interactions breaking and forming, but generally making a net favorable contribution to stability. However, there are also some entropic contributions that need to be considered. For example, when a molecule binds to the growing fibril, there is an entropy loss that depends on the concentration of the solution: the more diluted the solution, the greater the entropy loss. In addition, the growing fibril loses conformational entropy due to topological constraints. These losses are compensated by the massive enthalpic contributions arising from the newly established interactions, but also by an increase in solvent entropy upon burial of hydrophobic residues. Moreover, not all the residues in the peptide will be folded in the amyloid core but some of them will remain disordered in the flanking regions. It is possible that in their native structure, these residues have a more structured fold and so loss of structure during amyloid formation mitigates the loss of entropy in the core (Buell 2022; A. I. P. Taylor and Staniforth 2022).

It is thought that the formation of amyloids is not favored below specific concentrations, where the loss of entropy cannot be compensated by the hydrophobic effect or favorable interactions. While concentrations used *in vitro* for amyloid formation are sufficiently high so that concentration dependence is not detectable, it is intriguing how amyloid fibrils form *in vivo* for some proteins found at very low concentrations, such as A β . One possible explanation is that increased local concentration, for example in vesicles or surfaces, helps increase concentration above solubility allowing amyloid formation (Buell 2022).

3.2. Monomeric subunit

The monomeric subunit normally consists of a set of β -strands and turns with a compact conformation that is mostly stabilized by hydrophobic interactions between side chains. The core of the subunits is de-solvated and usually formed by a cluster of hydrophobic residues, an arrangement known as steric zipper (Figure 5) (Sawaya et al. 2021; A. I. P. Taylor and Staniforth 2022). For example, a recent study describes six steric zippers formed by the low complexity domain of TDP-43 and hypothesizes that they may be involved in the formation of pathogenic amyloid assemblies (Figure 3A) (Guenther et al. 2018).

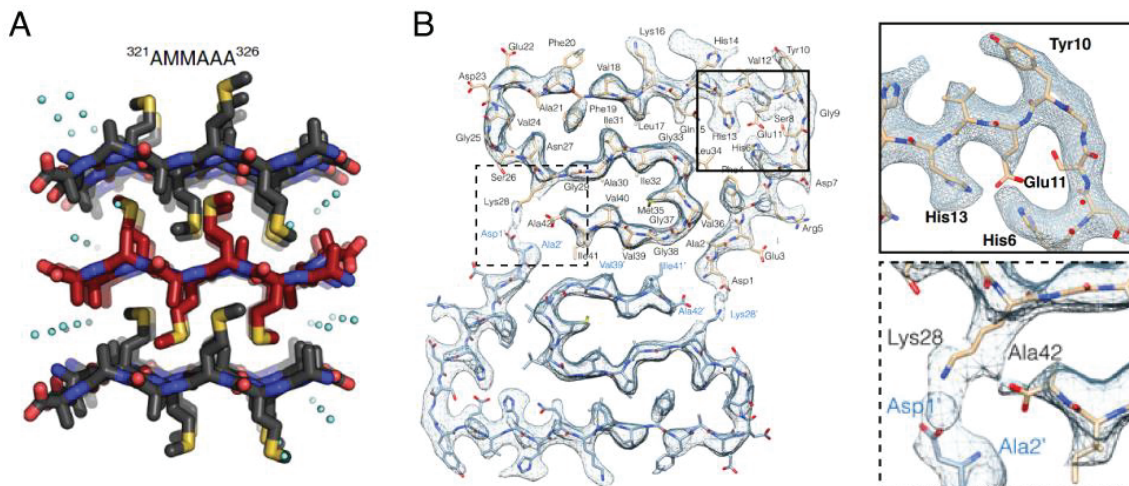


Figure 3. Interactions within monomers. **(A)** Top view of a steric zipper structure determined for an hexamer in the low complexity domain of TDP-43. Three β -sheets and two hydrophobic interfaces are shown. Adapted from (Gremer et al. 2017; Guenther et al. 2018). **(B)** Atomic model of a cross section for an A β 42 structure. Insets depict an intra-monomer (top inset) and an inter-monomer (bottom inset) salt bridges. Adapted from (Gremer et al. 2017; Guenther et al. 2018).

Other types of interactions also contribute to the folding of the monomeric subunits, such as salt bridges, either inside or outside steric zippers. Buried salt bridges, for example between residues H6-E11 and E11-H13 in one polymorph of A β 42 (Gremer et al. 2017) allow stabilization of turns inside the subunit (**Figure 3B**).

3.3. Stacked monomeric subunits constitute protofilaments

Monomeric subunits containing β -strands stack on top of each other to form a protofilament. Subunits adopt a nearly flat structure and orient the backbone hydrogen groups in parallel to the protofilament axis. By this means, each β -strand contributes to a global β -sheet all along the protofilament and backbone hydrogen bonds are the main interactions stabilizing subunit stacking (A. W. Fitzpatrick et al. 2011; Sawaya et al. 2021; A. I. P. Taylor and Staniforth 2022).

Subunits can be oriented in a parallel or antiparallel manner between them (**Figure 4A**). Parallel structures are normally in-register, meaning that identical residues are aligned one on top of the other. This arrangement is therefore stabilized by interactions such as π -stacking of aromatic residues (Gazit 2002) or amide ladders with hydrogen bonds between amide side chains, first observed for polyQs (Perutz et al. 1994). However, electrostatic interactions disfavor parallel in-register structures and they are better accommodated in antiparallel arrangements (Trovato et al. 2006). Now that >100 amyloid structures have been solved, it is possible to say that antiparallel arrangements are not as common as parallel, and that they are more associated with short sequences, where there are less constraints and less pressure to align amino acid residues. Some examples of antiparallel structures are the short peptide A β 11-25 (Petkova et al. 2004) or a polymorph of A β 40 D23N, carrying a mutation associated with disease. Antiparallel fibrils also appeared to be the toxic intermediates of the A β 40 aggregation process, but it was shown that these fibrils were metastable and ultimately switched to a parallel in-register conformation (Qiang et al. 2012).

There is an additional peculiar possible arrangement of subunits, the β -solenoid, which is well known for the heterokaryon incompatibility s (HET-s) prion from *Podospora anserina*. In this case, each subunit folds by itself in a left-handed β -helical manner forming a two-layered structure. Each layer contains three β -strands, resulting in protofilaments with three β -sheets of two stacked β -strands for each subunit. The HET-s structure resolves the unfavorable stacking of opposite charged residues by adopting a pseudo-in-register alignment (**Figure 4A**) (Wasmer et al. 2008).

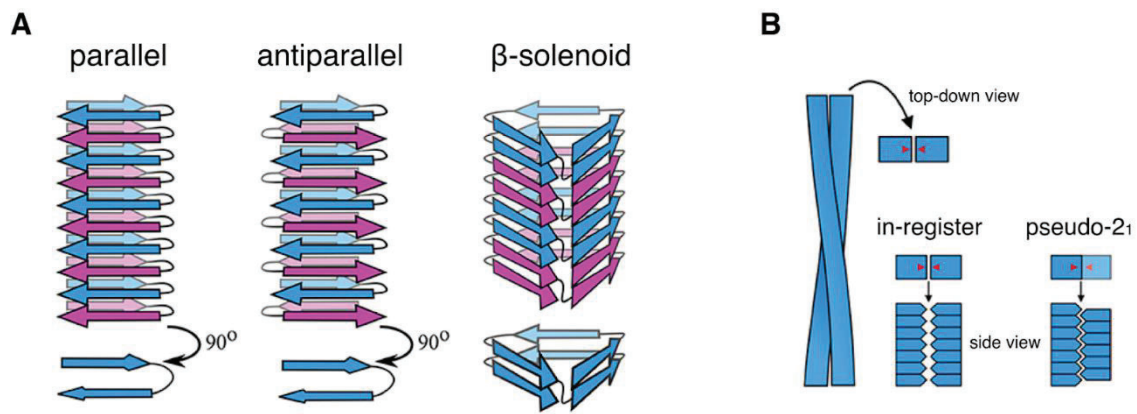


Figure 4. Possible arrangements **(A)** between stacked monomeric subunits or **(B)** between protofilaments. Adapted from (A. I. P. Taylor and Staniforth 2022).

3.4. Arrangement of protofilaments to form fibrils

Typically, two protofilaments twist around each other to form fibrils, although there are examples of fibrils with only one protofilament, such as the transthyretin fibrils purified from the tissue of a patient with hereditary amyloidosis and solved by cryoEM (Schmidt et al. 2019); and also fibrils with more than two protofilaments, for example recombinant Tau cryoEM structures with triple and quadruple fibril arrangements (Lövestam et al. 2022).

In the case of a fibril with two protofilaments, these can align in-register (also known as C_2 symmetry), meaning that facing monomeric subunits of each of the protofilaments are in the same plane, or have a pseudo-2₁ screw symmetry. In the latter case, one subunit from one protofilament is at ~ 2.4 Å far from its corresponding facing subunit at the other protofilament, given by the overall distance of ~ 4.8 Å between subunit stacks in each protofilament (**Figure 4B**) (A. I. P. Taylor and Staniforth 2022).

Inter-protofilament interactions at the interface are not especially strong but they occur in large numbers all along the length of the fibrils and so have an additivity effect. Indeed, interactions between monomers are typically stronger along the fibril (stacking monomers) rather than orthogonally (facing monomers, from different protofilaments). The interface between protofilaments is normally de-solvated, driven by the hydrophobic effect and van der Waals interactions (**Figure 5**) (Sawaya et al. 2021; A. I. P. Taylor and Staniforth 2022). A cryoEM structure for A β 42 showed that protofilaments are held together by an hydrophobic steric zipper between the side chains of V39 and I41. This structure is further stabilized by a salt bridge between D1 and K28 of opposite protofilaments (**Figure 3B**) (Gremer et al. 2017). Apart from salt bridges, hydrogen bonds are also common between protofilaments, for example, at the interface of a Tau structure three glycine residues form backbone-backbone hydrogen bonds (A. W. P. Fitzpatrick et al. 2017). The pseudo-2₁ symmetry allows both salt bridges and hydrogen bonds to form a zig-zag network that holds the two protofilaments together (A. I. P. Taylor and Staniforth 2022).

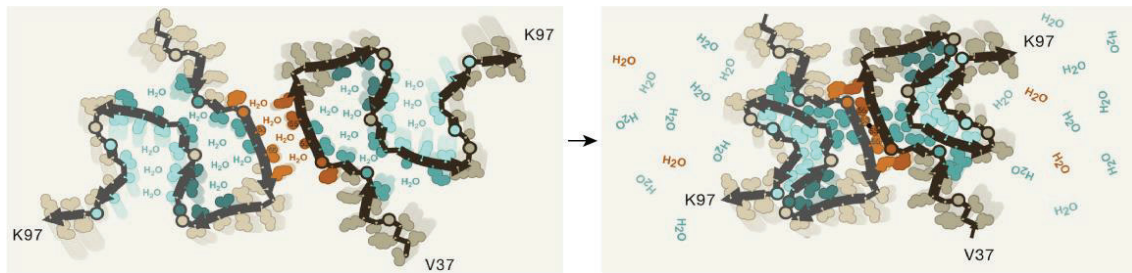


Figure 5. Steric zippers form by de-solvation within (blue) or between (orange) monomers, in regions with clustered hydrophobic residues. Adapted from (Sawaya et al. 2021).

3.5. Flanking regions

Currently, there are more than 80 amyloid structures for around 20 different proteins solved at near-atomic resolution (Sawaya et al. 2021). Thanks to these structures and a set of computational predictors, the aggregation-prone regions (APRs) and amyloid cores are identified and predicted quite accurately (Fernandez-Escamilla et al. 2004; Sormanni, Aprile, and Vendruscolo 2015). However, amyloidogenic proteins also contain large disordered regions flanking the APRs, also known as the ‘fuzzy coat’ (Tompa and Fuxreiter 2008). Due to their dynamic nature, these regions are not visible or result in low resolution densities in electron microscopy images. Therefore, despite having high quality structures for the amyloid cores, flanking regions have not been exhaustively studied and their structures remain elusive (Ulamec, Brockwell, and Radford 2020). Importantly, these regions can represent large fractions of the total sequence. For example, for α -syn, which has 140 amino acid residues, around 50-70% (depending on the polymorph structure) of the sequence is not part of the amyloid core (Guerrero-Ferreira et al. 2020).

Different types of modifications in these regions point at a crucial role for flanking regions in the process of amyloid formation: mutations linked to disease, post-translational modifications, chaperone interaction, nucleic acid or membrane binding involving flanking regions have all shown to enhance or suppress aggregation and be associated with cellular toxicity. For example, it has been proposed that flanking regions of aggregation hotspots disfavor aggregation by controlling protein expression levels. Inside the cell, there is a fine-tuned balance between protein expression and function, since high protein concentrations may help overcome energy barriers needed to form amyloids (Tartaglia et al. 2007). TDP-43 self-regulates protein expression by binding its own mRNA through an RNA-recognition motif (RRM1) flanking the C-terminal region forming the fibril core (Ayala et al. 2011). There are several examples of flanking regions binding to molecular chaperones and hence reducing aggregation, such as the N-terminal region of Tau binding to chaperone DnaJA2 (Mok et al. 2018) or 17 N-terminal residues of Htt forming a complex with Hsc70 and TriC (Monsellier et al. 2008). Small molecules can have a similar effect, such as dopamine, that by binding the C-terminal of α -syn drives the formation of off-pathway oligomers and inhibits fibril formation (Herrera et al. 2008).

Another strategy to avoid aggregation is the presence of gatekeeper residues, i.e. residues that prevent aggregation and so if mutated, aggregation increases. In many instances, gatekeepers are actually located in flanking regions, such as residue K35 located at the edge of the amyloid core of transthyretin (Sant’Anna et al. 2014) or 6 out of 7 gatekeeper residues in the disordered N-terminus of A β 42 (Seuma et al. 2021).

Flanking regions can also be responsible for increasing amyloid aggregation. Since many of them are intrinsically disordered, they are prone to mediate transient intra and inter-molecular interactions. For example, the negatively charged C-terminal region of α -syn can interact with the positively charged N-terminus, which otherwise protects the NAC region against aggregation (Stephens, Zacharopoulou, and Kaminski Schierle 2019). A β -hairpin structure outside the APR region protects Tau from aggregation but familial point mutations and alternative splicing disrupt this secondary structure and expose the APR (D. Chen et al. 2019). Similarly, mutations at the N-terminus of apo A-I disrupt a protective helix enhancing aggregation (Das et al. 2016). Flanking regions can also interact with molecules and surfaces, for example in the case of heterogeneous primary nucleation for α -syn (Galvagnion et al. 2015) or even interact with other amyloidogenic proteins: α -syn aggregation is enhanced in the presence of Tau and indeed, these two proteins co-aggregate in brains of patients with Dementia with Lewy Bodies (DLB) (Colom-Cadena et al. 2013).

3.6. Polymorphism

Amyloids from different peptide sequences share the generic cross- β architecture but differ in all other structural arrangements. Importantly, structures not only vary between different peptide sequences, but the same sequence can result in monomeric subunits, protofilaments and fibrils with different molecular structures and morphologies. This phenomenon, known as polymorphism, is at the basis of different ‘prion strains’, in which variants of the same protein result in a different phenotype (Uptain et al. 2001; Chiti and Dobson 2017; Iadanza et al. 2018; Sawaya et al. 2021).

Polymorphism may be a consequence of the generic ability of sequences to form amyloids, since the architecture is not given by the amino acid sequence itself but rather by the physicochemical properties it encodes. By this means, one sequence can assemble fibrils in multiple ways (Chiti and Dobson 2017; Iadanza et al. 2018; Sawaya et al. 2021). For example, there are 24 different structures for α -syn, 6 of which *ex vivo*, and all of them have conserved β -strands in the monomeric subunits and only differ in the interactions between monomers and their orientations (Guerrero-Ferreira et al. 2020).

Distinct polymorphs of the same protein have been often linked to different diseases. For example, the term ‘tauopathies’ includes about 25 different conditions associated with Tau aggregation and for at least 4 of them, each disease is associated with a different polymorph (Falcon, Zhang, Murzin, et al. 2018; Falcon, Zhang, Schweighauser, et al. 2018). However, more recently, it has also been shown that the same polymorph can be linked to distinct tauopathies. Thus, it is likely that each disease is defined by a specific set of polymorphs (Shi et al. 2021). To a lower extent, polymorphism also occurs between individuals with the same condition, which may concur to explain why patients with the same disease show different symptoms and disease phenotypes or why the load of amyloid deposits does not always correlate with disease severity (Qiang et al. 2017).

Whether a specific polymorph determines disease or conversely, a specific disease leads to conditions that shape protein conformation is still under debate. In the first scenario, the hypothesis of ‘polymorphism first’ assumes that all conformations are equally probable in all individuals and whichever polymorph emerges first, determines disease. In the case of ‘disease

first', it is hypothesized that the cellular environment provides disease-specific conditions that determine which polymorph emerges. However, a combination of both scenarios is also plausible, where disease-specific conditions select for a set of polymorphs that subsequently interact with cellular components in a disease-specific manner to affect phenotype (Sawaya et al. 2021).

4. Macroscopic and microscopic mechanisms in amyloid aggregation

Amyloid fibril formation is driven by a nucleation and growth process in which the soluble peptide assembles into aggregates. More specifically, nucleation is a probabilistic event in which molecules self-assemble *de novo* losing molecular degrees of freedom. This reaction is under kinetic control, meaning that free energy barriers - in this case a nucleation barrier - determine the outcome of the aggregation process, rather than the free energy of the final aggregates (Pellarin et al. 2010; T. Khan et al. 2018).

The number of fibrils formed as a function of time results in a sigmoidal kinetic curve. In a simplified polymerization model, monomers convert to nucleus during the lag phase and further addition of monomers allows rapid fibril growth during the exponential phase until the system reaches a final plateau or equilibrium phase (Ferrone 1999). However, at the microscopic level, amyloid fibril formation is a bit more complex, with various molecular mechanisms occurring at the same time (**Figure 6**). These mechanisms can be divided into 1) nucleation and secondary processes that increase the total number of aggregates (P), including primary nucleation, surface-catalyzed secondary nucleation and fibril fragmentation; and 2) growth processes that increase the overall aggregate mass (M), including fibril elongation and monomer dissociation (Meisl et al. 2017; Michaels et al. 2018).

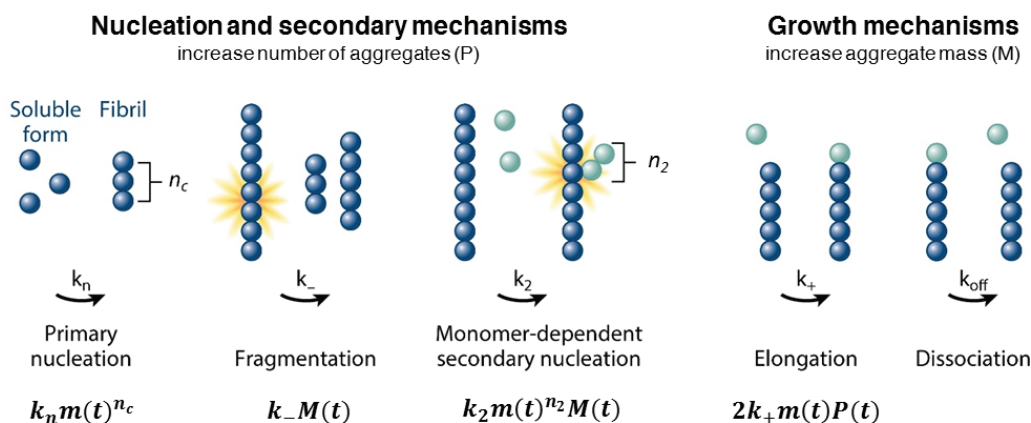


Figure 6. Schematic representation of the microscopic mechanisms of aggregation and their contributions to the overall rate of the reaction. k_n , k_- , k_2 , k_+ and k_{off} are the kinetic rate constants for primary nucleation, fragmentation, surface-catalyzed secondary nucleation, elongation and dissociation, respectively; and n_c and n_2 are the reaction orders for primary and secondary nucleation. Adapted from (Chiti and Dobson 2017).

4.1. Primary nucleation

The first step in protein aggregation is nucleation, also referred as primary or spontaneous nucleation. This process drives the protein phase transition of a homogeneous solution to a solution where the monomeric solution coexists with an aggregated phase.

In the nucleation process, a first critical nucleus is formed and as mentioned above, this is the most unstable species and has the highest Gibbs free energy in the reaction. Nucleation is a slow process with a high kinetic barrier which determines the duration of the lag time. After that, additional growth happens with lower free energy barriers (**Figure 7**) (Arosio, Knowles, and Linse 2015; Meisl et al. 2017).

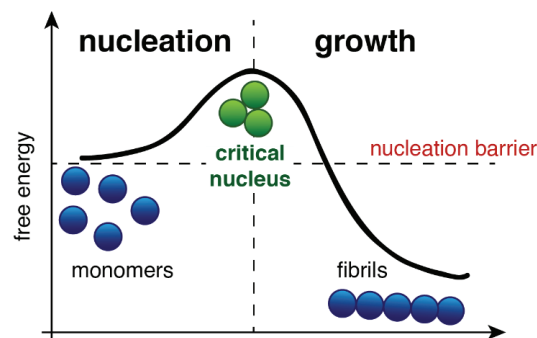


Figure 7. Free energy versus reaction progress for amyloid aggregation. The critical nucleus has the highest free energy in the system and the formation of fibrils is rate-limited by a high kinetic barrier.

According to classical nucleation theory (CNT), small clusters of monomers have high probability to dissociate back to single monomers due to the high interfacial energy towards water. However, with increasing size of the clusters, stabilization from interactions between monomers becomes more significant and the nucleus can further grow. CNT supports a one-step process in which conformational changes of the monomeric protein happen simultaneously with the nucleation step, and so nucleus enriched in β -sheets are formed directly in solution (Karthika, Radhakrishnan, and Kalaichelvi 2016).

Coarse-grained simulations suggest instead that a two-step process involving a pre-nucleation event is more likely to occur for amyloid nucleation at physiological conditions. By this means, soluble monomers would first assemble in disordered oligomers to then convert to a β fold. These oligomers would emerge from nonspecific interactions, help other peptides assemble and facilitate the conversion of other monomers into the β fold. Importantly, these studies also showed that oligomers preceding nucleation are of a very specific size and are only observed in rare fluctuations. All other oligomers dissociate back to monomers in solution after some time (Šarić et al. 2014; Šarić, Michaels, et al. 2016).

Overall, while CNT supports that the reaction order of nucleation is related to the size of the nucleus, simulations show that the reaction order is related to the proportion of species that promote the conformational conversion into the β fold. Amyloidogenic proteins have a wide range of structures in their native states, from α -helices, β -sheets or random coils, and so they must

undergo conformational rearrangements to adopt the amyloid fold. In this regard, a two-step process may be more plausible than the CNT single-step process, in which nucleus conversion and nucleation happen simultaneously (Šarić et al. 2014; Šarić, Michaels, et al. 2016).

4.2. Secondary nucleation mechanisms

Secondary nucleation mechanisms include surface-catalyzed secondary nucleation (in some cases also referred to simply as secondary nucleation), which is a monomer-dependent process; and fibril fragmentation, which is monomer-independent.

4.2.1. Surface-catalyzed secondary nucleation

Secondary nucleation, as primary nucleation, is a monomer-dependent process but in this case, the formation of nuclei is catalyzed by the surface of pre-formed aggregates. In specific cases, primary nucleation is also catalyzed by a surface, a process known as heterogeneous nucleation. One example is α -syn, which can bind and nucleate on a lipid bilayer (Galvagnion et al. 2015). However, it is important to distinguish between heterogeneous primary nucleation, occurring on external surfaces, from secondary nucleation, where the aggregates themselves act as a surface catalyzing aggregation (Törnquist et al. 2018).

While primary nucleation depends solely on available free monomer concentration, secondary nucleation represents an autocatalytic positive feedback loop in which existing aggregates control the amount of surface available for catalysis. The exponential increase of the number of aggregates with time is reflected in the shape of experimentally measured kinetics curve: a long lag phase is followed by a sharp and rapid increase of fibril mass (Meisl et al. 2017; Törnquist et al. 2018).

Coarse-grained simulations showed that the affinity of the peptide for the surface of the fibril is crucial for secondary nucleation and, similarly to primary nucleation, the peptide needs to undergo conformational changes (Šarić, Buell, et al. 2016). However, how, where and when these conformational changes and fibril formation happen during secondary nucleation is not fully understood. Many possible scenarios have been hypothesized and importantly, they do not necessarily have to be exclusive between them.

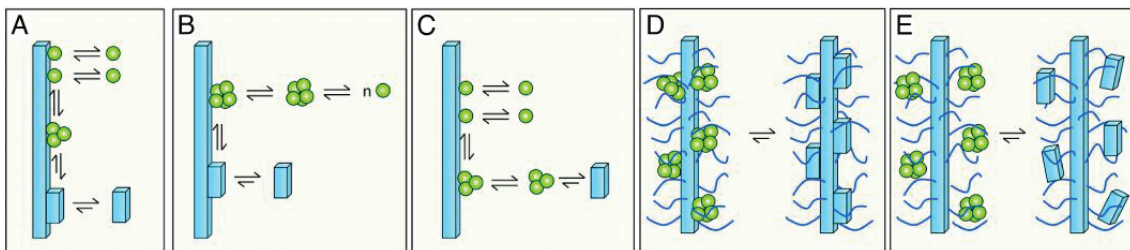


Figure 8. Possible mechanisms for surface-catalyzed secondary nucleation. **(A)** Monomers bind the fibrils, undergo a conformational change and then detach. **(B)** Instead of monomers, oligomers are the species binding the fibril surface and then change conformation. **(C)** Oligomers formed at the fibril surface detach before undergoing a conformational change, that happens later in solution. **(D)** Oligomers bind at the structured regions of the fibrils or **(E)** at the flanking disordered regions for subsequent conformational change. Adapted from (Törnquist et al. 2018).

For example, it is not known whether oligomers formed in solution are the species that bind the fibril surface (**Figure 8B**) or if monomers bind the surface to then associate into oligomers (**Figure 8A,C**). Then, oligomers need to undergo a conformational conversion to grow into fibrillar structures. This can happen either after detachment from the fibril (**Figure 8C**) or when they are still bound to it (**Figure 8A,B**), which would be more favorable for oligomers to adopt the same structure as the parent fibril. In fact, oligomers released in solution in the first scenario, with a different structure to that of the parent fibril, may represent non fibrillar toxic species (Törnquist et al. 2018).

It has also been debated where in the fibril surface secondary nucleation happens. Studies on A β 42 with molecular chaperones Brichos and clusterin revealed that while Brichos inhibits secondary nucleation, clusterin instead inhibits elongation by interaction with fibril ends, pointing at the fact that these two processes do not occur at the same surface locations (Cohen et al. 2015; Scheidt et al. 2019). It was also suggested that secondary nucleation does not happen in random locations but the fibril surface has catalytic sites with high specificity (Cukalevski et al. 2015). Importantly and as mentioned above, residues of the peptide chain that do not form the core, remain disordered and decorating the fibril surface (see section 3.5. *Flanking regions*). Nucleation could therefore happen at the structured core of the fibril, where residues have a well-defined positioning (**Figure 8D**) or at the more flexible and exposed parts (**Figure 8E**) (Ulamec, Brockwell, and Radford 2020).

4.2.2. Fibril fragmentation

By a process of fragmentation, existing aggregates break and generate smaller fragments, exposing new growth competent fibril ends where more monomeric peptides attach. This results in an exponential proliferation of fibril growth and at this point, the lag phase does not depend much on primary nucleation but rather on the time it takes for the first fibrils to multiply through fragmentation (Knowles et al. 2009; Knowles, Vendruscolo, and Dobson 2014).

Fragmentation, together with surface-catalyzed secondary nucleation, classifies as a secondary or multiplication mechanism. When these mechanisms are active, pre-formed aggregates self-replicate, which means there is an autocatalysis of new aggregates. This accelerates the rate of elongation, which subsequently speeds up the formation of other new aggregates through secondary processes. It has been recently shown that at least *in vitro*, self-replication is a common feature of many aggregating proteins, including both pathogenic and functional amyloids. Yet, self-replication time scales for pathogenic proteins are shorter than disease time scales, suggesting that self-replication is sufficiently quick to contribute to disease. Conversely, functional amyloids fall at the threshold where self-replication is not quick enough to be relevant to disease progression at biological timescales. Overall, self-replication and secondary mechanisms may be a common feature in aggregating proteins that have been fine-tuned by evolution (Meisl et al. 2022).

4.3. Growth mechanisms: elongation and dissociation

Elongation is the main process by which aggregate mass increases. During elongation, monomeric peptide is added to the ends of existing fibrils. Similarly to secondary nucleation, elongation can be described as a single-step or a two-step reaction. In the latter case, the monomer first attaches

to the fibril end in a monomer-concentration manner and then undergoes conformational rearrangement, which is independent from monomer concentration and can saturate (Scheibel, Bloom, and Lindquist 2004; Meisl et al. 2017).

The other process increasing aggregate mass is monomer dissociation, in which individual monomers are removed from the end of the fibrils and are diluted back to the monomeric solution. Since dissociation is slower than elongation, it is often neglected in the reaction kinetics (Michaels et al. 2018).

4.4. Global kinetic analysis

Studies on amyloid aggregation with well-designed experimental kinetics and mathematical models have provided a comprehensive and quantitative understanding of the microscopic mechanisms underlying macroscopic observations in amyloid fibril formation (Cohen et al. 2012). Experimentally, kinetics of aggregation can be monitored by tracking the increase in the total mass of aggregates as a function of time. This has been extensively done by using thioflavin T (ThT), a dye that specifically binds amyloids (Hellstrand et al. 2010). Macroscopic observations can be then transferred to the microscopic scale by using a generalized master equation that contains a set of nonlinear differential equations. The master equation describes how all elementary mechanisms involved in amyloid aggregation (**Figure 6**) jointly change the population of aggregates of a specific size (j) in function of time. The differential moment equations - which are a simplification of the master equation over j - for the aggregate mass concentration $M(t)$ and aggregate number $P(t)$ are defined as:

$$\begin{aligned}\frac{dP(t)}{dt} &= k_n m(t)^{n_c} + k_2 m(t)^{n_2} M(t) \\ \frac{dM(t)}{dt} &= 2k_+ m(t) P(t)\end{aligned}$$

where k_n , k_2 and k_+ are the kinetic rate constants for primary nucleation, secondary nucleation and elongation, respectively; and n_c and n_2 the reaction orders for primary and secondary nucleation, respectively.

An analytical solution to the master equation allows derivation of integrated rate laws, which can be directly compared to experimental curves (Knowles et al. 2009; Michaels et al. 2018). Just as experimental curves, integrated rate laws show a sigmoidal shape with an initial lag time, a rapid growth and a final plateau indicating monomer consumption. However, the lag and growth phases do not correspond simply to nucleus formation and fibril elongation events respectively, but each phase results from a combination of various microscopic mechanisms occurring simultaneously. By this means, one single curve is not sufficient to explain the interplay between mechanisms and so experimental data needs to be fitted globally by using a range of starting monomer concentrations (**Figure 9A**) (Knowles et al. 2009; Michaels et al. 2018).

Some parameters can be retrieved from the global fitting of integrated rate laws, for example, the reaction half-time ($t_{1/2}$), which is the time it takes for half of the initial monomer concentration $m(0)$ to be transformed into aggregates. The $t_{1/2}$ depends on the initial monomer concentration: $t_{1/2} \propto m(0)^\gamma$, where γ is the scaling exponent. Experimentally, γ is the slope of the $t_{1/2}$ plotted against the initial monomer concentration in a log-log plot, and its value gives insights into the underlying

mechanism dominating the reaction (**Figure 9B**) (Meisl et al. 2017). In order to successfully use integrated rate laws to fit experimental data, it is crucial to obtain high quality and reproducible measurements over a range of monomer concentrations, to have pure monomeric peptide without pre-formed nuclei, and to verify that ThT fluorescence measurements correlate linearly with the total mass of fibrils formed (Michaels et al. 2018).

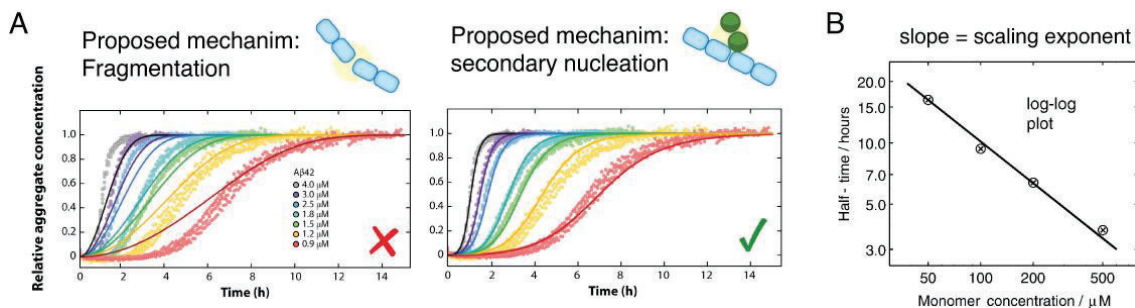


Figure 9. Global fitting of aggregation kinetics. **(A)** Aggregation is tracked experimentally (data points) by measuring aggregate concentration versus time for a range of initial monomer concentrations. Global fitting (lines) with a model that assumes fragmentation is the dominant mechanism of the reaction (left) fails to reproduce the experimental measurements. Instead, a model assuming secondary nucleation as the driving mechanism of aggregation (right), matches the data at all starting monomer concentrations. Adapted from (Michaels et al. 2018). **(B)** The scaling exponent is retrieved from the double logarithmic plot of the reaction half-time versus monomer concentration. Adapted from (Cohen et al. 2012).

This approach has been successfully applied to a set of proteins, for example it was shown that for both A β 42 and A β 40, the dominant mechanism for amyloid formation is a monomer-dependent secondary nucleation. However, for A β 40 it was also shown that the scaling exponent is not constant but depends on monomer concentration, indicating that the secondary nucleation process saturates at high concentrations, when the conversion and detachment of monomers from the fibril become rate-limiting (Cohen et al. 2013; Meisl et al. 2014).

Kinetic analysis has also been informative for perturbations in the system, such as mutations. For example, experimental kinetic measurements fitted with integrated rate laws showed that A β 42 E22G, a single point mutation associated with familial AD (fAD), is sufficient to switch the dominant mechanism of A β 42 aggregation, from a monomer-dependent to a monomer-independent secondary nucleation. Strikingly, introduction of a second mutation, A β 42 E22G/I31E, reverts the mechanism back to that of the wild type (WT) peptide (Bolognesi et al. 2014). This type of analysis also allows the investigation of the mechanism by which small molecules or chaperones target amyloid species. For example, it was shown that the molecular chaperone DNAJ B6 targets A β 42 and inhibits primary nucleation but proSP-C Brichos binds instead the surface of the fibrils, inhibiting secondary nucleation (Arosio et al. 2016).

4.5. Oligomeric species

Oligomers are prefibrillar species with great significance during amyloid formation, not only because they take part in the self-association process but also because they are associated with toxicity. At the beginning of the reaction and in the absence of fibrils, oligomers generate

exclusively through primary nucleation, since surface-catalyzed secondary nucleation requires pre-existing aggregates to initiate. Once a critical concentration of aggregates has formed, secondary nucleation overtakes primary nucleation as the major source of new oligomers. Oligomers can be precursors of amyloid fibrils (on-pathway) but also dead-end assemblies that do not further evolve into fibrils (off-pathway). Other mechanisms for amyloid formation such as fibril fragmentation also generate oligomers (Chiti and Dobson 2017).

Due to their nature - dynamic, transient and heterogeneous - oligomers have proved to be difficult to characterize. For both primary and secondary nucleation, simulations showed that the first oligomers that generate have little structure and need to undergo a conformational conversion to become growth-competent species (Šarić, Michaels, et al. 2016). This was proved for primary nucleation by single molecule studies with the prion protein Ure2p. The characterized oligomers had short life and their dissociation rate was higher than their conversion into growing fibrils. More rarely, a metastable oligomeric species could convert to a structurally different assembly and grow into fibrils (J. Yang et al. 2018).

Some populations of A β 40 and A β 42 oligomers have also been isolated, revealing that the content of β -sheet structure is related to the molecular weight, hence suggesting that oligomers are more structured with increasing numbers of molecules. In these studies, when various oligomers appear sequentially during aggregation, the first species are small and largely unstructured, and the ones containing β -sheet structures appear at later stages (**Figure 10**). Early oligomers show both antiparallel and parallel out-of-register arrangements, different from mature fibrils that normally have a parallel and in-register orientation, highlighting the need for early oligomers to undergo structural rearrangements. It is worth noting that some of the larger oligomers that have been isolated represent off-pathway species that dissociate back to monomers (Kayed et al. 2003; Chimon et al. 2007; Ahmed et al. 2010).

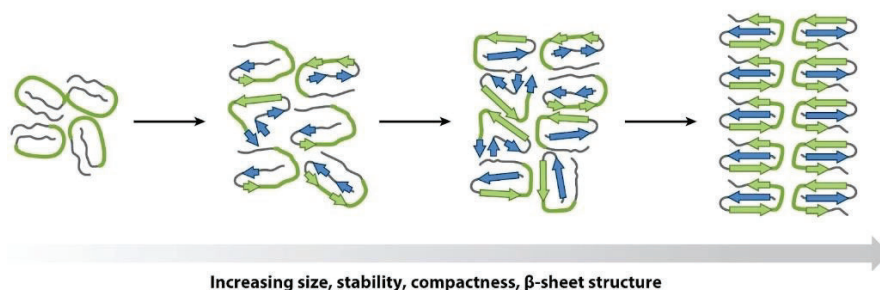


Figure 10. Oligomer structural rearrangements. Initial disordered and small species result in more ordered β -sheet structures compatible with fibril formation. Adapted from (Chiti and Dobson 2017).

For α -syn, two distinct types of oligomers (A and B) were found to form sequentially, but with a very slow transition between the two forms. Type A oligomers are smaller (less than 15 molecules), less compact and more disordered than the second population, type B, which instead show a substantial amount of β -sheet structure. Importantly, the second type of oligomers are more toxic to rat primary neurons, increasing the production of ROS (Cremades et al. 2012).

Why oligomers are toxic species is not fully understood, but it has been hypothesized that their toxicity is due to exposed hydrophobic patches on the surface (Bolognesi et al. 2010), which would otherwise remain buried in higher-order structures such as mature fibrils. These hydrophobic regions can potentially perturb cellular components. For example, it was shown in

primary neurons, that toxic type B α -syn oligomers can insert into lipid bilayers leading to an increase of Ca^{2+} and ROS, ultimately causing cell death (Fusco et al. 2017). It has also been shown that generally, smaller oligomers tend to be more toxic and that chaperones can induce further assembly into larger species, hence reducing toxicity. This suggests that toxicity from small oligomers may be due to exposed hydrophobic patches but also by a higher diffusion coefficient, which allows oligomers to diffuse and interact with cellular components more rapidly (Mannini et al. 2012).

Finally, although amyloid fibrils are not considered to be toxic *per se*, they are still species relevant to disease as they can serve as reservoirs for oligomers to be released upon fragmentation, act as surface catalysts to generate new oligomers or recruit and deplete key cellular components, altering protein homeostasis (Chiti and Dobson 2017).

5. Amyloid β

Aggregation of $\text{A}\beta_{42}$ into amyloid fibrils is a hallmark of AD. $\text{A}\beta$ is a cleavage product from the APP, a transmembrane protein with various cleavage sites. For example, the γ -secretase complex has target sites after amino acids 38, 40 or 42, resulting in $\text{A}\beta$ peptides of variable length. The cleavage site is highly relevant for the subsequent aggregation propensity of the resulting $\text{A}\beta$ peptide, with $\text{A}\beta_{42}$ known to be more amyloidogenic and responsible for cell toxicity than $\text{A}\beta_{40}$ (Haass and Selkoe 2007; Meisl et al. 2014). It has been shown that in cerebrospinal fluid (CSF) of healthy individuals, $\text{A}\beta_{40}$ is the most common isoform (~50%), followed by $\text{A}\beta_{38}$ (16%) and $\text{A}\beta_{42}$ (10%) (Kummer and Heneka 2014). In a mass spectrometry analysis, both $\text{A}\beta_{40}$ and $\text{A}\beta_{42}$ were observed in plaques in the cortex brain of both AD patients and healthy individuals. However, their presence in the hippocampus and cerebellum appeared to be exclusive to AD patients (Portelius et al. 2010).

Most cases of AD are sporadic (sAD) and thus of uncertain cause, but specific dominant mutations in $\text{A}\beta_{42}$ are known to cause familial forms of AD (fAD). In these cases, there is an early onset of the disease affecting individuals younger than 65 years (Hatami et al. 2017). Mutations are located either in APP - inside or outside the $\text{A}\beta$ 1-42 region - or in presenilin-1 (PS1) or presenilin-2 (PS2), which are part of the γ -secretase complex responsible for $\text{A}\beta$ cleavage. All mutations in PS1 and PS2 are known to enhance the production of $\text{A}\beta_{42}$ or to increase the $\text{A}\beta_{42}/\text{A}\beta_{40}$ ratio. There are a total of 18 known fAD mutations inside the $\text{A}\beta$ 1-42 region: 14 of them are single amino acid substitutions (A2V, H6R, D7N, D7H, E11K, K16Q, K16N, L17V, A21G, E22Q, E22K, E22G, D23N, V24M, L34V and A42T), one is a single amino acid deletion (E22 Δ) and one is a multi amino acid deletion (Δ 19-24). All but two of them (A2V and E22 Δ) show a dominant pattern of inheritance (Van Cauwenberghe, Van Broeckhoven, and Sleegers 2016; Hatami et al. 2017).

The number of mutations in $\text{A}\beta$ and mutations in PS1 and PS2 affecting $\text{A}\beta$ levels, highlight the causative link between $\text{A}\beta$ and AD. Despite several studies measuring the propensity of fAD variants to aggregate, there is no consensus on their effect on amyloid formation and the mechanism by which they cause disease. The main findings from the literature on fAD $\text{A}\beta_{42}$ mutations were summarized in (Seuma et al. 2021) and are included here in **Annex I**.

A

DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA
ARP1 (17-21) APR2 (29-42)

B

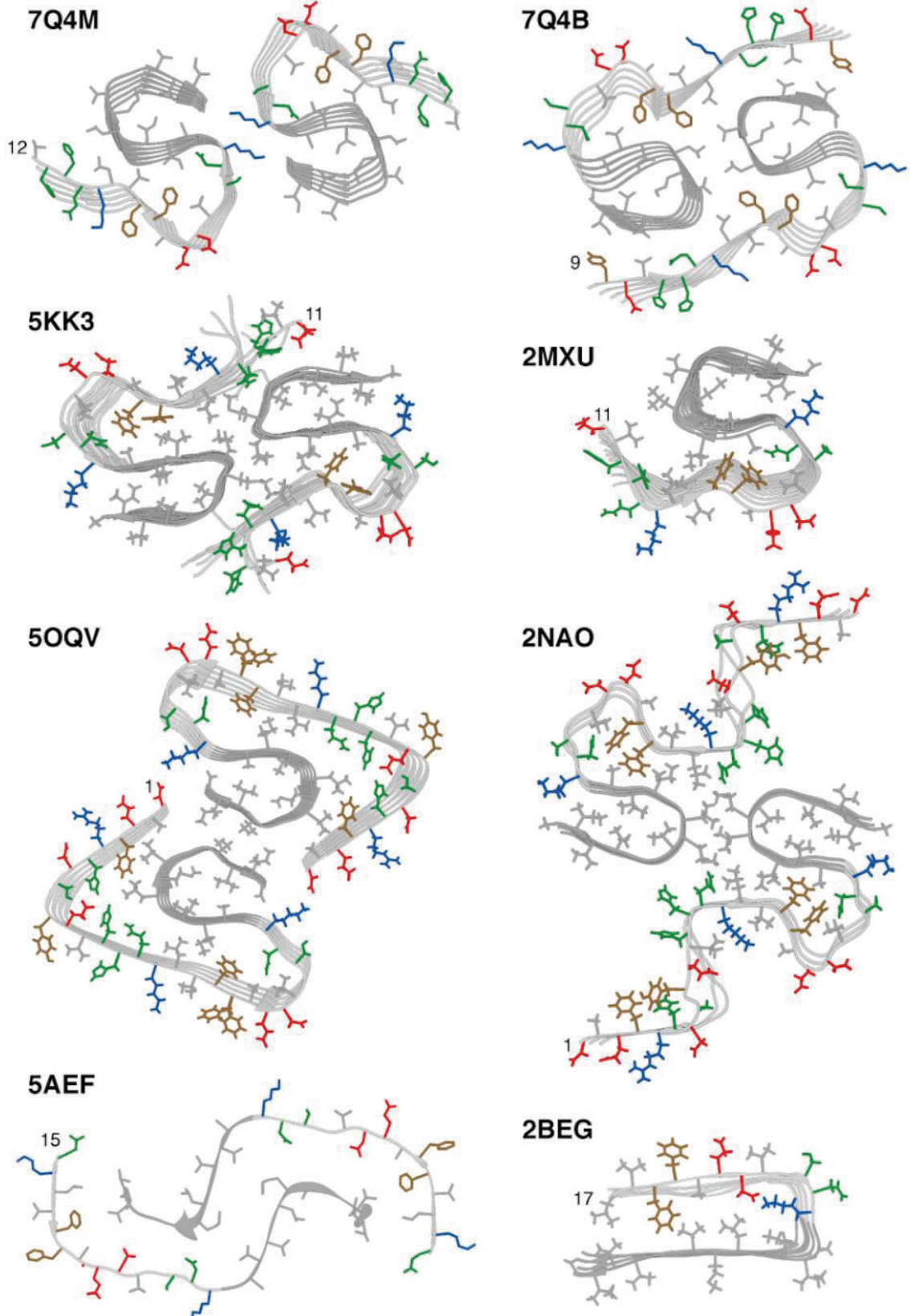


Figure 11. Structures for A β 42 (Lührs et al. 2005; Xiao et al. 2015; Schmidt et al. 2015; Colvin et al. 2016; Wälti et al. 2016; Gremer et al. 2017; Y. Yang et al. 2022)). **(A)** A β 42 amino acid sequence with APR as described in (Fernandez-Escamilla et al. 2004) depicted in yellow. **(B)** Structural models with the C-terminal region 29-42 depicted in darker gray. The first residue in each determined structure is indicated. **(A,B)** Color code indicates amino acid type, red: negatively charged, blue: positively charged, brown: aromatic, green: polar, black: glycine, gray: hydrophobic.

The amino acid sequence of A β 42 contains a distinguished C-terminal region at residues 29-42, composed exclusively by glycine and aliphatic residues and identified as an APR (Fernandez-Escamilla et al. 2004; van der Kant et al. 2022). At the N-terminus (residues 1-28), there is another predicted APR at residues 17-21. The rest of the N-terminus is composed mainly of polar and charged residues, with the first 1-11 residues remaining disordered in *ex vivo* A β 42 structures (**Figure 11A**) (Y. Yang et al. 2022).

Many structures have been determined for A β 42 fibrils by solid-state NMR and cryoEM (**Figure 11B and Table 1**) (Lührs et al. 2005; Xiao et al. 2015; Schmidt et al. 2015; Colvin et al. 2016; Wälti et al. 2016; Gremer et al. 2017; Y. Yang et al. 2022). In most structures, the monomeric subunit shows an S-shape with side chains of hydrophobic and aromatic residues - specific in each structure - forming two hydrophobic pockets and interacting with hydrogen bonds. The outer exposed surfaces are either hydrophobic, such as hydrophobic patches in structures 7Q4M and 5KK3; or hydrophilic, for example in structure 2NAO and in another surface of structure 5KK3, exposing side chains of polar and charged residues like E22 and D23. The arrangement of the two protofilaments (how monomers face each other) is very specific to each structure. The interface between protofilaments is hydrophobic in some structures (7Q4B or 5OQV) or electrostatic in some others, stabilized by a salt bridge between K28 and A42 (structure 7Q4M) or between K28 and D1 (structure 5OQV). In other cases, the salt bridge K28-A42 is instead stabilizing the monomeric subunit (structures 2NAO and 2MXU).

Table 1. Description of determined structures for A β 42 (Lührs et al. 2005; Xiao et al. 2015; Schmidt et al. 2015; Colvin et al. 2016; Wälti et al. 2016; Gremer et al. 2017; Y. Yang et al. 2022).

PDB ID	Source, method and residues determined	Main structural features
7Q4B	sAD patient, cryoEM Residues 9-42	Two protofilaments packing against each other with a pseudo-2 ₁ symmetry. The ordered core of each protofilament is G9-A42 and has 5 β -strands, with an N-terminal arm at G9-V18 and an S-shaped domain at F19-A42. The S-shaped domain is formed by two hydrophobic clusters, one with the N-terminal side chains F19, F20, V24 and I31; and the other with C-terminal side chains A30, I32, M35, V40 and A42. Between protofilaments, there is an hydrophobic interface with L34, V36, V39 and I41 on the outer surface of the S-shaped domain of one protofilament, with the side chains Y10, V12, Q15 and L17 on the N-terminal arm of the other protofilament. There are also hydrogen bonds between the side chains of H13 and H14; and E11 and H13.

7Q4M	fAD patient, cryoEM Residues 12-42	Two protofilaments with the ordered core at V12-A42, 4 β -strands in each subunit and the S-shaped domain comprising F20-A42. Side chain orientations are very similar to 7Q4B, with only a different orientation of peptides G25-S26 and V36-G37, flipped 180°. G25-S26 flip results in the expansion of the N-terminal hydrophobic cluster, accommodating the side chains of L17 and V18 instead of F19. The interface between the two protofilaments is smaller than 7Q4B: here it is formed by the opposite side of the S-shaped domain. The protofilaments pack against each other with a C_2 symmetry and are stabilized by an electrostatic interaction with K28 of one protofilament and the C-terminal carboxyl group of A42 of the other protofilament. The hydrophobic residues on the outer surfaces of the S-shape remain exposed, forming non-polar patches on the surface of filaments.
5KK3	Recombinant, spinning NMR Residues 11-42	Two protofilaments with the ordered core Q15-A42 and S-shaped monomeric subunits. Inside the hydrophobic pockets of the S-shaped domain, there are some intramolecular contacts: I41-G29, I41-K28, F19-I32, F20-V24 and F19-A30. The outer surface contains the hydrophilic side chains of E22, D23, K28 and S26. There are also two hydrophobic patches with V18, A21, V40 and A42. The salt bridge K28-A42 in this case is intramolecular. The protofilaments interact back-to-back, and at the interface there is the contact of M35 of one protofilament with Q15 and L17 of the other.
2MXU	Recombinant, ss NMR Residues 11-42	One protofilament with an S-shape and 3 β -strands at V12-V18, V24-G33 and V36-V40, with connecting loops at A21-S26 and G33-M35. The fragment 1-10 is disordered. There are (intramolecular) contacts between F20-A21, F20-V24, F19-A30, F19-I32 and F19-I31, in addition to the salt bridge K28-A42.
5OQV	Recombinant, cryoEM Residues 1-42	Two protofilaments with a pseudo- 2_1 symmetry and parallel in-register cross- β structure. The subunit forms an LS-shaped structure, with an L-shaped N-terminus and S-shaped C-terminus. The N-terminus is fully visible and part of the cross- β structure of the fibril. Three hydrophobic clusters stabilize the subunit: 1) A2, V36, F4, L34; 2) L17, I31, F19; 3) A30, I32, M35, V40. There is a salt bridge between D7-R5, E11-H6 and E11-H13. The salt bridges with E11 stabilize the kink in the N-terminal around residue Y10. The interaction between protofilament is not truly dimeric because subunits are stepwise shifted along the axis. Also, the subunit is not planar and the chain rises along the fibril axis, forming ends as 'groove' and 'ridge'. The interface between protofilaments is formed by the C-terminus, with interactions between V39 and I41 from each subunit. There is also a salt bridge between D1 and K28 between two different protofilaments.
2NAO	Synthetic, ss NMR Residues 1-42	Two protofilaments with C_2 symmetry and 5 in-register parallel β -strands, at A2-H6, Q15-V18, S26-K28, A30-I32 and V39-A42. Each subunit has an S-shape with two hydrophobic cores. Residues L17, F19, F20 and V24 from one strand interact with side chains of residues A30, I32 and L34. There is also an asparagine ladder with side chain N27 and a glutamine ladder with Q15. Residues F19 and F20 face the hydrophobic core with a non- β -strand like backbone conformation, while E22 and D23 are both

		exposed to the solvent. This segment and in particular F19-F20, are stacked off the main layer, almost reaching the next layer along the fibril axis. In the C-terminal hydrophobic core, I31 from one strand faces V36, V39 and I41 of the other strand. There is the salt bridge K28-A42 within each protofilament. At the interface between the two protofilaments, M35 interacts with both Q15 and L17, in addition to hydrophobic contacts between M35 and L34. The solvent-exposed surface has polar and charged side chains.
5AEF	Synthetic, cryoEM Residues 15-42	A dimer of two subunits with face-to-face packing at region G25-I41 and tilde-shaped. The two monomers interact by packing their hydrophobic C-terminal β -strands. There are three domains described, one central domain, with a zipper-like structural element, and two peripheral domains.
2BEG	Recombinant, NMR Residues 17-42	One protofilament with ordered core at V18-A42, with a β -strand-turn- β -strand arrangement, with the two in-register and parallel β -strands at V18-S26 and I31-A42. L17, F19 and A21 of one sheet mediate hydrophobic contacts with V36 and V40 of the other sheet. The loop N27-A30 connects the two strands, together with a salt bridge D23-K28. In addition, K28 interacts with I32 and L34.

6. Intrinsically disordered proteins and other types of phase transitions

Intrinsically disordered proteins (IDPs) constitute >30% of the human proteome (Alberti and Hyman 2016) and they contain multiple binding motifs that enable multivalent interactions with their partners (Martin et al. 2020). For example, some IDP sequences are rich in arginine and glycine repeats (RGG or RG) which allow binding to nucleic acids. In addition, some IDPs have low-complexity sequences and are enriched in Q, N, S and Y, resembling the composition of yeast prions. Thus, they are sometimes known as prion-like domains (PRDs) (Harrison and Shorter 2017).

Even though some IDPs form amyloids, such as A β 42 or α -syn, other IDPs - and especially PRDs - can also form other types of assemblies with less ordered structures, like biomolecular condensates. Indeed, intrinsically disordered regions (IDRs) and IDPs are enriched in proteins forming biomolecular condensates, which are membraneless cellular compartments that form by a process of condensation known as liquid-liquid phase separation. In this process, an initial homogeneous solution of macromolecules demixes into two distinct liquid phases: one that is enriched in certain macromolecules and the other that is depleted in the same macromolecules (Yoo, Triandafillou, and Drummond 2019). Phase separation has been linked to many biological processes, such as heterochromatin formation, transport between the nucleus and the cytoplasm, cellular organization, stress sensing and formation of cellular compartments such as the nucleoli. However, in many cases, whether phase separation is required for function is still under

debate (Alberti and Hyman 2016; Yoo, Triandafillou, and Drummond 2019; Alberti and Dormann 2019).

Apart from liquid-liquid phase separation, other phase transitions also exist, such as gelation or liquid-to-solid transitions. An example of the latter is the transition to amyloids. Yet, the physicochemical state of the protein does not necessarily determine its function or dysfunction. A ‘continuum model’ suggests that different IDPs adopt different material states which are tuned to its function and purpose, and so under physiological conditions, each of them populates a different region of this spectrum of states (Figure 12). Perturbations like mutations or environmental changes compromise the material state and condensation in various ways: by changing the mechanism of assembly, by altering the activity of a regulator of condensation, or by altering physicochemical conditions inside the cell. By this means, upon perturbation, proteins access other states in the continuum that may no longer be physiological but induce pathogenicity (Alberti and Hyman 2016).

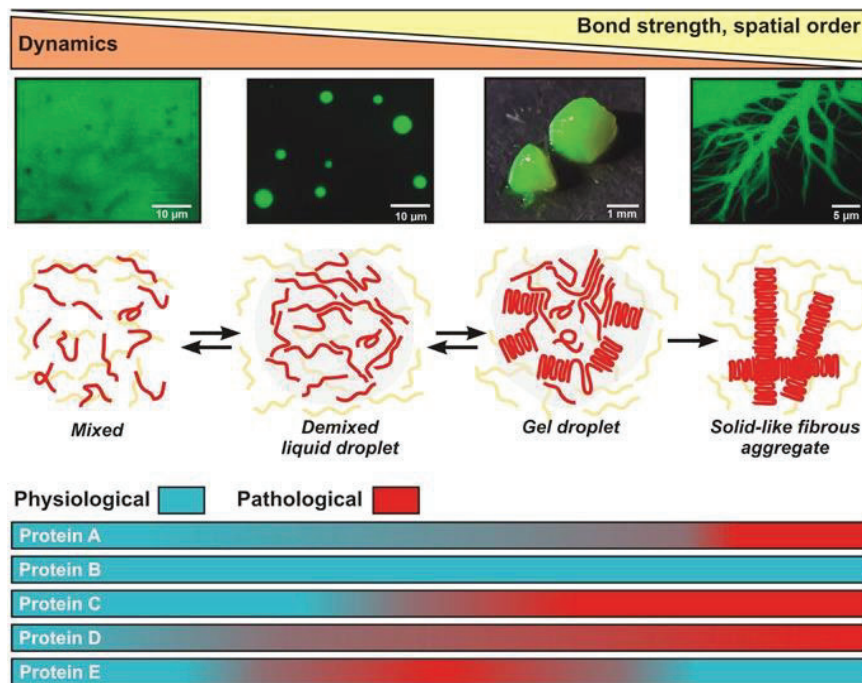


Figure 12. A continuum model for IDP physicochemical states. Under physiological conditions, each protein populates a specific material state, with specific dynamics and spatial order. Adapted from (Alberti and Hyman 2016).

Aberrant phase transitions have been linked to fatal diseases such as cancer and neurodegenerative conditions. A hallmark of all neurodegenerative diseases is the aggregation of the causal proteins, such as aggregates of TDP-43 or FUS in ALS and FTD (Ling, Polymenidou, and Cleveland 2013). However, it still remains unclear whether the aggregates are toxic *per se*, if they are only the final product of an aggregation process with alternative toxic species, or if cell death is due to loss of function of the corresponding protein (Fang et al. 2014; Bolognesi et al. 2019).

7. Variants of uncertain significance: the new challenge in human genetics

Recent advances in high-throughput technologies have facilitated the sequencing of human genomes and exomes at large-scale with a reduced cost, resulting in increasing amounts of data on human genetic variation. For example, the genome of >2,500 individuals from 26 different populations were sequenced by the 1000 Genomes Project, reporting a total of >88 million genetic variants, of which >84 million are single nucleotide polymorphisms (SNPs), 3.6 million are insertions and deletions (indels) and 60,000 are structural variants (1000 Genomes Project Consortium et al. 2015).

While this illustrates that genotyping is no longer a limitation, the main challenge nowadays in human genetics is to understand the phenotypic consequences of genetic variation: we still cannot accurately predict which and how genotypes result in specific phenotypes. The issue is even more alarming if we think that a typical genome differs from the reference genome at 4.1-5 million sites (1000 Genomes Project Consortium et al. 2015). Due to our inability to interpret how genetic variation links to function, most of the variants in the population are currently classified as VUS, i.e. variants of uncertain (clinical) significance (Fayer et al. 2021). The increasing amount of genomic data is not alleviating the so-called problem of the VUS: the number of unclassified variants is growing exponentially over time, currently accounting for the 70% of total variants in ClinVar, a repository for genomic variation and its association to human disease (Landrum et al. 2014; Fayer et al. 2021).

Traditionally, human genetics has focused on mutations displaying an observable trait or disease, biasing the attention to a small fraction of all the possible variants that exist in the population. One could argue that this strategy is sufficient to take clinical action but given the human population size and mutation rate, each single nucleotide change compatible with life is expected to occur in >50 individuals (Weile and Roth 2018). Hence, for a timely diagnosis, the effect of any variant should be measured ahead of ever being found in an individual. Even in some cases where pathogenic mutations are identified, there is still little understanding of the mechanism by which they cause disease. On top of that, many phenotypes are complex and emerge from the interplay between various genetic and environmental factors. In addition, many genes are pleiotropic and give rise to different phenotypes (Chesmore, Bartlett, and Williams 2018).

Most of the mutations described so far are located in coding genes, biased by the fact that studies have focused mainly on these regions. However, genome-wide association studies (GWAS) have discovered that >90% of SNPs associated with disease are located in non-coding regions, affecting splice sites, promoters and regulatory regions (Hindorff et al. 2009).

Multiplexed assays of variant effects (MAVEs) are emerging technologies that rely on massively parallel DNA synthesis and DNA sequencing and enable multiplexed construction of libraries and quantification of functional effects of thousands of mutations in parallel (Starita et al. 2017; Findlay 2021; Tabet et al. 2022). MAVEs have contributed to illuminating sequence-function relationships for some coding and non-coding regions, accelerating our understanding of VUS and anticipating diagnosis of disease outcomes. More broadly and beyond the scope of disease, mutagenesis is a fundamental tool to understand proteins at different levels: function, structure,

regulation and evolution. In this regard, MAVEs have the potential to build comprehensive maps of the impact of mutations that can guide the development of new therapeutic strategies or the engineering of synthetic proteins with improved or new functions. In addition, tracing evolutionary trajectories with MAVEs can reveal how adaptive mutations arise in a population under selective pressure.

7.1. Genetic variation beyond substitutions

Importantly, many mutagenesis studies have so far only focused on missense mutations - those that result from substituting the WT amino acid for another alternative amino acid. However, to build a more comprehensive picture of how mutations change proteins and alter gene regulation, other mutation types need to be considered, such as indels, recombination and splicing variants.

Other types of mutations matter because first off, they exist: for example indels are highly abundant in the human genome and are the second most common form of genetic variation after substitutions, accounting for 15-21% of polymorphisms (Mullaney et al. 2010). Second, they matter because they cause disease: 24% and 15% of Mendelian diseases are caused by small indels and alternative splicing, respectively (Jiang and Chen 2021; Stenson et al. 2017). Importantly, indels and other types of mutations have historically been under-reported because they have been missed by genotyping and classical sequence alignments. Finally, they also matter because they perturb proteins in different ways than substitutions do: while substitutions only alter the backbone of the protein, indels change the length of the amino acid sequence and hence the spatial positioning and distances between all other residues (Vetter et al. 1996). More generally, assessing different types of mutation and systematically comparing them also bring the opportunity to explore larger sequence spaces.

Scanning of indels is in its early days, but there are some pioneering studies addressing their impact on protein structure and function. For example, one of the first indels studies showed >30 variants in nine α -helices of T4 lysozyme with different outcomes, highlighting the plasticity of α -helices structures and how these structures may have changed during evolution (Vetter et al. 1996). Similarly, GFP was shown to be tolerant to many deletions in loops, helical elements and termini of strands. For one particular mutant, the authors also reported on an alternative folding process that would not have been accessible through substitution (Arpino et al. 2014). A recent study also revealed that deletions are more disruptive than insertions in a potassium channel (Macdonald et al. 2022). Other studies have mapped indels in bacteria (Gonzalez, Roberts, and Ostermeier 2019; Stephane Emond et al. 2020) and viruses (Ogden et al. 2019), and we have recently presented the first atlas of different mutation types in a human disease gene (Seuma, Lehner, and Bolognesi 2022).

8. Multiplexed assays of variant effects and deep mutational scanning

MAVEs is a broadly-used term that includes methods such as deep mutational scanning (DMS), massively parallel reporter assays (MPRAs) or saturation genome editing (SGE) (Gasparini,

Starita, and Shendure 2016; Weile and Roth 2018; Kinney and McCandlish, n.d.; Findlay 2021). Generally in DMS experiments and MAVEs, thousands of mutants can be tested in parallel in one single experiment that requires three main steps: 1) construction of a DNA library, 2) selection of variants for a specific phenotype and 3) deep sequencing of the library to link genotype to phenotype (**Figure 13**). While DMS normally refers to studies assessing protein function, MPRAs focus on regulatory regions. In both cases, variant libraries are introduced and over-expressed in plasmid constructs. Alternatively and thanks to emerging genetic editing tools such as CRISPR/Cas9, base editors or prime editing, it is now possible to integrate programmed variants directly into the genome. The great advantage of these technologies is the opportunity of studying genetic variation in the native chromosomal context, maintaining enhancers and distal factors that would otherwise be missed in plasmid libraries (Findlay et al. 2014; Erwood et al. 2022).

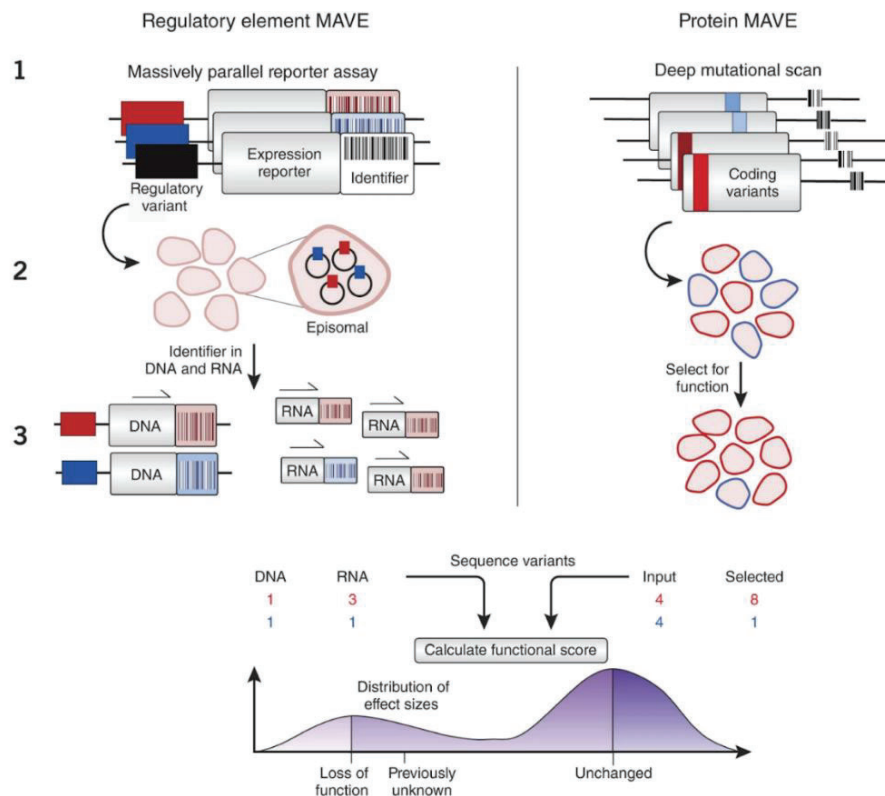


Figure 13. Main steps in MAVEs for either regulatory elements or proteins. 1) A library of variants of interest in a regulatory region or coding gene is constructed and barcoded, that is, each variant is associated to a specific identifier. 2) The library is introduced in a system model that undergoes a selection step in which variants are stratified by function or by their alteration in RNA expression 3) The library is recovered and the performance of each variant is quantified by deep sequencing. Adapted from (Gasperini, Starita, and Shendure 2016).

DMS was first described in 2010 by Fowler and Fields, when they used phage display to investigate the binding affinity of >600,000 variants of a human WW domain to its peptide ligand, coupled to high-throughput DNA sequencing. By this means, they pioneered the DMS technology, allowing functional characterisation of genotypic variation at scale (Fowler et al. 2010). Since then and during the last decade, DMS has been employed to address a range of questions in protein biology. Some studies have revealed structural insights such as new binding pockets or allosteric sites (McCormick et al. 2021; Faure et al. 2022), mapped key regions for

function (Melamed et al. 2013; Coyote-Maestas et al. 2022), discovered *in vivo* secondary structural elements (Bolognesi et al. 2019), described protein-protein interactions (Diss and Lehner 2018), tested protein stability (Matreyek et al. 2018) and tracked evolutionary trajectories (Domingo, Diss, and Lehner 2018; Starr et al., n.d.; Haddox et al. 2018). MPRA have been instead used to illuminate sequence-function relationships in gene regulatory sequences, like enhancers, promoters, splice sites or untranslated regions (Patwardhan et al. 2009; Julien et al. 2016; Maricque, Chaudhari, and Cohen 2018; Baeza-Centurion et al. 2020; Calderon et al. 2020).

Many of these studies have focused on assessing variant pathogenicity, with assays targeting disease genes or proteins. In fact, MAVEs have proved to be a useful tool to classify VUS and some of them have a striking pathogenicity predicting power. For example, a SGE study on the tumor suppressor gene *BRCA1* showed >96% accuracy in classifying pathogenic variants (Findlay et al. 2018). Other studies using yeast also accurately classified known disease variants and prioritized other current VUS as likely pathogenic, highlighting the potential use of simpler models such as yeast to report on disease mechanisms (Seuma et al. 2021; Sun et al. 2020).

Machine learning approaches have taken advantage of DMS datasets to build better predictors of pathogenicity (Griffith and Holehouse 2021; Høie et al. 2022), to infer generalizable models of genotype-phenotype maps (Tareen et al. 2022; Tonner, Pressman, and Ross 2022), to identify core regions in activation domains (Sanborn et al. 2021), to optimize antibodies (Bachas et al. 2022) or to infer the effect of untested variants in a gene (J. Zhou and McCandlish 2020), as well as extrapolating the predictions to new genes (Gray et al. 2018). In addition, there is an international community effort to collect MAVEs datasets in repositories such as MaveDB (Esposito et al. 2019), and organizations such as the Atlas of Variant Effects Alliance, with the mission of characterizing all variants in the human genome and providing accessible data and tools.

8.1. Different approaches for library construction

MAVEs require rationally designed libraries. Examples include libraries of single amino acid substitutions to study SNPs (Fowler et al. 2010; Melamed et al. 2013; Bolognesi et al. 2019; Starr et al. 2020), indels to explore various mutation types (Seuma, Lehner, and Bolognesi 2022; Macdonald et al. 2022), double or multiple amino acid substitutions to study combinatorial mutagenesis and epistasis (Olson, Wu, and Sun 2014; Diss and Lehner 2018; Schmiedel and Lehner 2019), scrambled versions of the WT amino acid sequence for positional dependencies (Sanborn et al. 2021; Staller et al. 2022), homologous sequences to map evolutionary trajectories (Domingo, Diss, and Lehner 2018), tiling sequences of long domains to identify hotspot regions (Sanborn et al. 2021), insertions of new protein motifs (Coyote-Maestas et al. 2019) or sequences covering specific physicochemical properties space (Staller et al. 2018).

Site saturation mutagenesis (i.e. substituting the WT codon of a specific position for all other possible codons) can be achieved by error-prone PCR. This simple and widely-used approach relies on a polymerase that introduces mismatches during PCR amplification at a certain frequency (Wilson and Keefe 2001). Although the frequency of error/kb can be tuned by adjusting PCR conditions, the polymerase will most likely only introduce one single nucleotide change in each codon, hence limiting the number of possible amino acid changes. Similarly, another polymerase that introduces frame-shift mutations can be used in PCR amplification to obtain a

library with insertions and deletions (Stéphane Emond et al. 2008). Mutagenesis and mutated positions can be more directly controlled by using designed PCR primers with degenerate codons, such as NNK or NNS. These two codons codify for virtually all codons while ensuring a reduced number of stop codons. One example is the inverse PCR method, that uses pairs of PCR primers pointing in opposite directions and a degenerate codon at the 5' of the forward primer. Upon PCR linearisation and re-ligation, the plasmid incorporates a new mutation in the targeted position (Jain and Varadarajan 2014). However, pairs of oligos targeting each position have to be used in independent PCR reactions, limiting the throughput of library construction. Nicking mutagenesis (Wrenbeck et al. 2016) and PFunkel (Firnberg and Ostermeier 2012) methods use instead a pool of oligos with degenerate codons at different positions in one-pot PCR reaction. In both cases, a single strand DNA template is first mutated with mutagenic oligos, followed by synthesis of the complementary strand. While nicking mutagenesis relies on restriction enzymes to degrade WT template strands, PFunkel uses uracil-containing templates that can be degraded by uracil DNA glycosylase. Other strategies that do not rely on PCR amplification to incorporate mutations make use of synthetic oligos or oligo pools (Macdonald et al. 2022). These can be directly cloned inside the linearised vector provided they are designed with flanking restriction sites for digestion, or with regions that have homology with the vector for recombination.

These are only a few of all possible strategies to build mutational libraries, with variable mutagenesis efficiency, coverage, positional bias, mutation type outcomes and cost, so that using one method or another depends on the experimental purpose. For example, custom oligo pools are normally expensive and the synthesis quality drops with sequence length. Yet, they enable the construction of comprehensive libraries that would otherwise be very difficult to obtain by means of other methods, such as libraries encompassing indels, scrambled versions of the sequence or multiple mutated positions. The methods used in our work are summarized in **Table 2**.

Table 2. Comparison of different approaches for library construction.

Method	Type of mutagenesis	Advantages	Disadvantages
Error-prone PCR	Substitutions at the nucleotide level	Rapid, low cost, high throughput	Difficult to control mutation rate, high representation of the WT sequence
Inverse PCR	Substitutions, at both nucleotide and amino acid level	Low cost	Low-throughput
Nicking mutagenesis	Substitutions, at both nucleotide and amino acid level	Rapid, high throughput	Expensive reagent, high representation of the WT sequence
Synthetic oligo pools	Customizable (substitutions, insertions, deletions, scrambled sequences)	Rapid, high-throughput, obtain any type of mutation and combinatorial libraries	Expensive, quality drops with sequence length

8.2. Engineering selection

Tailored selection assays have been used in DMS and MAVEs, in order to interrogate variants for a specific phenotype. One important aspect in this step is to evaluate and tune the discriminating power of the assay, ensuring a phenotypic dynamic range in which variants with divergent outcomes can be selected. To do so, individual variants with a prior known phenotype can be used as controls (Kowalsky et al. 2015).

Cell-based assays rely on selectable phenotypes, such as cell growth or fitness: over generations, cells carrying detrimental variants will disappear in the population while those carrying a neutral or even beneficial variant will be enriched (Findlay et al. 2018; Bolognesi et al. 2019; Gersing et al. 2022). Other strategies that rely on cell survival are those that interrogate protein-protein interactions, such as yeast two-hybrid (Y2H) or protein-fragment complementation assays (PCA) (Diss and Lehner 2018; Faure et al. 2022). Here, cell growth is determined by the strength of the interaction between two putative interacting proteins, that - upon binding - reconstitute a functional transcriptional factor or enzyme. More complex phenotypes such as cell shape or protein colocalization are also selectable thanks to recent advances in microscopy technologies (Hasle et al. 2020; Schraivogel et al. 2022).

Libraries can also be selected by auxotrophic or fluorescent reporters. Although it also applies to interrogating proteins, this is particularly useful for MPRA with DNA or RNA libraries. For example, variants in promoter regions can be selected by fluorescence-activated cell sorting (FACS) if the library is fused upstream of a GFP reporter construct (X. Li et al., n.d.; Matreyek et al. 2018). The amount of fluorescent signal is a readout of the ability of the promoter variant to regulate transcription. Quantification of RNA levels has also been used in MPRA to study regulatory elements or splicing events (Patwardhan et al. 2009; Calderon et al. 2020).

It is worth noting that *in vivo* MAVEs have been applied in different model organisms. Yeast and bacteria models have the advantage of being cheap, highly scalable and with fast generation times. For example, yeast has been used in many functional complementation assays, where human genes can rescue the deletion of their yeast orthologs. By this means, the endogenous yeast gene is replaced by a library of variants of a human gene and the relative fitness changes are interpreted as a direct effect of each variant. However, only about 200 human genes relevant to disease are known to be amenable for functional complementation assays in yeast (Sun et al. 2016). In some cases, specific phenotypes may be limited to only one model organism. For example, the nucleation assay used in our work for amyloidogenic proteins is exclusive for yeast since it is based on a yeast prion switching phenotype (Seuma et al. 2021). In other cases, more complex models such as mammalian cell lines may be more suitable to better mimic human phenotypes. It was recently shown that some human genes with great relevance to disease are essential in haploid human cell lines (e.g. HAP1) and indeed, this model system was used to test pathogenicity in 13 exons of the *BRCA1* gene, a tumor suppressor gene (Findlay et al. 2018).

Alternatively to *in vivo* methods, *in vitro* display technologies can also be used to test protein variants for their ability to bind interactors. One example is phage display, where the protein is displayed on the surface of the phages and subjected to several rounds of binding to its interactor. Variants weakly binding to the interactor are washed away and depleted in the final pull (Fowler et al. 2010).

A more extended overview of selection assays that have been already used or have great potential to be transferred to DMS can be found in (Seuma and Bolognesi 2022), included here in **Annex II**.

8.3. Using DMS to track amyloid nucleation

In section 1. *Protein folding* we reviewed that upon mutation, some proteins can adopt multiple conformations and undergo a process of self-assembly. DMS approaches that select for specific biophysical states are therefore suitable to understand the impact of mutations on the equilibrium between the different states.

Two DMS approaches with different selections have been used to map the mutational landscape of the A β 42 peptide. The first strategy reports on solubility: A β 42 is fused to dihydrofolate reductase (DHFR) and only soluble variants of A β 42 allow the enzyme to remain soluble and functional in the presence of its competitive inhibitor methotrexate. All those A β 42 aggregating variants inactivate DHFR and cause cell toxicity (Gray et al. 2019). The second approach tracks instead amyloid nucleation (see section 4. *Macroscopic and microscopic mechanisms in amyloid aggregation* for more details on amyloid nucleation). In this case, A β 42 is fused to the nucleation domain of Sup35 (Sup35N), a well-known yeast prion that functions as a translation termination factor. Importantly, Sup35 loses its function upon switching from soluble to amyloid conformation in the prion state (denoted as *[PSI+]*). This conformational change occurs naturally at very low frequency due to a high kinetic barrier but can be induced by seeding of another prion or amyloidogenic protein. This induced loss-of-function phenotype is exploited in selection as follows: nucleation of A β 42 variants is rate-limiting for the nucleation of Sup35N, which induces *[PSI+]* switching of endogenous Sup35p and a read-through of a premature stop codon in the adenine gene. Hence, growth in the absence of adenine allows selection of nucleating variants of A β 42 (**Figure 14**) (Chandramowliswaran et al. 2018; Seuma et al. 2021; Seuma, Lehner, and Bolognesi 2022).

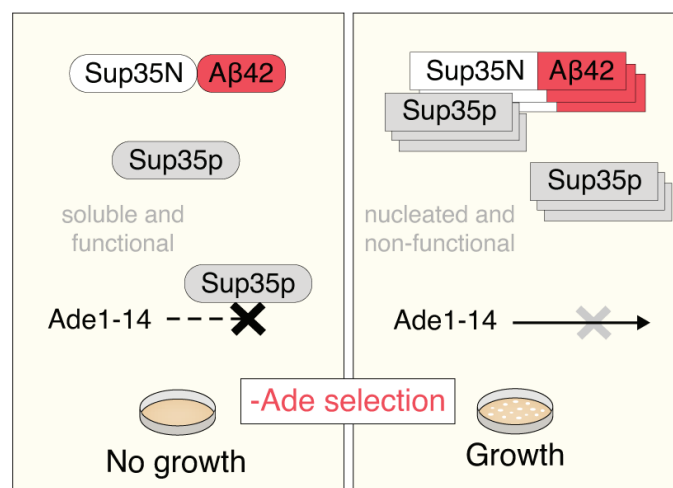


Figure 14. Amyloid nucleation selection assay. A β 42 fused to Sup35N seeds aggregation of Sup35p, causing a read-through of a premature stop in the *ade1* reporter gene and hence allowing growth in a selective medium lacking adenine.

8.4. Massively parallel sequencing

The last step of DMS experiments is the high-throughput sequencing of DNA extracted from the population before and after selection. Each cell carries a DNA sequence that encodes for the protein variant that is being expressed and selected for. Therefore, by sequencing the DNA, each phenotype is linked to a specific genotype. The performance of each variant inside the population is scored by calculating the frequency of reads in the output (after selection) over the input (before selection). This quantitative measurement is also known as the enrichment score and is standardly normalized to the score of the WT sequence (Kowalsky et al. 2015).

Many DMS studies have used paired-end Illumina sequencing, with primers annealing to the constant regions of the plasmid library (Kowalsky et al. 2015). However, this approach has coverage and read length restrictions, impeding the targeting of long sequences. To overcome this limitation, each variant can be tagged with a short barcode. By this means, and with previous barcode-variant association by long-read PacBio sequencing, only the barcodes need to be deep sequenced for library quantification (Starr et al. 2020).

Computational pipelines such as DiMSum (Faure et al. 2020) and Enrich2 (Rubin et al. 2018) have been developed to support the analysis of DMS data from the processing of the raw sequencing data, to the estimation of enrichment scores and diagnosis of data quality.

8.5. Inferring protein structure using DMS

Mutations within a protein are assumed to act independently and have additive effects, meaning that the outcome of a double mutant is the sum of the effects of the corresponding single mutants. However, when mutations have a non-independent effect, they are called epistatic and a genetic interaction is detected. It is assumed that at least some of the residues in direct structural contact within a protein will result in non-independent effect when mutated, i.e. variants at these positions are strongly interacting (**Figure 15**) (Domingo, Baeza-Centurion, and Lehner 2019).

Relying on this idea, DMS data, with hundreds or thousands of mutation measurements, has been used to determine structural elements in proteins (Schmiedel and Lehner 2019; Rollins et al. 2019). By mapping genetic interactions all along the protein, the resulting patterns can report on secondary structures that are present in the context the protein is being selected for. This approach can be used not only in globular domains but also to explore *in vivo* conformations of disordered proteins that are otherwise very difficult to resolve by traditional structural methods (Bolognesi et al. 2019).

Double mutant cycles (DMC) have been used to study the energetic coupling between residues, decades before DMS (Carter et al. 1984; Horovitz and Fersht 1990; Ackermann et al. 1998). In a classic DMC, two residues are mutated separately and in combination and the free energy of a specific process - for example, protein folding - is measured. The energetic coupling is then calculated as the difference between the expected and the measured energies. This approach can be used to study energetic interactions in proteins and complexes of known structure, but also to characterize conformations that are inaccessible by traditional structural approaches, such as transition states in protein folding (Horovitz and Fersht 1990; Horovitz 1996).

Although the specific methods to quantify phenotypes and throughput differ, quantification of genetic interactions via DMS and DMC rely on the same principles and have the same goal: to identify energetically coupled residues. Thus, combining DMS and DMC will allow quantification of the free energy of a specific process for thousands of protein variants at scale.

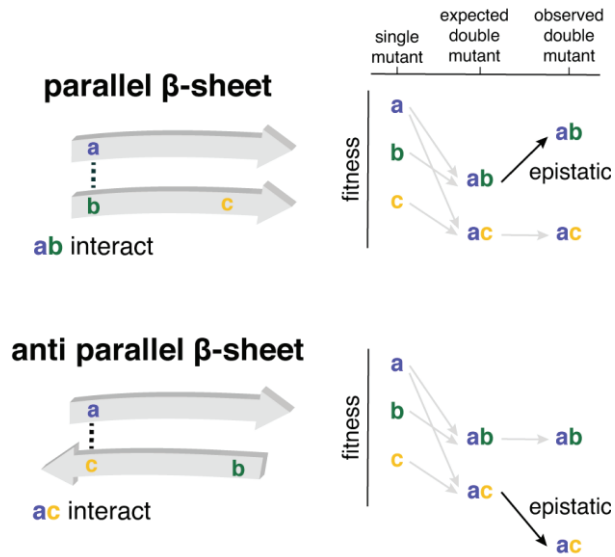


Figure 15. Residues in close structural proximity are likely to be epistatic.

Objectives

- Develop a deep mutational scanning assay that tracks amyloid aggregation for thousands of protein variants in parallel.
- Map the effect of all possible A β 42 mutations on amyloid aggregation.
- Systematically compare the effect of different types of mutations on amyloid aggregation, including substitutions, insertions, deletions and truncations.
- Decipher the underlying mechanism by which A β 42 Alzheimer's disease mutations drive aberrant aggregation.
- Map the effect of mutations on toxicity in a disordered domain of TDP-43 by using deep mutational scanning.
- Use genetic interactions to identify critical structural contacts for protein function and dysfunction.

Results

Summary

The results section of this thesis comprises three manuscripts. The first one, *The genetic landscape for amyloid beta fibril nucleation accurately discriminates familial Alzheimer's disease mutations* (see Chapter I) represents the first comprehensive map of how mutations alter the propensity of any protein to form amyloid. We used deep mutational scanning (DMS) to quantify the effect of >14,000 single and double amino acid substitutions on the nucleation of amyloid β 42 peptide (A β 42). Our data reveals a modular organization of the impact of mutations on A β 42, with most mutations at the C-terminus disrupting nucleation but with a more moderate effect at the N-terminus. We also identify 7 gatekeeper residues at the N-terminus that prevent amyloid nucleation. In addition, we map the effect of all dominant mutations associated with familial AD (fAD). Unlike computational predictors or previous experimental assays done before, nucleation scores accurately discriminate fAD variants. Moreover, the agreement of nucleation scores with human genetics suggests that fAD is a nucleation disease.

The second manuscript is currently under revision and available as a preprint, entitled *An atlas of amyloid aggregation: the impact of substitutions, insertions, deletions and truncations on amyloid beta fibril nucleation* (see Chapter II) and represents the first systematic comparison of mutation types in any human gene. Here, we used Deep Indel Mutagenesis (DIM) to generate the first atlas of amyloid aggregation including different types of mutations beyond substitutions, such as insertions, deletions and truncations. This work illustrates how the effect of mutations is not easily predictable and that important differences exist among different mutation types. The dataset provides fundamental mechanistic insights into the amyloid formation process and identifies variants of all types beyond substitutions that accelerate nucleation and are likely pathogenic.

The third manuscript *The mutational landscape of a prion-like domain* (see Chapter III), represents the first comprehensive mutational scanning of a prion-like domain. In this case, we developed a DMS toxicity assay to quantify the effect of mutations in the prion-like domain of TDP-43. This dataset reveals that TDP-43 toxicity in yeast can be explained mainly on the basis of hydrophobicity of its primary sequence, with increased hydrophobicity decreasing toxicity for the cell. We show that toxic variants form liquid-like and dynamic condensates, in contrast to those non-toxic variants forming insoluble cytoplasmic assemblies. Moreover, liquid condensates localize close to the cell nucleus, in agreement with the idea that toxicity may come from interfering with transport. We reason that toxicity from TDP-43 liquid condensates may be prevented by titrating the protein into solid aggregates. In addition, we quantify genetic interactions and identify two secondary structural elements that form *in vivo*, revealing how disordered regions can actually be partially structured *in vivo*.

Thesis director report

The first chapter of the results takes the form of the article:

Seuma, M., Faure, A. J., Badia, M., Lehner, B., & Bolognesi, B. (2021). **The genetic landscape for amyloid beta fibril nucleation accurately discriminates familial Alzheimer's disease mutations.** eLife, 10, e63364. <https://doi.org/10.7554/eLife.63364>

This work was published in eLife, in the Q1 with an impact factor of 8.713 (2022). The student is the first author and performed the experiments, analyzed the data (together with A.F. and M.B.) and wrote the manuscript (together with B.L. and B.B).

The second chapter of the results takes the form of the article:

Seuma, M., Lehner, B., Bolognesi, B. (2022). **An atlas of amyloid aggregation: the impact of substitutions, insertions, deletions and truncations on amyloid beta fibril nucleation.** bioRxiv 2022.01.18.476804; <https://doi.org/10.1101/2022.01.18.476804>

At the time of writing this thesis (September 2022), this work is under minor revisions in Nature Communications, in the Q1 with an impact factor of 17.694 (2022). A preprint is currently available in bioRxiv. The student is the first author and performed all experiments and analyzed the data. The student also wrote the manuscript together with B.L. and B.B.

The third chapter of the results takes the form of the article:

Bolognesi, B., Faure, A. J., Seuma, M., Schmiedel, J. M., Tartaglia, G. G., & Lehner, B. (2019). **The mutational landscape of a prion-like domain.** Nature communications, 10(1), 4162. <https://doi.org/10.1038/s41467-019-12101-z>

This work was published in Nature Communications, in the Q1 with an impact factor of 17.694 (2022). The student is the second author and performed the experiments, together with B.B.

In addition, Annex II contains the review article:

Seuma, M., & Bolognesi, B. (2022). **Understanding and evolving prions by yeast multiplexed assays.** Current opinion in genetics & development, 75, 101941. <https://doi.org/10.1016/j.gde.2022.101941>

This work was published in Current opinion in genetics & development, in the Q1 with an impact factor of 5.578 (2022). The student wrote the review together with B.B.

BOLOGNESI
BENEDETTA
MARTA -
Y2129836G

Digitally signed
by BOLOGNESI
BENEDETTA
MARTA -
Y2129836G
Date: 2022.09.22
08:45:49 +02'00'

Chapter I. The genetic landscape for amyloid beta fibril nucleation accurately discriminates familial Alzheimer's disease mutations

The genetic landscape for amyloid beta fibril nucleation accurately discriminates familial Alzheimer's disease mutations

Mireia Seuma¹, Andre J Faure², Marta Badia¹, Ben Lehner^{2,3,4*}, Benedetta Bolognesi^{1*}

¹Institute for Bioengineering of Catalonia (IBEC), The Barcelona Institute of Science and Technology, Barcelona, Spain; ²Center for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain; ³Universitat Pompeu Fabra (UPF), Barcelona, Spain; ⁴ICREA, Pg. Lluís Companys, Barcelona, Spain

Abstract Plaques of the amyloid beta (A β) peptide are a pathological hallmark of Alzheimer's disease (AD), the most common form of dementia. Mutations in A β also cause familial forms of AD (fAD). Here, we use deep mutational scanning to quantify the effects of >14,000 mutations on the aggregation of A β . The resulting genetic landscape reveals mechanistic insights into fibril nucleation, including the importance of charge and gatekeeper residues in the disordered region outside of the amyloid core in preventing nucleation. Strikingly, unlike computational predictors and previous measurements, the empirical nucleation scores accurately identify all known dominant fAD mutations in A β , genetically validating that the mechanism of nucleation in a cell-based assay is likely to be very similar to the mechanism that causes the human disease. These results provide the first comprehensive atlas of how mutations alter the formation of any amyloid fibril and a resource for the interpretation of genetic variation in A β .

***For correspondence:**

ben.lehner@crg.eu (BL);
bbolognesi@ibecbarcelona.eu
(BB)

Competing interests: The authors declare that no competing interests exist.

Funding: See page 15

Received: 23 September 2020

Accepted: 01 February 2021

Published: 01 February 2021

Reviewing editor: Patrik Verstreken, KU Leuven, Belgium

© Copyright Seuma et al. This article is distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use and redistribution provided that the original author and source are credited.

Introduction

Amyloid plaques consisting of the amyloid beta (A β) peptide are a pathological hallmark of Alzheimer's disease (AD), the most common cause of dementia and a leading global cause of morbidity with very high societal and economic impact (*Ballard et al., 2011; World Health Organization, 2012*). Although most cases of AD are sporadic and of uncertain cause, rare familial forms of the disease also exist (*Campion et al., 1999*). These inherited forms of dementia typically have earlier onset and are caused by high penetrance mutations in multiple loci, including in the amyloid precursor protein (*APP*) gene, which encodes the protein from which A β is derived by proteolytic cleavage (*O'Brien and Wong, 2011*). Several mutations in *PSEN1* and *PSEN2*, the genes coding for the secretases performing sequential cleavage of *APP*, also lead to autosomal dominant forms of AD. The two most abundant isoforms of A β generated upon cleavage are 42 and 40 amino acids (aa) in length, with the longer A β peptide associated with increased aggregation in vitro and cellular toxicity (*Meisl et al., 2014; Sandberg et al., 2010*). The amyloid state is a thermodynamically low energy state but, both in vitro and in vivo, the spontaneous formation of amyloids is normally very slow because of the high kinetic barrier of fibril nucleation (*Knowles et al., 2014*). The process of nucleation generates oligomeric A β species that have been hypothesized to be particularly toxic to cells and that then grow into fibrils (*Michaels et al., 2020; Bolognesi et al., 2010; Cleary et al., 2005*).

Fourteen different mutations in the A β peptide have been reported to cause familial Alzheimer's disease (fAD), with all but two having a dominant pattern of inheritance (*Weggen and Behr, 2012; Van Cauwenberghe et al., 2016*). However, it is not clear why these particular mutations cause fAD

eLife digest Alzheimer's disease is the most common form of dementia, affecting more than 50 million people worldwide. Despite more than 400 clinical trials, there are still no effective drugs that can prevent or treat the disease. A common target in Alzheimer's disease trials is a small protein called amyloid beta. Amyloid beta proteins are 'sticky' molecules. In the brains of people with Alzheimer's disease, they join to form first small aggregates and then long chains called fibrils, a process which is toxic to neurons.

Specific mutations in the gene for amyloid beta are known to cause rare, aggressive forms of Alzheimer's disease that typically affect people in their fifties or sixties. But these are not the only mutations that can occur in amyloid beta. In principle, any part of the protein could undergo mutation. And given the size of the human population, it is likely that each of these mutations exists in someone alive today.

Seuma et al. reasoned that studying these mutations could help us understand the process by which amyloid beta forms new aggregates. Using an approach called deep mutational scanning, Seuma et al. mutated each point in the protein, one at a time. This produced more than 14,000 different versions of amyloid beta. Seuma et al. then measured how quickly these mutants were able to form aggregates by introducing them into yeast cells.

All the mutations known to cause early-onset Alzheimer's disease accelerated amyloid beta aggregation in the yeast. But the results also revealed previously unknown properties that control how fast aggregation occurs. In addition, they highlighted a number of positions in the amyloid beta sequence that act as 'gatekeepers'. In healthy brains, these gatekeepers prevent amyloid beta proteins from sticking together. When mutated, they drive the protein to form aggregates.

This comprehensive dataset will help researchers understand how proteins form toxic aggregates, which could in turn help them find ways to prevent this from happening. By providing an 'atlas' of all possible amyloid beta mutations, the dataset will also help clinicians interpret any new mutations they encounter in patients. By showing whether or not a mutation speeds up aggregation, the atlas will help clinicians predict whether that mutation increases the risk of Alzheimer's disease.

(*Weggen and Beher, 2012; Van Cauwenberghe et al., 2016*), and these 14 mutations represent only 3.7% of the possible 378 single nucleotide changes that can occur in A β . As for nearly all disease genes, therefore, the molecular mechanism by which mutations cause the disease remains unclear and the vast majority of possible mutations in A β are variants of uncertain significance (VUS). This makes the clinical interpretation of genetic variation in this locus a difficult challenge (*Starita et al., 2017; Gelman et al., 2019*). Moreover, given the human mutation rate and population size, it is likely that nearly all of these possible variants in A β actually exist in at least one individual currently alive on the planet (*Conrad et al., 2011*). A comprehensive map of how all possible mutations affect the formation of A β amyloids and how these changes relate to the human disease is therefore urgently needed.

More generally, amyloid fibrils are associated with many different human diseases (*Knowles et al., 2014*), but how mutations alter the propensity of proteins to aggregate into amyloid fibrils is not well understood and there has been no large-scale analysis of the effects of mutations on the formation of any amyloid fibril. Here, we address this shortcoming by quantifying the rate of fibril formation for >14,000 variants of A β . This provides the first comprehensive map of how mutations alter the propensity of any protein to form amyloid fibrils. The resulting data provide numerous mechanistic insights into the process of A β fibril nucleation. Moreover, they also accurately classify all the known dominant fAD mutations, validating the clinical relevance of a simple cell-based model and providing a comprehensive resource for the interpretation of clinical genetic data.

Results

Deep mutagenesis of A β

To globally quantify the impact of mutations on the nucleation of A β fibrils, we used an in vivo selection assay in which the nucleation of A β is rate-limiting for the aggregation of a second amyloid, the yeast prion [PSI⁺] encoded by the *sup35* gene (Chandramowlishwaran *et al.*, 2018). Aggregation of Sup35p causes read-through of UGA stop codons, allowing growth-based selection using an auxotrophic marker containing a premature termination codon (Figure 1A and Figure 1—figure supplement 1A). We generated a library containing all possible single nucleotide variants of A β 42 fused to the nucleation (N) domain of Sup35p and quantified the effect of mutations on the rate of nucleation in triplicate by selection and deep sequencing (Faure *et al.*, 2020; see Materials and methods). The selection assay was highly reproducible, with enrichment scores for aa substitutions strongly correlated between replicates (Figure 1B and Figure 1—figure supplement 1B).

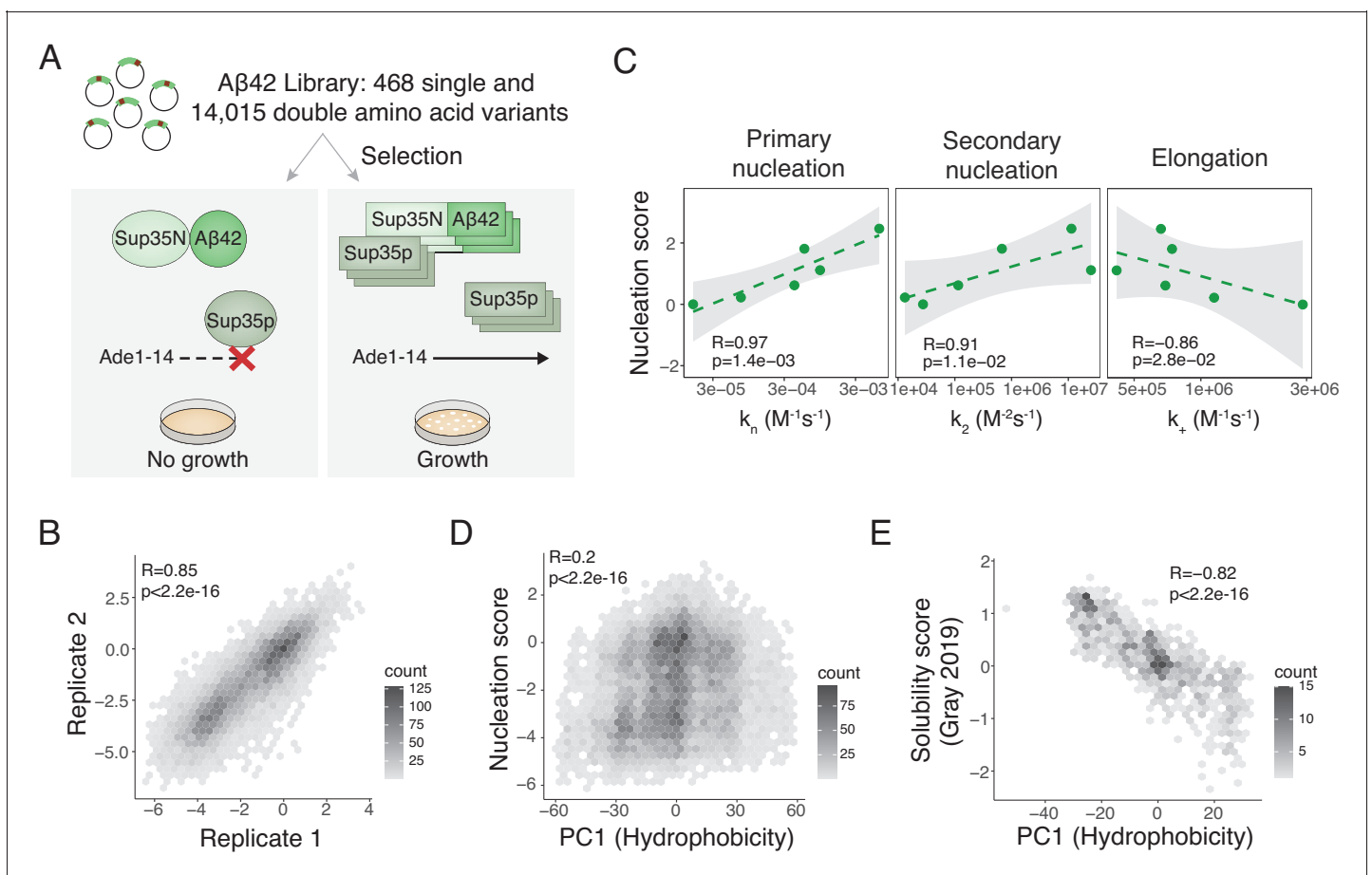


Figure 1. Deep mutagenesis of amyloid beta (A β) nucleation. (A) In vivo A β selection assay. A β fused to the Sup35N domain seeds aggregation of endogenous Sup35p causing a read-through of a premature UGA in the *Ade1-14* reporter, allowing the cells to grow in medium lacking adenine. (B) Correlation of nucleation scores for biological replicates 1 and 2 for single and double amino acid (aa) mutants. Pearson correlation coefficient and p-value are indicated (Figure 1—figure supplement 1B) $n = 10,157$ genotypes. (C) Correlation of nucleation scores with in vitro primary and secondary nucleation and elongation rate constants (Yang *et al.*, 2018). Weighted Pearson correlation coefficient and p-value are indicated. (D) Nucleation scores as a function of principal component 1 (PC1) aa property changes (Bolognesi *et al.*, 2019) for single and double aa mutants ($n = 14,483$ genotypes). Weighted Pearson correlation coefficient and p-value are indicated. (E) Solubility scores (Gray *et al.*, 2019) as a function of PC1 changes (Bolognesi *et al.*, 2019) for $n = 895$ single and double mutants. Pearson correlation coefficient and p-value are indicated.

The online version of this article includes the following source data and figure supplement(s) for figure 1:

Figure supplement 1. Reproducibility of the assay and correlation with in vitro fibril nucleation.

Figure supplement 1—source data 1. Raw colony counts from independent testing of the strains expressing the variants reported in Figure 1—figure supplement 1A.

In vivo nucleation scores are highly correlated with in vitro rates of amyloid nucleation

Comparing our in vivo enrichment scores to the qualitative effects of 16 mutations analysed in vitro across 10 previous publications validated the assay, with mutational effects matching the effects on in vitro nucleation previously reported for 14 A β variants out of 16. (*Supplementary file 1*). Moreover, the in vivo scores correlate extremely well with the rate of nucleation of A β variants in positions 21, 22, 23 (*Yang et al., 2018; Törnquist et al., 2018; Figure 1C* and *Figure 1—figure supplement 1C*). We henceforth refer to the in vivo enrichment scores as ‘nucleation scores’ (NS).

Two mechanisms of in vivo A β aggregation

A prior deep mutational scan quantified the effects of mutations on the abundance of A β fused to an enzymatic reporter (*Gray et al., 2019*). These ‘solubility scores’ do not predict the effects of mutations on A β nucleation (*Figure 1—figure supplement 1D*). Previously we identified a principal component of aa properties (principal component 1 [PC1], related to changes in hydrophobicity) that predicts the aggregation and toxicity of the amyotrophic lateral sclerosis (ALS) protein TDP-43 when it is expressed in yeast (*Bolognesi et al., 2019*). PC1 is also not a good predictor of A β nucleation (*Figure 1D*) but it does predict the previously reported changes in A β solubility (*Figure 1E*), suggesting that A β is aggregating by a similar process to TDP-43 in this alternative selection assay (*Gray et al., 2019*) but by a different mechanism in the nucleation selection.

Nucleation scores for 14,483 A β variants

The distribution of mutational effects for A β nucleation has a strong bias towards reduced nucleation, with 56% of single aa substitutions reducing nucleation but only 16% increasing it (Z-test, false discovery rate [FDR] = 0.1, *Figure 2A*). Moreover, mutations that decrease nucleation in our dataset typically have a larger effect than those that increase it, with many mutations reducing nucleation to the background rate observed for A β variants containing premature termination codons (*Figure 2A*).

In addition to covering all aa changes obtainable through single nt mutations, our mutagenesis library was designed to contain a substantial fraction of double mutants. In total, we quantified the impact of 14,015 double aa variants of A β . Double mutants were even more likely to reduce nucleation, with 63% decreasing and only 5.5% increasing nucleation (Z-test, FDR = 0.1; *Figure 2B*). Therefore, mutations more frequently decrease rather than increase A β nucleation.

A β has a modular mutational landscape

Inspecting the heatmap of mutational effects for aa changes at all positions in A β reveals strong biases in the locations of mutations that increase and decrease nucleation (*Figure 2C and D*, and *Figure 2—figure supplement 1A*). Mutations that decrease nucleation are highly enriched in the C-terminus of A β , whereas mutations that increase nucleation are enriched in the N-terminus (*Figure 2E*). Indeed, >84% of mutations in the C-terminus (residues 27-42) reduce nucleation and only 9.6% increase it (FDR = 0.1). In contrast, the effects of mutations are smaller (*Figure 2F*) and also more balanced in the first 26 aa of the peptide, with 38.6% decreasing and 20% increasing nucleation (FDR = 0.1).

These differences in the direction and strength of mutational effects between the N- and C-terminal regions of A β suggest a modular organization of the peptide. This modularity is also reflected in the primary sequence of A β , which has a hydrophobic C-terminus and a more polar and charged N-terminus (eight out of nine charged residues in A β are found before residue 24 and the peptide consists entirely of hydrophobic residues from position 29) (*Figures 2C* and *3A*). Consistent with this modular organization, mutations in the few hydrophobic residues in the N-terminus have effects that are more similar to mutations in polar residues in the N-terminus rather than in hydrophobic residues in the C-terminus. Similarly, mutations in the most C-terminal charged residue (K28) frequently strongly reduce nucleation, just as they do in the adjacent hydrophobic positions (*Figure 3A*).

Gatekeeper residues act as anti-nucleators

Considering the entire A β peptide, there are only seven positions in which mutations are not more likely to decrease rather than increase nucleation (FDR = 0.1; *Figure 2D*). Strikingly, these positions,

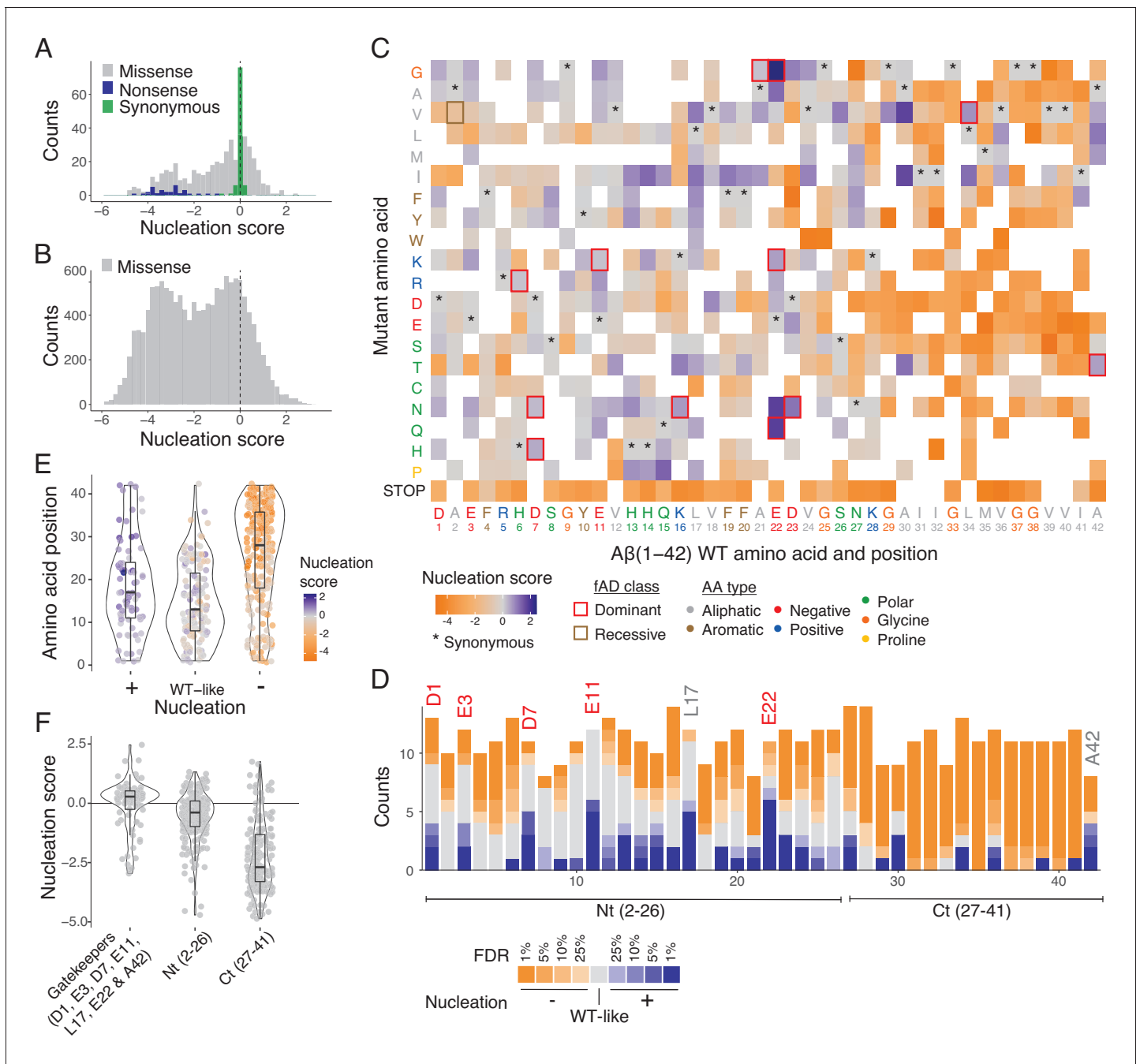


Figure 2. Modular organization of mutational effects in amyloid beta ($A\beta$). (A and B) Nucleation scores distribution for single (A) and double (B) amino acid (aa) mutants. $n = 468$ (missense), $n = 31$ (nonsense), $n = 90$ (synonymous) for singles, and $n = 14,015$ (missense) for doubles. Vertical dashed line indicates wild-type (WT) score (0). (C) Heatmap of nucleation scores for single aa mutants. The WT aa and position are indicated in the x-axis and the mutant aa is indicated on the y-axis, both coloured by aa class. Variants not present in the library are represented in white. Synonymous mutants are indicated with '*' and familial Alzheimer's disease (fAD) mutants with a box, coloured by fAD class. (D) Number of variants significantly increasing (blue) and decreasing (orange) nucleation at different false discovery rates (FDRs). Gatekeeper positions (D1, E3, D7, E11, L17, E22, and A42) are indicated on top of the corresponding bar and coloured on the basis of aa type. The N-terminal and C-terminal definitions are indicated on the x-axis. Gatekeeper positions are excluded from the N-terminal and C-terminal classes. (E) Aa position distributions for variants that increase (+), decrease (-), or have no effect on nucleation (WT-like) (FDR < 0.1). (F) Nucleation score distributions for the three clusters of positions defined on the basis of nucleation: Nt (2-26), Ct (27-41), and gatekeeper positions (clusters are mutually exclusive). Horizontal line indicates WT nucleation score (0). Nonsense (stop) mutants were only included in A and C. Boxplots represent median values and the lower and upper hinges correspond to the 25th and 75th percentiles, respectively. Whiskers extend from the hinge to the largest value no further than $1.5 \times IQR$ (interquartile range). Outliers are plotted individually or omitted when the boxplot is plotted together with individual data points or a violin plot.

Figure 2 continued on next page

Figure 2 continued

The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. Mutational effects in amyloid beta (A β).

which we refer to as ‘gatekeepers’ of nucleation (Rousseau *et al.*, 2006; Pedersen *et al.*, 2004), include five of the six negatively charged residues in A β . The sixth gatekeeper is an unusual hydrophobic residue in the N-terminus, L17, where seven mutations increase nucleation and only one decreases it (FDR = 0.1; Figure 2D). The final aa of the peptide, A42, also has an unusual distribution of mutational effects that is different to the rest of the C-terminus, with four mutations increasing and three mutations decreasing nucleation (FDR = 0.1; Figure 2D).

Taken together, on the basis of mutational effects, we therefore distinguish the following mutually exclusive positions in A β : the C-terminus (aa 27-41) where the majority of mutations strongly decrease nucleation, the N-terminus (aa 2-26) where mutations have smaller and more balanced effects, and seven gatekeeper residues (D1, E3, D7, E11, D22, L17, A42) where mutations frequently increase nucleation. We consider each of these classes below.

Mutations in the N- and C-terminal regions

Mutations in the C-terminus nearly all decrease nucleation (Figure 3A). This is consistent with the C-terminus forming part of the tightly packed amyloid core of all known structural polymorphs of both A β 42 (Colvin *et al.*, 2016; Meier *et al.*, 2017; Wälti *et al.*, 2016; Xiao *et al.*, 2015; Gremer *et al.*, 2017; Lühns *et al.*, 2005; Schmidt *et al.*, 2015) and A β 40 (Kollmer *et al.*, 2019; Lu *et al.*, 2013; Qiang *et al.*, 2012; Sgourakis *et al.*, 2015; Paravastu *et al.*, 2008; Schütz *et al.*, 2015). Consistent with this, we quantified the nucleation of three C-terminal fragments of the peptide (aa 22-42, 24-42, 27-42) and found that they nucleate similarly or better than full length A β (Figure 3—figure supplement 1C). Mutations to polar and charged residues in this region nearly all decrease nucleation, but so too do most changes to other hydrophobic residues (Figure 3B), suggesting specific side chain packing in this region is important for nucleation. The relative effects of different mutations are only partially captured by changes in hydrophobicity (Figure 3F; Pearson correlation coefficient, R = 0.45) and by predictors of aggregation potential (Figure 3—figure supplement 1A). Only a few mutations in this region increase nucleation: substitutions to isoleucine at positions 30, 34, and 39; mutations to valine at positions 29, 30, and 34; a change to threonine at position 30; changes to leucine and methionine at 36; and a mutation to phenylalanine at position 41 (FDR = 0.1).

Mutations in the N-terminus of A β have a more balanced effect on nucleation, and these effects are not well predicted by either hydrophobicity or predictors of aggregation potential (Figure 3—figure supplement 1B,D and E). The effects of introducing particular aa are, however, biased, with the introduction of asparagine, isoleucine, and valine most likely to increase nucleation (Figure 3C and Figure 3—figure supplement 2). As at the C-terminus, the introduction of negative charged residues typically strongly reduces nucleation (Figure 3B and C). However, in contrast to what is observed in the C-terminus (Figure 3B), the effects of introducing positive charge are less severe (Figure 3C). Interestingly, the effects of mutations to proline, isoleucine, valine, and threonine in the N-terminus depend on the position in which they are made: mutations in the first 12 residues typically decrease nucleation, whereas mutations in the next four to nine residues increase nucleation (Figure 3E). The conformational rigidity of proline and the beta-branched side chains of isoleucine, valine, and threonine that disfavour helix formation suggest that disruption of a secondary structure in this region may favour nucleation. Interestingly, this same region was highlighted as the part of the peptide remaining most disordered across different states of the solution ensemble of A β in molecular dynamics simulations, with the same region also making extensive long-range contacts in different states of the kinetic ensemble (Löhr *et al.*, 2021).

The role of charge in limiting A β nucleation

At five of six negatively charged positions in A β , mutations frequently increase nucleation (Figures 2D and 3A). Moreover, the introduction of negative charge at other positions strongly decreases nucleation (Figure 3A), suggesting that negatively charged residues act as gatekeepers

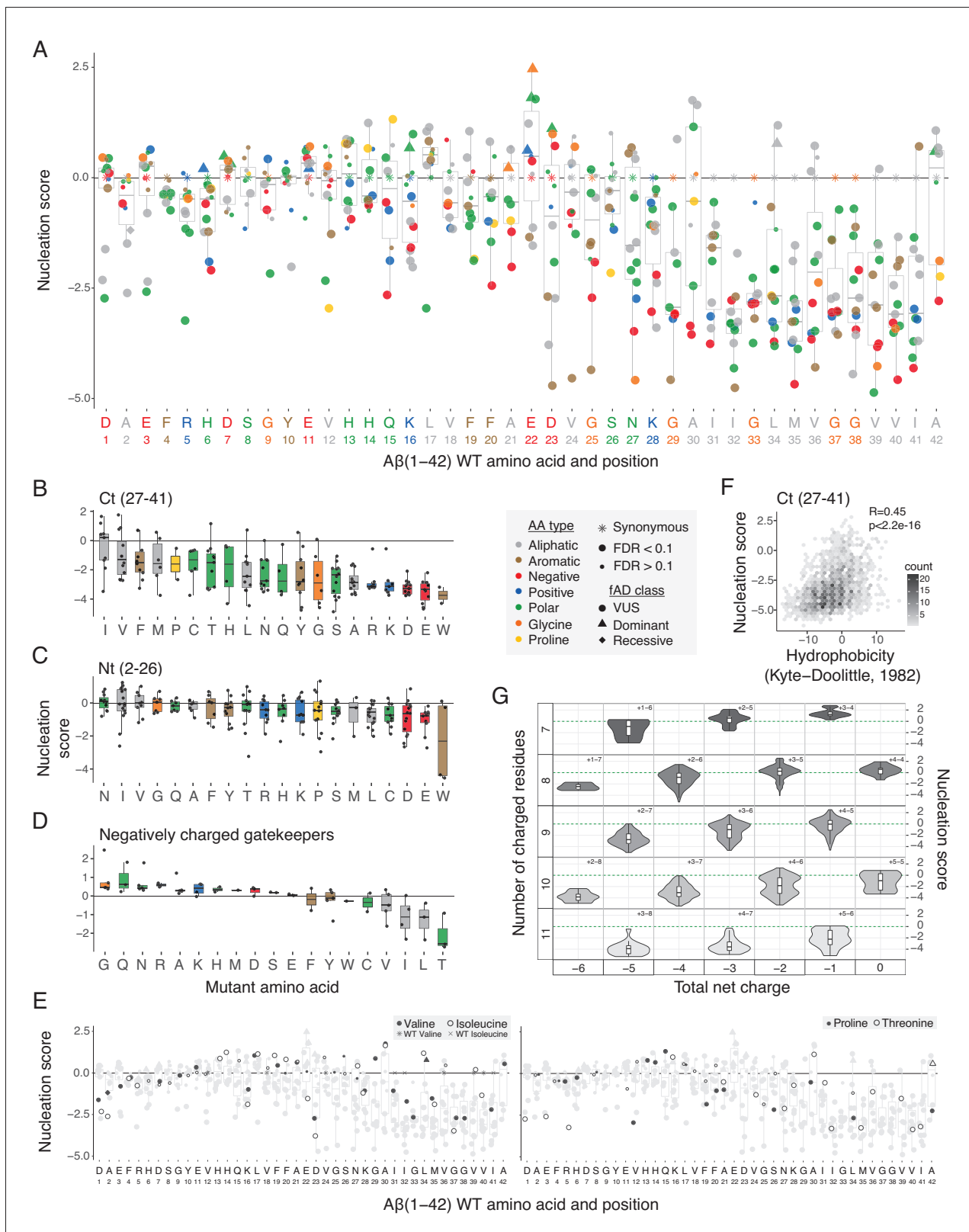


Figure 3. Determinants of amyloid beta (Aβ) nucleation. (A) Effect of single aa mutants on nucleation for each Aβ position. The wild-type (WT) aa and position are indicated on the x-axis and coloured on the basis of aa type. The horizontal line indicates the WT nucleation score (0). (B to D) Effect of each mutant aa on nucleation for the Ct (27-41) (B), the Nt (2-26) (C), and the negatively charged gatekeeper positions (D1, E3, D7, E11, and E22) (D). The three position clusters are mutually exclusive. Colour code indicates aa type. The horizontal line is set at the WT nucleation score (0). (E) Effect on

Figure 3 continued on next page

Figure 3 continued

nucleation for single aa mutations to proline, threonine, valine, and isoleucine. Mutations to other aa are indicated in grey. The horizontal line indicates WT nucleation score (0). Point size and shape indicate false discovery rate (FDR) and familial Alzheimer's disease (fAD) class, respectively (see legend). (F) Nucleation scores as a function of hydrophobicity changes (Kyte and Doolittle, 1982) for single and double aa mutants in the Ct (27-41) cluster. Only double mutants with both mutations in the indicated position-range were used. Weighted Pearson correlation coefficient and p-value are indicated. (G) Nucleation score distributions arranged by the number of charged residues (y-axis) and the total net charge (x-axis) for single and double aa mutants in the full peptide (1-42). Only polar, charged, and glycine aa types were taken into account, for both WT and mutant residues. Colour gradient indicates the total number of charged residues. Numbers inside each cell indicate the number of positive and negative residues. The horizontal line indicates WT nucleation score (0). Boxplots represent median values and the lower and upper hinges correspond to the 25th and 75th percentiles, respectively. Whiskers extend from the hinge to the largest value no further than 1.5*IQR (interquartile range). Outliers are plotted individually or omitted when the boxplot is plotted together with individual data points or a violin plot.

The online version of this article includes the following source data and figure supplement(s) for figure 3:

Figure supplement 1. Determinants of amyloid beta (A β) nucleation.

Figure supplement 1—source data 1. Raw colony counts from indepednet testing of the strains expressing the N-terminal truncated variants reported in **Figure 3—figure supplement 1C**.

Figure supplement 2. Effect of mutations to each specific amino acid (aa) on amyloid beta (A β) nucleation.

(Pedersen et al., 2004; Rousseau et al., 2006) to limit nucleation (**Figure 3D** and **Figure 3—figure supplement 1D**). In contrast, mutations in the three positively charged residues (R5, K16, K28) mostly decrease nucleation (**Figure 2D**). Mutating the negatively charged gatekeepers to the polar aa glutamine and asparagine, to positively charged residues (arginine and lysine), or to small side chains (glycine and alanine) increases nucleation (**Figure 3D**). Mutating the same positions to hydrophobic residues typically reduces nucleation (**Figure 3D**). This is consistent with a model in which the negative charge at these positions acts to limit nucleation, but that the overall polar and unstructured nature of the N-terminus must be maintained for effective nucleation.

To further investigate the role of charge in controlling A β nucleation, we extended our analyses to the double mutants. Including double mutants allows the net charge of A β to vary over a wider range and it also allows comparison of the nucleation of peptides with the same net charge but a different total number of charged residues (e.g., a net charge of -3 can result from a negative/positive aa composition of 6/3, as in wild-type A β , or compositions of 7/4, 5/2, etc.). Considering all mutations between charged and polar residues or glycine reveals that, although reducing the net charge of the peptide from -3 progressively increases nucleation (**Figure 3G**), the total number of charged residues is also important: for a given net charge, nucleation is increased in peptides containing fewer charged residues of any sign (**Figure 3G** and **Figure 3—figure supplement 1F and G**). Thus, both the overall charge and the number of charged residues control the rate of A β nucleation.

Hydrophobic gatekeeper residues

In addition to the five negatively charged gatekeeper residues, mutations most frequently increase nucleation of A β in two specific hydrophobic residues: L17 and A42 (**Figure 2C and D**). At position 17, changes to polar, aromatic, and aliphatic aa all increase nucleation, as does the introduction of a positive charge and mutation to proline. Only a mutation to cysteine reduces nucleation (**Figure 2C**). This suggests a specific role for leucine at position 17 in limiting nucleation, perhaps as part of a nucleation-limiting secondary structure suggested by the mutational effects of proline, isoleucine, valine, and threonine in this region (**Figure 3E**).

Finally, the distribution of mutational effects at position 42 differs from that in the rest of the hydrophobic C-terminus of A β , with mutations most often increasing nucleation (**Figure 2D**; FDR = 0.1). The mutations that increase nucleation are all to other aliphatic residues (**Figures 2C and 3A**). The distinction of position 42 is interesting because of the increased toxicity and aggregation propensity of A β 42 compared to the shorter A β 40 APP cleavage product (Meisl et al., 2014; Sandberg et al., 2010).

Nucleation scores accurately discriminate fAD mutations

To investigate how nucleation in the cell-based assay relates to the human disease, we considered all the mutations in A β known to cause fAD. In total, there are 11 mutations in A β reported to cause dominantly inherited fAD and one additional variant of unclear pathogenicity (H6R) (Janssen et al.,

2003). These 12 known disease mutations are not well discriminated by commonly used computational variant effect predictors (Figure 4 and Figure 4—figure supplement 1A) or by computational predictors of protein aggregation and solubility (Figure 4 and Figure 4—figure supplement 1B). They are also poorly predicted by the previous deep mutational scan of A β designed to quantify changes in protein solubility, suggesting the disease is unrelated to the biophysical process quantified in this assay (Gray et al., 2019; Figure 4—figure supplement 1C).

In contrast, the scores from our in vivo nucleation assay accurately classify the known dominant fAD mutations, with all 12 mutations increasing nucleation (Figure 4, area under the receiver operating characteristic curve, ROC–AUC = 0.9, two-tailed Z-test, $p < 2.2 \times 10^{-16}$). This suggests the biophysical events occurring in this simple cell-based assay are highly relevant to the development of the human disease.

Consistent with the overall mutational landscape, the known fAD mutations are also enriched in the N-terminus of A β (Figure 2C). In some positions the known fAD mutations are the only mutation or one of only a few mutations that can increase nucleation. For example, based on our data, K16N is likely to be one of only two fAD mutations in position 16. However, in other positions, there are several additional variants that increase nucleation as much as the known fAD mutation. At position 11, for example, there are five mutations with a NS higher than the known E11K disease mutation (Figure 2C and D). Overall, our data suggest there are likely to be many additional dominant fAD mutations beyond the 12 that have been reported to date (Supplementary file 2).

In addition to the 12 known dominant fAD mutations, two additional variants in A β have been suggested to act recessively to cause fAD (Di Fede et al., 2009; Tomiyama et al., 2008). One of these variants is a codon deletion (E22 Δ) and is not present in our library. The other variant, A2V, does not have a dominant effect on nucleation in our assay (Figure 2C), consistent with a recessive pattern of inheritance and a different mechanism of action, such as reduced β -cleavage and increased A β 42 generation, as previously proposed (Benilova et al., 2014). More generally, of the hundreds of aa changes possible in the peptide, our data prioritize 63 as candidate fAD variants (Supplementary file 2); 131 variants are likely to be benign, and 262 reduce A β nucleation and so may even be protective. These include variants already reported in the gnomAD database of human genetic variation (Figure 4—figure supplement 1D). With the currently available data for patients

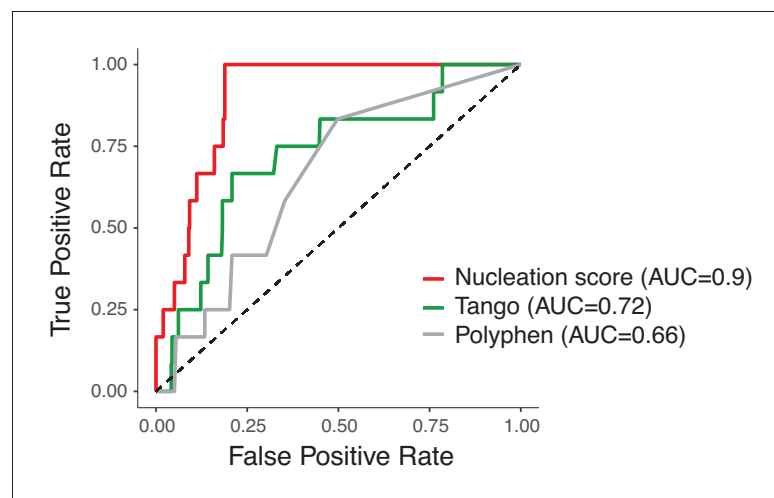


Figure 4. Amyloid beta (A β) nucleation accurately discriminates dominant familial Alzheimer’s disease (fAD) variants. Receiver operating characteristic (ROC) curves for 12 fAD mutants versus all other single aa mutants in the dataset. Area under the curve (AUC) values are indicated in the legend. Diagonal dashed line indicates the performance of a random classifier. The nucleation scores and categories for all fAD variants are reported in **Supplementary file 1**.

The online version of this article includes the following figure supplement(s) for figure 4:

Figure supplement 1. Discrimination of familial Alzheimer’s disease (fAD) variants by aggregation and variant effect predictors.

carrying fAD mutations, we could not observe a correlation between NS and disease age-of-onset (Ryman et al., 2014; Figure 4—figure supplement 1E).

Discussion

Taken together, the data presented here provides the first large-scale analysis of how mutations promote and prevent the aggregation of an amyloid. The results reveal a modular organization for the impact of mutations on the nucleation of A β . Moreover, they show that the rate of nucleation in a cell-based assay identifies all of the mutations in A β that cause dominant fAD. The dataset therefore provides a useful resource for the future clinical interpretation of genetic variation in A β .

A majority of mutations in the C-terminal core of A β disrupt nucleation, consistent with specific hydrophobic contacts in this region being required for nucleation. In contrast, mutations that increase nucleation are enriched in the polar N-terminus with mutations in negatively charged gatekeeper residues and the L17 gatekeeper being particularly likely to accelerate aggregation. Indeed, decreasing both the net charge of the peptide and the total number of charged residues increases nucleation.

Little is known about the structure of A β during fibril nucleation, but the results presented here are in general consistent with the nucleation transition state resembling the known mature fibril structures of A β where the C-terminal region of the peptide is located in the amyloid core and the N-terminus is disordered and solvent exposed (Figure 5 and Figure 5—figure supplements 1 and 2). Although the N-terminus is not required for nucleation, it does affect the process when present and most mutations that accelerate nucleation are located in the N-terminus. Interestingly, the effects of mutations in residues immediately before position 17 suggest that the formation of a structural element in this region may interfere with nucleation.

That accelerated nucleation is a common cause of fAD is also supported by the effects of mutations in APP outside of A β and by the effects of mutations in PSEN1 and PSEN2. These mutations destabilise enzyme-substrate complexes, increasing the production of the longer A β peptides that more effectively nucleates amyloid formation (Szaruga et al., 2017; Veugelen et al., 2016). In addition, A β 42 oligomers are hypothesised to be more toxic (Michaels et al., 2020; Bolognesi et al., 2010). It is possible that the effects of some of the mutations reported here on nucleation are also mediated by a change in the concentration of A β rather than by an increase in a kinetic rate parameter. Some of the variants evaluated here may have additional effects, for example, altering cleavage of APP. Future work will be needed to test these hypotheses.

Comparing our results to the effects of mutations on A β solubility quantified in a previous high-throughput analysis (Gray et al., 2019) provides evidence that, in the same type of cell (yeast), A β can aggregate in at least two different ways. Moreover, the different performance of the two sets of scores from these datasets in classifying fAD mutations suggests that one of these aggregation processes (quantified by the nucleation assay employed here) is likely to be very similar to the aggregation that occurs in the human brain in fAD. The other pathway of aggregation (quantified by the solubility assay; Gray et al., 2019), however, is less obviously related to the human disease, because mutations that cause fAD do not consistently affect it. This second aggregation pathway is, at least to a large extent, driven by changes in hydrophobicity, similar to what we previously reported for the aggregation in yeast of the ALS protein, TDP-43 (Bolognesi et al., 2019).

More generally, our results highlight how the combination of deep mutational scanning and human genetics can be a general 'genetic' strategy to quantify the disease relevance of biological assays. Many in vitro and in vivo assays are proposed as 'disease models' in biomedical research with their relevance often justified by how 'physiological' the assays seem or how well phenotypes observed in the model match those observed in the human disease. The range of phenotypes that can be assessed and their similarity to the pathology of AD human brains are appealing features of many animal models of AD and many important insights have been derived – and will continue to be derived – from animal models (Sasaguri et al., 2017). However, there are applications where animal models cannot be realistically used, for example, for high-throughput compound screening for drug discovery and for testing hundreds or thousands of genetic variants of unknown significance. For these applications, in vitro or cell-based (Pimenova and Goate, 2020; Veugelen et al., 2016) assays are required and an important challenge is to evaluate the 'disease relevance' of different assays. Our study highlights an approach to achieve this, which is to use the complete set of known disease-

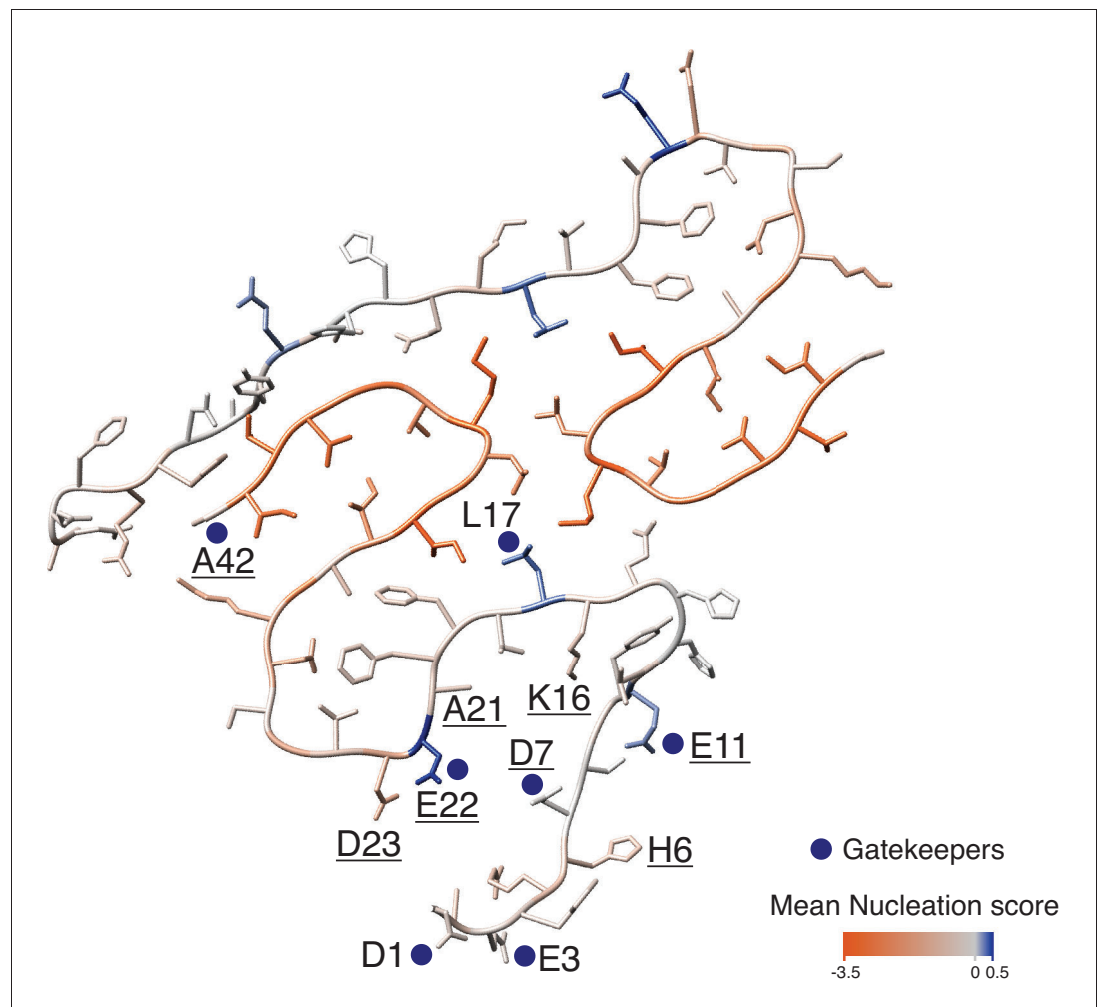


Figure 5. Mutational landscape of the amyloid beta ($A\beta$) amyloid fibril. Average effect of mutations visualized on the cross-section of an $A\beta$ amyloid fibril (PDB accession 5KK3; [Colvin et al., 2016](#)). Nucleation gatekeeper residues and known familial Alzheimer's disease (fAD) mutations positions are indicated by the wild-type (WT) aa identity on one of the two monomers; gatekeepers are indicated with blue dots and fAD are underlined. A single layer of the fibril is shown and the unstructured N-termini (aa 1-14) are shown with different random coil conformations for the two $A\beta$ monomers. See [Figure 5—figure supplement 2](#) for alternative $A\beta_{42}$ amyloid polymorphs.

The online version of this article includes the following figure supplement(s) for figure 5:

Figure supplement 1. Modular organization of $A\beta_{42}$ and $A\beta_{40}$ polymorphs.

Figure supplement 2. Modular organization of mutational effects and gatekeepers visualized on $A\beta_{42}$ polymorphs.

causing mutations to quantify the 'genetic agreement' between an assay and a disease. Thus, although the yeast-based assay that we employed here might typically be dismissed as 'non-physiological,' 'artificial,' or 'lacking many features important for a neurological disease,' unbiased massively parallel genetic analysis provides very strong evidence that it is reporting on biophysical events that are extremely similar to – or the same as – those that cause the human disease. Indeed, one could argue that this simple system is now better validated as a model of fAD than many others, including animal models where the effects of only one or a few mutations (including control mutations) have ever been tested. Similarly strong agreement between mutational effects in a cellular assay and the set of mutations already known to cause a disease is observed for other diseases ([Starita et al., 2017](#); [Gelman et al., 2019](#)), suggesting the generality of this approach.

We suggest therefore that the combination of deep mutational scanning and human genetics provides a general strategy to quantify the disease relevance of in vitro and cell-based assays. We encourage that deep mutagenesis should be employed early in discovery programmes to ‘genetically validate’ (or invalidate) the relevance of assays for particular diseases. The concordance between mutational effects in an assay and a disease is an unbiased metric that can be used to prioritize between different assays. Quantifying the ‘genetic agreement’ between an assay and a disease will help prevent time and resources being wasted on research that actually has little relevance to a disease.

Finally, the strikingly consistent effects of the dominant fAD mutations in our assay further strengthen the evidence that fAD is a ‘nucleation disease’ ultimately caused by an increased rate of amyloid nucleation (Aprile et al., 2017; Cohen et al., 2018; Knowles et al., 2009). This accelerated nucleation can be caused by the direct effects of mutations in A β — such as those quantified here — or by changes in upstream factors (Szaruga et al., 2017). If this hypothesis is correct, then nucleation is the key bioph step to target to prevent or treat AD. We suggest that the ‘genetic validation’ of assays by mutational scanning and comparison to sets of known disease-causing mutations will be increasingly important in assay development and drug discovery pipelines.

Materials and methods

Plasmid library construction

The plasmid P_{CUP1}-Sup35N-A β 42 used in this study was a kind gift from the Chernoff lab (Chandramowlishwaran et al., 2018).

The A β coding sequence and two flanking regions of 52 bp and 72 bp, respectively, upstream and downstream of A β were amplified (primers MS_01 and MS_02, **Supplementary file 3**) by error-prone PCR (Mutazyme II DNA polymerase, Agilent). Thirty cycles of amplification and 0.01 ng of initial template were used to obtain a mutagenesis rate of 16 mutations/kb, according to the manufacturer’s protocol. The product was treated with DpnI (FastDigest, Thermo Scientific) for 2 hr and purified by column purification (MinElute PCR Purification Kit, Qiagen). The fragment was digested with EcoRI and XbaI restriction enzymes (FastDigest, Thermo Scientific) for 1 hr at 37°C and purified from a 2% agarose gel (QIAquick Gel Extraction Kit, Qiagen). In parallel, the P_{CUP1}-Sup35N-A β 42 plasmid was digested with the same restriction enzymes to remove the WT A β sequence, treated with alkaline phosphatase (FastAP, Thermo Scientific) for 1 hr at 37°C to dephosphorylate the 5’ ends, and purified from a 1% agarose gel (QIAquick Gel Extraction Kit, Qiagen).

Mutagenised A β was then ligated into the linearised plasmid in a 5:1 ratio (insert:vector) using a ligase treatment (T4, Thermo Scientific) overnight. The reaction was dialysed with a membrane filter (Merck Millipore) for 1 hr, concentrated 4x, and transformed in electrocompetent *Escherichia coli* cells (10-beta Electrocompetent, NEB). Cells were recovered in SOC medium and plated on LB with ampicillin. A total of 4.1 million transformants were estimated, ensuring that each variant of the library was represented more than 10 times; 50 ml of overnight *E. coli* culture was harvested to purify the A β plasmid library with a midi prep (Plasmid Midi Kit, Qiagen). The resulting library contained 29.9% of WT A β , 23.8% of sequences with 1 nt change, and 21.8% of sequences with 2 nt changes.

Large-scale yeast transformation

Saccharomyces cerevisiae [psi-pin-] (MATa *ade1-14 his3 leu2-3,112 lys2 trp1 ura3-52*) strain (also provided by the Chernoff lab) was used in all experiments in this study (Chandramowlishwaran et al., 2018).

Yeast cells were transformed with the A β plasmid library starting from an individual colony for each transformation tube. After an overnight pre-growth culture in YPDA medium at 30°C, cells were diluted to OD₆₀₀ = 0.3 in 175 ml YPDA and incubated at 30°C 200 rpm for ~5 hr. When cells reached the exponential phase, they were harvested, washed with milliQ, and resuspended in sorbitol mixture (100 mM LiOAc, 10 mM Tris pH 8, 1 mM EDTA, 1M sorbitol). After a 30 min incubation at room temperature (RT), 5 μ g of plasmid library and 175 μ l of ssDNA (UltraPure, Thermo Scientific) were added to the cells. PEG mixture (100 mM LiOAc, 10 mM Tris pH 8, 1 mM EDTA pH 8, 40% PEG3350) was also added and cells were incubated for 30 min at RT and heat-shocked for 15 min at

42°C in a water bath. Cells were harvested, washed, resuspended in 350 ml recovery medium (YPD, sorbitol 0.5M, 70 mg/L adenine) and incubated for 1.5 hr at 30°C 200 rpm. After recovery, cells were resuspended in 350 ml -URA plasmid selection medium and allowed to grow for 50 hr. Transformation efficiency was calculated for each tube of transformation by plating an aliquote of cells in -URA plates. Between 1 and 2.5 million transformants per tube were obtained. Two days after transformation, the culture was diluted to $OD_{600} = 0.02$ in 1 l -URA medium and grown until the exponential phase. At this stage, cells were harvested and stored at -80°C in 25% glycerol.

Selection experiments

Three independent replicate selection experiments were performed. Tubes were thawed from the -80°C glycerol stocks and mixed proportionally to the number of transformants in a 1 l total -URA medium at $OD_{600} = 0.05$. A minimum of 3.7 million yeast transformants were used for each replicate to ensure the coverage of the full library and reaching therefore a 10x coverage of each variant.

Once the culture reached the exponential phase, cells were resuspended in 1 l protein inducing medium (-URA, 20% glucose, 100 μM Cu_2SO_4) at $OD_{600} = 0.05$. As a result, each variant was represented at least 100 times at this stage. After 24 hr the input pellets were collected by centrifuging 220 ml of cells and stored at -20°C for later DNA extraction (input pellets). In parallel, 18.5 million cells of the same culture underwent selection, with a starting coverage of at least 50 copies of each variant in the library. For selection, cells were plated on -ADE-URA selective medium in 145 cm^2 plates (Nunc, Thermo Scientific) and let grow for 7 days at 30°C. Colonies were then scraped off the plates and recovered with PBS 1x to be centrifuged and stored at -20°C for later DNA extraction (output pellets).

For individual testing of specific variants, cells were plated on -URA (control) and -ADE-URA (selection) plates in three independent replicates. Individual growth was calculated as the percentage of colonies growing -ADE-URA relative to colonies growing in -URA.

DNA extraction

The input and output pellets (three replicates, six tubes in total) were thawed and resuspended in 2 ml extraction buffer (2% Triton-X, 1% SDS, 100 mM NaCl, 10 mM Tris pH 8, 1 mM EDTA pH 8), and underwent two cycles of freezing and thawing in an ethanol-dry ice bath (10 min) and at 62°C (10 min). Samples were then vortexed together with 1.5 ml of phenol:chloroform:isoamyl 25:24:1 and 1.5 g of glass beads (Sigma). The aqueous phase was recovered by centrifugation and mixed again with 1.5 ml phenol:chloroform:isoamyl 25:24:1. DNA precipitation was performed by adding 1:10 V of 3M NaOAc and 2.2 V of 100% cold ethanol to the aqueous phase and incubating the samples at -20°C for 1 hr. After a centrifugation step, pellets were dried overnight at RT.

Pellets were resuspended in 1 ml resuspension buffer (10 mM Tris pH 8, 1 mM EDTA pH 8) and treated with 7.5 μl RNase A (Thermo Scientific) for 30 min at 37°C. The DNA was finally purified using 75 μl of silica beads (QIAEX II Gel Extraction Kit, Qiagen), washed and eluted in 375 μl elution buffer.

DNA concentration in each sample was measured by quantitative PCR, using primers (MS_03 and MS_04, **Supplementary file 3**) that anneal to the origin of replication site of the plasmid at 58°C.

Sequencing library preparation

The library was prepared for high-throughput sequencing in two rounds of PCR (Q5 High-Fidelity DNA Polymerase, NEB). In PCR1, the A β region was amplified for 15 cycles at 68°C with frame-shifted primers (MS_05 to MS_18, **Supplementary file 3**) with homology to Illumina sequencing primers; 300 million of molecules were used for each input or output sample. The products of PCR1 were purified with an ExoSAP-IT treatment (Affymetrix) and a column purification step (QIAquick PCR Purification Kit) and then used as the template of PCR2. This PCR was run for 10 cycles at 62°C with Illumina indexed primers (MS_19 to MS_25, **Supplementary file 2**) specific for each sample (three inputs and three outputs). The six samples were then pooled together equimolarly. The final library sample was purified from a 2% agarose gel with silica beads (QIAEX II Gel Extraction Kit, Qiagen); 125 bp paired-end sequencing was run on an Illumina HiSeq2500 sequencer at the CRG Genomics Core Facility.

Data processing

FastQ files from paired-end sequencing of the A β library before ('input') and after selection ('output') were processed using a custom pipeline (<https://github.com/lehner-lab/DiMSum>). DiMSum (Faure et al., 2020) is an R package that uses different sequencing processing tools such as FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) (for quality assessment), Cutadapt (Martin, 2011) (for constant region trimming), and USEARCH (Edgar, 2010) (for paired-end read alignment). Sequences were trimmed at 5' and 3', allowing an error rate of 0.2 (i.e., read pairs were discarded if the constant regions contained more than 20% mismatches relative to the reference sequence). Sequences differing in length from the expected 126 bp or with a Phred base quality score below 30 were discarded. As a result of this processing, around 150 million total reads passed the filtering criteria.

At this stage, unique variants were aggregated and counted using Starcode (<https://github.com/gui11aume/starcode>). Variants containing indels and nonsynonymous variants with synonymous substitutions in other codons were excluded. The result is a table of variant counts which can be used for further analysis.

For downstream analysis, variants with less than 50 input reads in any of the replicates were excluded and only variants with a maximum of two aa mutations were used.

Nucleation scores and error estimates

On the basis of variant counts, the DiMSum pipeline (Faure et al., 2020; <https://github.com/lehner-lab/DiMSum>) was used to calculate nucleation scores (NS) and their error estimates. For each variant in each replicate NS was calculated as:

$$\text{Nucleation score} = ES_i - ES_{wt}$$

where $ES_i = \log(F_i \text{ OUTPUT}) - \log(F_i \text{ INPUT})$ for a specific variant and $ES_{wt} = \log(F_{wt} \text{ OUTPUT}) - \log(F_{wt} \text{ INPUT})$ for A β WT.

DiMSum models measurement error of NS by assuming that variants with similar counts in input and output samples have similar errors. Based on errors expected from Poisson-distributed count data, replicate-specific additive and multiplicative (one each for input and output samples) modifier terms are fit to best describe the observed variance of NS across all variants simultaneously.

After error calculation, NS were merged by using the error-weighted mean of each variant across replicates and centered using the error-weighted means frequency of synonymous substitutions arising from single nt changes. Merged NS and NS for each independent replicate, as well as their associated error estimates, are available in **Supplementary file 4**.

Nonsense (stop) mutants were excluded for the analysis except when indicated (**Figure 2A and C** and **Figure 2—figure supplement 1A**).

K-medoids clustering

We used K-medoids, or the partitioning around medoids algorithm, to cluster the matrix of single aa variant NS estimates by residue position with the number of clusters estimated by optimum average silhouette width, for values of K in [1,10]. The silhouette width is a measure of how similar each object (in this case residue position) is to its own cluster. In order to take into account uncertainty in NS estimates in the determination of the optimum number of clusters, we repeated this analysis after random resampling from the NS (error) distributions of each single aa variant ($n = 100$). Based on this clustering, we defined the N-terminus as aa 2-26 and the C-terminus as aa 27-41 (**Figure 2—figure supplement 1B**). Seven positions where as many (or more) single mutations increase as decrease nucleation were defined as 'gatekeepers' (D1, E3, D7, E11, L17, E22, A42) and excluded from the N- and C-terminus classes. Only those positions where most mutations are significantly different from WT (FDR = 0.1) were considered for the definition of gatekeepers.

Aa properties, aggregation, and variant effect predictors

Nucleation scores were correlated with aa properties and scores from aggregation, solubility, and variant effect prediction algorithms. Pearson correlations were weighted based on the error terms associated with the NS of each variant using the R package 'weights.' The aa property features were retrieved from a curated collection of numerical indices representing various aa physicochemical and

biochemical properties (<http://www.genome.jp/aaindex/>). We also used a principal component of these aa properties from a previous work (PC1; [Bolognesi et al., 2019](#)) that relates strongly to changes in hydrophobicity. For each variant (single and double aa mutants), the values of a specific aa property represent the difference between the mutant and the WT scores.

For the aggregation and solubility algorithms (Tango [[Fernandez-Escamilla et al., 2004](#)], Zyggregator [[Tartaglia and Vendruscolo, 2008](#)], CamSol [[Sormanni et al., 2015](#)], and Waltz [[Oliveberg, 2010](#)]), individual residue-level scores were summed to obtain a score per aa sequence. We then calculated the log value for each variant relative to the WT score (single and double aa mutants for Tango, Zyggregator, CamSol and single aa mutants for Waltz). For the variant effect predictors (Polyphen [[Adzhubei et al., 2013](#)] and CADD [[Rentzsch et al., 2019](#)]), we also calculated the log value for each variant (only single aa mutants) but in this case values were scaled relative to the lowest predicted score.

fAD, gnomAD, and Clinvar variants

The table of fAD mutations used in this study was taken from <https://www.alzforum.org/mutations/app>. Allele frequencies of APP variants were retrieved from gnomAD ([Karczewski, 2020](#)) (<https://gnomad.broadinstitute.org/>) and the clinical significance of variants was taken from their Clinvar ([Landrum et al., 2014](#)) classification (<https://www.ncbi.nlm.nih.gov/clinvar>).

ROC curves were built and AUC values were obtained using the 'pROC' R package.

PDB structures

The coordinates of the following PDB structures were used for **Figure 5, Figure 5—figure supplements 1 and 2**: 5OQV, 2NAO, 5KK3, 2BEG, 2MXU, 5AEF, 6SHS, 2LMN, 2LMP, 2LNQ, 2MVX, 2M4J, 2MPZ ([Gremer et al., 2017](#); [Colvin et al., 2016](#); [Wälti et al., 2016](#); [Lührs et al., 2005](#); [Xiao et al., 2015](#); [Schmidt et al., 2015](#); [Kollmer et al., 2019](#); [Lu et al., 2013](#); [Qiang et al., 2012](#); [Sgourakis et al., 2015](#); [Schütz et al., 2015](#)).

Acknowledgements

Work in the lab of BB is supported by the Spanish Ministry of Science, Innovation and Universities through the project RTI2018-101491-A-I00 (MICIU/FEDER), by the CERCA Program/Generalitat de Catalunya and by funding from the Agència de Gestió d'Ajuts Universitaris i de Recerca (2019FI_B 01311) to MS. Work in the lab of BL is supported by a European Research Council (ERC) Consolidator grant (616434), the Spanish Ministry of Science, Innovation and Universities (BFU2017-89488-P and SEV-2012-0208), the Bettencourt Schueller Foundation, Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR, 2017 SGR 1322.), and the CERCA Program/Generalitat de Catalunya. We acknowledge the support of the Spanish Ministry of Science and Innovation to the EMBL partnership and the Centro de Excelencia Severo Ochoa. We thank the Chernoff lab for kindly providing strains and plasmids and the CRG Genomics core facility for their assistance with sequencing.

Additional information

Funding

Funder	Grant reference number	Author
Ministerio de Ciencia e Innovación	RTI2018-101491-A-I00	Benedetta Bolognesi
Ministerio de Ciencia e Innovación	BFU2017-89488-P	Ben Lehner
H2020 European Research Council	616434	Ben Lehner
Agència de Gestió d'Ajuts Universitaris i de Recerca	SGR 1322	Ben Lehner
Agència de Gestió d'Ajuts Universitaris i de Recerca	2019FI_B 01311	Mireia Seuma

Fondation Bettencourt Schuel-
ler Prize

Ben Lehner

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Mireia Seuma, Conceptualization, Formal analysis, Validation, Investigation, Visualization, Methodology, Writing - original draft, Writing - review and editing; Andre J Faure, Software, Investigation, Visualization, Methodology; Marta Badia, Investigation; Ben Lehner, Conceptualization, Supervision, Funding acquisition, Writing - original draft, Writing - review and editing; Benedetta Bolognesi, Conceptualization, Supervision, Methodology, Funding acquisition, Writing - original draft, Writing - review and editing

Author ORCIDs

Mireia Seuma  <https://orcid.org/0000-0002-6140-4530>

Andre J Faure  <https://orcid.org/0000-0002-4471-5994>

Marta Badia  <https://orcid.org/0000-0002-9712-9163>

Ben Lehner  <https://orcid.org/0000-0002-8817-1124>

Benedetta Bolognesi  <https://orcid.org/0000-0002-6632-947X>

Decision letter and Author response

Decision letter <https://doi.org/10.7554/eLife.63364.sa1>

Author response <https://doi.org/10.7554/eLife.63364.sa2>

Additional files

Supplementary files

- Supplementary file 1. Table listing the impact on aggregation rates for 16 A β 42 variants for which these measurements could be retrieved from the literature. For the same variants, the table also reports nucleation scores, as quantified in this study, and the qualitative agreement or disagreement with the previously published data.
- Supplementary file 2. Table listing the mutations in A β 42 that significantly increase nucleation score and that are therefore proposed as novel familial Alzheimer's disease (fAD) candidates. For each mutation, the corresponding nucleation score (NS) is reported.
- Supplementary file 3. List of oligonucleotides used in this study.
- Supplementary file 4. Processed data required to make all analyses and figures in this paper. Read counts, nucleation scores, and associated error terms are reported for each A β 42 variant in each replicate. See sheet one for a deeper explanation of headers.
- Transparent reporting form

Data availability

Raw sequencing data and the processed data table (Supplementary file 4) have been deposited in NCBI's Gene Expression Omnibus (GEO) as record GSE151147. All code used for data analysis is available at <https://github.com/BEBlab/abeta> (copy archived at <https://archive.softwareheritage.org/swh:1:rev:86e1e1be4ee6eb97c1c00b0bd53f98f4e4ea807f/>).

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Seuma M, Faure A, Badia M, Lehner B, Bolognesi B	2020	The genetic landscape for amyloid beta fibril nucleation accurately discriminates familial Alzheimer's disease mutations	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE151147	NCBI Gene Expression Omnibus, GSE151147

References

- Adzhubei I**, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics* **76**:76. DOI: <https://doi.org/10.1002/0471142905.hg0720s76>
- Aprile FA**, Sormanni P, Perni M, Arosio P, Linse S, Knowles TPJ, Dobson CM, Vendruscolo M. 2017. Selective targeting of primary and secondary nucleation pathways in A β 42 aggregation using a rational antibody scanning method. *Science Advances* **3**:e1700488. DOI: <https://doi.org/10.1126/sciadv.1700488>
- Ballard C**, Gauthier S, Corbett A, Brayne C, Aarsland D, Jones E. 2011. Alzheimer's disease. *The Lancet* **377**: 1019–1031. DOI: [https://doi.org/10.1016/S0140-6736\(10\)61349-9](https://doi.org/10.1016/S0140-6736(10)61349-9)
- Benilova I**, Gallardo R, Ungureanu A-A, Castillo Cano V, Snellinx A, Ramakers M, Bartic C, Rousseau F, Schymkowitz J, De Strooper B. 2014. The Alzheimer disease protective mutation A2T modulates kinetic and thermodynamic properties of Amyloid- β (A β) Aggregation. *Journal of Biological Chemistry* **289**:30977–30989. DOI: <https://doi.org/10.1074/jbc.M114.599027>
- Bolognesi B**, Kumita JR, Barros TP, Esbjorner EK, Luheshi LM, Crowther DC, Wilson MR, Dobson CM, Favrin G, Yerbury JJ. 2010. ANS binding reveals common features of cytotoxic amyloid species. *ACS Chemical Biology* **5**: 735–740. DOI: <https://doi.org/10.1021/cb1001203>, PMID: 20550130
- Bolognesi B**, Faure AJ, Seuma M, Schmiedel JM, Tartaglia GG, Lehner B. 2019. The mutational landscape of a prion-like domain. *Nature Communications* **10**:4162. DOI: <https://doi.org/10.1038/s41467-019-12101-z>, PMID: 31519910
- Campion D**, Dumanchin C, Hannequin D, Dubois B, Belliard S, Puel M, Thomas-Anterion C, Michon A, Martin C, Charbonnier F, Raux G, Camuzat A, Penet C, Mesnage V, Martinez M, Clerget-Darpoux F, Brice A, Frebourg T. 1999. Early-onset autosomal dominant Alzheimer disease: prevalence, genetic heterogeneity, and mutation spectrum. *The American Journal of Human Genetics* **65**:664–670. DOI: <https://doi.org/10.1086/302553>, PMID: 10441572
- Chandramowlishwaran P**, Sun M, Casey KL, Romanyuk AV, Grizel AV, Sopova JV, Rubel AA, Nussbaum-Krammer C, Vorberg IM, Chernoff YO. 2018. Mammalian amyloidogenic proteins promote prion nucleation in yeast. *Journal of Biological Chemistry* **293**:3436–3450. DOI: <https://doi.org/10.1074/jbc.M117.809004>
- Cleary JP**, Walsh DM, Hofmeister JJ, Shankar GM, Kuskowski MA, Selkoe DJ, Ashe KH. 2005. Natural oligomers of the amyloid-beta protein specifically disrupt cognitive function. *Nature Neuroscience* **8**:79–84. DOI: <https://doi.org/10.1038/nn1372>, PMID: 15608634
- Cohen SIA**, Cukalevski R, Michaels TCT, Šarić A, Törnquist M, Vendruscolo M, Dobson CM, Buell AK, Knowles TPJ, Linse S. 2018. Distinct thermodynamic signatures of oligomer generation in the aggregation of the amyloid- β peptide. *Nature Chemistry* **10**:523–531. DOI: <https://doi.org/10.1038/s41557-018-0023-x>, PMID: 29581486
- Colvin MT**, Silvers R, Ni QZ, Can TV, Sergeyev I, Rosay M, Donovan KJ, Michael B, Wall J, Linse S, Griffin RG. 2016. Atomic resolution structure of monomorphic a β 42 amyloid fibrils. *Journal of the American Chemical Society* **138**:9663–9674. DOI: <https://doi.org/10.1021/jacs.6b05129>, PMID: 27355699
- Conrad DF**, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV, Zilversmit M, Cartwright R, Rouleau GA, Daly M, Stone EA, Hurles ME, Awadalla P, 1000 Genomes Project. 2011. Variation in Genome-Wide mutation rates within and between human families. *Nature Genetics* **43**:712–714. DOI: <https://doi.org/10.1038/ng.862>, PMID: 21666693
- Di Fede G**, Catania M, Morbin M, Rossi G, Suardi S, Mazzoleni G, Merlin M, Giovagnoli AR, Prioni S, Erbetta A, Falcone C, Gobbi M, Colombo L, Bastone A, Beeg M, Manzoni C, Francescucci B, Spagnoli A, Cantù L, Del Favero E, et al. 2009. A recessive mutation in the APP gene with dominant-negative effect on amyloidogenesis. *Science* **323**:1473–1477. DOI: <https://doi.org/10.1126/science.1168979>, PMID: 19286555
- Edgar RC**. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**:2460–2461. DOI: <https://doi.org/10.1093/bioinformatics/btq461>, PMID: 20709691
- Faure AJ**, Schmiedel JM, Baeza-Centurion P, Lehner B. 2020. DiMSum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. *Genome Biology* **21**:207. DOI: <https://doi.org/10.1186/s13059-020-02091-3>, PMID: 32799905
- Fernandez-Escamilla AM**, Rousseau F, Schymkowitz J, Serrano L. 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnology* **22**:1302–1306. DOI: <https://doi.org/10.1038/nbt1012>, PMID: 15361882
- Gelman H**, Dines JN, Berg J, Berger AH, Brnich S, Hisama FM, James RG, Rubin AF, Shendure J, Shirts B, Fowler DM, Starita LM, Brotman Baty Institute Mutational Scanning Working Group. 2019. Recommendations for the collection and use of multiplexed functional data for clinical variant interpretation. *Genome Medicine* **11**:85. DOI: <https://doi.org/10.1186/s13073-019-0698-7>, PMID: 31862013
- Gray VE**, Sitko K, Kameni FZN, Williamson M, Stephany JJ, Hasle N, Fowler DM. 2019. Elucidating the molecular determinants of a β aggregation with deep mutational scanning. *G3: Genes, Genomes, Genetics* **9**:3683–3689. DOI: <https://doi.org/10.1534/g3.119.400535>, PMID: 31558564
- Gremer L**, Schölzel D, Schenk C, Reinartz E, Labahn J, Ravelli RBG, Tusche M. 2017. Fibril structure of Amyloid- β (1–42) by cryoelectron microscopy. *Science* **9**:eaao2825. DOI: <https://doi.org/10.1126/science.aao2825>
- Janssen JC**, Beck JA, Campbell TA, Dickinson A, Fox NC, Harvey RJ, Houlden H, Rossor MN, Collinge J. 2003. Early onset familial Alzheimer's disease: Mutation frequency in 31 families. *Neurology* **60**:235–239. DOI: <https://doi.org/10.1212/01.WNL.0000042088.22694.E3>, PMID: 12552037
- Karczewski KJ**. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *bioRxiv*. DOI: <https://doi.org/10.1101/531210>

- Knowles TP**, Waudby CA, Devlin GL, Cohen SI, Aguzzi A, Vendruscolo M, Terentjev EM, Welland ME, Dobson CM. 2009. An analytical solution to the kinetics of breakable filament assembly. *Science* **326**:1533–1537. DOI: <https://doi.org/10.1126/science.1178250>, PMID: 20007899
- Knowles TP**, Vendruscolo M, Dobson CM. 2014. The amyloid state and its association with protein misfolding diseases. *Nature Reviews Molecular Cell Biology* **15**:384–396. DOI: <https://doi.org/10.1038/nrm3810>, PMID: 24854788
- Kollmer M**, Close W, Funk L, Rasmussen J, Bsoul A, Schierhorn A, Schmidt M, Sigurdson CJ, Jucker M, Fändrich M. 2019. Cryo-EM structure and polymorphism of $\alpha\beta$ amyloid fibrils purified from Alzheimer's brain tissue. *Nature Communications* **10**:4760. DOI: <https://doi.org/10.1038/s41467-019-12683-8>, PMID: 31664019
- Kyte J**, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **157**:105–132. DOI: [https://doi.org/10.1016/0022-2836\(82\)90515-0](https://doi.org/10.1016/0022-2836(82)90515-0), PMID: 7108955
- Landrum MJ**, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research* **42**:D980–D985. DOI: <https://doi.org/10.1093/nar/gkt1113>, PMID: 24234437
- Löhr T**, Kohlhoff K, Heller GT, Camilloni C, Vendruscolo M. 2021. A kinetic ensemble of the Alzheimer's $A\beta$ peptide. *Nature Computational Science* **1**:71–78. DOI: <https://doi.org/10.1038/s43588-020-00003-w>
- Lu JX**, Qiang W, Yau WM, Schwieters CD, Meredith SC, Tycko R. 2013. Molecular structure of β -Amyloid fibrils in Alzheimer's Disease Brain Tissue. *Cell* **154**:1257–1268. DOI: <https://doi.org/10.1016/j.cell.2013.08.035>, PMID: 24034249
- Lührs T**, Ritter C, Adrian M, Riek-Loher D, Bohrmann B, Döbeli H, Schubert D, Riek R. 2005. 3d structure of Alzheimer's amyloid-beta(1-42) fibrils. *PNAS* **102**:17342–17347. DOI: <https://doi.org/10.1073/pnas.0506723102>, PMID: 16293696
- Martin M**. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**:10–12. DOI: <https://doi.org/10.14806/ej.17.1.200>
- Meier BH**, Riek R, Böckmann A. 2017. Emerging structural understanding of amyloid fibrils by Solid-State NMR. *Trends in Biochemical Sciences* **42**:777–787. DOI: <https://doi.org/10.1016/j.tibs.2017.08.001>, PMID: 28916413
- Meisl G**, Yang X, Hellstrand E, Frohm B, Kirkegaard JB, Cohen SI, Dobson CM, Linse S, Knowles TP. 2014. Differences in nucleation behavior underlie the contrasting aggregation kinetics of the $\alpha\beta 40$ and $\alpha\beta 42$ peptides. *PNAS* **111**:9384–9389. DOI: <https://doi.org/10.1073/pnas.1401564111>, PMID: 24938782
- Michaels TCT**, Šarić A, Curk S, Bernfur K, Arosio P, Meisl G, Dear AJ, Cohen SIA, Dobson CM, Vendruscolo M, Linse S, Knowles TPJ. 2020. Dynamics of oligomer populations formed during the aggregation of Alzheimer's $A\beta 42$ peptide. *Nature Chemistry* **12**:445–451. DOI: <https://doi.org/10.1038/s41557-020-0452-1>, PMID: 32284577
- O'Brien RJ**, Wong PC. 2011. Amyloid precursor protein processing and Alzheimer's disease. *Annual Review of Neuroscience* **34**:185–204. DOI: <https://doi.org/10.1146/annurev-neuro-061010-113613>, PMID: 21456963
- Oliverberg M**. 2010. Waltz, an exciting new move in amyloid prediction. *Nature Methods* **7**:187–188. DOI: <https://doi.org/10.1038/nmeth0310-187>, PMID: 20195250
- Paravastu AK**, Leapman RD, Yau WM, Tycko R. 2008. Molecular structural basis for polymorphism in Alzheimer's beta-amyloid fibrils. *PNAS* **105**:18349–18354. DOI: <https://doi.org/10.1073/pnas.0806270105>, PMID: 19015532
- Pedersen JS**, Christensen G, Otzen DE. 2004. Modulation of S6 fibrillation by unfolding rates and gatekeeper residues. *Journal of Molecular Biology* **341**:575–588. DOI: <https://doi.org/10.1016/j.jmb.2004.06.020>, PMID: 15276845
- Pimenova AA**, Goate AM. 2020. Novel presenilin 1 and 2 double knock-out cell line for in vitro validation of PSEN1 and PSEN2 mutations. *Neurobiology of Disease* **138**:104785. DOI: <https://doi.org/10.1016/j.nbd.2020.104785>, PMID: 32032730
- Qiang W**, Yau WM, Luo Y, Mattson MP, Tycko R. 2012. Antiparallel β -sheet architecture in Iowa-mutant β -amyloid fibrils. *PNAS* **109**:4443–4448. DOI: <https://doi.org/10.1073/pnas.1111305109>, PMID: 22403062
- Rentsch P**, Witten D, Cooper GM, Shendure J, Kircher M. 2019. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research* **47**:D886–D894. DOI: <https://doi.org/10.1093/nar/gky1016>, PMID: 30371827
- Rousseau F**, Serrano L, Schymkowitz JW. 2006. How evolutionary pressure against protein aggregation shaped chaperone specificity. *Journal of Molecular Biology* **355**:1037–1047. DOI: <https://doi.org/10.1016/j.jmb.2005.11.035>, PMID: 16359707
- Ryman DC**, Acosta-Baena N, Aisen PS, Bird T, Danek A, Fox NC, Goate A, Frommelt P, Ghetti B, Langbaum JBS, Lopera F, Martins R, Masters CL, Mayeux RP, McDade E, Moreno S, Reiman EM, Ringman JM, Salloway S, Schofield PR, et al. 2014. Symptom onset in autosomal dominant Alzheimer disease: a systematic review and meta-analysis. *Neurology* **83**:253–260. DOI: <https://doi.org/10.1212/WNL.0000000000000596>
- Sandberg A**, Luheshi LM, Söllvander S, Pereira de Barros T, Macao B, Knowles TP, Biverstål H, Lendel C, Ekholm-Petterson F, Dubnovitsky A, Lannfelt L, Dobson CM, Hård T. 2010. Stabilization of neurotoxic Alzheimer amyloid-beta oligomers by protein engineering. *PNAS* **107**:15595–15600. DOI: <https://doi.org/10.1073/pnas.1001740107>, PMID: 20713699
- Sasaguri H**, Nilsson P, Hashimoto S, Nagata K, Saito T, De Strooper B, Hardy J, Vassar R, Winblad B, Saido TC. 2017. APP mouse models for Alzheimer's disease preclinical studies. *The EMBO Journal* **36**:2473–2487. DOI: <https://doi.org/10.15252/embj.201797397>, PMID: 28768718
- Schmidt M**, Rohou A, Lasker K, Yadav JK, Schiene-Fischer C, Fändrich M, Grigorieff N. 2015. Peptide dimer structure in an $A\beta(1-42)$ fibril visualized with cryo-EM. *PNAS* **112**:11858–11863. DOI: <https://doi.org/10.1073/pnas.1503455112>, PMID: 26351699

- Schütz AK**, Vagt T, Huber M, Ovchinnikova OY, Cadalbert R, Wall J, Güntert P, Böckmann A, Glockshuber R, Meier BH. 2015. Atomic-resolution three-dimensional structure of amyloid β fibrils bearing the Osaka mutation. *Angewandte Chemie International Edition* **54**:331–335. DOI: <https://doi.org/10.1002/anie.201408598>, PMID: 25395337
- Sgourakis NG**, Yau WM, Qiang W. 2015. Modeling an in-register, parallel "iowa" $\alpha\beta$ fibril structure using solid-state NMR data from labeled samples with rosetta. *Structure* **23**:216–227. DOI: <https://doi.org/10.1016/j.str.2014.10.022>, PMID: 25543257
- Sormani P**, Aprile FA, Vendruscolo M. 2015. The CamSol method of rational design of protein mutants with enhanced solubility. *Journal of Molecular Biology* **427**:478–490. DOI: <https://doi.org/10.1016/j.jmb.2014.09.026>, PMID: 25451785
- Starita LM**, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, Seelig G, Shendure J, Fowler DM. 2017. Variant interpretation: functional assays to the rescue. *The American Journal of Human Genetics* **101**:315–325. DOI: <https://doi.org/10.1016/j.ajhg.2017.07.014>, PMID: 28886340
- Szaruga M**, Munteanu B, Lismont S, Veugelen S, Horré K, Mercken M, Saido TC, Ryan NS, De Vos T, Savvides SN, Gallardo R, Schymkowitz J, Rousseau F, Fox NC, Hopf C, De Strooper B, Chávez-Gutiérrez L. 2017. Alzheimer's-Causing Mutations Shift $A\beta$ Length by Destabilizing γ -Secretase- $A\beta_n$ Interactions. *Cell* **170**:443–456. DOI: <https://doi.org/10.1016/j.cell.2017.07.004>, PMID: 28753424
- Tartaglia GG**, Vendruscolo M. 2008. The zyggregator method for predicting protein aggregation propensities. *Chemical Society Reviews* **37**:1395–1401. DOI: <https://doi.org/10.1039/b706784b>, PMID: 18568165
- Tomiyama T**, Nagata T, Shimada H, Teraoka R, Fukushima A, Kanemitsu H, Takuma H, Kuwano R, Imagawa M, Ataka S, Wada Y, Yoshioka E, Nishizaki T, Watanabe Y, Mori H. 2008. A new amyloid beta variant favoring oligomerization in Alzheimer's-type dementia. *Annals of Neurology* **63**:377–387. DOI: <https://doi.org/10.1002/ana.21321>, PMID: 18300294
- Törnquist M**, Michaels TCT, Sanagavarapu K, Yang X, Meisl G, Cohen SIA, Knowles TPJ, Linse S. 2018. Secondary nucleation in amyloid formation. *Chemical Communications* **54**:8667–8684. DOI: <https://doi.org/10.1039/C8CC02204F>
- Van Cauwenberghe C**, Van Broeckhoven C, Sleegers K. 2016. The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Genetics in Medicine* **18**:421–430. DOI: <https://doi.org/10.1038/gim.2015.117>, PMID: 26312828
- Veugelen S**, Saito T, Saido TC, Chávez-Gutiérrez L, De Strooper B. 2016. Familial Alzheimer's disease mutations in presenilin generate amyloidogenic $A\beta$ peptide seeds. *Neuron* **90**:410–416. DOI: <https://doi.org/10.1016/j.neuron.2016.03.010>, PMID: 27100199
- Wälti MA**, Ravotti F, Arai H, Glabe CG, Wall JS, Böckmann A, Güntert P, Meier BH, Riek R. 2016. Atomic-resolution structure of a disease-relevant $A\beta(1-42)$ amyloid fibril. *PNAS* **113**:E4976–E4984. DOI: <https://doi.org/10.1073/pnas.1600749113>, PMID: 27469165
- Weggen S**, Behr D. 2012. Molecular consequences of amyloid precursor protein and presenilin mutations causing autosomal-dominant Alzheimer's disease. *Alzheimer's Research & Therapy* **4**:9. DOI: <https://doi.org/10.1186/alzrt107>, PMID: 22494386
- World Health Organization**. 2012. *WHO Dementia: A Public Health Priority*: WHO. https://www.who.int/mental_health/publications/dementia_report_2012/en/.
- Xiao Y**, Ma B, McElheny D, Parthasarathy S, Long F, Hoshi M, Nussinov R, Ishii Y. 2015. $A\beta(1-42)$ fibril structure illuminates self-recognition and replication of amyloid in Alzheimer's disease. *Nature Structural & Molecular Biology* **22**:499–505. DOI: <https://doi.org/10.1038/nsmb.2991>, PMID: 25938662
- Yang X**, Meisl G, Frohm B, Thulin E, Knowles TPJ, Linse S. 2018. On the role of sidechain size and charge in the aggregation of $A\beta_{42}$ with familial mutations. *PNAS* **115**:E5849–E5858. DOI: <https://doi.org/10.1073/pnas.1803539115>, PMID: 29895690

Chapter II. An atlas of amyloid aggregation: the impact of substitutions, insertions, deletions and truncations on amyloid beta fibril nucleation

An atlas of amyloid aggregation: the impact of substitutions, insertions, deletions and truncations on amyloid beta fibril nucleation

Authors

Mireia Seuma¹, Ben Lehner^{2,3,4*}, Benedetta Bolognesi^{1*}

Affiliations

¹Institute for Bioengineering of Catalonia (IBEC), The Barcelona Institute of Science and Technology, Baldiri Reixac 10-12, 08028, Barcelona Spain

²Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Doctor Aiguader 88, 08003, Barcelona, Spain

³Universitat Pompeu Fabra (UPF), Barcelona, Spain.

⁴ICREA, Pg. Lluís Companys 23, Barcelona 08010, Spain

*Corresponding authors

Email: bbolognesi@ibecbarcelona.eu, ben.lehner@crg.eu

Abstract

Multiplexed assays of variant effects (MAVEs) guide clinical variant interpretation and reveal disease mechanisms. To date, MAVEs have focussed on a single mutation type - amino acid (AA) substitutions - despite the diversity of coding variants that cause disease. Here we use Deep Indel Mutagenesis (DIM) to generate the first comprehensive atlas of diverse variant effects for a disease protein, the amyloid beta (A β) peptide that aggregates in Alzheimer's disease (AD) and is mutated in familial AD (fAD). The atlas identifies known fAD mutations and reveals many variants beyond substitutions accelerate A β aggregation and are likely to be pathogenic. Truncations, substitutions, insertions, single- and internal multi-AA deletions differ in their propensity to enhance or impair aggregation, but likely pathogenic variants from all classes are highly enriched in the polar N-terminus of A β . This first comparative atlas highlights the importance of including diverse mutation types in MAVEs and provides important mechanistic insights into amyloid nucleation.

Introduction

Amyloid fibrils are the hallmarks of more than 50 human diseases, including Alzheimer's disease (AD), Parkinson's disease, frontotemporal dementia, amyotrophic lateral sclerosis and systemic amyloidoses¹. Mutations in the proteins that aggregate in the common forms of neurodegeneration also cause rare familial neurodegenerative diseases. For example, amyloid plaques of the amyloid beta (A β) peptide are a pathological hallmark of AD and specific dominant mutations in A β also cause familial Alzheimer's disease (fAD)^{2,3}.

The structures of many amyloid fibrils have now been determined⁴ including those of A β fibrils extracted post-mortem from AD patient brains⁵. In these fibrils, the peptide adopts an S-shaped fold from residue 19 to 42, with the aliphatic C-terminus 29-42 packed as the inner core of the amyloid fibril. A more exposed N-terminal arm connects this to the first part of the peptide which remains unstructured in mature fibrils (residues 1-9 in sporadic and 1-11 in fAD). Despite these high-resolution structures, the mechanism by which fibrils form in the first place - the nucleation reaction - is still poorly understood, even though this is the fundamental process that needs to be understood and targeted to prevent amyloid diseases^{6,7}. Moreover, we have only a superficial understanding of how specific mutations accelerate the process of amyloid nucleation to cause familial diseases. In A β , several of the known fAD mutations are at residues outside the structured amyloid core⁸. Amongst other consequences, this makes the clinical interpretation of genetic variants challenging, with the vast majority of mutations identified in aggregating proteins classified as variants of uncertain significance (VUSs)⁹.

Multiplexed assays of variant effects (MAVEs)¹⁰ use cell-based or *in vitro* selection assays to build comprehensive atlases of variant effects (AVEs)¹¹ to guide the clinical interpretation of VUSs¹¹. This approach, which is also called deep mutational scanning (DMS), uses massively parallel DNA synthesis, selection and deep sequencing to quantify the relative activities of variants in a functional assay¹². Applied to disease genes, DMS can also reveal disease mechanisms and it can be used to genetically-validate the relevance of cellular and *in vitro* disease models to human disease¹³. For example, we recently adopted a cell-based assay¹⁴ to allow massively parallel quantification of variant effects on protein aggregation. Measuring the effects of single nucleotide changes in A β revealed that the assay both accurately quantifies the rate of amyloid fibril nucleation and that it identifies all of the dominant substitutions known to cause fAD¹⁵.

To date MAVE experiments¹¹ have focussed on a single type of mutation - amino acid (AA) substitutions - and have largely ignored additional forms of genetic variation. Insertions and deletions (indels), in particular, are an abundant and important class of genetic variation in protein coding regions known to cause many human genetic diseases^{16,17}, with small indels (<21 bp) causing approximately 24% of Mendelian diseases^{18,19}. Indels are a fundamentally different perturbation to a protein sequence to substitutions: whereas substitutions only alter AA side chains, indels are backbone mutations that change the length of the polypeptide chain and so may be expected to have more severe effects²⁰. However, despite their importance, there has been very little systematic quantification of the effects of indels in proteins²¹⁻²³, particularly in disease genes, and many computational methods for predicting variant effects simply ignore

them²⁴. To our knowledge, a systematic comparison of the effects of AA substitutions, insertions and deletions is lacking for any human disease gene.

Here we address this fundamental shortcoming in human genetics by providing the first comprehensive comparison of the effects of substitutions, insertions and deletions in a human disease gene.

The resulting AVE quantifies the effects of diverse sequence changes on the aggregation of A β and is the first dataset that can be used to guide the clinical interpretation of different types of mutation in a human disease gene. It reveals that many mutations beyond substitutions accelerate the aggregation of A β and so are likely to be pathological. The atlas identifies the two deletions known to cause fAD, but reveals that they are only two of the many insertions and deletions that are likely to be pathological. The atlas also provides fundamental mechanistic insight into the process of amyloid nucleation, illustrating the power of deep indel mutagenesis (DIM) to illuminate sequence-to-activity relationships.

Results

Deep Indel Mutagenesis of amyloid beta

To quantify and contrast the effects of diverse genetic variants on the aggregation of the 42 AA form of A β (A β 42), which is the most abundant component of amyloid plaques in AD, we performed Deep Indel Mutagenesis (DIM) by synthesizing a library containing all possible single AA substitutions (n=798), all possible single AA insertions (n=780), all single AA deletions (n=37), all internal multi-AA deletions ranging in length from 2-39 AA (n=731), and all progressive truncations from the N-terminus, C-terminus or both, removing 2-39 AA (n=817, Fig. 1a).

We quantified the effects of these different classes of variants in a cell-based selection assay where the aggregation of A β nucleates the aggregation of an endogenous protein, a process required for growth in selective conditions (Fig. 1a)^{14,15}. After selection, the enrichment of each variant in the library was quantified by deep sequencing^{15,25}. The resulting enrichment scores are reproducible between replicates (Supplementary Fig. 1a) and correlate well with previous measurements (R=0.82, Supplementary Fig. 1b) as well as with the effects of variants quantified individually (R=0.89, Supplementary Fig. 1c). In addition, and as previously reported¹⁵, the enrichment scores correlate linearly with the *in vitro* measured kinetic rate constants of A β amyloid fibril nucleation (R=0.96, Supplementary Fig. 1d)^{15,26}, so we refer to them as “nucleation scores”.

Contrasting the impact of substitutions, deletions and insertions in a disease gene

The resulting dataset provides the first opportunity to comprehensively compare the effects of different types of mutation - substitutions, insertions, deletions and truncations - in a human disease gene. Focussing on single AA changes, the most frequent mutational effect is reduced aggregation, with 43% of substitutions, 44% of insertions, and 37% of deletions having lower nucleation scores (NS) than wild-type (WT) A β (false discovery rate, FDR=0.1, NS- variants, Fig. 1b,d). The effects of multi-AA deletions are stronger, with 60% of internal multi-AA deletions and 97% of multi-AA truncations from one or both ends reducing nucleation (FDR=0.1, Fig. 1b,d).

Many variants beyond substitutions accelerate A β aggregation

Variants in A β identified in families with fAD accelerate A β aggregation, consistent with a gain-of-function mechanism^{15,27}. Unlike computational methods to predict aggregation or variant effects, the experimental nucleation scores accurately classify fAD variants (Supplementary Fig. 1e). In total, there are 307 variants in our library (10%) that accelerate A β aggregation (FDR=0.1, NS+ variants): 108 substitutions, 77 insertions, 5 single AA deletions, 104 internal multi-AA deletions and 13 truncations (Fig. 1f,g and Supplementary Table 1). There are thus many variants beyond substitutions that accelerate the aggregation of A β .

All types of variant that promote aggregation are strongly enriched in the N-terminus

The primary sequence of A β consists of an N-terminal region enriched in charged and polar residues (AA 1-28, two thirds of the peptide) and a C-terminal region composed entirely of aliphatic residues and glycines (AA 29-42, one third of the peptide) (Fig. 1a).

For all classes of mutation, variants that reduce nucleation are strongly enriched in the aliphatic C-terminus of A β : 60% of substitutions, 54% of insertions, 78% of single AA deletions, 85% of the internal multi-AA deletions and all truncations that reduce nucleation occur in this hydrophobic region (FDR=0.1; Fig. 1c,f and Supplementary Fig. 1f). Indeed for all mutation types, the majority of variants in this region impair nucleation: 76% of substitutions, 76% of insertions, all single AA deletions, 94% of internal multi-AA deletions and all truncations (Fig. 1e).

In contrast, variants that accelerate nucleation are strongly enriched in the polar N-terminus. In total, 87% of variants that accelerate nucleation (267/307, FDR=0.1, NS+ variants) are located in the polar N-terminus (Fig. 1f). This contrasts to just 18% of variants that reduce nucleation (Fig. 1f). This strong enrichment is true for all mutation types: 85% of substitutions, 82% of insertions, 90% of multi-AA deletions, all single AA deletions, and all truncations that accelerate nucleation occur in the N-terminus (Fig. 1c,f,g). Very few variants in the aliphatic C-terminus increase nucleation: only 16 substitutions (6%) and 14 insertions (5%), while none of the single AA deletions do so. Similarly, no C-terminal truncations accelerate nucleation and only one internal multi-AA deletion in the C-terminus does so (Fig. 1e-g).

Mutation classes differ in their propensity to promote or prevent amyloid nucleation

The different classes of mutation do, however, vary in how likely they are to increase or decrease nucleation when they occur in the same region. The type of mutation most likely to accelerate nucleation is N-terminal truncations, with 50% increasing nucleation and no N-terminal truncation reducing nucleation (FDR=0.1, Fig. 1e). More internal multi-AA deletions in the N-terminus increase than decrease nucleation (28% vs. 19%), as do more single AA deletions (19% vs 11%). In contrast, single AA substitutions in the N-terminus are more likely to decrease (26%) than increase (18%) nucleation, as are insertions (30% decrease and 12% increase) (FDR=0.1, Fig. 1e).

In summary, the DIM data reveals that there are many mutations beyond single AA substitutions that accelerate A β aggregation and so are potentially pathogenic (Supplementary Table 1). Moreover, they show that, for all mutation types, the vast majority of variants that accelerate nucleation are located in the polar and charged N-terminal region of A β . However, the different classes of mutation have very different distributions of mutational effects in the N-terminus: whereas single AA substitutions and insertions in the N-terminus are more likely to decrease nucleation than increase it, the opposite is true for single AA deletions, internal multi-AA deletions

and N-terminal truncations: these mutation classes more often enhance nucleation than impair it, suggesting they are particularly likely to be pathogenic if they occur.

AA preferences in the N-terminus: polar, positive, small and P residues promote nucleation

Considering all positions, the effect of substituting in an AA is moderately correlated to the effect of inserting the same AA before or after the same position ($R=0.49$ and $R=0.51$, respectively, Fig. 2e). This relationship is, however, partly driven by the distinct impact of mutations at the N and the C-terminus (Supplementary Fig. 3). Thus, although the consequences of insertions and substitutions are related, they are also clearly distinct, as is also revealed by comparing their effects at each individual residue (Supplementary Fig. 5a) and their average effects across all residues (Supplementary Fig. 4a).

Considering the N and C-terminal regions separately, the average effects of inserting or substituting in AAs in any positions are strongly related ($R=0.91$ and $R=0.73$ for residues 1-28 and 29-42, respectively, Fig. 2f and Supplementary Fig. 4b,c). Both substituting and inserting polar residues (especially, N,H,T,Q) into the N-terminus frequently promotes aggregation, as does adding positively charged residues (K,R). Interestingly, substituting in or inserting G or A into the N-terminus also frequently increases nucleation as does adding a P, particularly in the second half of the N-terminus (Fig. 2a,b,d and Supplementary Figs. 2a,b and 6-8). Since P residues are unlikely to be tolerated in the core of structured fibrils, their effect in promoting nucleation may be via changes in the ensemble of soluble $A\beta^{28}$, rather than due to changes in the fibril transition state. For example, adding P might impair the formation of a transient secondary structure that - in the WT ensemble - acts to prevent nucleation.

Overall, these enrichments for polar, positively charged, small and P residues are very different to the sequence preferences used by computational methods to predict protein aggregation²⁹⁻³², and these methods indeed perform very poorly for predicting the effects of mutations in the N-terminus of $A\beta$ (Supplementary Fig. 9).

AA preferences in the N-terminus: increased hydrophobicity and negatively charged residues reduce aggregation

Also inconsistent with the expectations of predictive methods, the substitutions and insertions in the N-terminus that most often reduce aggregation are additions of hydrophobic (W,L,F,M,I,Y,V) and negatively charged (D,E) AAs (Fig. 2a,b,d and Supplementary Fig. 2a,b). Consistent with this, individually deleting negatively charged residues and L17 from the N-terminus often increases nucleation (Fig. 2c), as does substituting away from these same AAs (Supplementary Fig. 2c).

However, there are many exceptions to these general trends, highlighting the importance of generating the full mutational matrix. For example, W insertions and substitutions to W mostly promote aggregation in residues 1-12 but nearly always impair aggregation at positions 13-28

(Fig. 2a,b and Supplementary Fig. 2a,b). In addition, many substitutions to V and I in positions 13-20 strongly increase aggregation, as do many hydrophobic insertions after residue 13 between two histidines. At particular positions the impact of substitutions or insertions can also be quite distinct: for example substitutions of F19 and F20 rarely increase nucleation and only for mutations to hydrophobic AA, while many insertions between F19 and F20, increase nucleation, especially of charged and polar residues (Fig. 2a,b, and Supplementary Fig. 2a,b). At other residues the preferences are more similar: for example both substitutions to and insertions of G at residues 22 and 23 increase nucleation resulting in some of the fastest nucleating variants in the library (Fig. 2a,b and Supplementary Figs. 2a,b and 13), suggesting that increasing flexibility or reducing side chain volume in this region favors nucleation.

Thus, although simple rules can predict mutational effects to some extent, the comprehensive A β data suggests full experimental datasets and new computational methods will be required for the clinical interpretation of variants in aggregating proteins.

In summary, the role of the N-terminal two thirds of A β in promoting and preventing amyloid nucleation must be very different to that of the C-terminus that forms the hydrophobic core of A β fibrils. To our knowledge, no existing mechanistic models can satisfactorily account for mutational effects in this region^{33,34}. In mature A β fibrils derived from patients⁵ and formed *in vitro*^{5,35-39}, part or all of the N-terminus remains unstructured (Fig. 5 and Supplementary Fig. 14). Changes in aggregation caused by mutations in this region could be due to their effects on the ensemble of soluble A β . Alternatively, the N-terminus could participate directly in the nucleation reaction, establishing interactions in the nucleation transition state.

The Osaka mutation (Δ E22) is the fastest nucleating single AA deletion

To date, only one single AA deletion has been reported in families with fAD: deletion of residue E22, named the Osaka mutation after the city in which it was first identified⁴⁰. Strikingly, our data shows that the Osaka mutation is the single AA deletion that most enhances the nucleation of A β (Fig. 2c). However, an additional 4 single AA deletions promote nucleation (FDR=0.1), suggesting that they may also be pathogenic. All of these deletions are in the N-terminus of the peptide (Fig. 2c).

The Uppsala mutation (Δ 19-24) lies in a hotspot of internal multi-AA deletions that promote A β nucleation

After we generated this dataset, the first internal multi-AA deletion in A β that causes fAD was reported⁴¹. This deletion, referred to as the Uppsala mutation, removes AA 19-24. The Uppsala mutation strongly promotes nucleation in our dataset (Fig. 3a,b). However, the comprehensive DIM atlas also reveals that there are an additional 103 internal multi-AA variants that promote A β aggregation (FDR=0.1; Fig. 3a,b, Supplementary Fig. 10a and Supplementary Table 1). Strikingly, however, the Uppsala mutation is located in the center of a hotspot region where many different deletions accelerate aggregation (Fig. 3a-d). In total, 35 multi-AA deletions removing some or all of residues 17-27 increase the nucleation of A β (FDR=0.1, Fig. 3d and Supplementary Fig. 10a). This suggests that there are potentially many more pathogenic deletions that remain to

be discovered that remove residues in this central hotspot region, as well as additional pathogenic deletions throughout the N-terminus (Supplementary Table 1).

The multi-AA deletion hotspot is centered on the negatively charged residues E22 and D23 (Fig. 3a-d and Supplementary Fig. 10a). Many substitutions at these two positions also accelerate nucleation (Fig. 2a) as does the individual deletion at position E22 (Osaka mutation, Fig. 2c). However, not all internal multi-AA deletions that remove E22 or D23 increase aggregation, with deletions starting from positions 4,5,12 and 13 that remove E22 or D23 failing to accelerate nucleation (Fig. 3a). In these cases, a negatively charged residue is relocated to the immediate proximity - one or two residues away - of the core (AA 29-42, Fig. 3a and Supplementary Fig. 10b), where they likely compensate for the loss of negative charge.

The importance of charge in mediating the effects of multi-AA deletions is also suggested by a cluster of deletions in the first 15 residues of A β that accelerate nucleation (Fig. 3a and Supplementary Fig. 10a). This region contains four of the negatively charged residues in A β (D1, E3, D7 and E11) with many substitutions of these residues also accelerating aggregation (Fig. 2a). The matrix of internal multi-AA deletions further reveals that deletions that remove D1 have higher NSs than those that keep it; the same is true for D7 (Fig. 3a and Supplementary Fig. 10a). This segment is unstructured in nearly all mature A β fibril polymorphs³⁵⁻³⁹, including those in AD brains⁵ (Fig. 5 and Supplementary Fig. 14), yet diverse types of mutation in this region strongly increase aggregation.

Mutations in the aliphatic core that accelerate aggregation

The vast majority of mutations of any type within the aliphatic C-terminus (AA 29-42) of A β strongly disrupt nucleation (Fig. 1c,e,f). Indeed all insertions in the 33-38 stretch disrupt nucleation, suggesting that this may constitute the inner core of the nucleation transition state (Fig. 2b and Supplementary Fig. 2b).

However, there are some variants in the C-terminus that increase nucleation: 16 substitutions, 14 insertions, one internal multi-AA deletion within the C-terminus and nine multi-AA deletions that involve C-terminal residues (FDR=0.1, Fig. 1f,g). The substitutions in the C-terminus that accelerate nucleation are enriched at A30 and A42. At position 42, mutations to L promote nucleation, as do changes to C, T and N (Fig. 2a and Supplementary Fig. 2a). Among these, only A42T is a known fAD variant (Supplementary Table 4). At position 34 and 36, 4 substitutions to alternative hydrophobic AAs promote nucleation, suggesting that the L and V side chains may not be optimal in the nucleation transition state. L34V is also a known fAD variant (Supplementary Table 4). Insertions that promote nucleation are also enriched at specific positions. Polar insertions at position 32, flanking G33, may favor a turn, and polar, aromatic and hydrophobic insertions at positions 39, 41 and 42 (Fig. 2b and Supplementary Fig. 2b) at the end of the core may be more easily accommodated by minor structural rearrangements.

The only deletion within the core that accelerates nucleation is the removal of G33 and L34, although the individual deletion of each residue disrupts nucleation (Fig. 3a and Supplementary

Fig. 10a). It is possible that adjustments in the core can accommodate removal of these two residues by the formation of a similar structural polymorph. Finally, nine internal multi-AA deletions that bridge the N and C-terminus increase nucleation (FDR=0.1, Fig. 3a,e and Supplementary Fig. 10a). These deletions remove aliphatic core residues but replace them with a similar number of aliphatic residues from a more N-terminal segment of the peptide (Fig. 3a,e). It is likely that these internal multi-AA deletions are therefore creating alternative aliphatic cores that nucleate to form the same or similar structural polymorphs as full length A β . We find that these alternative cores that increase nucleation have a specific range of core lengths, with the hydrophobic stretch spanning from 13 to 16AA, very similar to the 14 AA length in the WT peptide (Fig. 3e,f and Supplementary Fig. 11a).

Positive charge promotes the nucleation of a minimal A β core

The DIM dataset shows that progressively removing AAs from the N-terminus of A β generates many peptides that aggregate faster than the full 42AA isoform, with 13/27 N-terminal truncations promoting nucleation (Fig. 4a,b and Supplementary Fig. 12). Such N-terminally truncated fragments of A β have been detected in AD patients^{42,43} (Supplementary Table 2) and our data suggests that environmental triggers, infections or genomic alterations that increase their production are likely to accelerate A β aggregation and so may be causally important in familial and sporadic AD.

In contrast, all N-terminal truncations that remove at least one residue of the aliphatic core (AA 29-42) very strongly reduce aggregation, further highlighting the critical requirement for this region in nucleation (Fig. 4a and Supplementary Fig. 12). Strikingly, however, the aliphatic core alone nucleates very slowly (FDR=0.1, Fig. 4a). The addition of residue 28 to this minimal core dramatically accelerates nucleation, with the 15AA peptide consisting of residues 28-42 actually being the fastest nucleating N-terminally truncated form of A β (Fig. 4a,b). This minimal A β core nucleates faster than full-length A β (Fig. 4a) and is too short to form the S-shaped amyloid fibrils polymorph observed in AD plaques⁵ and so likely adopts a smaller C-shaped polymorph with two main strands facing each other. The rapid nucleation of this 15AA peptide is particularly striking given the observation that all multi-AA deletions of more than 23AA prevent nucleation (Fig. 3a).

Residue 28 is a lysine and many of the other faster nucleating N-terminally truncated peptides also have positively charged residues at or close to their N-termini (Fig. 4b). Moreover, internal multi-AA deletions that remove K28 but that still nucleate often have a positively charged residue at the N-terminus of the core (Supplementary Figs. 10c and 11b). We therefore tested the hypothesis that it is the addition of a positively charged residue that accelerates nucleation of the minimal aliphatic core of A β . Adding the positively charged residues K or R to the N-terminus of the A β core (AA 29-42) strongly accelerated nucleation (Fig. 4c). In contrast, adding the negatively charged residues D or E did not (Fig. 4c). The addition of a single positively charged residue is therefore sufficient to dramatically accelerate the aggregation of the aliphatic core of A β . It is possible that positive residues, but not negative ones, at position 28 engage in a salt bridge with the carboxyl group at the C-terminus of the peptide to promote nucleation⁴⁴.

Residue 42 is required for fast nucleation

In contrast to the effects of N-terminal truncations, removing even a single AA from the C-terminus of A β strongly reduces nucleation (Fig. 4a and Supplementary Fig. 12). That A42 plays an important role in the nucleation of A β is consistent with previous reports that A β 42 aggregates faster than A β 40⁶. However, position 42 does not need to be an A: multiple substitutions and multiple insertions before position 42 (Fig. 2a,b and Supplementary Fig. 2a,b) either do not disrupt nucleation or actually accelerate it. This suggests that the requirement for position 42 may therefore primarily be a steric one, for example to position a free carboxyl terminus in the nucleation transition state⁴⁴.

Discussion

We have presented here the first systematic comparison of the effects of substitutions, insertions and deletions in a human disease gene. The resulting dataset shows that the consequences of AA insertions, deletions and truncations are not trivial to predict from the effects of substitutions, highlighting the importance of including Deep Indel Mutagenesis (DIM) when constructing an atlas of variant effects (AVE)¹¹ for the interpretation of clinical genetic variants.

The dataset provides a comprehensive AVE for A β aggregation that can be used to guide the future clinical interpretation of variants as they are discovered. The atlas reveals that many variants beyond substitutions accelerate the aggregation of A β and so are likely to be pathogenic. The identification of 307 variants that accelerate aggregation (Supplementary Fig. 13 and Supplementary Table 1) in this very short 42AA peptide highlights the potentially enormous diversity of disease-causing variants in the human genome. For example, the A β AVE reveals that the Uppsala mutation (Δ 19-24) is just one of many internal multi-AA deletions in a central hotspot region of A β that accelerate aggregation; these additional deletions are also likely to be pathogenic, as are multiple additional single AA deletions in the N-terminus and many N-terminal truncations of the peptide.

The substitutions, insertions, deletions and truncations that accelerate nucleation are all strongly enriched in the polar N-terminal region of A β (Fig. 5 and Supplementary Fig. 14). The different classes differ, however, in their distributions of mutational effects, with substitutions and insertions in the N-terminus more likely to impair rather than enhance aggregation but single and multi-AA deletions more likely to enhance rather than impair it. N-terminal truncations of A β are particularly likely to accelerate nucleation, raising the intriguing possibility that increased production of N-terminally truncated forms of A β triggered by environmental exposures, pathogens or genetics might be an important cause of familial and sporadic AD.

The DIM dataset also provides substantial mechanistic insight into amyloid nucleation. The vast majority of mutations of any type in the aliphatic C-terminus of A β strongly reduce aggregation, which is consistent with this region forming the core of all known mature A β fibril polymorphs^{5,35-39}. The exquisite sensitivity of this region to mutation suggests very strong structural constraints for amyloid nucleation. Only a few specific substitutions and insertions are tolerated, with these concentrated in residues at the end of the peptide which may more easily accommodate different side chains. In addition, the only internal multi-AA deletions that still nucleate despite removing residues from the C-terminus core are those which replace the missing residues with a similar number of aliphatic AA from a more N-terminal region of the peptide. The 14AA aliphatic core of A β , however, nucleates very poorly unless a positively charged residue is added at the N-terminus. We speculate that this charge may help solubilise this very hydrophobic peptide, prevent an alternative off-pathway non-amyloid aggregation process or participate directly in the nucleation transition state, for example by the formation of a salt bridge with the carboxy-terminus⁴⁴.

In contrast, mutations in the polar N-terminus have much more diverse effects, with 267 variants in this region accelerating aggregation. Much of this region remains unstructured in mature A β fibrils, including those in AD amyloid plaques (Fig. 5 and Supplementary Fig. 14)^{5,35–39}. Mutational effects in this region are not predicted by existing computational methods and they are not obviously interpretable using current mechanistic models of amyloid nucleation: polar, small and positively charged residues as well as P tend to increase nucleation whereas hydrophobic and negatively charged residues tend to decrease it. However, these preferences can also be quite different at individual sites and in sub-regions of the N-terminus, further highlighting the importance of generating complete datasets for the interpretation of clinical variants.

We speculate that mutations in the polar N-terminus of A β may control the rate of nucleation either because of effects on the ensemble of soluble A β or because the region participates in the transition state of the nucleation reaction in an as yet undefined manner. Future work should aim to distinguish between these possibilities. The very strong enrichment of disease-causing and aggregation-promoting mutations in the polar N-terminus of A β makes understanding how polar extensions promote and prevent amyloid aggregation one of the highest priority goals in AD and amyloid research.

References

1. Chiti, F. & Dobson, C. M. Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade. *Annu. Rev. Biochem.* **86**, 27–68 (2017).
2. Campion, D. *et al.* Early-onset autosomal dominant Alzheimer disease: prevalence, genetic heterogeneity, and mutation spectrum. *Am. J. Hum. Genet.* **65**, 664–670 (1999).
3. O'Brien, R. J. & Wong, P. C. Amyloid precursor protein processing and Alzheimer's disease. *Annu. Rev. Neurosci.* **34**, 185–204 (2011).
4. Iadanza, M. G., Jackson, M. P., Hewitt, E. W., Ranson, N. A. & Radford, S. E. A new era for understanding amyloid structures and disease. *Nat. Rev. Mol. Cell Biol.* **19**, 755–773 (2018).
5. Yang, Y. *et al.* Cryo-EM structures of amyloid- β 42 filaments from human brains. *Science* **375**, 167–172 (2022).
6. Meisl, G. *et al.* Differences in nucleation behavior underlie the contrasting aggregation kinetics of the A β 40 and A β 42 peptides. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 9384–9389 (2014).
7. Knowles, T. P. J., Vendruscolo, M. & Dobson, C. M. The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.* **15**, 384–396 (2014).
8. Weggen, S. & Behr, D. Molecular consequences of amyloid precursor protein and presenilin mutations causing autosomal-dominant Alzheimer's disease. *Alzheimers. Res. Ther.* **4**, 9 (2012).
9. Starita, L. M. *et al.* Variant Interpretation: Functional Assays to the Rescue. *Am. J. Hum. Genet.* **101**, 315–325 (2017).
10. Esposito, D. *et al.* MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol.* **20**, 223 (2019).
11. Members, A. A. F. The Atlas of Variant Effects (AVE) Alliance: understanding genetic variation at nucleotide resolution. (2021).
12. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801 (2014).
13. Manolio, T. A. *et al.* Bedside Back to Bench: Building Bridges between Basic and Clinical Genomic Research. *Cell* **169**, 6–12 (2017).
14. Chandramowliswaran, P. *et al.* Mammalian amyloidogenic proteins promote prion nucleation in yeast. *J. Biol. Chem.* **293**, 3436–3450 (2018).
15. Seuma, M., Faure, A., Badia, M., Lehner, B. & Bolognesi, B. The genetic landscape for amyloid beta fibril nucleation accurately discriminates familial Alzheimer's disease mutations. *Elife* **10**, e63364 (2021).
16. Montgomery, S. B. *et al.* The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.* **23**, 749–761 (2013).
17. Lin, M. *et al.* Effects of short indels on protein structure and function in human genomes. *Sci. Rep.* **7**, 1–9 (2017).
18. Mullaney, J. M., Mills, R. E., Pittard, W. S. & Devine, S. E. Small insertions and deletions (INDELs) in human genomes. *Hum. Mol. Genet.* **19**, R131–6 (2010).
19. Stenson, P. D. *et al.* The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-

- generation sequencing studies. *Hum. Genet.* **136**, 665–677 (2017).
20. Vetter, I. R. *et al.* Protein structural plasticity exemplified by insertion and deletion mutants in T4 lysozyme. *Protein Sci.* **5**, 2399–2415 (1996).
 21. Gonzalez, C. E., Roberts, P. & Ostermeier, M. Fitness Effects of Single Amino Acid Insertions and Deletions in TEM-1 β -Lactamase. *J. Mol. Biol.* **431**, 2320–2330 (2019).
 22. Emond, S. *et al.* Accessing unexplored regions of sequence space in directed enzyme evolution via insertion/deletion mutagenesis. *Nat. Commun.* **11**, 1–14 (2020).
 23. Arpino, J. A. J., Reddington, S. C., Halliwell, L. M., Rizkallah, P. J. & Jones, D. D. Random single amino acid deletion sampling unveils structural tolerance and the benefits of helical registry shift on GFP folding and structure. *Structure* **22**, 889–898 (2014).
 24. Shin, J.-E. *et al.* Protein design and variant prediction using autoregressive generative models. *Nat. Commun.* **12**, 1–11 (2021).
 25. Faure, A. J., Schmiedel, J. M., Baeza-Centurion, P. & Lehner, B. DiMSum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. *Genome Biol.* **21**, 207 (2020).
 26. Thacker, D. *et al.* The role of fibril structure and surface hydrophobicity in secondary nucleation of amyloid fibrils. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 25272–25283 (2020).
 27. Hatami, A., Monjazebe, S., Milton, S. & Glabe, C. G. Familial Alzheimer's Disease Mutations within the Amyloid Precursor Protein Alter the Aggregation and Conformation of the Amyloid- β Peptide. *J. Biol. Chem.* **292**, 3172–3185 (2017).
 28. Löhr, T., Kohlhoff, K., Heller, G. T., Camilloni, C. & Vendruscolo, M. A kinetic ensemble of the Alzheimer's A β peptide. *Nature Computational Science* **1**, 71–78 (2021).
 29. Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J. & Serrano, L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* **22**, 1302–1306 (2004).
 30. Tartaglia, G. G. & Vendruscolo, M. The Zyggregator method for predicting protein aggregation propensities. *Chem. Soc. Rev.* **37**, 1395–1401 (2008).
 31. Sormanni, P., Aprile, F. A. & Vendruscolo, M. The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol.* **427**, 478–490 (2015).
 32. Oliveberg, M. Waltz, an exciting new move in amyloid prediction. *Nat. Methods* **7**, 187–188 (2010).
 33. Törnquist, M. *et al.* Secondary nucleation in amyloid formation. *Chem. Commun.* **54**, 8667–8684 (2018).
 34. Michiels, E. *et al.* Entropic Bristles Tune the Seeding Efficiency of Prion-Nucleating Fragments. *Cell Rep.* **30**, 2834–2845.e3 (2020).
 35. Colvin, M. T. *et al.* Atomic Resolution Structure of Monomorphic A β 42 Amyloid Fibrils. *J. Am. Chem. Soc.* **138**, 9663–9674 (2016).
 36. Lührs, T. *et al.* 3D structure of Alzheimer's amyloid-beta(1-42) fibrils. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 17342–17347 (2005).
 37. Gremer, L. *et al.* Fibril structure of amyloid- β (1-42) by cryoelectron microscopy. *Science* **9**, eaao2825–9 (2017).
 38. Wälti, M. A. *et al.* Atomic-resolution structure of a disease-relevant A β (1-42) amyloid fibril. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E4976–84 (2016).
 39. Xiao, Y. *et al.* A β (1-42) fibril structure illuminates self-recognition and replication of

- amyloid in Alzheimer's disease. *Nat. Struct. Mol. Biol.* **22**, 499–505 (2015).
40. Tomiyama, T. *et al.* A new amyloid beta variant favoring oligomerization in Alzheimer's-type dementia. *Ann. Neurol.* **63**, 377–387 (2008).
 41. Pagnon de la Vega, M. *et al.* The Uppsala APP deletion causes early onset autosomal dominant Alzheimer's disease by altering APP processing and increasing amyloid β fibril formation. *Sci. Transl. Med.* **13**, (2021).
 42. Cabrera, E. *et al.* A β truncated species: Implications for brain clearance mechanisms and amyloid plaque deposition. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1864**, 208–225 (2018).
 43. Dunys, J., Valverde, A. & Checler, F. Are N- and C-terminally truncated A β species key pathological triggers in Alzheimer's disease? *J. Biol. Chem.* **293**, 15419–15428 (2018).
 44. Das, A., Korn, A., Carroll, A., Carver, J. A. & Maiti, S. Application of the Double-Mutant Cycle Strategy to Protein Aggregation Reveals Transient Interactions in Amyloid- β Oligomers. *J. Phys. Chem. B* **125**, 12426–12435 (2021).
 45. Yang, X. *et al.* On the role of sidechain size and charge in the aggregation of A β 42 with familial mutations. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E5849–E5858 (2018).
 46. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7**, Unit7.20 (2013).
 47. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
 48. Gray, V. E. *et al.* Elucidating the Molecular Determinants of A β Aggregation with Deep Mutational Scanning. *G3* **9**, 3683–3689 (2019).
 49. Bolognesi, B. *et al.* The mutational landscape of a prion-like domain. *Nat. Commun.* **10**, 4162 (2019).
 50. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).
 51. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
 52. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

Acknowledgments

M.S. is supported by a fellowship from Agència de Gestió d'Ajuts Universitaris i de Recerca (2019FI_B 01311). Work in the lab of BB and BL is supported by the la Caixa Research Foundation project 'DeepAmyloids' (LCF/PR/HR21/52410004). Work in the lab of BB is also supported by the Spanish Ministry of Science, Innovation and Universities (RTI2018-101491-A-I00 (MICIU/FEDER)) and the CERCA Program/Generalitat de Catalunya. Work in the lab of BL is also supported by a European Research Council (ERC) Advanced Grant ('Mutanomics' 883742), the Spanish Ministry of Science, Innovation and Universities (PID2020-118723GB-I00), the Bettencourt Schueller Foundation, the AXA Research Foundation, Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR, 2017 SGR 1322) and the CERCA Program/Generalitat de Catalunya. We also acknowledge the support of the Spanish Ministry of Science and Innovation to the EMBL partnership and the Centro de Excelencia Severo Ochoa. We thank the Chernoff lab for providing strains and plasmids and the CRG Genomics Core Technology for sequencing. We also thank Andre Faure and Marta Badia for advice on data analysis, Leire Moriones for assistance with validation experiments and Xavier Salvatella for discussion.

Author contributions

M.S. performed all experiments and analyses. M.S., B.L. and B.B. designed the experiments and analyses and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Figure legends

Figure 1. Deep indel mutagenesis of A β

a A β coding sequence colored by AA class (red: negative, blue: positive, green: polar, gray: aliphatic, brown: aromatic, dark gray: glycine) and schematics of the *in vivo* selection assay. A β , fused to sup35N, seeds aggregation of sup35p causing a read-through of a premature stop codon in the *ade2* reporter gene allowing growth in medium lacking adenine. **b** Distribution of nucleation scores for each class of mutations. Dashed lines indicate WT nucleation score (0). **c** Distributions of nucleation scores for mutations in different regions: N-terminus (AA 1-28), C-terminus (AA 29-42) or both (AA 1-42). **d,e** Frequency of variants increasing or decreasing nucleation at different FDRs for the full peptide (**d**) and for each peptide region (**e**). **f** Frequency and total counts of each mutation type for variants increasing (NS+), decreasing (NS-) or having no effect (WT-like) at FDR=0.1. **g** Number and type of variants increasing nucleation (NS+) for each peptide region.

Figure 2. Single AA variant atlases

a Heatmap of nucleation scores for single AA substitutions. The WT AA and position are indicated in the x-axis and the mutant AA is indicated in the y-axis. Variants not present are indicated in gray, synonymous mutants with '*' and fAD mutants with a black box. Non-nucleating variants (with no NS, see methods) are indicated with '-'. The distribution of nucleation scores for each position is summarized in the violin plots below the heatmap. **b** Heatmap of nucleation scores for single AA insertions at each position. **c** Effect of single AA deletions. The horizontal line indicates the WT nucleation score (0). Vertical error bars indicate 95% confidence interval of the mean. **d** Frequency of AA increasing or decreasing nucleation (FDR=0.1) upon substitutions (top) or insertion (bottom) for each peptide region. **e** Correlation of nucleation scores for substitutions to each AA at each position and insertions of the same AA before (top) or after (bottom) that position. Color indicates peptide region (N and C-terminus). **f** Correlation of average nucleation scores for each AA, for insertions and substitutions at the N-terminus (top) and at the C-terminus (bottom). Color indicates AA type. Pearson correlation coefficients are indicated in (**e**) and (**f**).

Figure 3. Internal multi-AA deletion atlas

a Matrix of nucleation scores for deletions. The dashed-line black square depicts the hotspot of deletion effects (consecutive deleted positions where NS+ frequency > 1/2 max(NS+ frequency, i.e. deletions starting at positions 17-23 and ending at positions 22-27) and the yellow dots indicate deletions removing residues in both the N and C-terminus that increase NS (see (**e**)). Variants not present are represented in gray and non-nucleating variants (with no NS, see methods) are indicated with '-'. **b** Effect on nucleation of deletions of 1-6AA length. The WT AA and position are indicated in the x-axis. The black squares indicate fAD variants: Osaka (E22 Δ) and Uppsala (Δ 19-24). Color code as in (**a**). **c** Frequency of variants increasing nucleation (NS+), decreasing nucleation (NS-) or with no difference from WT (WT-like) at FDR=0.1, for sequences with a specific first deleted position (i.e. each column in the matrix), last deleted position (i.e. each row in the matrix) or missing a specific residue, at the N-terminus (AA 1-28). **d** AA sequence for variants inside the hotspot of deletion effects with significantly increased NS (FDR=0.1). **e** AA sequence for variants with internal multi-AA deletions removing residues from both the N and C-

terminus, with significantly increased NS (FDR=0.1). AA coloured by AA class. Color code as in (d). f Nucleation scores of variants with putative alternative aliphatic cores of different lengths. The horizontal line indicates the WT nucleation score (0). Vertical red line indicates WT core length (14AA). Variants displaying alternative cores were defined as those internal multi-AA deletions removing residues in both the N and C-terminus that replace part of the C-terminus with exclusively aliphatic residues (n=64).

Figure 4. Positive charge accelerates nucleation of a minimal A β core

a Effect of N-terminal (top) and C-terminal (bottom) truncations on nucleation. Vertical error bars indicate 95% confidence interval of the mean NS. **b** AA sequences of N-terminal truncations that increase nucleation at FDR=0.1. AA are coloured by class. **c** Effect of adding a positively or negatively charged residue at the N or C-terminus of the A β core (AA 29-42). Nucleation quantified as percentage of colonies in medium lacking adenine vs. medium containing adenine. One-way ANOVA with Dunnett's multiple comparisons test. * p<0.05, **p<0.01, *** p<0.001.

Figure 5. Mutational effects visualized on fAD A β fibril structures

In fibrils extracted from the brains of fAD patients, A β 42 adopts an S-shaped structure at the C-terminus with an N-terminal arm linking to an unstructured region indicated by the dashed line (PDB: 7Q4M)⁵. **a** Single AA substitutions, single AA insertions and N-terminal multi-AA deletions: color intensity indicates the percentage of NS+ (blue) or NS- (red) mutations at each position or losing each position (for multi-AA deletions) (FDR=0.1). **b** Single AA deletions and N-terminal truncations: color intensity depicts the nucleation score of each single AA deletion or of the N-terminal truncation starting at that position. White depicts positions that are not mutated in each dataset.

Supplementary figures and tables

Supplementary Figure 1. Reproducibility and assay validation

a Correlation of nucleation scores for three biological replicates ($n_{1-2}=2,951$, $n_{1-3}=2,984$, $n_{2-3}=2,950$ genotypes). **b** Correlation of nucleation scores measured for the synthetic library used in this study and a previous library generated by error-prone PCR ($n=423$ common variants)¹⁵. **c** Correlation of nucleation scores measured in the competition experiment or individually for selected variants ($n=10$). Vertical and horizontal error bars indicate 95% confidence intervals of mean NS. Pearson correlation coefficients are indicated in **(a-c)**. **d** Correlation of nucleation scores with *in vitro* primary and secondary nucleation rate constants⁴⁵. Weighted Pearson correlation coefficients are indicated. **e** Receiver operating characteristic (ROC) curves for 12 of all the single AA substitutions described as dominant fAD variants (H6R, D7N, D7H, E11K, K16Q, A21G, E22Q, E22K, E22G, D23N, L34V and A42T) versus all other single AA substitutions present in the dataset ($n_{\text{non-fAD}}=739$) for two DMS datasets (Nucleation score and Solubility score), aggregation predictors (Tango, Zyggregator, Waltz, Camsol²⁹⁻³²) and variant effect predictors (Polyphen and CADD^{46,47}). Area under the curve (AUC) values are indicated. Diagonal dashed line indicates the performance of a random classifier. **f** Number and type of variants increasing nucleation (NS-, FDR=0.1) for each peptide region.

Supplementary Figure 2. Mutational effects of single AA substitutions and insertions

a Heatmap of nucleation scores FDR categories for single AA substitutions. The WT AA and position are indicated in the x-axis and the mutant AA is indicated in the y-axis. Variants not present are represented in gray. Synonymous mutants are indicated with '*' and fAD mutants with a black box. **b** Heatmap of nucleation scores FDR categories for single AA insertions. **c** Frequency of increasing or decreasing nucleation (FDR=0.1) single AA substitutions upon substituting specific WT AA, for each peptide region.

Supplementary Figure 3. Mutational effects of single AA substitutions and insertions

a,b Clustering of single AA mutation nucleation scores by mutated residue identity and position. Position is indicated in the x-axis; AA insertions were considered after **(a)** or before **(b)** each position. Mutations are indicated in the y-axis and labeled with an 's' for substitutions or an 'i' for insertions, followed by the substituted or inserted AA. 'del' indicates single AA deletion of that position.

Supplementary Figure 4. Comparing the mutational effects of single AA variants

a Correlation of average nucleation scores for each position, for single AA insertions before or after a specific position and single AA substitutions (left) or single AA deletions (middle), and for single AA deletions and single AA substitutions (right) at the corresponding position. Color code indicates peptide region (N-terminus, AA 1-28, or C-terminus, AA 29-42). **b** Correlation of average nucleation scores for each AA, for single AA deletions and single AA substitutions (top row), single AA insertions and single AA substitutions (middle row) and single AA insertions and single AA deletions (bottom row); and for the full peptide (left column), the N-terminus (AA 1-28, middle column) or the C-terminus (AA 29-42, right column). **c** Correlation of average nucleation scores for each AA, for the C and the N-terminus, for single AA substitutions (left), single AA insertions

(middle) and single AA deletions (right). AA labels are coloured by AA class in (b) and (c). Pearson correlation coefficients are indicated. Dashed lines indicate the WT nucleation score (0).

Supplementary Figure 5. Comparing the mutational effects of single AA substitutions and insertions

a,b Correlation of nucleation scores at each position arranged by each AA type, between single AA substitutions and single AA insertions before (a) or after (b) the corresponding position. Pearson correlation coefficients are indicated. Dashed lines indicate the WT nucleation score (0).

Supplementary Figure 6. Comparing the mutational effects of single AA substitutions and insertions

a,b Correlation of nucleation scores for each AA type arranged by position, between single AA substitutions and single AA insertions before (a) or after (b) the corresponding position. Pearson correlation coefficients are indicated. Dashed lines indicate the WT nucleation score (0). Color code indicates AA position.

Supplementary Figure 7. Mutational effects of substituting in specific AAs

The wild-type (WT) AA and position are indicated on the x-axis and coloured on the basis of their effect (NS+ or NS-) and FDR category. The horizontal line indicates the WT nucleation score (0).

Supplementary Figure 8. Mutational effects of inserting specific AAs

The wild-type (WT) AA and position are indicated on the x-axis and coloured on the basis of their effect (NS+ or NS-) and FDR category. The horizontal line indicates the WT nucleation score (0).

Supplementary Figure 9. Evaluation of mutational effect and aggregation predictors

a Correlation of nucleation scores with the predictions of aggregation predictors (Tango, Zyggregator, Waltz and Camsol)²⁹⁻³², variant effect predictors (CADD, Polyphen)^{46,47}, solubility scores⁴⁸, PC1⁴⁹ and hydrophobicity⁵⁰ for single AA mutations, at the N-terminus (left) or the C-terminus (right). Pearson correlation coefficients are indicated. Dashed lines indicate the WT nucleation score (0). **b** Receiver operating characteristic (ROC) curves for classifying increasing nucleation variants (NS+, FDR=0.1) for single AA mutations, at the N and C-terminus, for aggregation predictors²⁹⁻³², variant effect predictors^{46,47}, solubility scores⁴⁸, PC1⁴⁹ and hydrophobicity⁵⁰. Area under the curve (AUC) values are indicated. Diagonal dashed line indicates the performance of a random classifier.

Supplementary Figure 10. Multi-AA deletions

a Heatmap of nucleation scores FDR categories for multi-AA deletions. The WT AA and position of the first and last residues deleted are indicated in the x-axis and y-axis, respectively. The black squares indicate fAD variants: Osaka (E22Δ) and Uppsala (Δ19-24). Variants not present are represented in gray. **b** Effect on nucleation of variants that delete E22, D23N or both. The distance the closest negative residue (D,E) - if present - to the C-terminus (AA 29-42) is shown in the x-axis. Variants with no negative residues are also shown (no D/E). Shape indicates the identity of the residue and color code indicates FDR=0.1 category. The horizontal line indicates the WT nucleation score (0). **c** Generation of new N-terminus sequences flanking the Aβ core. Nucleation score distributions of each AA at each position for deletions at the N-terminus. Distance from the

C-terminus (AA 29-42) is indicated in the x-axis, as well as WT AA and position. Color of the violin plot indicates median nucleation score for each distribution.

Supplementary Figure 11. Multi-AA deletion variants

a AA sequence for variants with internal multi-AA deletions located at both N and C-terminus, with significantly decreased nucleation (FDR=0.1). **b** AA sequence for deletions at the N-terminus removing residue K28, with positive nucleation score (NS>0).

Supplementary Figure 12. N- and C-terminal truncations

a,b Heatmap of nucleation scores (**a**) or FDR categories (**b**) for truncations from one or both ends of the peptide. The WT AA and position of the first and last residues of the resulting peptide are indicated in the x-axis and y-axis, respectively.

Supplementary Figure 13. Top nucleating sequences in the library

AA sequence for 1% variants with highest NS (all FDR=0.1) in the library. AA are coloured by AA class.

Supplementary Figure 14. Impact of diverse classes of mutations along the structure of A β 42 fibrils from sporadic AD brains

The impact of all mutations of all classes is summarized over the structure of A β 42 fibrils (PDB: 7Q4B)⁵. In fibrils extracted from sporadic AD brains, A β 42 adopts a S-shaped structure at the C-terminus with an N-terminal arm linking to an unstructured region. **a** Single AA substitutions, single AA insertions and N-terminal multi-AA deletions: Color intensity indicates the percentage of NS+ (blue) or NS- (red) mutations at each position or losing each position (for multi-AA deletions) (FDR=0.1). **b** Single AA deletions and N-terminal truncations: Color intensity depicts the nucleation score of each single AA deletion or of the N-terminal truncation starting at that position. White depicts positions that are not mutated in each dataset.

Supplementary Table 1

List of candidate fAD pathogenic variants with increased nucleation (FDR=0.1).

Supplementary Table 2

List of N-terminal A β truncations reported in the literature and their corresponding nucleation score and category.

Supplementary Table 3

List of oligonucleotides used in this study.

Supplementary Table 4

Processed data required to reproduce the analysis and figures in this paper, with read counts, nucleation scores, FDR category, associated error terms and associated pathogenicity.

Material and Methods

Library design

The designed library contains a total of 3,164 unique A β 42 variants, with all single AA substitutions at each position (n=798), all single AA insertions at all positions (n=780), all deletions ranging from 1 to 39 AA in size in all positions (n=768), sequences truncated from either one or both ends of the peptide with a minimum peptide length of 3 AA and maximum peptide length of 40 AA (n=817), and the A β 42 WT sequence (n=1).

Plasmid Library construction

The synthetic library was synthesized by Twist Bioscience and consisted of an A β 42 variant region of 9 nt to 129 nt, flanked by 25 nt upstream and 21 nt downstream constant regions. 10ng of the library were amplified by PCR (Q5 high-fidelity DNA polymerase, NEB) for 12 cycles with primers annealing to the constant regions (primers MS_01-02, Supplementary Table 3), according to the manufacturer's protocol. The product was then purified by column purification (MinElute PCR Purification Kit, Qiagen). In parallel, the P_{CUP1}-Sup35N-A β 42 plasmid was linearised by PCR (Q5 high-fidelity DNA polymerase, NEB) with primers that remove the WT A β 42 sequence (primers MS_03-04, Supplementary Table 3). The product was purified from a 1% agarose gel (QIAquick Gel Extraction Kit, Qiagen).

The library was then ligated into 100ng of the linearised plasmid in a 5:1 (insert:vector) ratio by a Gibson approach with 3h of incubation followed by dialysis for 45 min on a membrane filter (MF-Millipore 0,025 μ m membrane, Merck). The product was transformed into 10-beta Electrocompetent E.coli (NEB), by electroporation with 2.0kV, 200 Ω , 25 μ F (BioRad GenePulser machine). Cells were recovered in SOC medium for 30 min and grown overnight in 30 ml of LB ampicillin medium. A small amount of cells were also plated in LB ampicillin plates to assess transformation efficiency. A total of 50,000 transformants were estimated, meaning that each variant in the library is represented >15 times. 5ml of overnight culture were harvested to purify the A β 42 library with a mini prep (QIAprep Miniprep Kit, Qiagen).

Yeast transformation

Saccharomyces cerevisiae [psi-pin-] (MATa ade1-14 his3 leu2-3,112 lys2 trp1 ura3-52) strain was used in all experiments in this study.

Yeast cells were transformed with the A β 42 plasmid library in three biological replicates. An individual colony was grown overnight in 25ml YPDA medium at 30 C and 200 rpm. Cells were diluted in 150 ml to OD₆₀₀ =0.25 and grown for 4-5 h. When cells reached the exponential phase (OD~0.7-0.8), they were splitted in 10 transformation tubes of 15 ml each. Each tube was treated as follows: cells were harvested at 400 g for 5 min, washed with milliQ and resuspended in 1 ml YTB (100 mM LiOAc, 10 mM Tris pH 8.0, 1 mM EDTA). They were harvested again and resuspended in 72 μ l YTB. 100 ng of plasmid library were added to the cells, together with 8 μ l of salmon sperm DNA (UltraPure, Thermo Scientific) previously boiled, 60 μ l of dimethyl sulfoxide (Merck) and 500 μ l of YTB-PEG (100 mM LiOAc, 10 mM Tris pH 8.0, 1 mM EDTA, 40% PEG 3350). Heat-shock was performed at 42 C for 14 min in a thermo block. Finally, cells were harvested and resuspended in 300 ml plasmid selection medium (-URA, 20% glucose), pooling together the 10 transformation tubes and allowing them to grow for 50 h at 30 C and 200 rpm. A small amount of cells were also plated in plasmid selection medium to assess transformation efficiency. A total of 118,125, 152,000 and 139,500 transformants were

estimated for each biological replicate respectively, meaning that each variant in the library is represented >37 times.

After 50 h, cells were diluted in 25 ml plasmid selection medium to OD =0.02 and grown exponentially for 15 h. Finally, the culture was harvested and stored at -80 C in 25 % glycerol.

Selection experiments

In vivo selection assays were performed in five technical replicates for each biological replicate. For each technical replicate, cells were thawed from -80 C in 20 ml plasmid selection medium at OD=0.05 and grown until exponential for 15 h. At this stage, cells were harvested and resuspended in 20 ml protein induction medium (-URA, 20% glucose, 100 uM Cu₂SO₄) at OD=0.05. After 24 h the 4x 5ml input pellets were collected and 1 million cells/replicate were plated on -ADE-URA selection medium in 145 cm² plates (Nunc, Thermo Scientific). Plates were incubated at 30 C for 7 days inside an incubator. Finally colonies were scraped off the plates with PBS 1x and harvested by centrifugation to collect the output pellets. Both input and output pellets were stored at -20 C for later DNA extraction.

DNA extraction and sequencing library preparation

One input and one output pellets for each technical and biological replicate (2x5x3 samples) were resuspended in 0.5 ml extraction buffer (2% Triton-X, 1% SDS, 100mM NaCl, 10mM Tris-HCl pH8, 1mM EDTA pH8). They were then freeze for 10 min in an ethanol-dry ice bath and heated for 10 min at 62 C. This cycle was repeated twice. 0.5 ml of phenol:chloroform:isoamyl (25:24:1 mixture, Thermo Scientific) was added together with glass beads (Sigma). Samples were vortexed for 10 min and centrifuged for 30 min at 4000 rpm. The aqueous phase was then transferred to a new tube, and mixed again with phenol:chloroform:isoamyl, vortexed and centrifuged for 45 min at 4000 rpm. Next, the aqueous phase was transferred to another tube with 1:10V 3M NaOAc and 2.2V cold ethanol 96% for DNA precipitation. After 30min at -20 C, samples were centrifuged and pellets were dried overnight. The following day, pellets were resuspended in 0.3 ml TE 1X buffer and treated with 10ul RNase A (Thermo Scientific) for 30 min at 37 C. DNA was finally purified using 10 ul of silica beads (QIAEX II Gel Extraction Kit, Qiagen) and eluted in 30 ul elution buffer. Plasmid concentrations were measured by quantitative PCR with SYBR green (Merck) and primers annealing to the origin of replication site of the P_{CUP1}-Sup35N-Aβ42 plasmid at 58 C for 40 cycles (primers MS_05-06, Supplementary Table 3).

The library for high-throughput sequencing was prepared in a two-step PCR (Q5 high-fidelity DNA polymerase, NEB). In PCR1, 50 million of molecules were amplified for 15 cycles with frame-shifted primers with homology to Illumina sequencing primers (primers MS_07-20, Supplementary Table 3). The products were purified with ExoSAP treatment (Affymetrix) and by column purification (MinElute PCR Purification Kit, Qiagen). They were then amplified for 10 cycles in PCR2 with Illumina indexed primers (primers MS_21-37, Supplementary Table 3). The six samples of each technical replicate were pooled together equimolarly and the final product was purified from a 2% agarose gel with 20 ul silica beads (QIAEX II Gel Extraction Kit, Qiagen).

The library was sent for 125 bp paired-end sequencing in an Illumina HiSeq2500 sequencer at the CRG Genomics core facility. In total, >426 million paired-end reads were obtained, which is between 7-20 million per sample (i.e. input or output for a specific technical and biological replicate), representing >2200x read coverage for each designed variant in the library.

Individual variant testing

Selected A β 42 variants for individual testing were obtained by PCR linearisation (Q5 high-fidelity DNA polymerase, NEB) with mutagenic primers (primers MS_38-47, Supplementary Table 3). PCR products were treated with Dpn1 overnight and transformed in DH5 α competent E.coli. Plasmids were purified by mini prep (QIAprep Miniprep Kit, Qiagen) and transformed into yeast cells using one transformation tube of the transformation protocol described above. All constructions were verified by Sanger sequencing.

Yeast cells expressing individual variants were grown overnight in plasmid selection medium (-URA 20% glucose). They were then diluted to OD 0.05 in protein induction medium (-URA 20% glucose 100 μ M Cu $_2$ SO $_4$) and grown for 24h. Cells were plated on -URA (control) and -ADE-URA (selection) plates in three independent replicates, and allowed to grow for 7 days at 30 C. Adenine growth was calculated as the percentage of colonies in -ADE-URA relative to colonies in -URA.

Data processing

FastQ files from paired end sequencing of the A β 42 library were processed using the DiMSum pipeline (<https://github.com/lehner-lab/DiMSum>)²⁵, an R package that wraps sequencing processing tools, such as FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) for quality assessment; Cutadapt⁵¹ for constant region trimming; and VSEARCH⁵² for read alignment. 5' and 3' constant regions were trimmed, allowing a maximum of 20% of mismatches relative to the reference sequence. Sequences with a Phred base quality score below 30 were discarded. At this stage, around 370 million reads passed the filtering criteria.

Unique variants were then aggregated and counted using Starcode (<https://github.com/gui11aume/starcode>). Non-designed variants were also discarded for further analysis, as well as variants with less than 10 input reads in any of the replicates and variants resulting from one single nt change with less than 1000 input reads. Estimates from DiMSum²⁵ were used to choose the filtering thresholds.

Nucleation scores and error estimates

The DiMSum package (<https://github.com/lehner-lab/DiMSum>)²⁵ was also used to calculate nucleation scores (NS) and their error estimates for each variant in each biological replicate as:

$$\text{Nucleation score} = ES_i - ES_{WT}$$

Where $ES_i = \log(F_i \text{ OUTPUT}) - \log(F_i \text{ INPUT})$ for a specific variant and

$$ES_{WT} = \log(F_{WT} \text{ OUTPUT}) - \log(F_{WT} \text{ INPUT}) \text{ for A}\beta\text{42 WT.}$$

NSs for each variant were merged across biological replicates using error-weighted mean and centered to the WT A β 42 NS. All NS and associated error estimates are available in Supplementary Table 4.

Data analysis

Variants in the library

NS was obtained for 3,087 unique A β 42 variants, which were splitted into mutation classes: 751 single AA substitutions, 763 single AA insertions, 37 single AA deletions, 729 internal multi-AA deletions, 817 truncations (from one or both ends) and WT A β 42.

In addition, nine variants (2 single AA substitutions, 6 single AA insertions and one multi-AA deletion) were classified as non-nucleating but do not have an associated NS (ie. they have input reads but no output reads) and are indicated as such in Figs. 2a,b and Fig. 3a. Each variant is assigned to one mutation class: deletions from position 1 or 42 are classified as truncations and not deletions, and deletions of positions 1 and 42 are classified as single AA deletion and not as truncations. Multiple mutation classes can be combined for visualization or analysis (e.g. truncations and single deletions are included in the deletions matrix in Fig. 3a).

We assign to single AA insertions the position of the inserted AA (e.g. an insertion between positions 1 and 2 is an insertion at position 2). In the case of insertions between positions 28 and 29 (i.e. between the N and C-terminus), they are insertions at position 29 but considered N-terminal mutations.

Different mutations can result in the same coding sequence (e.g. H13 Δ and H14 Δ , or DAEDVGSNKGAIIGLMVGGVVIA, which is Δ 2-20, Δ 3-21 and Δ 4-22). This is the case for single AA insertions, single and multi-AA deletions. In general, they are only considered as one coding variant but considered multiple times for visualization or if the analysis is position-specific, in figures: Figs. 2b-f, 3a,b and 5, and Supplementary Figs. 2-6 and 10a.

Aggregation and variant effect predictors

For the aggregation predictors (Tango, Zyggregator, Waltz, Camsol²⁹⁻³¹), individual residue-level scores were summed to obtain a score per single AA mutation sequence. We then calculated the log value for each variant relative to the WT score. For the variant effect predictors (Polyphen and CADD^{46,47}), we also calculated the log value for each single AA substitutions variant but in this case values were scaled relative to the lowest predicted score.

We also used an hydrophobicity scale⁵⁰ and a principal component from a previous study (PC1⁴⁹) that relates strongly to changes in hydrophobicity. For each single AA substitution variant, the values of a specific AA property represent the difference between the mutant and the WT scores.

ROC analysis

ROC curves and AUC values were built and obtained using the 'pROC' R package. The table of fAD mutations was taken from <https://www.alzforum.org/mutations/app>. The nucleation scores and categories for all fAD variants, as well as the criteria used to consider them as fAD, are reported in Supplementary Table 4.

Data availability

Raw sequencing data and the processed data table (Supplementary Table 4) are deposited in NCBI's Gene Expression Omnibus (GEO) as GSE193837. All scripts used for downstream analysis and to reproduce all figures are in <https://github.com/BEBlab/DIM-abeta>.

Fig. 1

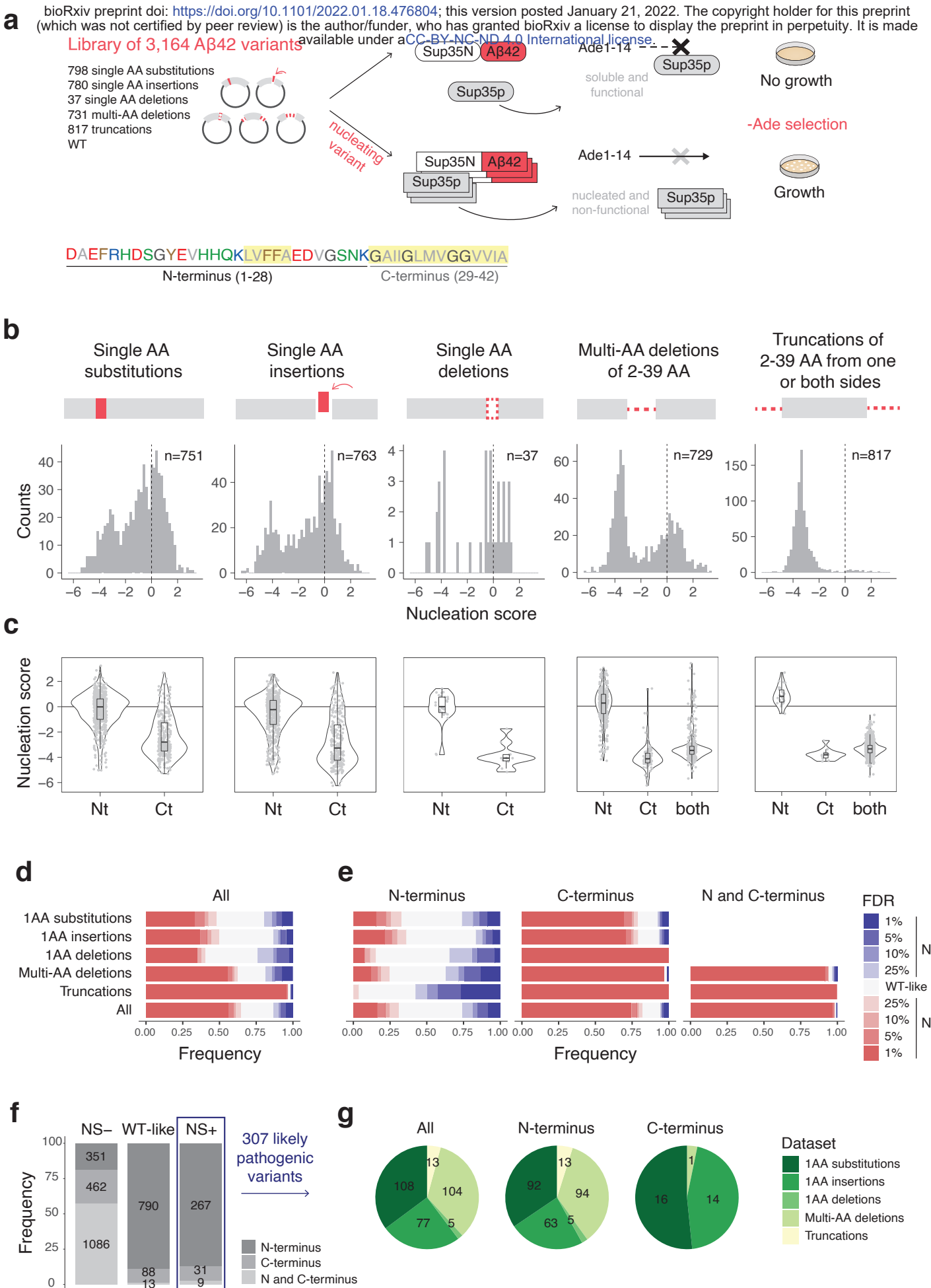


Fig. 2

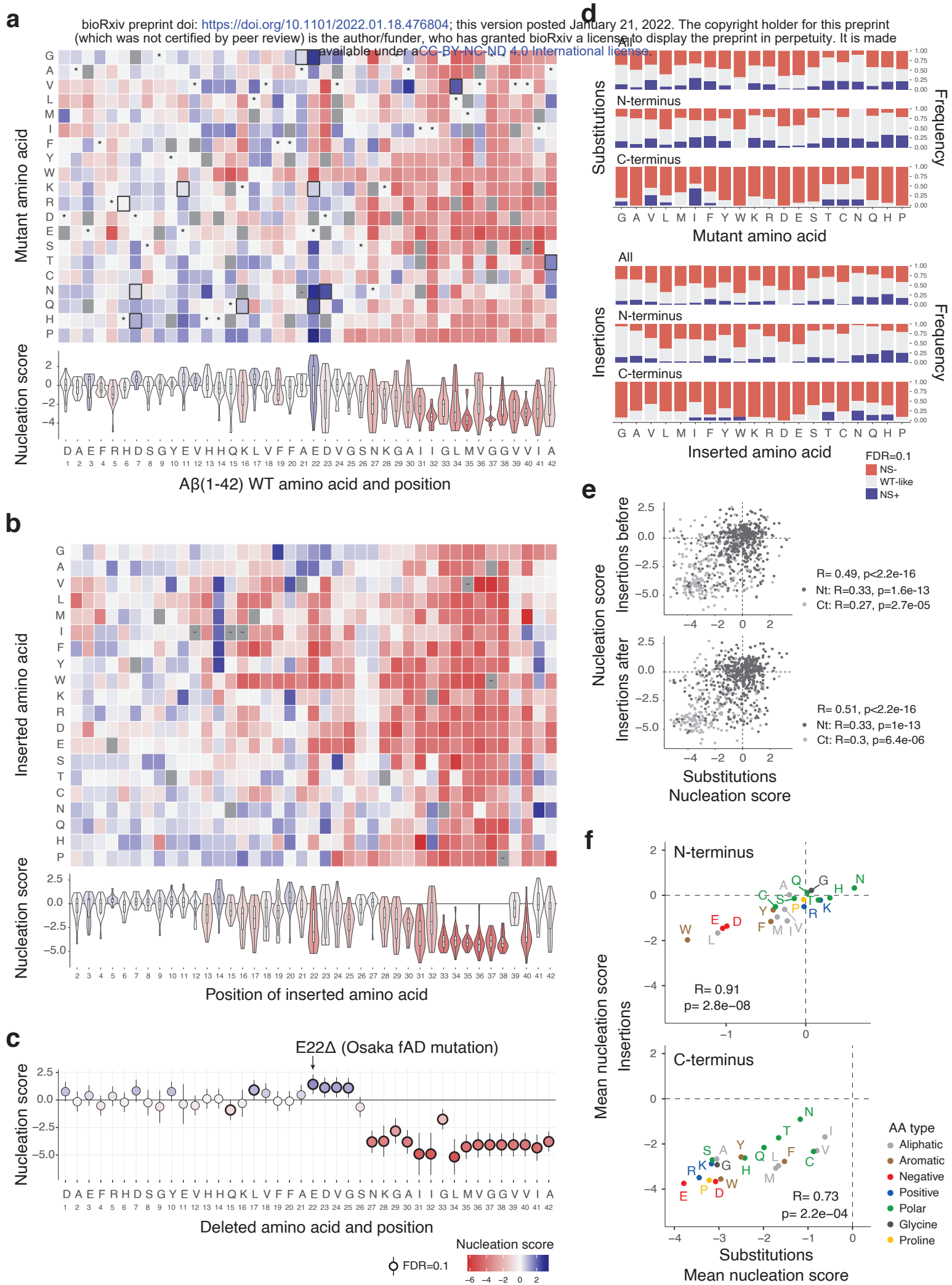


Fig. 3

bioRxiv preprint doi: <https://doi.org/10.1101/2022.01.18.476804>; this version posted January 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

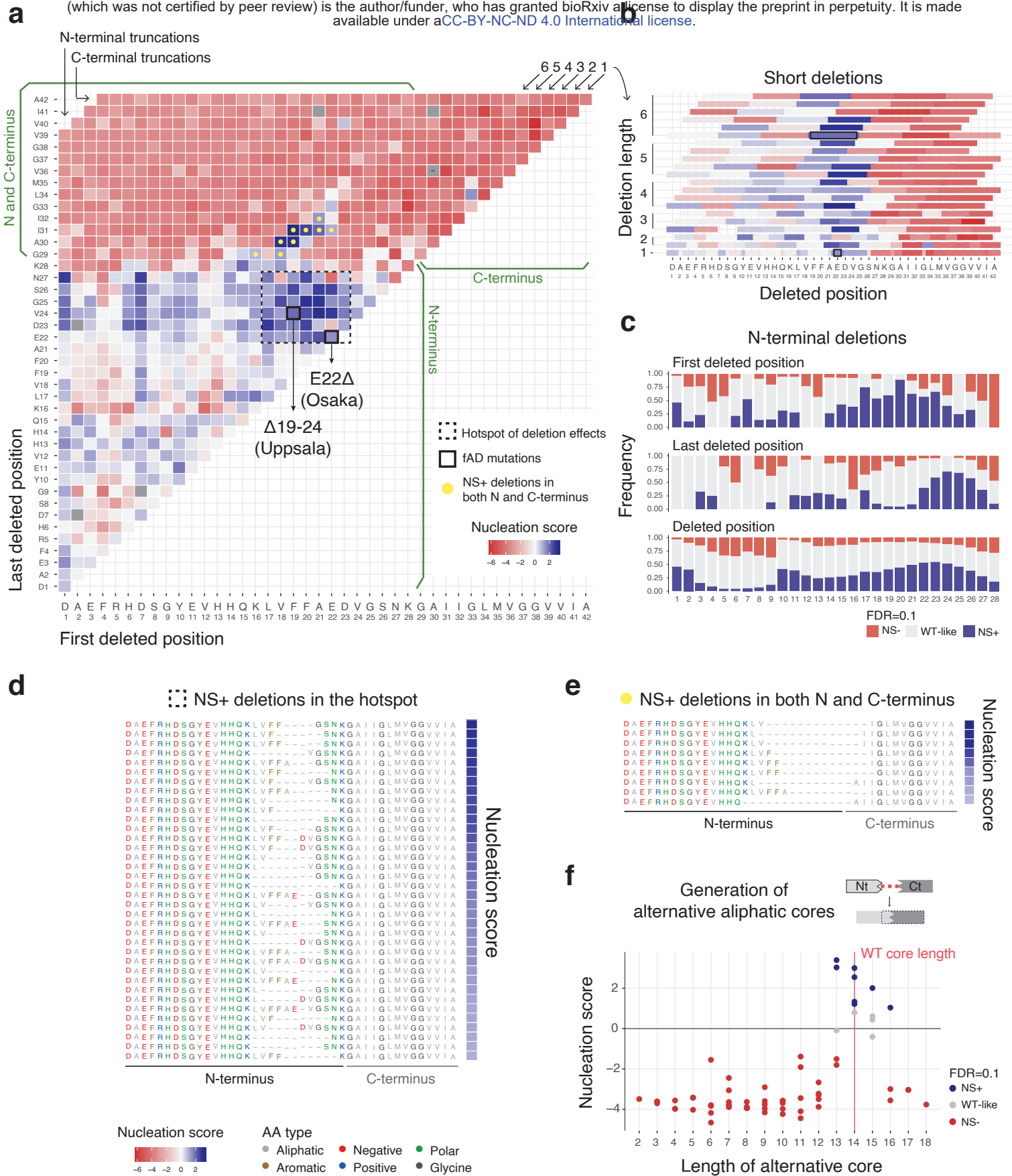


Fig. 4

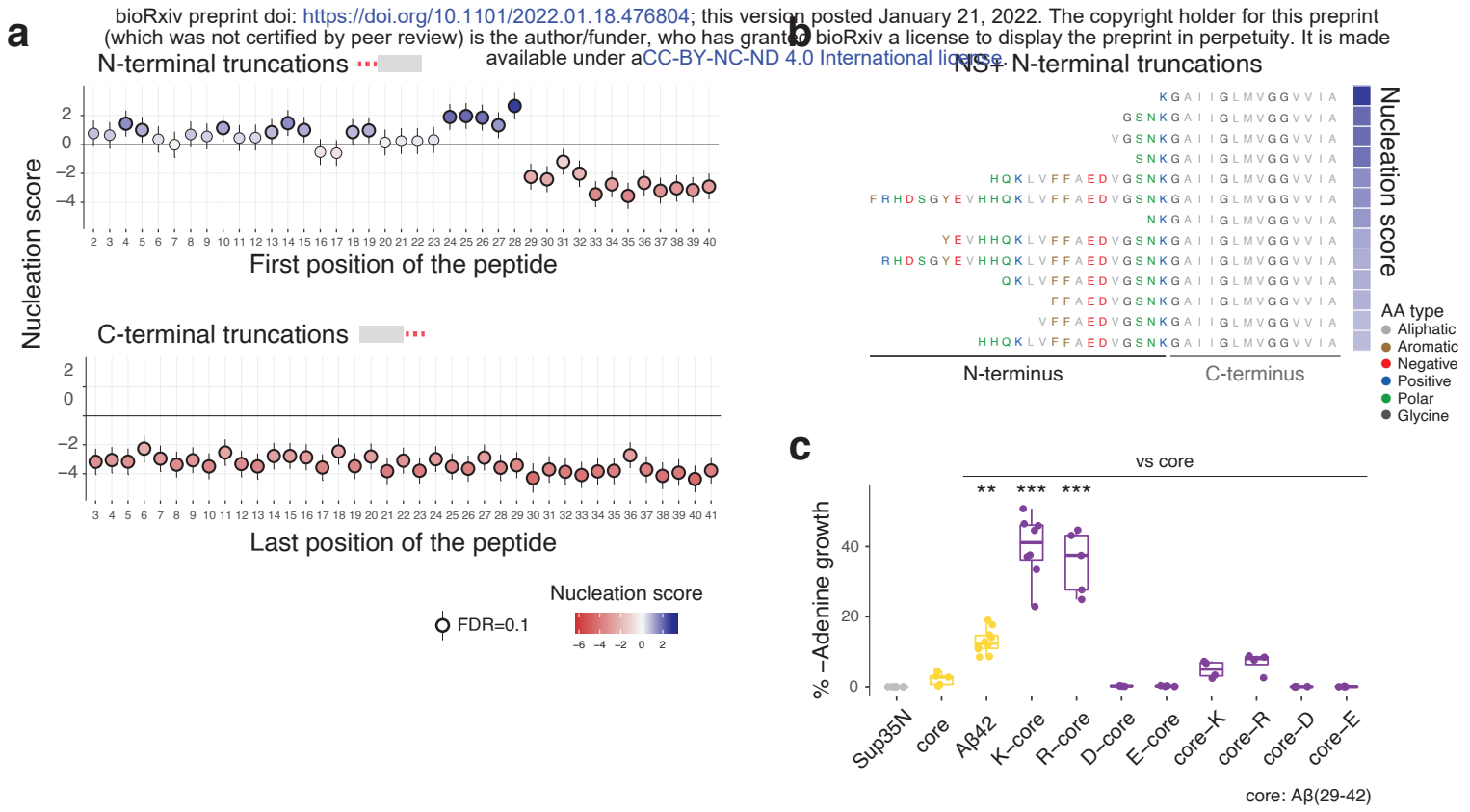
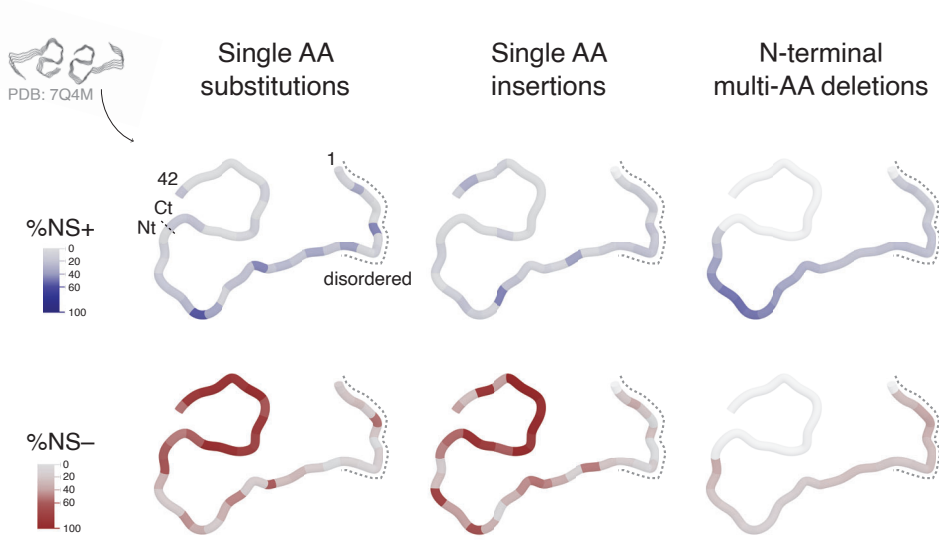


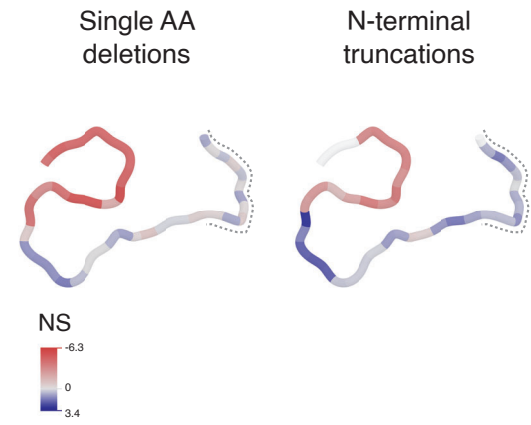
Fig. 5

bioRxiv preprint doi: <https://doi.org/10.1101/2022.01.18.476804>; this version posted January 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/).

a



b





Chapter III. The mutational landscape of a prion-like domain

ARTICLE

<https://doi.org/10.1038/s41467-019-12101-z>

OPEN

The mutational landscape of a prion-like domain

Benedetta Bolognesi^{1,2,6}, Andre J. Faure ^{1,6}, Mireia Seuma^{1,2}, Jörn M. Schmedel¹,
Gian Gaetano Tartaglia^{1,3,4,5} & Ben Lehner ^{1,3,4}

Insoluble protein aggregates are the hallmarks of many neurodegenerative diseases. For example, aggregates of TDP-43 occur in nearly all cases of amyotrophic lateral sclerosis (ALS). However, whether aggregates cause cellular toxicity is still not clear, even in simpler cellular systems. We reasoned that deep mutagenesis might be a powerful approach to disentangle the relationship between aggregation and toxicity. We generated >50,000 mutations in the prion-like domain (PRD) of TDP-43 and quantified their toxicity in yeast cells. Surprisingly, mutations that increase hydrophobicity and aggregation strongly decrease toxicity. In contrast, toxic variants promote the formation of dynamic liquid-like condensates. Mutations have their strongest effects in a hotspot that genetic interactions reveal to be structured *in vivo*, illustrating how mutagenesis can probe the *in vivo* structures of unstructured proteins. Our results show that aggregation of TDP-43 is not harmful but protects cells, most likely by titrating the protein away from a toxic liquid-like phase.

¹Center for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Doctor Aiguader 88, 08003 Barcelona, Spain. ²Institute of Bioengineering of Catalonia (IBEC), The Barcelona Institute of Science and Technology, Barcelona, Spain. ³Universitat Pompeu Fabra (UPF), Barcelona, Spain. ⁴Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain. ⁵Department of Biology ‘Charles Darwin’, Sapienza University of Rome, P.le A. Moro 5, Rome 00185, Italy. ⁶These authors contributed equally: Benedetta Bolognesi, Andre J. Faure. Correspondence and requests for materials should be addressed to B.B. (email: bbolognesi@ibecbarcelona.eu) or to B.L. (email: ben.lehner@crg.eu)

The conversion of specific proteins into insoluble aggregates is a hallmark of many neurodegenerative disorders, including Alzheimer's, Parkinson's, Huntington's, and Amyotrophic Lateral Sclerosis (ALS) with dominantly inherited mutations in aggregate-forming proteins causing rare familial forms of these diseases^{1–6}. However, both in humans and in animal models, there is often only a weak association between the presence of aggregates and disease progression^{7–9}. Indeed, multiple therapeutic approaches that reduce the formation of aggregates have failed at different stages of development^{10–12}. On the other hand, there is increasing evidence that alternative protein assemblies generated during or in parallel to the aggregation process may be toxic^{13–17}. Despite evidence that cellular damage may be induced either before, after or independent of the formation of insoluble aggregates, the latter are still widely assumed to be pathogenic in many neurodegenerative diseases^{18,19}.

For many proteins, aggregation depends critically on intrinsically disordered regions with a low sequence complexity resembling that of infectious yeast prions. These prion-like domains (PRDs) are also enriched in proteins that can form liquid-like cellular condensates^{20–22} through liquid-demixing. This is a concentration-dependent process through which proteins can separate into two coexisting liquid phases and it has been extensively characterized both in vitro and in the cytoplasm²³. In several proteins PRDs are necessary and sufficient for liquid-liquid demixing^{23,24}. At least in vitro, insoluble aggregates can nucleate from more liquid phases^{24–26}, leading to the suggestion that liquid de-mixed states can mature into pathological aggregates¹⁹.

Disordered regions²⁷ and low-complexity sequences²⁸ are also enriched in dosage-sensitive proteins that are toxic when their concentration is increased. At least for one model protein that has been tested, however, it is the formation of a concentration-dependent liquid-like phase—not aggregation—that causes cellular toxicity²⁸. Similarly, the toxicity of two mutant forms of the prion Sup35 could be explained only on the basis of their ability to populate a non-aggregate, liquid-like state^{20,29}.

Cytoplasmic aggregates of the TAR DNA-binding protein 43 (TDP-43) are a hallmark of ALS, present in 97% of post-mortem samples^{2,30}. TDP-43 aggregates are also present at autopsy in nearly all cases of frontotemporal dementia (FTD) that lack tau-containing inclusions (about half of all cases of FTD which is the second most common dementia)³¹. TDP-43 aggregates also represent a hallmark of inclusion body myopathy, and a secondary pathology in Alzheimer's, Parkinson's, and Huntington's disease^{31–33}. However, TDP-43 aggregates are also observed—albeit at low frequency—in control samples³⁴ and, in vitro, TDP-43 can form both amyloid aggregates and liquid condensates^{35–39}. Mutations in TDP-43 cause ~5% of familial ALS (fALS) cases^{8,40}, with these mutations reported to interfere with nuclear-cytoplasmic transport, RNA processing, splicing, and protein translation^{7,41–46}. However, despite extensive investigation, the molecular form of the protein that causes cellular toxicity is still unknown^{7,47}.

We reasoned that systematic ('deep') mutagenesis could be an unbiased approach to identify and investigate the toxic species of proteins^{48–50}. A map of which amino acid (AA) changes increase or decrease the toxicity of a protein to a cell should, if sufficiently comprehensive, clarify both the properties of the protein and its in vivo conformational states associated with toxicity⁵¹. The effects of a small number of mutations on TDP-43 toxicity or aggregation have been previously reported^{15,35,52–55}. However, on the basis of a handful of mutations, the relationship between aggregation and toxicity is far from clear.

Here we show by quantifying the effects of >50,000 mutations in the PRD of TDP-43 that increasing hydrophobicity and aggregation strongly reduce the toxicity of this protein in yeast. Moreover, mutations that increase the toxicity of TDP-43 actually

promote the formation of dynamic liquid-like cytoplasmic condensates. Mutations have their strongest effects in a central 'hotspot' region of the PRD TDP-43. The patterns of genetic interactions in double mutants in this region reveal that this 'unstructured' region is actually structured in vivo. Our results illustrate how deep mutagenesis can be used to probe the sequence-function relationships and the in vivo structures of 'disordered' proteins. We propose that aggregation of TDP-43 is not harmful but actually protects cells, most likely by titrating protein from a toxic liquid-like phase.

Results

Deep mutagenesis of the TDP-43 prion-like domain. We used error-prone oligonucleotide synthesis to comprehensively mutate the PRD of TDP-43. We introduced the library into *Saccharomyces cerevisiae* cells, induced expression and used deep sequencing before and after induction to quantify the relative effects of each variant on growth in three biological replicates (Fig. 1a). After quality control and filtering (Supplementary Fig. 1a and c), the dataset quantifies the relative toxicity of 1,266 single and 56,730 double amino acid (AA) changes in the PRD with high reproducibility (Fig. 1b, Supplementary Fig. 1d and e). The toxicity scores also correlate very well with the toxicity of the same variants retested in the absence of competition (Fig. 1c).

The toxicity of both single and double mutants has a tri-modal distribution (Fig. 1d, Supplementary Fig. 2a and c), with 18,023 variants more toxic and 16,152 variants less toxic than wild-type (WT) TDP-43 (*t*-test false discovery rate, FDR = 0.05). The dataset therefore allows us to investigate how mutations both increase and decrease toxicity. Very interestingly, ALS TDP-43 mutations increase toxicity, with a strong bias towards moderate effects (*t*-test, *p*-value = 0.005) (Fig. 1d, Supplementary Fig. 2d).

Mutation effects are largest in a central hotspot of the PRD.

Plotting the mean toxicity of all mutations at each position in the sequence reveals a 31 AA hotspot (312–342) where the effects of mutations are strongest (Fig. 1e). The variance in toxicity per position is also the highest within this hotspot, with mutations both strongly increasing and decreasing toxicity (Fig. 1e). A heatmap of the toxicity of all of the single mutations also clearly reveals this hotspot, with most mutations of strong positive or negative effect falling within this 31 AA window (Fig. 1f). Equally strikingly, mutations to the same AA but in different positions within the hotspot often have very similar effects (Fig. 1f). In particular, mutations to charged and polar residues increase toxicity throughout the hotspot and mutations to hydrophobic AAs decrease toxicity (Fig. 1f).

Hydrophobicity and aggregation potential predict toxicity.

To more systematically identify features associated with changes in toxicity we made use of all 53,468 variants carrying one or two AA substitutions (excluding STOP codon variants). We used principal components analysis (PCA) to reduce the redundancy in a list of over 350 AA physicochemical properties (Supplementary Fig. 3) and linear regression to quantify how well changes in these physicochemical properties predict changes in the toxicity of TDP-43. A principal component very strongly related to hydrophobicity is the most predictive feature of toxicity, explaining 66% of the variance in toxicity of all 8,040 mutants within the 312–342 hotspot and 51% of the variance in toxicity of all genotypes (Fig. 2a). With the same approach, we tested the performance of established predictors of protein aggregation, intrinsic disorder and other properties. None of them are as predictive as hydrophobicity (Fig. 2b). Importantly, after controlling for hydrophobicity, additional features such as

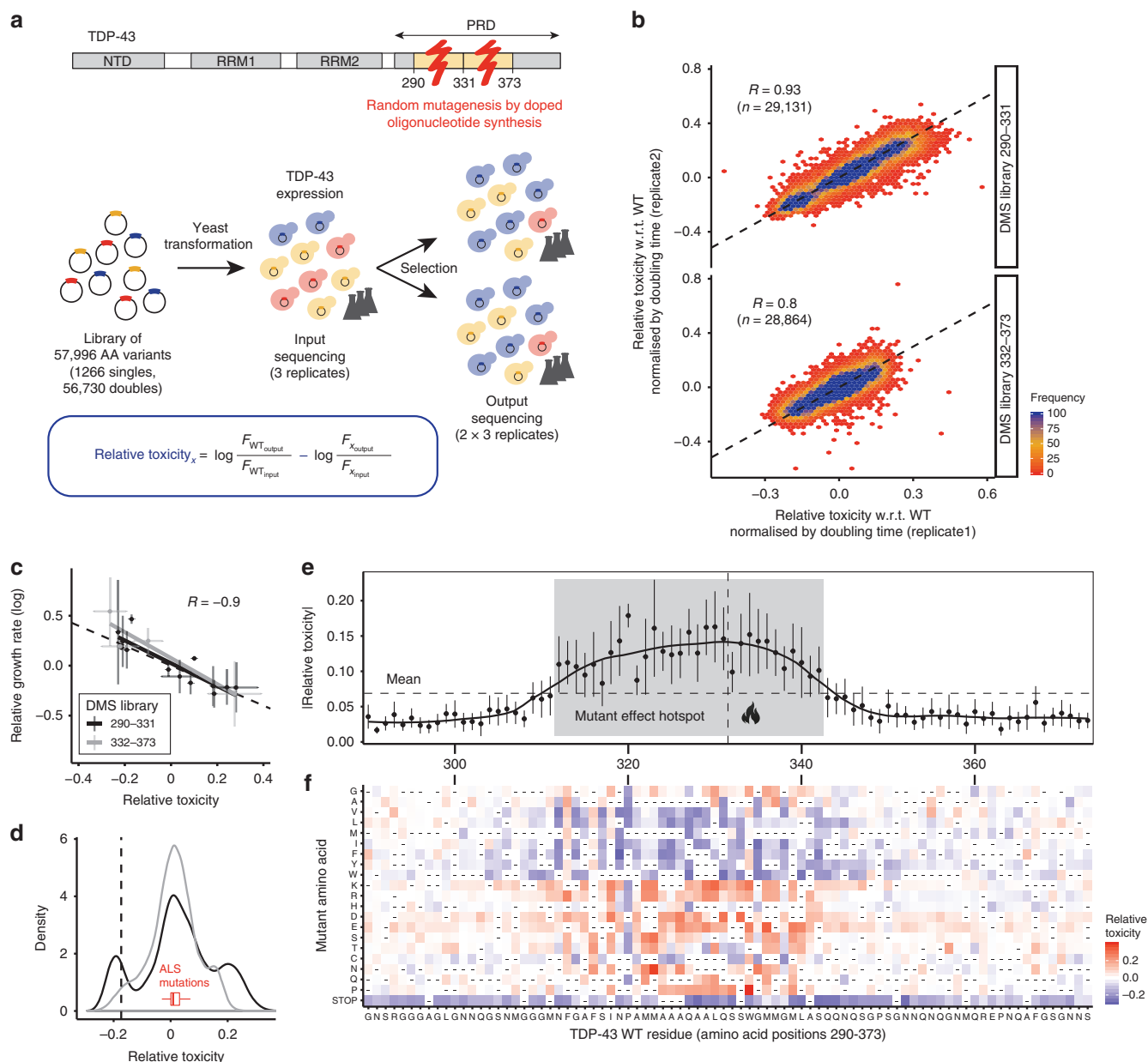


Fig. 1 Deep mutational scanning (DMS) of the prion-like domain (PRD) of TDP-43. **a** Domain structure of TDP-43 and DMS experimental protocol: For each library, three independent selection experiments were performed. In each experiment one input culture was split into two cultures for selection upon induction of TDP-43 expression (6 outputs total). Relative toxicity of variants was calculated from changes of output to input frequencies relative to WT. **b** Correlation of toxicity estimates between replicates 1 and 2 for single and double amino acid (AA) mutants shown separately for each library (290–332; 332–373). The Pearson correlation coefficients (*R*) are indicated. Toxicity correlations between all replicates are shown in Supplementary Fig. 1d, e. **c** Comparison of toxicity from pooled selections and individually measured growth rates for selected variants. Vertical and horizontal error bars indicate 95% confidence intervals of mean growth rates and toxicity estimates respectively. Linear fits of the data are shown separately for each library and Pearson correlation (*R*) after pooling data from both libraries is indicated. **d** Toxicity distribution of single and double mutants, shown separately for each library (colour key as in panel **c**). WT variant has toxicity of zero, mean toxicity of variants with single STOP codon mutation is indicated by dashed vertical line. The red boxplot depicts the distribution of toxicity estimates for all human disease mutations (including sporadic and familial ALS mutations). Outliers are not depicted but are reported in Supplementary Fig. 2d, e. **e** Absolute toxicity of single mutants stratified by position. Error bars indicate 95% confidence intervals of mean (per-position) toxicity estimates. A local polynomial regression (loess) over toxicity estimates of all single mutants is shown. The vertical dashed line indicates the boundary between the two DMS libraries. The horizontal dashed line indicates the mean absolute toxicity of all single mutants. The mutant effect “hotspot” (mean per-position |toxicity| > mean |toxicity|) is highlighted in grey. **f** Heatmap showing single mutant toxicity estimates. The vertical axis indicates the identity of the substituted (mutant) AA. Heatmap cells of variants not present in the library are denoted by “-”

charge and aromaticity do not predict toxicity (Fig. 2d, e, Supplementary Fig. 4a) with aggregation potential accounting for an additional 4% of variance in the hotspot (Fig. 2f, g).

That increased hydrophobicity and aggregation potential are strongly associated with reduced toxicity across >50,000 genotypes

was unexpected given previous work that reported an increased number of intracellular aggregates for a set of TDP-43 variants toxic to yeast⁵⁴ and the widely-held view that aggregation is harmful to cells^{42,52,56}. We therefore further investigated the effects of mutants that alter the hydrophobicity and toxicity of TDP-43.

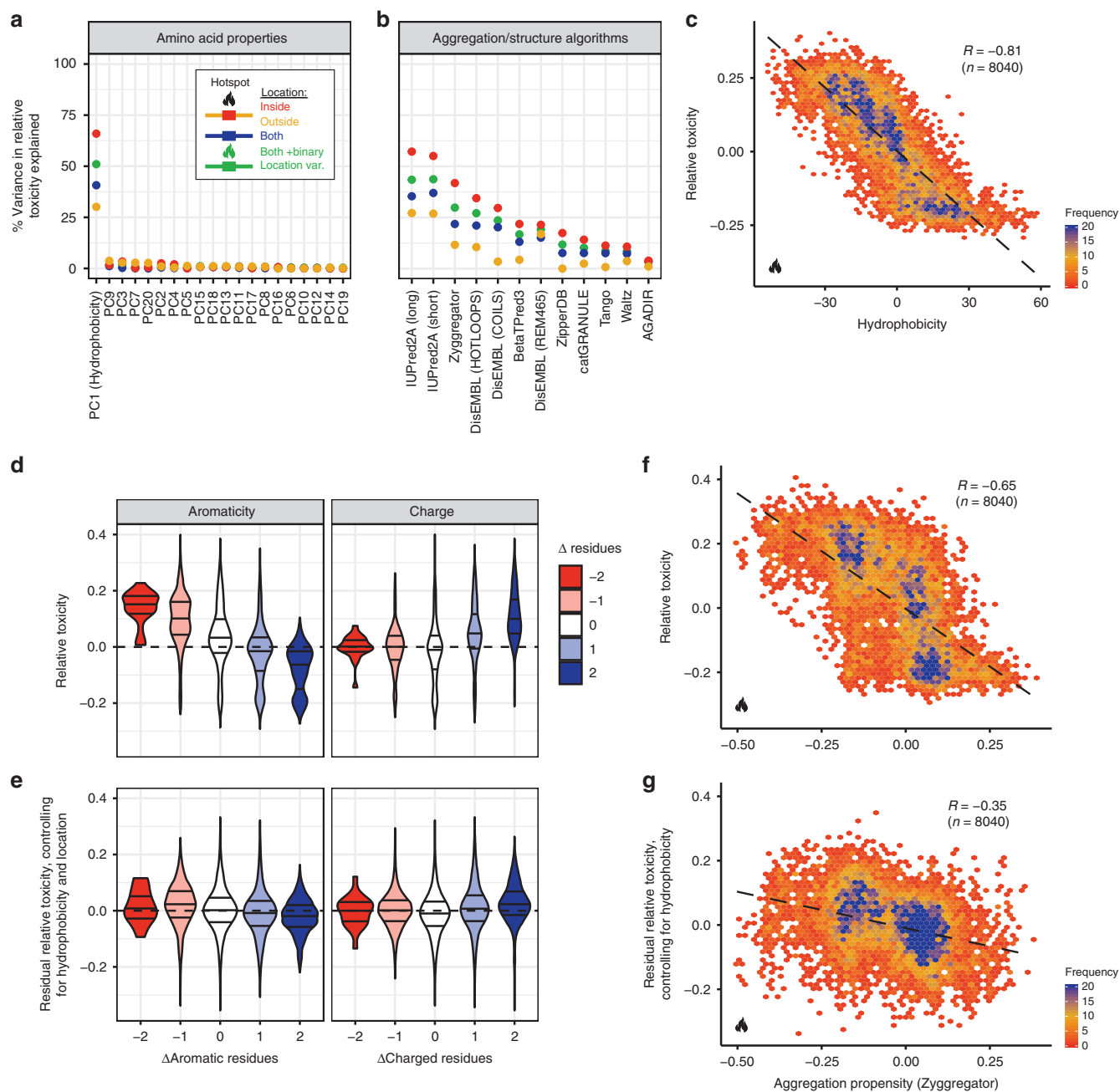
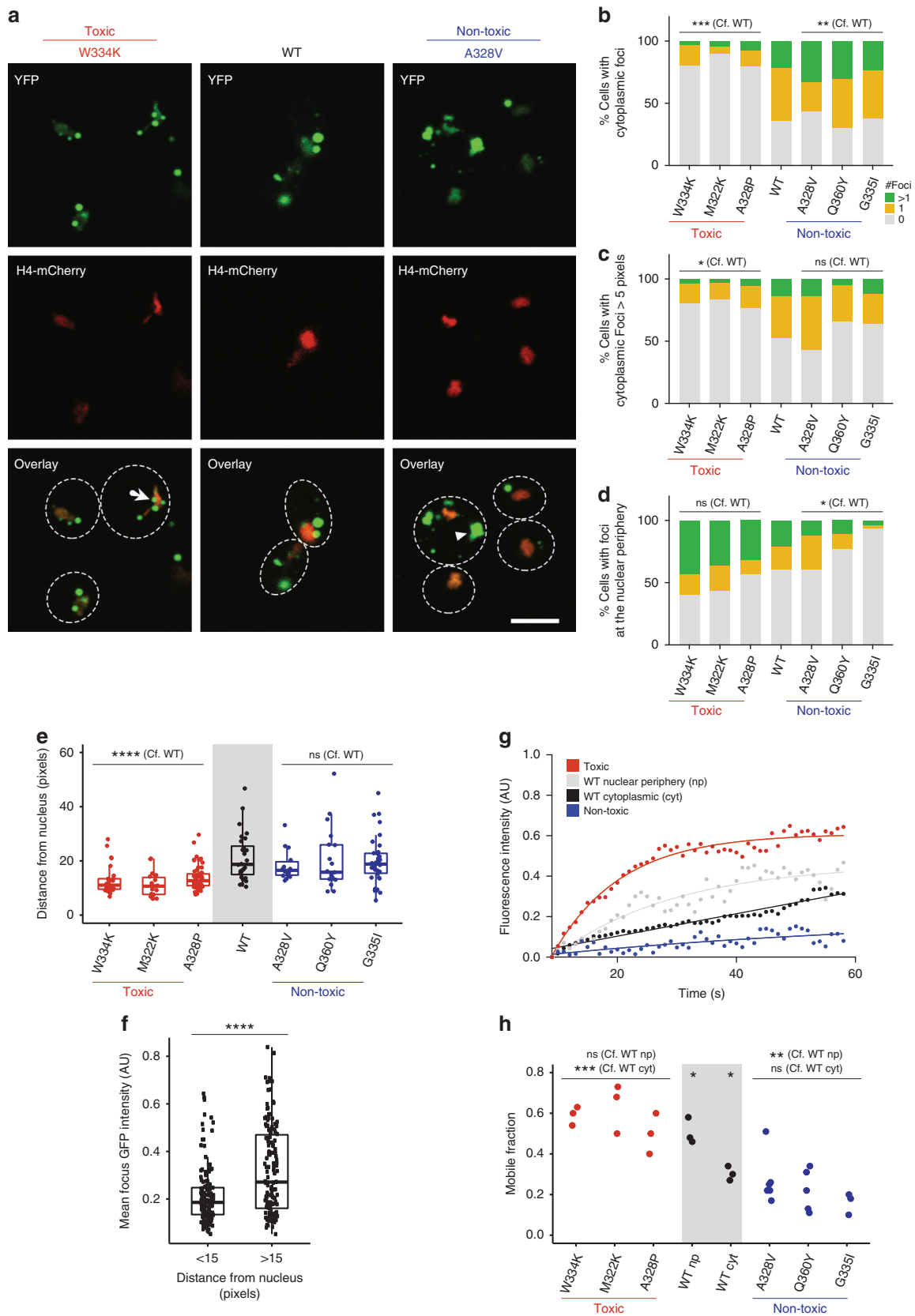


Fig. 2 Changes in hydrophobicity are highly predictive of TDP-43 cellular toxicity. **a** Percentage variance of toxicity explained by linear regression models predicting single and double mutant variant toxicity from changes in AA properties upon mutation (PCs, principal components of a collection of AA physico-chemical properties). Different regression models were built for different subsets of the data. Simple linear regression models for all variants (blue) or only variants inside (red) or outside (yellow) the hotspot region. And a regression model using all variants and including a binary location variable (inside/outside hotspot) as well as an interaction term between binary location variable and the indicated AA property feature (green). **b** Percentage variance of toxicity explained by linear regression models predicting variant toxicity using scores from aggregation/structure algorithms (see Methods). Colour key shown in panel (a). See also Supplementary Fig. 4. **c** Toxicity of variants with single or double mutations within the hotspot region as a function of hydrophobicity changes (PC1) induced by mutation. The Pearson correlation (R) before binning is indicated. See also Supplementary Fig. 9a. **d** Toxicity distributions of single and double mutants stratified by the change in the number of aromatic (H,F,W,Y,V) or charged residues (R,D,E,K) relative to the WT sequence. Horizontal axis as in panel (e). **e** Distribution of residual toxicity after controlling for the effect of hydrophobicity and location on toxicity (green regression model in panel a) stratified by the number of aromatic (H,F,W,Y,V) or charged (R,D,E,K) AAs. **f** Single and double mutant variant toxicity as a function of changes in aggregation propensity (Zygggregator). Only variants occurring within the toxicity hotspot are depicted. The Pearson correlation (R) before binning is indicated. **g** Toxicity as a function of aggregation propensity after controlling for hydrophobicity (red regression model in panel a). Only variants occurring within the toxicity hotspot are depicted. The Pearson correlation (R) before binning is indicated. See also Supplementary Fig. 9b



Two classes of cytoplasmic TDP-43 foci. WT TDP-43 localizes to both the nucleus and to the cytoplasm of yeast cells^{54,55} (Fig. 3a). In the nucleus, TDP-43 is diffuse, but in the cytoplasm it forms *puncta*, consistent with previous observations^{41,57}. We observe that cytoplasmic WT TDP-43 forms two types of

assemblies: small foci in the nuclear periphery and larger foci detached from the nucleus (Fig. 3a, c). We find that mutations that decrease TDP-43 hydrophobicity and increase TDP-43 toxicity increase the number of the small foci at the nuclear periphery and reduce the number of large distal foci (Fig. 3b, c, f,

Fig. 3 Mutations leading to formation of solid-like aggregates rescue toxicity. **a** Representative fluorescence microscopy images of yeast cells expressing indicated YFP-tagged TDP-43 variants (W334K TDP-43 = toxic, A328V TDP-43 = non-toxic). H4-mCherry marks nuclei (red). Contrast was enhanced equally for the green and red channels in all images. **b** Percentage of cells with cytoplasmic foci (Cells scored: $n[\text{toxic}] = 219$, $n[\text{WT}] = 30$, $n[\text{non-toxic}] = 213$). Fisher's Exact test. **c** Percentage of cells with cytoplasmic foci with size over 5 pixels automatically detected by CellProfiler. Fisher's Exact test. (Cells scored: $n[\text{toxic}] = 167$, $n[\text{WT}] = 23$, $n[\text{non-toxic}] = 167$). **d** Percentage of cells with foci at the nuclear periphery (Cells scored: $n[\text{toxic}] = 219$, $n[\text{WT}] = 30$, $n[\text{non-toxic}] = 213$). Fisher's exact test. **e** Distance of foci from nucleus center for toxic (red), non-toxic (blue), and WT (black) TDP-43. Boxplots represent median values, interquartile ranges and Tukey whiskers with individual data points superimposed. Kruskal Wallis with Dunn's multiple comparisons test ($n = >20$ foci/variant). **f** Average fluorescence intensity of foci localized closer (<15 pixels, $n=147$) or further (>15 pixels, $n = 138$) from the nucleus. Boxplots represent median values, interquartile ranges and Tukey whiskers with individual data points superimposed. Mann-Whitney test. **g** Representative individual fluorescence recovery traces for variants reported in panel (e). Lines are the result of a single exponential fitting. **h** Mobile Fraction as calculated by fitting FRAP traces for toxic (red), non-toxic (blue) and WT (black) TDP-43. Each point results from fitting an individual trace. One-way ANOVA with Tukey's multiple comparisons test. Images were taken on cells growing from at least 3 independent starting colonies. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$. Scale bar = 5 μM . Source data are provided as a Source Data file

Supplementary Fig. 5a). TDP-43 mutations reported in ALS (Supplementary Fig. 2e) also increase the number of foci at the nuclear periphery compared to WT TDP-43 (Supplementary Fig. 6a, b). In contrast, mutations that increase hydrophobicity and reduce toxicity reduce the number of small nucleus-associated foci and increase the number of large distal foci (Fig. 3b, c, f, Supplementary Fig. 5a).

Toxic mutations promote dynamic liquid-like condensates. We used fluorescence recovery after photobleaching (FRAP) to characterize the dynamics of TDP-43 variants in the different foci. The large cytoplasmic foci formed by non-toxic variants show little exchange of TDP-43 molecules with the soluble cytoplasmic pool. In contrast, the small foci localized at the nuclear periphery can exchange more protein with the cytoplasm, consistent with a more liquid-like state (Fig. 3d, e). Such differences in dynamics have been described also for distinct types of misfolded protein compartments⁵⁸. Both types of compartments co-localize with the yeast chaperone Hsp104 (Supplementary Fig. 7a). The large immobile TDP-43 foci are also brighter than the small dynamic ones (Fig. 3g), similar to what has been observed for Huntingtin variants that partition between immobile bright assemblies and liquid-like dimmer ones⁵⁹. The non-toxic TDP-43 variants also have a higher protein concentration quantified by Western blotting (Supplementary Fig. 5b).

Taken together, these results suggest that mutations that increase the hydrophobicity of TDP-43 result in a re-localization of the protein away from small and dynamic, liquid-like foci at the nuclear periphery to large and more solid aggregates in the cytoplasm. A reduction in hydrophobicity has the opposite effect.

Genetic interactions reveal the hotspot structure in vivo. The hotspot region of the TDP-43 PRD (AA 312–342) is a conserved region^{35,36}, with hydrophobicity more similar to the globular domains of TDP-43 than to the surrounding hydrophilic disordered regions (Fig. 4b). The hotspot is contained within a region (311–360) that was previously shown to be sufficient for both in vitro aggregation and the formation of cytoplasmic foci³⁵. Fragments from within this region have previously been shown to have the potential to form different types of secondary structures in vitro. More specifically, nuclear magnetic resonance (NMR) spectroscopy of the PRD revealed that residues 321–342 can adopt an α -helical structure in certain conditions^{35,36,47} and four different 6–11 AA peptides from the region could form cross- β amyloid or amyloid-like fibrils whose structures were determined by X-ray crystallography⁵². However, it is unknown whether any of these structures exist in vivo for full-length TDP-43.

We have shown recently that the pattern of genetic (epistatic) interactions between mutations in a protein can report on the

secondary structure of that molecule when it is performing the function that is being selected for^{51,60}. In particular, when a sequence forms an α -helix, the side chains of residues separated by 3–4 AA are close in space and similarly oriented so that mutations in these AA interact similarly with mutations in the rest of the protein. In contrast, in a β -strand, the side chains of residues separated by 2 AA are close and similarly oriented and so make similar genetic interactions with other mutations (Fig. 4a)⁶¹.

We used the 52,272 double mutants (excluding STOP codon variants) in our dataset to identify pairs of mutations that genetically interact. We first identified pairs of mutations that had unexpectedly high or low toxicity (<5 th and >95 th percentile of the expected toxicity distribution, negative and positive epistasis for growth rate, respectively). We then quantified the similarity of epistasis enrichment profiles between pairs of positions and compared these patterns to those expected for α -helices and β -strands, scoring significance by randomization⁵¹ (Fig. 4a).

This revealed that the patterns of epistasis in our dataset are consistent with two secondary structure elements forming inside the PRD in vivo: a β -strand at residues 311–316 and an α -helix at residues 324–331 (Fig. 4c). The β -strand identified by the epistasis analysis coincides with one of the peptides in the TDP-43 PRD that, in vitro, can form cross- β structures⁵² typical of protein aggregates (Fig. 4d). The crystals of this specific peptide consist of a non-conventional β -strand termed a low-complexity aromatic-rich kinked segment (LARKS)⁶². In this in vitro structure, Phe 313 and Phe 316 face the same side of the sheet, whereas in a canonical sheet the side chains of odd and even residues face opposite sides. Strikingly, this non-canonical contact between Phe 313 and Phe 316 is also identified by the in vivo epistasis analysis, with a similarity in interaction profile ranking amongst the top two residue pairs in this region. In addition, the contact between Phe 316 and Ala 315, which again is compatible with a LARKS but not with a canonical β -strand has the highest predicted contact score among neighbouring residues (Fig. 4d). The predicted contact map built on the basis of in vivo epistatic interactions strikingly matches the Protein Data Bank (PDB) structure for LARKS 312–317 (Fig. 4d, Supplementary Fig. 8).

On the other hand, the genetic interactions of mutations in the 324–330 region match those expected for an α -helix (Fig. 4e). This region is part of the portion (321–342) of TDP-43 that can transiently and cooperatively fold into an α -helix in vitro^{36,47,63}. This helix is stabilized by inter-molecular contacts and its self-interaction was proposed to seed liquid-demixing in vitro. Amyloid fibrils can grow from the liquid de-mixed state and circular dichroism spectroscopy revealed that the helix can transition to a β -sheet over time, compatible with the process of aggregation^{35,63}. On the basis of epistasis, the top scoring predicted contacts in this region are between residues separated

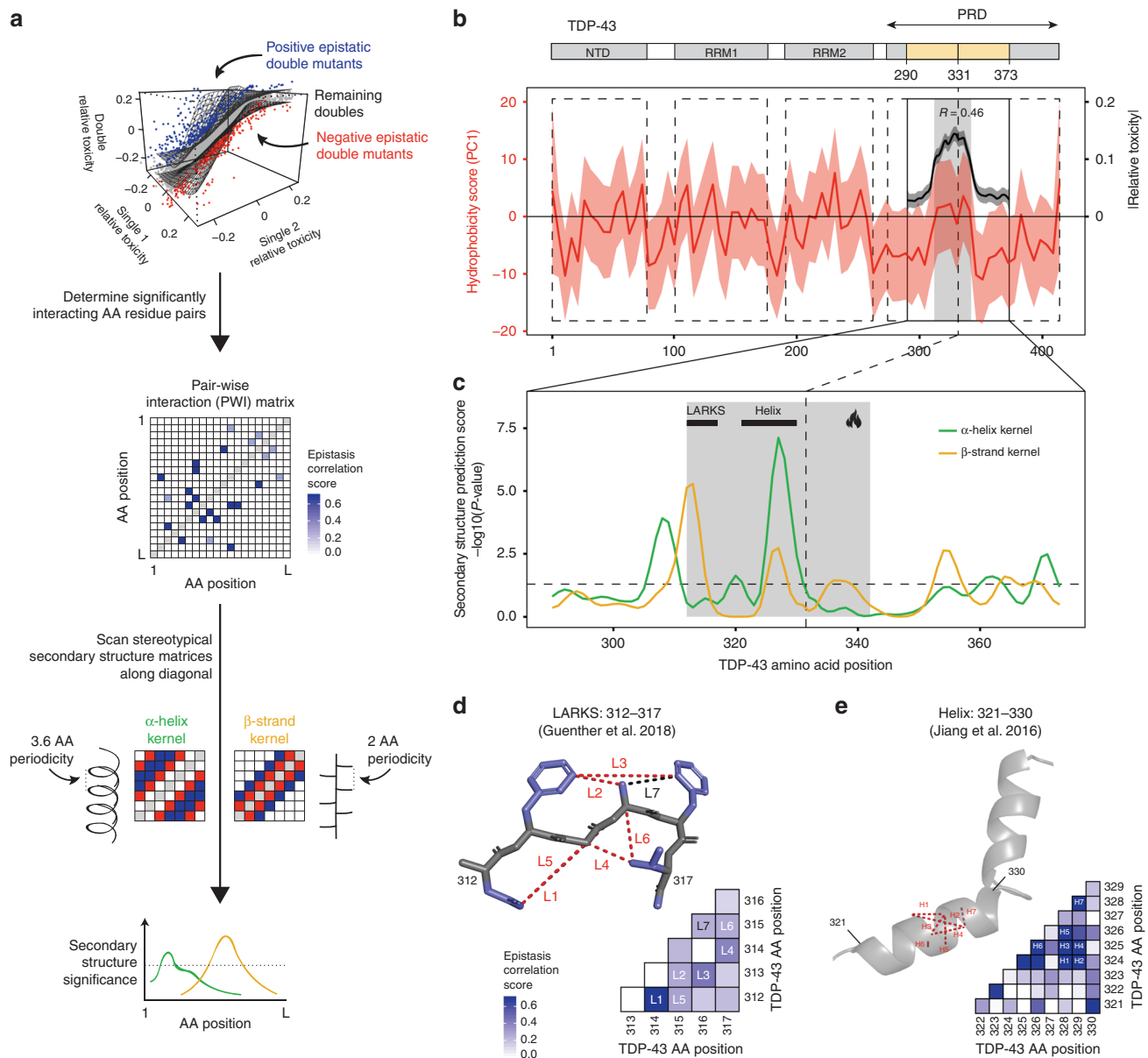


Fig. 4 Correlated patterns of epistasis predict secondary structural elements within the PRD of TDP-43. **a** Schematic representation of the computational strategy to identify *in vivo* secondary structures. Double mutant variants are classified as epistatic if they are more (95th percentile) or less (5th percentile) toxic than other variants with similar single mutant toxicities (top). A pair-wise interaction (PWI) matrix of epistasis correlation scores is then constructed by quantifying the similarity of a pair of positions' interactions with all other mutated positions in the protein. The epistasis correlation scores along the diagonal of the PWI matrix are then tested for agreement with the stereotypical periodicity of α -helix and β -strand, using two-dimensional kernels (bottom), to calculate the likelihood of adjacent positions forming secondary structures. **b** Local polynomial regression (loess) of hydrophobicity (PC1) of the WT TDP-43 sequence with 95% confidence interval. For reference, smoothed toxicity estimates in the mutated positions within the PRD are shown. The Pearson correlation coefficient (R) between hydrophobicity and mean toxicity effects of single mutants at each position before smoothing is indicated. **c** Secondary structure predictions from epistasis correlation scores for α -helix and β -strand kernels based on the strategy described in panel a. Black bars annotate previously described structural features: LARKS, low-complexity aromatic-rich kinked segment (312–317)⁵²; Helix (321–330)³⁵. The dashed horizontal line indicates the nominal significance threshold $P = 0.05$. **d** Epistatic interactions in region 312–317 are consistent with positions of similar side-chain orientations interacting in a previously reported *in vitro* LARKS structure. Epistasis correlation matrix and top seven epistasis correlation score interactions annotated on the LARKS reference structure (monomer from PDB entry 5whn, <https://www.rcsb.org/structure/5WHN>). Dashed lines on structure connect interacting residues at minimal distance between side chain heavy atoms. Side chain atoms are depicted in blue. **e** Epistatic interactions in region 321–330 are consistent with positions of similar side-chain orientations interacting in an α -helix. Epistasis correlation matrix and top seven epistasis correlation score interactions annotated on the Helix reference structure (monomer from PDB entry 5whn, <https://www.rcsb.org/structure/5WHN>)

by 3–4 AA such as Ala 324 and Ala 328, or Ala 325 and Ala 328, consistent with interactions between side chains of an α -helix (Fig. 4e).

The pattern of *in vivo* epistatic interactions between mutations in TDP-43 therefore is compatible with a model in which two of the secondary structures that have previously been observed

in vitro for fragments of TDP-43 actually form in vivo in the full-length protein.

Discussion

Specific protein aggregates have long been recognized as the hallmarks of many neurodegenerative diseases^{4–6,52,64}. However, whether these aggregates are the cause of these diseases, non-pathological by-products, or a protective mechanism is still very unclear and hotly debated^{13–16}. Indeed, although it is often assumed to be the case, it is not even clear whether aggregates are the cause of toxicity when aggregating proteins are expressed in simpler cellular systems^{54,55}. We reasoned that deep mutagenesis might be an effective approach to resolve this question.

In this study, we have tested this approach using the ALS protein TDP-43 that both aggregates and causes toxicity in the model eukaryote, *S. cerevisiae*. Quantifying the effects of >50,000 mutants of TDP-43 revealed unequivocally that increasing the hydrophobicity and aggregation of TDP-43 strongly reduces the toxicity of this protein in yeast cells. Consistently, mutations that reduce hydrophobicity and the aggregation potential of TDP-43 increase the toxicity of the protein. Although they reduced the formation of large, solid aggregates, mutations that increase toxicity promote the formation of alternative foci—dynamic, liquid-like TDP-43 condensates clustered at the nuclear periphery. We propose therefore that aggregation reduces the toxicity of TDP-43 to yeast cells because it titrates TDP-43 away from this toxic liquid-like phase (Fig. 5a).

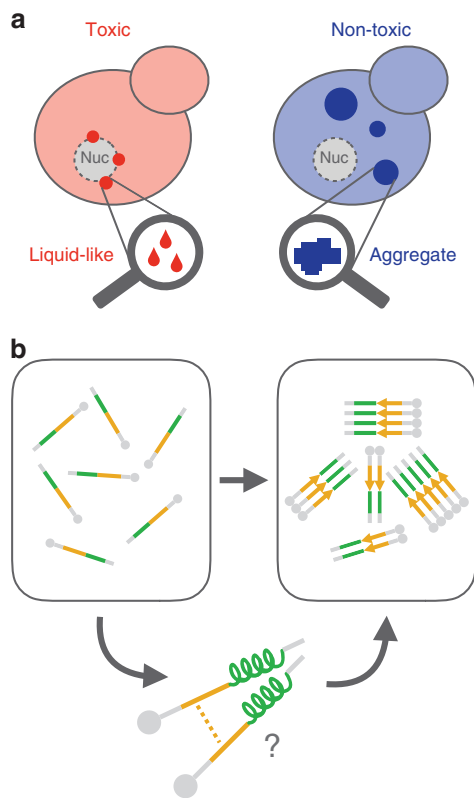


Fig. 5 Model of how AA changes determine toxicity of TDP-43. **a** Mutations that promote formation of insoluble cytoplasmic aggregates decrease TDP-43 toxicity, while mutations that cause the protein to stall in a liquid de-mixed phase increase its toxicity to the cell. **b** Secondary structure elements, within the toxicity hotspot 312–342, promote the aggregation process of TDP-43, with a transient helix forming on pathway to β -rich aggregates

That TDP-43 aggregates are protective rather than toxic is consistent with previous work in multiple systems, including the rescue of toxicity by the accumulation of RNA lariats that sequester TDP-43 into large aggregates⁶⁵. Moreover, in mammalian cells, liquid de-mixed TDP-43 was recently shown to recruit the nuclear pore component Nup62 and the importin- α transporter, resulting in nuclear transport impairment and toxicity⁴⁴. Thus, although it still remains to be established whether aggregation of TDP-43 is also protective in mammalian cells and neurons, it seems likely that this will be the case. The observation that all recurrent fALS mutations increase the toxicity of TDP-43 in yeast and by a similar magnitude (Supplementary Fig. 2d) is very striking and suggests that the yeast system may indeed capture molecular mechanisms relevant to the human disease. Indeed, given the late age of onset of ALS, it is particularly interesting that the fALS mutations are all moderate effect mutations when expressed in yeast, as it may be the case that the more toxic variants of TDP-43 are embryonic lethal in humans.

More generally, our results demonstrate that deep mutagenesis is a powerful approach for determining the sequence-function relationships of intrinsically disordered proteins, including probing their in vivo structures. Mutations had their strongest effects within a central hotspot region of the TDP-43 PRD. Our recently developed approach⁵¹ that uses the patterns of genetic interactions in double mutants to report on structural contacts reveals that this ‘unstructured’ hotspot region is very likely to be structured in vivo with the formation of these secondary structures altering the toxicity of the protein. Indeed, secondary structure elements within this region have been shown to be important for the phase separation and aggregation of fragments of TDP-43 in vitro^{35,36,52}. A parsimonious model based on previous in vitro work^{35,36,47} is that the helix forms first in the pathway of aggregation towards a β -rich species (Fig. 5b). Consistent with this, destabilizing mutations, such as any substitution of Phe 313 and Phe 316 in the LARKS, or the introduction of proline into the 324–330 helix, increase toxicity (Fig. 1f).

The conformations of ‘unstructured’ proteins are notoriously difficult to study and the interactions between mutations in double mutants provide a general method to probe the in vivo structures of these proteins whenever a selection assay is available. We envisage that this approach can be adopted to study the functions, toxicity, and in vivo structures of other intrinsically disordered proteins, including the many other proteins implicated in neurodegenerative diseases.

Our conclusions derived from deep mutagenesis of TDP-43 are also consistent with observations for other genes, such as the reduced toxicity of SOD-1 variants that increase aggregation^{16,66} and the increased survival of neurons containing Huntingtin inclusion bodies⁶⁷. They are also consistent with increasing evidence that insoluble aggregates are not pathogenic in multiple other neurodegenerative diseases^{64,68,69}, and with the clinical failure of therapeutic approaches that reduce the occurrence of aggregates^{10,12,70–72}.

Indeed, if insoluble aggregates generally titrate proteins away from alternative toxic phases, interactions and functions, then promoting rather than alleviating aggregation might be the more appropriate therapeutic goal in neurodegenerative diseases.

Methods

Yeast strains and plasmids. *Saccharomyces cerevisiae* S288C BY4741 (*MATa his3 Δ 1 leu2 Δ 0 met15 Δ 0 ura3 Δ 0*) was used in all experiments. Plasmid pRS416 containing TDP-43 or TDP-43-YFP under control of the Gal1 promoter was purchased from Addgene⁵⁴. Mutagenesis for the characterization of TDP-43 variants was performed through PCR linearization with specifically-designed primers (Supplementary Data 1, primers: BB_1 to BB_6). The resulting products were then either treated with DpnI or purified from a 1% agarose gel with a QIAquick Gel Extraction Kit (Qiagen) and transformed into *E. coli* DH5 α competent cells

(Invitrogen) for plasmid purification and validation through Sanger sequencing. The plasmid used in the co-localization assays contains RNQ1-mCherry under control of the Gal1 promoter was a kind gift from the Rick Gardner lab. Genes coding for the other proteins for which co-localization was tested were cloned in this plasmid by gap-repair.

Library construction. Two 186 nt oligonucleotides were purchased from TriLink. Each consisted of a 'doped' region of 126 nt, corresponding to TDP-43 AA 290–331 or AA 332–373, flanked by 30 nt of the WT TDP-43 sequence on each side. Each position in the mutated area, was doped with an error rate of 1.59%. The target frequency for each library was 27.0% for single mutants and 27.3% for double mutants. With this approach, the WT sequence was represented with a frequency of 13.3%. Although a barcoding strategy⁷³ could have improved the robustness of sequencing reads, we estimated that the impact of misreads due to the direct sequencing approach here employed would sum up to less than two additional counts per double nucleotide variant attributable to sequencing error (see Variants Toxicity and Error Estimates). Each oligonucleotide was amplified by PCR (Q5 High-Fidelity DNA Polymerase, NEB) for 15 cycles, purified using an E-gel electrophoresis system (Agarose 2%) followed by column purification with a MinElute PCR Purification Kit (Qiagen). In order to introduce the doped sequence in the full-length TDP-43 sequence, the purified oligonucleotide was cloned into 100 ng of linearized pRS416 Gal TDP-43 by a Gibson approach (Supplementary Data 1, primers BB_7 to BB_10). The product was then transformed into 10-beta Electrocompetent *E. coli* (NEB), by electroporation in a Bio-Rad GenePulser machine (2.0 kV, 200 Ω , 25 μ F). Cells were recovered in SOC medium (NEB) for 30 min and plated on LB with ampicillin. A total of $\sim 2.7 \times 10^6$ transformants were estimated. The plasmid library was purified with a GeneJET Plasmid Midiprep Kit (Thermo Scientific).

Yeast transformation and selection experiments. Yeast cells were transformed with the TDP-43 doped plasmid in 4 independent biological replicates for each library. One single colony was grown overnight in 30 ml YPDA medium at 30 °C for each replica. Cells were diluted to 0.3 optical density at a wavelength of 600 nm (OD600) in 175 ml of YPDA and incubated for 4 h at 30 °C. Cells were then harvested, washed, re-suspended in 8.575 mL SORB (100 mM LiOAc, 10 mM Tris pH 8.0, 1 mM EDTA, 1 M sorbitol) and incubated for 30 min at room temperature. For the transformation, 10 mg per mL of salmon sperm DNA and 3.5 μ g TDP-43 plasmid library were used. Cells were mixed to 100 mM LiOAc, 10 mM Tris-HCl pH 8.0, 1 mM EDTA/NaOH pH 8.0 and 40% PEG 3350. Heat-shock was performed for 20 min at 42 °C. YPD with 0.5 M sorbitol was used to recover the cells, incubating them for 1 h at 30 °C. After recovery, cells were resuspended in SC-URA 2% raffinose medium, while an aliquote was plated to calculate transformation efficiency.

After ~ 50 h of growth, cells were diluted in SC-URA 2% raffinose medium and grown for 4.5 generations. At this stage, 400 mL of each replica were harvested, washed, split into two tubes and frozen at -20 °C for later extraction of input DNA. To induce plasmid expression, for each replicate two cultures were diluted in SC-URA 2% galactose medium. After 5–6 generations, 2×400 mL for each replicate were harvested to obtain output pellets for DNA extraction.

DNA extraction and library preparation. Input and Output pellets were resuspended in 1.5 mL extraction buffer (2% Triton-X, 1% SDS, 100 mM NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0). Two cycles of freezing in an ethanol-ice bath and heating at 62 °C were performed. Deproteinization was performed using 25:24:1 phenol-chloroform-isoamyl alcohol and glass beads. After centrifugation, the aqueous phase, containing the DNA, was recovered and treated again with phenol-chloroform-isoamyl alcohol. The samples were incubated for 30 min at -20 °C with 1:10 V 3 M NaOAc and 2.2 V 100% ethanol for DNA precipitation. At this stage and after centrifugation for 30 min, the pellets were dried overnight at room temperature. RNA was eliminated by incubation with RNase 10 mg per mL for 30 min at 37 °C. DNA purification was achieved with a QIAEX II Gel Extraction Kit (Qiagen) and DNA was eluted in 375 μ L of elution buffer. DNA concentration was measured by q-PCR, with primers annealing to the Ori site of the pRS416 plasmid (Supplementary Data 1, primers BB_11, BB_12).

The TDP-43 library was then prepared for deep sequencing by PCR amplification in two steps using Q5 High-Fidelity DNA Polymerase (NEB). In step 1, 300 million plasmids were amplified for 15 cycles using frame-shifted adaptor primers with partial homology to standard Illumina sequencing primers (Supplementary Data 1, primers BB_13 to BB_47). Samples were treated with ExoSAP (Affymetrix) and purified with QIAEX II kit (Qiagen). PCR products from the first step were used as templates in the second PCR step, where indexed Illumina primers (Supplementary Data 1, primers TS_HT_D7X_7 to TS_HT_D7X_95) were used for a 10 cycles amplification. DNA concentration was then quantified by means of a Quant-iT[™] PicoGreen[®] dsDNA Assay Kit (Promega). All replicates were pooled together in an equimolar ratio. Finally, the pooled sequencing library was run on a 2% agarose gel, purified and sent for 125 base-pair (bp) paired-end Illumina sequencing at the CRG Genomics Unit.

Individual growth rate measurements. Yeast cells expressing selected TDP-43 variants were grown overnight in SC-URA 2% raffinose non-inducing medium and diluted to 0.2 OD600 until exponential phase. Then they were diluted to 0.1 OD600 in SC-URA 2% galactose to assess growth in inducing conditions. Growth was monitored by measuring OD600 in a 96-well plate at 10 min intervals inside an Infinite M200 PRO microplate reader (Tecan). Plates were kept constantly shaking at 30 °C. Growth curves were fitted in order to extrapolate growth rates that correspond to the maximum slope of the linear range of the LN(OD600) curve over time.

Equipment and settings. Imaging was performed by using a Confocal TCS SP8 and a Confocal TCS SP5 (Leica) equipped with PMT detectors both for fluorescence and transmitted light images. AOBs beam-splitter systems are in place on both instruments. 63X oil immersion objectives and the LAS AF software were used for all imaging. YFP fluorescence was excited with a 488 nm laser, while mCherry fluorescence with a 561 nm laser. Ranges for emission detection were 495–554 and 637–670 nm respectively. Image depth is 8-bit in all cases and pixel size equals 120.4 nm. The LUT is linear and covers the full range of the data.

Fluorescence microscopy and image analysis. Yeast cells expressing TDP-43 selected variants were grown in SC-URA 2% raffinose non-inducing medium and then transferred to SC-URA 2% galactose medium to induce protein expression for 8 h. They were then imaged under a Confocal TCS SP8 microscope (Leica). Counting of foci was conducted both manually and by automated pipelines using the CellProfiler software where quantification of fluorescent intensity was tracked for each focus. The coordinates of the center of each focus and nucleus were derived from CellProfiler and used to calculate distances using a custom R script (pipelines available at https://github.com/lehner-lab/tardbpdms_cellprofiler_scripts).

Fluorescence recovery after photobleaching. Yeast cells expressing TDP-43 selected variants were grown in SC-URA 2% raffinose non-inducing medium and then transferred to SC-URA 2% galactose medium to induce protein expression for 8 h. The cells were immobilized to an 8-well cover slide by Concanavalin-A-mediated cell adhesion. Cells were then imaged under a Confocal TCS SP5 microscope (Leica) where bleaching was achieved with 488 Laser Power at 70% for three frames (1.3 s per frame) while fluorescence recovery was recorded for 50 frames. The curves were then fitted to a single exponential, following normalization, with the EasyFrap package⁷⁴.

Protein extraction and western blotting. Single yeast colonies were grown overnight in non-inducing medium and then diluted to 0.2 OD600 in galactose medium to induce protein expression for ~ 8 h. At this stage, 6×10^7 cells were collected and re-suspended in 200 μ L EtOH and 2.5 μ L PMSF. Samples were vortexed with glass beads for 15 min at 4 °C and frozen overnight at -80 °C. The samples were dried in a speed vacuum for 20 min and resuspended in 200 μ L solubilizing buffer (20 mM Tris HCl pH 6.8, 2% SDS). After boiling for 5 min, the lysate fraction was run on a NuPAGE 4–12% Bis-Tris gels (Novex) and transferred to PVDF membranes in an iBlot (Invitrogen). Membranes were blocked with 5% milk powder in TBS-T and incubated overnight at 4 °C with primary antibodies: anti-GFP mouse antibody (Santa Cruz sc-9996) and anti-PGK1 mouse antibody (Novex 459250) diluted 1:1000 and 1:5000 in 2.5% powder milk respectively. Secondary antibody anti-proteinG was incubated for 1 h at room temperature. Proteins were detected with an enhanced chemi-luminescence system (Millipore Luminata) and visualized using an Amersham Imager 600 (GE Healthcare).

Sequencing data pre-processing. FastQ files from paired-end sequencing of replicate deep mutational scanning (DMS) libraries before ('input') and after selection ('output') were processed using a custom pipeline (<https://github.com/lehner-lab/DiMSum>, manuscript in prep.). DiMSum is an R package that wraps common biological sequence processing tools including FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) (for quality assessment), cutadapt (for demultiplexing and constant region trimming), USEARCH⁷⁵ (for paired-end read alignment) and the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). First, 5' constant regions were trimmed, but read pairs were discarded if 5' constant regions contained more than 20% mismatches to the reference sequence. Read pairs were aligned (reads that did not match the expected 126 bp length were discarded) and Phred base quality scores of aligned positions were calculated using USEARCH. Reads that contained base calls with Phred scores below 30 (290–331 DMS library) or below 25 (332–373 DMS library) were discarded. Approximately five and seven million reads passed these filtering criteria in each sample corresponding to the 290–331 and 332–373 libraries respectively. Finally, unique variants were counted and merged into a single table of variant counts (aggregated across technical output replicates) per DMS library. One out of four input replicates (and all associated output samples) from each DMS library were discarded due to considerably lower correlations with the other replicates (Supplementary Fig. 1a, b).

Variant toxicity and error estimates. All analyses of toxicity were performed on variants with a maximum of two AA mutations, but no synonymous mutations in other codons. Firstly, sample-wise counts for variants identical at the AA level were aggregated. For each replicate selection, relative toxicity of variants was calculated from variant counts in input ($F_{x_{input}}$) and output ($F_{x_{output}}$) samples as

$$\text{Relative toxicity}_x = \text{ES}_{\text{WT}} - \text{ES}_x \quad (1)$$

where $\text{ES}_x = \ln \frac{F_{x_{output}}}{F_{x_{input}}}$ and ES_{WT} represents the WT enrichment score. Uncertainty of toxicity values was estimated as a combination of expected Poisson error based on read counts and error between replicate selections as:

$$\epsilon_x = \sqrt{\frac{1}{F_{x_{input}}} + \frac{1}{F_{x_{output}}} + \frac{1}{F_{\text{WT}_{input}}} + \frac{1}{F_{\text{WT}_{output}}} + \epsilon_r^2}. \quad (2)$$

Here, ϵ_r , the error between replicate selections, is estimated from the variance of toxicity estimates across replicates for variants whose expected count-based Poisson error approaches zero. Toxicity estimates and associated errors per replicate selection were also normalized by the replicate-specific number of cell doublings during selection to yield relative growth rates per generation.

In ‘doped’ variant libraries, individual double mutants are represented less frequently than single mutants or the WT sequence and due to this under-representation toxic double mutants (that are depleted due to slower growth during selection) are often not observed in the output samples (Supplementary Fig. 1c). To calculate toxicity estimates for such double mutants and avoid skewed marginal toxicity distributions due to these drop-out events, we used a Bayesian approach to estimate toxicity of double mutants based on a prior, i.e., toxicity distributions of highly represented doubles that originate from single mutants with similar toxicity estimates⁵¹. These corrected toxicity estimates show improved heteroscedasticity and reduced variance, especially for under-represented double mutants (Supplementary Fig. 1c).

Variant toxicity distributions were first normalized between replicate selections of the same DMS library to have equal standard deviations. Then toxicity estimates of each variant across replicate selections were merged by taking the error-weighted mean across replicate selections. Finally, distributions of merged toxicity estimates from each DMS library were centred on the error-weighted means of toxicity of single codon synonymous (silent) variants in each DMS library and scaled such that the error-weighted means of single STOP codon variants coincided for both DMS libraries (Supplementary Fig. 2a and c). Furthermore, we removed low confidence variants supported by an average of less than ten input reads from all downstream analyses. Merged and normalized toxicity estimates, as well as toxicity estimates from independent replicates before merging and normalisation, are available in Supplementary Data 3 and 4 respectively.

The impact of misreads (i.e. sequencing errors) was evaluated by measuring the per base error frequency in the WT sequence 10 bp upstream and 10 bp downstream of the mutagenized (doped) region. The frequency of an incorrect base call in these regions is 0.0001 ($\text{sd} = 6 \times 10^{-5}$) for the 290–331 library and 0.0004 ($\text{sd} = 4 \times 10^{-4}$) with little variability depending on the wild-type base. By multiplying these frequencies by the length of the doped region we calculated the probability of a misread in each variant (0.0126 for the 290–331 library and 0.0504 for the 332–373 library). Single nucleotide substitutions account for $\sim 2 \times 10^6$ reads in a typical input sample of the 290–331 library, of which we estimate 98.74% to be “true” single nucleotide variants on the basis of a 0.0126 misread probability. Therefore, we estimate an additional 2×10^4 misreads originate from single nucleotide variants ($2 \times 10^4 = 0.0126/0.9874 \times 2 \times 10^6$). In the 126 bp mutagenized region, a total of $7875 \times 3 \times 3 = \sim 7 \times 10^4$ possible double nucleotide variants exist, since each base in each pair can be mutated to one of the three other nucleotides. We therefore estimated that, even in a scenario in which single nucleotide variants are solely distributed among all possible double nucleotide variants, the additional count due to sequencing errors in the 290–331 library would be ~ 0.5 as it follows from the estimated additional 2.6×10^4 misreads over a total of 7×10^4 possible doubles. Similarly, additional counts due to sequencing errors would not reach 2 even in the 332–373 library, where the misread frequency was higher (0.0004).

Linear regression models to predict variant toxicity. We used simple linear regression to predict variant toxicity from (i) a collection of AA property features, (ii) a panel of scores from aggregation/structure algorithms and (iii) location with respect to the toxicity hotspot.

The AA property features were derived from a PCA of a curated collection of numerical indices representing various physicochemical and biochemical properties of AAs (<http://www.genome.jp/aaindex/>). From a total of 539 indices, we retained 379 high confidence indices with no missing values (including five additional indices absent from the original database; see Supplementary Data 2). Results of PCA and selected variable loadings on the normalized matrix are shown in Supplementary Fig. 3. For single mutant variants, AA property feature values represent the difference between the WT and mutant PC scores.

Similarly, aggregation, disorder, structure and other feature values for single mutant variants represent the difference between scores obtained using WT and single mutant AA sequences. AGADIR, *cat*GRANULE and Tango provide a single

score per AA sequence. Unless a single score per AA sequence was provided (i.e. AGADIR, *cat*GRANULE, Tango), individual residue-level scores were summed to obtain a score per AA sequence (i.e. BetaTPred3, DISEMBL, IUPred2A, Waltz, ZipperDB, Zyggregator). The entire PRD AA sequence was supplied to AGADIR and all unique six-mers to ZipperDB. For the remainder, the full-length AA sequence was used.

Variants inside the hotspot were defined as those with mutant residue positions in the range of 312–342. Change in absolute charge (regardless of sign) is shown in Fig. 2d, e, because this feature is more predictive of toxicity than change in charge itself. For double mutant variants, we summed the feature values of the constituent singles for both AA property and aggregation/structure algorithm features. Regression models were built using either (i) all variants, restricting variants to those occurring either (ii) inside or (iii) outside the toxicity hotspot (for double mutants both mutations have to occur either inside or outside the hotspot region), or (iv) including a binary location variable (0: one/all outside, 1: one inside, one outside, 2: one/all inside toxicity hotspot) and a third term indicating the interaction between location and the AA property or aggregation/structure algorithm feature.

Predicting secondary structure from epistasis. Epistasis is the non-independence of mutation effects, i.e., the toxicity of double mutants is different from that expected given the toxicity of their constituent single mutant variants. We have previously shown that epistasis between double mutants can result from structural interactions within proteins and therefore can be used to infer secondary and tertiary structural features^{51,60}. In brief, double mutants were classified as epistatic if they had more extreme toxicity values (below 5th percentile or above 95th percentile) than other double mutants with similar single mutant toxicities, which was estimated from non-parametric surface fits of double mutant toxicity as a function of a two-dimensional single mutant toxicity space (Fig. 4a).

Double mutants close to the lower or upper measurement range limits (where the power to detect significant epistasis is reduced) were excluded from epistasis quantification. We calculated position-pair enrichments for epistatic double mutants resulting in a pair-wise enrichment matrix. Diagonal entries on this matrix were imputed as column-wise mean enrichments. An epistasis correlation score matrix was then derived from this enrichment matrix by calculating the partial correlation of epistasis interaction profiles (columns of the enrichment matrix) between all pairs of positions. The rationale for the correlation score is that structurally close positions within a protein should have similar epistatic interactions with all other positions in the protein. Calculating partial correlations additionally removes transitive interactions and was found to be superior over epistasis enrichments in estimating secondary structures⁵¹.

Secondary structure propensities were calculated by testing for agreement of epistasis correlation score patterns with the stereotypical periodicities of an α -helix and β -strand, using two-dimensional kernels at each position along the diagonal of the epistasis correlation score matrix⁵¹. Significance of secondary structure propensities was assessed by comparison to propensities derived from 10^4 randomized epistasis correlation score matrices.

Similarly, LARKS structure propensities were calculated using PDB-structure derived contact matrices based on a minimal side-chain heavy atom distance of 4.5 Å (Supplementary Fig. 8) for both WT (PDB 5WHN [<https://www.rcsb.org/structure/5WHN>]) and mutant sequences (PDB 5WHP [<https://www.rcsb.org/structure/5WHP>]) and 5WKB [<https://www.rcsb.org/structure/5WKB>]). Contact matrix values were normalised to have zero sum. Association score matrix values were normalised to have mean of zero and unit variance. Significance of LARKS structure propensities was assessed by comparison to propensities derived from 10^4 randomized epistasis correlation score matrices, where randomization was restricted to within-LARKS interactions, i.e., distances compatible with a six-mer.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available from the corresponding author upon request. Raw sequencing data and the processed data table (Supplementary Data 3) have been deposited in NCBI’s Gene Expression Omnibus (GEO) and are accessible through the GEO Series accession number GSE128165. The source data underlying Fig. 3 and Supplementary Figs. 5 and 6 are provided as a Source Data file.

Code availability

All software code and custom scripts are available on GitHub: <https://github.com/lehner-lab/DIMSum> for raw read processing, <https://github.com/lehner-lab/tardbpdms> for all downstream analyses and to produce all figures, and https://github.com/lehner-lab/tardbpdms_cellprofiler_scripts for CellProfiler pipelines.

Received: 20 May 2019 Accepted: 15 August 2019

Published online: 13 September 2019

References

- Buratti, E. Functional Significance of TDP-43 Mutations in Disease. *Adv. Genet.* **91**, 1–53 (2015).
- Ling, S.-C., Polymenidou, M. & Cleveland, D. W. Converging mechanisms in ALS and FTD: disrupted RNA and protein homeostasis. *Neuron* **79**, 416–438 (2013).
- Chiti, F., Stefani, M., Taddei, N., Ramponi, G. & Dobson, C. M. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* **424**, 805–808 (2003).
- Eisenberg, D. & Jucker, M. The amyloid state of proteins in human diseases. *Cell* **148**, 1188–1203 (2012).
- Chiti, F. & Dobson, C. M. Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.* **75**, 333–366 (2006).
- Polymeropoulos, M. H. et al. Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science* **276**, 2045–2047 (1997).
- Arnold, E. S. et al. ALS-linked TDP-43 mutations produce aberrant RNA splicing and adult-onset motor neuron disease without aggregation or loss of nuclear TDP-43. *Proc. Natl Acad. Sci. USA* **110**, E736–E745 (2013).
- Taylor, J. P., Brown, R. H. & Cleveland, D. W. Decoding ALS: from genes to mechanism. *Nature* **539**, 197–206 (2016).
- Gordon, D. et al. Single-copy expression of an amyotrophic lateral sclerosis-linked TDP-43 mutation (M337V) in BAC transgenic mice leads to altered stress granule dynamics and progressive motor dysfunction. *Neurobiol. Dis.* **121**, 148–162 (2019).
- De Strooper, B. Lessons from a failed γ -secretase Alzheimer trial. *Cell* **159**, 721–726 (2014).
- Karran, E., Mercken, M. & Strooper, B. D. The amyloid cascade hypothesis for Alzheimer's disease: an appraisal for the development of therapeutics. *Nat. Rev. Drug Discov.* **10**, 698 (2011).
- Mitsumoto, H., Brooks, B. R. & Silani, V. Clinical trials in amyotrophic lateral sclerosis: why so many negative trials and how can trials be improved? *Lancet Neurol.* **13**, 1127–1138 (2014).
- Bolognesi, B. et al. ANS binding reveals common features of cytotoxic amyloid species. *ACS Chem. Biol.* **5**, 735–740 (2010).
- Cremades, N. et al. Direct observation of the interconversion of normal and toxic forms of α -synuclein. *Cell* **149**, 1048–1059 (2012).
- Fang, Y.-S. et al. Full-length TDP-43 forms toxic amyloid oligomers that are present in frontotemporal lobar dementia-TDP patients. *Nat. Commun.* **5**, 4824 (2014).
- Zhu, C., Beck, M. V., Griffith, J. D., Deshmukh, M. & Dokholyan, N. V. Large SOD1 aggregates, unlike trimeric SOD1, do not impact cell viability in a model of amyotrophic lateral sclerosis. *Proc. Natl Acad. Sci. USA* **115**, 4661–4665 (2018).
- Escusa-Toret, S., Vonk, W. I. M. & Frydman, J. Spatial sequestration of misfolded proteins by a dynamic chaperone pathway enhances cellular fitness during stress. *Nat. Cell Biol.* **15**, 1231–1243 (2013).
- Mateju, D. et al. An aberrant phase transition of stress granules triggered by misfolded protein and prevented by chaperone function. *EMBO J.* **36**, 1669–1687 (2017).
- Alberti, S. & Hyman, A. A. Are aberrant phase transitions a driver of cellular aging? *Bioessays* **38**, 959–968 (2016).
- Khan, T. et al. Quantifying nucleation in vivo reveals the physical basis of prion-like phase behavior. *Mol. Cell* **71**, 155–168.e7 (2018).
- Guo, L. et al. Nuclear-import receptors reverse aberrant phase transitions of RNA-binding proteins with prion-like domains. *Cell* **173**, 677–692.e20 (2018).
- Franzmann, T. M. et al. Phase separation of a yeast prion protein promotes cellular fitness. *Science* **359**, eaao5654 (2018).
- Wang, J. et al. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell* **174**, 688–699.e16 (2018).
- Patel, A. et al. A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. *Cell* **162**, 1066–1077 (2015).
- Molliex, A. et al. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell* **163**, 123–133 (2015).
- Wegmann, S. et al. Tau protein liquid-liquid phase separation can initiate tau aggregation. *EMBO J.* **37**, e98049 (2018).
- Vavouri, T., Semple, J. I., Garcia-Verdugo, R. & Lehner, B. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell* **138**, 198–208 (2009).
- Bolognesi, B. et al. A concentration-dependent liquid phase separation can cause toxicity upon increased protein expression. *Cell Rep.* **16**, 222–231 (2016).
- Halfmann, R. et al. Opposing effects of glutamine and asparagine govern prion formation by intrinsically disordered proteins. *Mol. Cell* **43**, 72–84 (2011).
- Neumann, M. et al. Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science* **314**, 130–133 (2006).
- Higashi, S. et al. Concurrence of TDP-43, tau and alpha-synuclein pathology in brains of Alzheimer's disease and dementia with Lewy bodies. *Brain Res.* **1184**, 284–294 (2007).
- Schwab, C., Arai, T., Hasegawa, M., Yu, S. & McGeer, P. L. Colocalization of transactivation-responsive DNA-binding protein 43 and Huntingtin in inclusions of Huntington disease. *J. Neuropathol. Exp. Neurol.* **67**, 1159–1165 (2008).
- Amador-Ortiz, C. et al. TDP-43 immunoreactivity in hippocampal sclerosis and Alzheimer's disease. *Ann. Neurol.* **61**, 435–445 (2007).
- Uchino, A. et al. Incidence and extent of TDP-43 accumulation in aging human brain. *Acta Neuropathol. Commun.* **3**, 35 (2015).
- Jiang, L.-L. et al. Structural transformation of the amyloidogenic core region of TDP-43 protein initiates its aggregation and cytoplasmic inclusion. *J. Biol. Chem.* **288**, 19614–19624 (2013).
- Conicella, A. E., Zerze, G. H., Mittal, J. & Fawzi, N. L. ALS mutations disrupt phase separation mediated by α -helical structure in the TDP-43 low-complexity C-terminal domain. *Struct./Fold. Des.* **24**, 1537–1549 (2016).
- Sun, Y. & Chakrabartty, A. Phase to phase with TDP-43. *Biochemistry* **56**, 809–823 (2017).
- Schmidt, H. B., Barreau, A. & Rohatgi, R. Decoding and recoding phase behavior of TDP43 reveals that phase separation is not required for splicing function. Preprint at <https://www.biorxiv.org/content/10.1101/548339v1> (2019).
- Babinchak, W. M. et al. The role of liquid-liquid phase separation in aggregation of the TDP-43 low complexity domain. *J. Biol. Chem.* (2019). <https://doi.org/10.1074/jbc.RA118.007222>
- Sreedharan, J. et al. TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. *Science* **319**, 1668–1672 (2008).
- Chou, C.-C. et al. TDP-43 pathology disrupts nuclear pore complexes and nucleocytoplasmic transport in ALS/FTD. *Nat. Neurosci.* **21**, 228–239 (2018).
- McGurk, L. et al. Poly(ADP-Ribose) Prevents pathological phase separation of TDP-43 by promoting liquid demixing and stress granule localization. *Mol. Cell* **71**, 703–717 (2018).
- Coyne, A. N. et al. Fragile X protein mitigates TDP-43 toxicity by remodeling RNA granules and restoring translation. *Hum. Mol. Genet.* **24**, 6886–6898 (2015).
- Gasset-Rosa, F. et al. Cytoplasmic TDP-43 de-mixing independent of stress granules drives inhibition of nuclear import, loss of nuclear TDP-43, and cell death. *Neuron* **102**, 339–357 (2019).
- D'Alton, S. et al. Divergent phenotypes in mutant TDP-43 transgenic mice highlight potential confounds in TDP-43 transgenic modeling. *PLoS ONE* **9**, e86513 (2014).
- Mann, J. R. et al. RNA binding antagonizes neurotoxic phase transitions of TDP-43. *Neuron* **102**, 321–338 (2019).
- Jiang, L.-L. et al. Two mutations G335D and Q343R within the amyloidogenic core region of TDP-43 influence its aggregation and inclusion formation. *Sci. Rep.* **6**, 23928 (2016).
- Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801 (2014).
- Staller, M. V. et al. A high-throughput mutational scan of an intrinsically disordered acidic transcriptional activation domain. *Cell Syst.* **6**, 444–455.e6 (2018).
- Ravarani, C. N. et al. High-throughput discovery of functional disordered regions: investigation of transactivation domains. *Mol. Syst. Biol.* **14**, e8190 (2018).
- Schmiedel, J. M. & Lehner, B. Determining protein structure using deep mutagenesis. *Nat. Genet.* **51**, 1177–1186 (2019).
- Guenther, E. L. et al. Atomic structures of TDP-43 LCD segments and insights into reversible or pathogenic aggregation. *Nat. Struct. Mol. Biol.* **25**, 463–471 (2018).
- Mompeán, M. et al. Structural evidence of amyloid fibril formation in the putative aggregation domain of TDP-43. *J. Phys. Chem. Lett.* **6**, 2608–2615 (2015).
- Johnson, B. S. et al. TDP-43 is intrinsically aggregation-prone, and amyotrophic lateral sclerosis-linked mutations accelerate aggregation and increase toxicity. *J. Biol. Chem.* **284**, 20329–20339 (2009).
- Johnson, B. S., McCaffery, J. M., Lindquist, S. & Gitler, A. D. A yeast TDP-43 proteinopathy model: exploring the molecular determinants of TDP-43 aggregation and cellular toxicity. *Proc. Natl Acad. Sci. USA* **105**, 6439–6444 (2008).
- Tamaki, Y. et al. Elimination of TDP-43 inclusions linked to amyotrophic lateral sclerosis by a misfolding-specific intrabody with dual proteolytic signals. *Sci. Rep.* **8**, 6030 (2018).
- Farrarwell, N. E. et al. Distinct partitioning of ALS associated TDP-43, FUS and SOD1 mutants into cellular inclusions. *Sci. Rep.* **5**, 13416 (2015).
- Kaganovich, D., Kopito, R. & Frydman, J. Misfolded proteins partition between two distinct quality control compartments. *Nature* **454**, 1088–1095 (2008).

59. Peskett, T. R. et al. A liquid to solid phase transition underlying pathological Huntingtin Exon1 aggregation. *Mol. Cell* **70**, 588–601.e6 (2018).
60. Rollins, N. J. et al. Inferring protein 3D structure from deep mutation scans. *Nat. Genet.* **51**, 1170–1176 (2019).
61. Toth-Petroczy, A. et al. Structured states of disordered proteins from genomic sequences. *Cell* **167**, 158–170.e12 (2016).
62. Hughes, M. P. et al. Atomic structures of low-complexity protein segments reveal kinked β sheets that assemble networks. *Science* **359**, 698–701 (2018).
63. Lim, L., Wei, Y., Lu, Y. & Song, J. ALS-causing mutations significantly perturb the self-assembly and interaction with nucleic acid of the intrinsically disordered prion-like domain of TDP-43. *PLoS Biol.* **14**, e1002338 (2016).
64. Chiti, F. & Dobson, C. M. Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade. *Annu. Rev. Biochem.* **86**, 27–68 (2017).
65. Karmakola, M. et al. Inhibition of RNA lariat debranching enzyme suppresses TDP-43 toxicity in ALS disease models. *Nat. Genet.* **44**, 1302–1309 (2012).
66. Weisberg, S. J. et al. Compartmentalization of superoxide dismutase 1 (SOD1G93A) aggregates determines their toxicity. *Proc. Natl Acad. Sci. USA* **109**, 15811–15816 (2012).
67. Arrasate, M., Mitra, S., Schweitzer, E. S., Segal, M. R. & Finkbeiner, S. Inclusion body formation reduces levels of mutant huntingtin and the risk of neuronal death. *Nature* **431**, 805–810 (2004).
68. Collinge, J. Mammalian prions and their wider relevance in neurodegenerative diseases. *Nature* **539**, 217–226 (2016).
69. Knowles, T. P. J., Vendruscolo, M. & Dobson, C. M. The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.* **15**, 384–396 (2014).
70. Karran, E. & De Strooper, B. The amyloid cascade hypothesis: are we poised for success or failure? *J. Neurochem.* **139**, 237–252 (2016).
71. Chiò, A. et al. Lithium carbonate in amyotrophic lateral sclerosis: lack of efficacy in a dose-finding trial. *Neurology* **75**, 619–625 (2010).
72. Dupuis, L. et al. A randomized, double blind, placebo-controlled trial of pioglitazone in combination with riluzole in amyotrophic lateral sclerosis. *PLoS ONE* **7**, e37885 (2012).
73. Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S. & Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nat. Methods* **12**, 203–206 (2015).
74. Rapsomaniki, M. A. et al. easyFRAP: an interactive, easy-to-use tool for qualitative and quantitative analysis of FRAP data. *Bioinformatics* **28**, 1800–1801 (2012).
75. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

Acknowledgements

Work in B.L.'s lab was supported by a European Research Council (ERC) Consolidator grant (616434), the Spanish Ministry of Economy and Competitiveness (BFU2017-89488-P), the AXA Research Fund, the Bettencourt Schueller Foundation, and Agencia

de Gestio d'Ajuts Universitaris i de Recerca (AGAUR, SGR-831). G.G.T.'s lab was supported by the European Research Council (RIBOMYLOME_309545) and the Spanish Ministry of Economy and Competitiveness (BFU2014-55054-P and BFU2017-86970-P). We acknowledge support from the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013-2017', the EMBL Partnership, and the CERCA Program/Generalitat de Catalunya. We thank Pablo Baeza Centurión, Xavier Salvatella, Alexandros Armaos and Benjamin Lang for discussion and assistance and the Eisenberg lab for help with the ZipperDB analysis.

Author contributions

B.B. and B.L. conceived the project and designed the experiments; B.B. and M.S. performed the experiments; A.J.F., B.B. and J.M.S. performed analyses of sequences; A.J.F. and J.M.S. analysed the genetic interactions and structures; G.G.T. initiated, designed and carried out the original computational analysis of physicochemical properties; B.B., A.J.F. and B.L. wrote the manuscript with input from all authors.

Additional information

Supplementary Information accompanies this paper at <https://doi.org/10.1038/s41467-019-12101-z>.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at <http://ngp.nature.com/reprintsandpermissions/>

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019

Discussion

Predicting mutational effects in amyloid forming sequences

To build a comprehensive map for the impact of mutations on amyloid nucleation, we successfully developed a novel approach that couples a yeast phenotypic assay to deep mutational scanning (DMS) and allows testing thousands of protein variants at scale. We used this approach for the amyloid β peptide (A β 42), one of the most well-known mammalian amyloidogenic proteins linked to Alzheimer's Disease (AD). Our work represents the first mutational scanning of the aggregation of any amyloid protein and the first systematic comparison of mutation types in any human gene.

The majority of single amino acid substitutions in the hydrophobic C-terminus decrease nucleation, consistent with all proposed A β structures where this region is forming the amyloid core of the fibrils. However, not only mutations to polar and charged residues in the core disrupt nucleation, but also changes to aliphatics and aromatics. In fact, there are only a few hydrophobic changes that are tolerated in this region, such as valine or isoleucine at positions 29, 30 or 34, in addition to some polar residues, for example cysteine at positions 29, 38 and 42, or asparagine at positions 33 and 42. Overall, this suggests a very specific arrangement of the side chains for the packing of the core with only a handful of solutions for proper nucleation, which would not have been easily predicted. At the N-terminus, mutations have a more balanced effect but similarly to the C-terminus, they are very position-specific. For example, increasing hydrophobicity disrupts nucleation when mutating the first ten residues, but after position 11 some aliphatic amino acids such as valine and isoleucine are tolerated. In addition, many polar, proline residues and positive - but not negative - charges increase nucleation. Interestingly, the same amino acid type does not always have the same impact at each position, for example, while aspartic acid increases nucleation at positions 18 and 19, its counterpart glutamic acid decreases it.

Moreover and beyond substitutions, with our approach we could also assess and compare the impact on nucleation of other types of mutations in A β 42, such as insertions, deletions and truncations. We found that the modular effect of mutations is common across all classes, with a general trend showing that mutations at the C-terminus are more disruptive than at the N-terminus, the latter having a substantial amount of increasing-nucleation variants. Beyond this, the impact of the different classes of mutations is not easily predictable, and the effect of mutations of similar amino acid identity and position do not always correlate between mutation types. For example, the effect of substitutions and insertions only correlates when averaging the effect of mutations of specific amino acid types for the N and C-terminus separately or at specific positions, but not when taking each individual residue identity and position all along the sequence. In addition, some regions of the peptide are more tolerant to specific types of mutations. This is well illustrated with the 33-38 region at the C-terminus, where virtually all insertions and deletions decrease nucleation but a handful of substitutions are tolerated. Another example is that while at the last four residues of the peptide some substitutions and insertions increase nucleation, no deletions or truncations are tolerated. In some other regions, all mutation types have similar effects on nucleation. For example, the region 17-27, defined as the hotspot of deletion effects with 35 aggregating large deletions, also contains variants of all other mutation types that increase nucleation, such as single deletions of residues 22-26 or substitutions and insertions at positions 22 and 23.

Overall, the amino acid preferences for nucleation are difficult to predict especially because they are very position and region specific along the peptide sequence. Indeed, existing computational tools poorly predict the outcome of mutations on nucleation and their association to disease, with only mutational effects at the C-terminus partially captured by hydrophobicity and predictors of aggregation, such as Zyggregator or Tango. Moreover, these algorithms are trained and - up to this work - tested only on single amino acid substitutions. We showed they also have little predictive power for insertions and single deletions, and that they cannot assess mutations affecting multiple residues - with proven biological and clinical relevance - because they rely on sequence length (**Figure 16**). This highlights the urge to include various types of genetic variants in the design of mutational libraries and generate experimental data that can be then used to build better computational predicting tools. More broadly, testing the performance of all possible variants in a protein results in the full distribution of possible mutational effects. This is crucial to put the impact of specific variants into the context of the whole phenotypic space, rather than comparing them simply to another variant such as the wild type (WT), as it has been historically done.

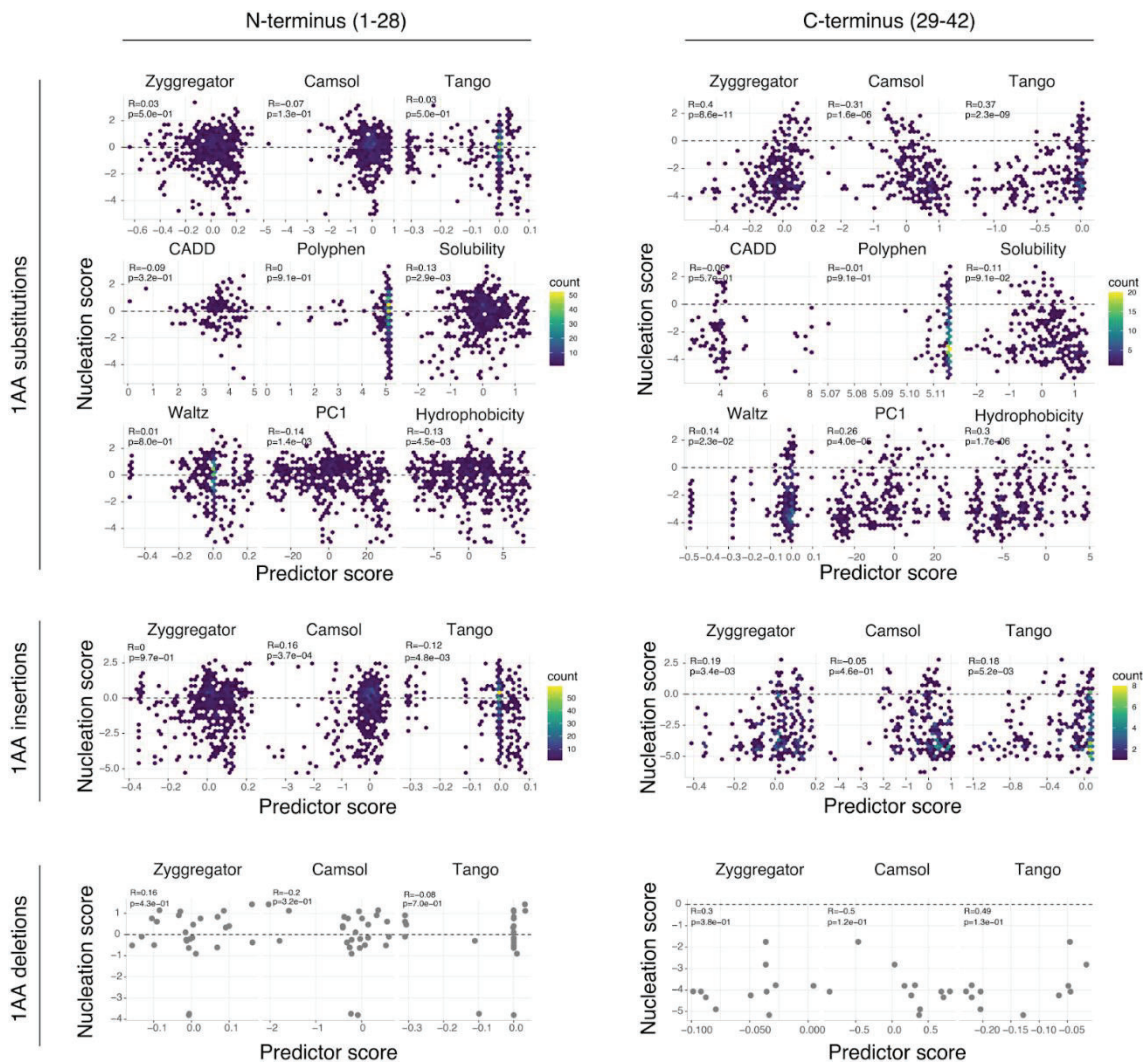


Figure 16. Evaluation of mutational effect and aggregation predictors. Correlation of nucleation scores with the predictions of aggregation predictors (Tango, Zyggregator, Waltz and Camsol) (Fernandez-Escamilla et al. 2004; Tartaglia and Vendruscolo 2008; Sormanni, Aprile, and Vendruscolo 2015; Oliveberg 2010), variant effect predictors (CADD, Polyphen) (Rentzsch et al.

2019; Adzhubei, Jordan, and Sunyaev 2013), solubility scores (Gray et al. 2019), PC1 (Bolognesi et al. 2019) and hydrophobicity (Kyte and Doolittle 1982) for single amino acid mutations, at the N-terminus (left) or the C-terminus (right). Pearson correlation coefficients are indicated. Dashed lines indicate the WT nucleation score (0).

Two aggregation-prone regions (APRs) have been identified by Tango, one of the most widely used computational tools for protein aggregation. One APR is located at residues 17-21 at the N-terminus and the other comprises the C-terminal 29-42 residues. Despite both being hydrophobic stretches, our data show their tolerance to mutations is completely different. While APR1 behaves very similarly to the rest of the N-terminus 1-28, with many mutations increasing nucleation, APR2 is very sensitive and intolerant to changes. This further highlights that the behavior of amyloid proteins is not readily predictable from their primary sequence.

Structural biology has substantially contributed to our understanding of A β 42 fibrils, for example, by determining the existence of the amyloid core with a very similar structure in all known polymorphs. However, these studies do not capture the specific amino acid preferences at the different positions and are not able to determine flexible and disordered regions such as the N-terminus of A β 42, which we showed has high implications in nucleation and disease. Overall, we envision that experimental deep mutagenesis of disordered sequences is a powerful tool to gain mechanistic insights of protein regions that remain unsolved by traditional structural methods.

Discriminating disease variants and prioritizing disease candidates

The dataset of single amino acid substitutions contains all 16 dominant familial AD (fAD) mutations known to date, and all of them show increased nucleation relative to WT A β 42. Indeed, nucleation scores accurately classify all fAD variants, meaning they can be used to predict the outcome of variants of unknown significance (VUS). In addition, the agreement of our data with human genetics suggests that fAD is very likely to be a nucleation disease. The power of nucleation scores to classify fAD variants is further confirmed by assessing other types of genetic variants: the only single amino acid deletion, E22 Δ , and the only multi amino acid deletion, Δ 19-24, associated with fAD, both show increased nucleation score.

Some of the fAD mutations were described while we were performing these experiments, for example, K16Q, L17V or Δ 19-24 mutations were found only after 2020. Traditionally, new A β 42 fAD associated mutations are reported only once they are seen in a patient and a bit alarmingly, the rate at which they are being discovered is not any better than in the early 90s when the first cases were described (**Figure 17**).

Our assay overcomes this limitation and can contribute to the classification of candidate pathogenic variants by providing a nucleation score for all possible mutations in the peptide at scale. In total, our approach prioritizes 307 variants as candidate fAD, 108 of them are substitutions, 77 insertions, 5 single amino acid deletions, 13 truncations and 104 multi amino acid deletions; revealing that all mutation types beyond substitutions are likely to cause disease.

Strikingly, 267 of the total 307 are located at the N-terminus of A β 42, the region already containing 16 out of the 18 fAD mutations.

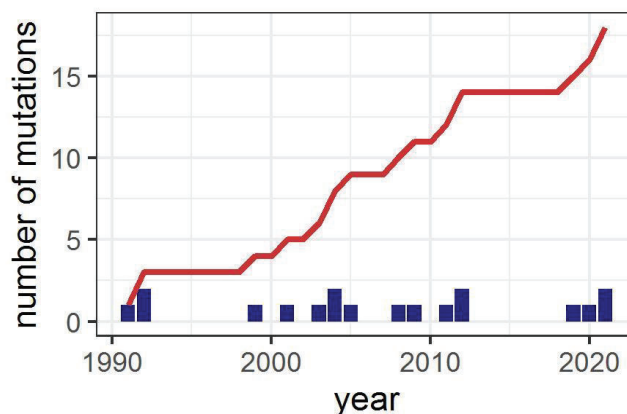


Figure 17. Number of A β 42 mutations associated with fAD discovered with time. Red line indicates the cumulative number of mutations (Data from APP | Alzforum. (n.d.). Retrieved August 30, 2022, from <https://www.alzforum.org/mutations/app>)

It is worth mentioning that not all A β 42 mutations have a dominant pattern of inheritance, but E22 Δ and A2V are recessive. While E22 Δ shows the same effect as dominant fAD mutations, an increase in nucleation, A2V instead decreases nucleation in our assay.

E22 Δ , known as the Japanese mutation, was originally found in three individuals: one homozygous with AD, one heterozygote with mild cognitive impairment (MCI) and another heterozygous but healthy individual. Only the homozygous individual showed AD, but since one heterozygous individual manifested MCI (which may precede dementia in some cases), the authors suggested an incomplete penetrance behavior of the mutation that may act in a dose-dependent manner (Tomiya et al. 2008). One limitation of our approach is that homozygous versus heterozygous genotypes cannot be assessed. However, the incomplete penetrance behavior suggested for E22 Δ could be approached by running the experiment at different protein expression levels, since the library is expressed under a concentration-dependent inducible copper promoter (data not shown).

A2V, the other mutation with a recessive pattern of inheritance, is known to alter the processing of APP, increasing the production of A β (Di Fede et al. 2009). Our assay cannot assess this mechanism because the library construct contains only the A β 42 fragment and so cleavage and processing steps are omitted. The same study suggests that pathogenicity may also come from the A2V A β 42 peptide due to increased aggregation relative to the WT. Yet, this was only tested in the A β 40 isoform background, which is well-known to be less aggregation-prone than A β 42. In our assay, A β 40 shows decreased nucleation score, highlighting the importance of the last two residues for proper nucleation. Therefore, we reasoned that the A2V mutation - or any other mutation - may have completely a different outcome in the A β 40 or A β 42 backgrounds and indeed, for a dataset of single amino acid substitutions tested in both backgrounds (n=155), we see a very poor correlation of nucleation scores (r=0.27, data not shown).

Another aspect in our assay that generates debate is the use of yeast as a disease model. Generally, more complex systems such as animal models are considered more appropriate for reproducing human disease phenotypes, especially when it comes to neurodegenerative diseases. However, some of the common AD mouse models, for example, rely on two or three overexpressed transgenes to obtain an observable phenotype, which does not always fully recapitulate that of human disease (Sasaguri et al. 2017) and thus we argue that this is not necessarily a particularly physiological model. In yeast, by assessing large numbers of sequence variants or experimental conditions in a simpler and scalable mode, we have the possibility of statistically testing whether the outcome of (all) disease variants can be distinguished from that of the overall set of variants. This work is therefore a tangible example of how a yeast assay can become more powerful than many other assays that were assumed to be more ‘physiological’ in the first place: whichever mechanism is occurring inside the yeast cell, it may be the same or very similar to that causing human fAD. Overall, we believe that any assay in any model, from simple cell-based assays to animal models, should be genetically validated (or invalidated) for its relevance to disease by using clinical genetic data.

Structural insights from deep mutagenesis

Various structures have been proposed for A β 42, including two from AD human brains, and a set of structures for fibrils formed by the A β 40 isoform. Having in mind how polymorphic amyloid proteins are, it is very likely that differences in experimental conditions, even if subtle, have a great impact on the final arrangement of the protofilaments and fibrils. However, the vast majority of A β 42 structures (6 out of 8) share some fundamental features, such as an S-shape fold of the monomeric subunit from residue 9 (starting position varies on each polymorph) to residue 42, and a tightly packed amyloid core at residues 29-42. The first residues of the peptide remain instead undetermined and likely disordered in 4 of these 6 structures. For simplicity, here we focus the discussion on the two *ex vivo* models for A β 42 fibrils in disease.

The two models for A β 42 fibrils, type I and type II, have been associated with AD with different patterns of inheritance. Type I fibrils were more abundant in individuals with sporadic AD (sAD, 87%, 100% and 77% of type I fibrils in three patients, respectively), while type II were more abundant in individuals with familial AD (fAD, 100% and 76% in two patients, respectively). Here, individuals with fAD did not have any mutation inside the A β 42 region, but one in the *APP* gene downstream of A β 42 and the other in the *PSEN1* gene, meaning that the amino acid sequence encoding type I and type II fibrils both correspond to WT A β 42. Type I fibrils were also found in fAD patients and vice versa, although to a lower extent. Five additional individuals with other amyloidogenic conditions (aging-related tau astrogliopathy, ARTAG; Parkinson’s disease, PD; dementia with Lewy bodies, DLB; frontotemporal dementia, FTD; and pathological aging, PA) showed a 100% of type II fibrils (Y. Yang et al. 2022). This raises the question of how different types of fibrils are formed in sAD versus fAD and other conditions, and whether polymorphs determine disease phenotypes or conversely, whether one specific polymorph emerges as a consequence of disease-specific conditions. The fact that type II fibrils are involved in many different diseases suggests that the disease is responsible for establishing specific environmental conditions in which the same type of fibrils emerge. However, that type I fibrils are specific to one unique disease also suggests that protein conformation is responsible for the disease outcome.

In any case, it is plausible that the two structures coexist and indeed, type I fibrils were also found in fAD patients and vice versa, although to a lower extent (Y. Yang et al. 2022).

The β fold forming the amyloid core for both types of fibrils is almost identical. It is known that different polymorphs typically share a common β fold and differ on the arrangement of more exposed flanking regions. In the amyloid core, there is a very tight and specific arrangement of side chains, as our mutational data supports. For example, the insertions dataset identifies a stretch between residues 33-38 that cannot accommodate any additional side chain. Moreover, all alternative amyloid cores discovered in the internal deletions dataset indicate that residues from position 33 are essential and that alternative stretches re-arranging the core can only have the same, or a very similar length to that of the WT core. In both types of fibrils, the hydrophobic pocket at the C-terminus is formed by side chains of residues A30, I32, M35 and V40. A42 is also part of the hydrophobic core in type I fibrils but forming a salt bridge with K28 in the inter-monomer interface in type II. The other hydrophobic pocket in the S-shape is different for the two structures due to a twist of the backbone at residues G25-S26 in type II, resulting in a bigger intra-monomer interface but impeding an inter-monomer interface. In type I, F19 and F20 form the intra-monomer interface, and Y10, V12, Q15 and L17 of one monomer, which also face inside, form the inter-monomer interface with residues L34, V36, V39 and I41 of the other monomer. In type II, due to the backbone flip, L17 and V18 are part of the intra-monomer hydrophobic pocket and Q15 remains solvent-exposed. Y10 and V12 are not determined in type II and so their orientation is unclear but, in any case, Q15 and L17 are not available for an inter-monomer interface. Therefore, in type II the inter-monomer interface is not hydrophobic but formed by a salt bridge between residues K28 of one monomer and A42 of the other monomer, which in this case is not part of the hydrophobic C-terminal pocket. The hydrophobic C-terminal side chains that are not part of the hydrophobic pocket remain exposed to the solvent (**Figure 18**).

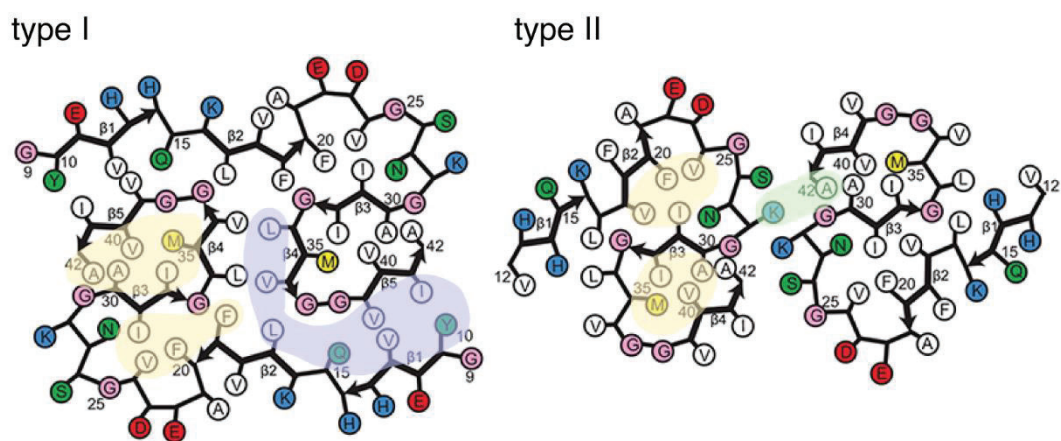


Figure 18. Schematics of type I and type II fibrils structures with depicted interfaces. Yellow: hydrophobic intra-monomer interfaces; blue: hydrophobic inter-monomer interface in type I; green: electrostatic inter-monomer interface in type II.

Here, we are assuming that folding of the amyloid core precedes the arrangement of the rest of the monomer (for example the formation of a second hydrophobic pocket) as well as the arrangement of the dimer. Folding of the amyloid core is mainly driven by the hydrophobic effect and may be one of the first steps in the overall aggregation process, determining which residues and regions remain available for the subsequent formation of inter-monomer interfaces. In this

line, folded or partially folded monomers would exist in solution to then form pre-fibrillar assemblies, either stacking one on top of the other, or creating facing dimers with an inter-monomer interface. Another possibility though, is that interactions for the formation of the dimer interface are required for the folding of the monomer and its amyloid core. However, it seems unlikely that inter-monomer interfaces as different as those in type I and type II, drive the formation of such similar amyloid cores. In addition, elongation, which takes place once the first nuclei are formed by nucleation, may also contribute to the folding of the monomer. By this means, the free ends of pre-existing fibrils may template the folding of monomers that eventually adopt the same structure as the peptides already incorporated in the fibril.

Our data shows that a positive residue in front of the C-terminal region (i.e., K28 in the sequence) represents the minimal core necessary for nucleation. This could agree with all previous arguments, as a charged residue could keep the hydrophobic core soluble facilitating a conformation-specific folding of the core, rather than amorphous aggregation. It could also drive the formation of the dimer interface, as the salt bridge in type II fibrils; or even stabilize the monomer with a salt bridge with A42 in the same monomer, as seen for some other structures (5KK3, 2MXU and 2NAO).

We also show that not only K28 in front of the amyloid core influences nucleation, but all the N-terminal region plays a critical role in modulating nucleation. For all types of mutations, the majority of mutations that increase nucleation cluster in this region. One extreme example are truncations, where only those at the N-terminus increase nucleation, while all those affecting the C-terminus decrease it. As mentioned above, there is a specific stretch at the N-terminus between residues 17 and 27 where many variants increase nucleation. In the context of the fibril structures, these residues are forming the second hydrophobic pocket of the S-shape. Residues E22 and D23 are two negative charges in this region and in the two structures they are both exposed to the solvent, even if they are one next to the other. This disposition facilitates the looping in the S-shape and indeed, E22 has the angles in a glycine-specific region of the Ramachandran plot, suggesting it is well accommodated in a turn (**Figure 19**). We also define E22 and D23 - together with other negatively charged residues at the N-terminus - as gatekeepers of nucleation, meaning that mutating or removing them accelerates nucleation. Up to six fAD mutations, including substitutions, one single amino acid deletion and one large internal deletion, affect these two positions. Overall, this suggests that altering this loop and the hydrophobic pocket increases nucleation in many cases and has direct implications in disease.

Disruption of the N-terminus may lead to other conformations, for example, exposing flanking regions that recruit new monomeric protein and accelerate formation of new aggregates by secondary nucleation. However, there are other possible scenarios for how the N-terminus controls nucleation rates. For example, the N-terminus could be important for the ensemble of soluble peptides or establish specific interactions at the transition state of nucleation.

Although the two types of fibrils discussed here may co-exist, it is possible that one is kinetically more favorable or thermodynamically more stable than the other, or that one is more causative of disease. Our current mutational datasets cannot determine whether we are tracking only one type of structure or a combination of both, since we cannot distinguish whether a mutation impacts only the monomer or the dimer interface. For example, a mutation from leucine to valine at position 17, resulting in increased hydrophobicity, could favor the formation of the hydrophobic core in type II structure but the inter-monomer interface in type I.

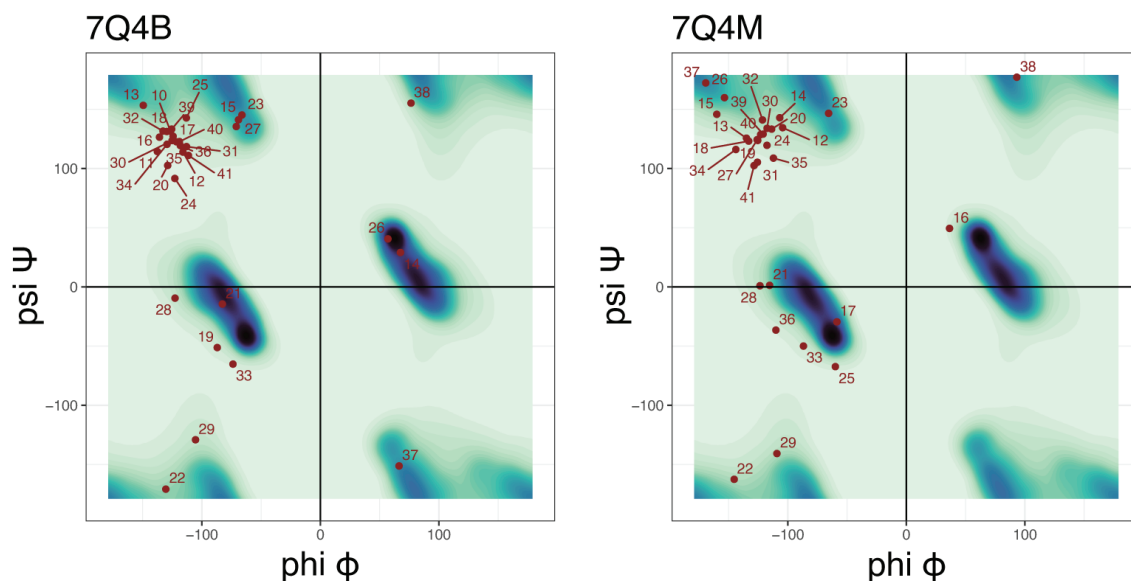


Figure 19. Ramachandran plots for glycine residues with mapping of psi (Ψ) and phi (ϕ) angles in type I (7Q4B) and type II (7Q4M) A β 42 structures. Background density data from 500 proteins, obtained from (Lovell et al. 2003).

Aggregation kinetics and the transition state

We proved our DMS approach is tracking nucleation by correlating the scores with *in vitro* measurements of aggregation kinetics for a handful of variants for which rate constants were accurately measured and available. We find that our scores strongly agree with both primary and secondary nucleation. Secondary nucleation is the microscopic mechanism known to be rate determining for the WT A β 42 aggregation reaction, but at least for the subset of variants correlated here, we cannot distinguish whether our assay is tracking either only one or both types of nucleation. In case there were any substantial differences inside the yeast cell, additional *in vitro* kinetic measurements for more variants would be required to decouple them.

Computer simulations and experimental studies on oligomeric species suggest that the first oligomers formed during nucleation are disordered and emerge from nonspecific interactions. Eventually, they undergo conformational changes and become more ordered species, most likely containing a substantial amount of β -sheet structures that make them compatible with pre-formed fibrils. However, there are still many unanswered questions regarding how nucleation happens at the molecular level. For example, it is not known whether the minimal nucleus at the transition state corresponds to an early or a late oligomer; whether it is disordered or already contains some structural elements and contacts. In case the minimal nucleus is ordered to some extent, it is not known how many molecules are needed and how they are arranged, whether they stack forming a pre-protofilament or instead they face forming a dimeric complex. Thus, whether the structures adopted by oligomeric species during the aggregation process resemble or differ from those of mature fibrils, it is not fully understood.

By measuring nucleation - the rate-limiting step in the aggregation reaction with a high kinetic barrier - we assume we are reporting on the species formed at the transition state. We do not know what these species are in terms of structures, size or morphology. It is plausible to think they are oligomers with an extent of structure that allows them to act as seeds for further aggregation, but we currently do not have a way to measure this *in vivo*. What we know is that mutational impact suggests that species in the transition state have - to some extent - a similar structure to that of mature fibrils, with a highly packed C-terminal amyloid core. More plainly, we use the mature fibril structures to rationalize the effect of mutations because they are the only currently available structural models for A β 42.

Yeast prion biology also helps us speculate further through which mechanism nucleation increases. What is known for yeast prion proteins and thus for Sup35, used as a reporter in our assay, is that arrangement of different amyloid structures results in different phenotypes, known as ‘prion strains’. By this means, amyloid fibrils of the Sup35 protein with a less rigid amyloid core are more easily fragmented by molecular chaperones and produce more oligomeric seeds. This results in increased prion proliferation, increased stop codon read-through capacity and better growth in a medium lacking adenine. Therefore, it is possible that our assay is measuring oligomeric seeds that result from fibril fragmentation, with very similar structures to those of mature fibrils. In addition, that fibrils with less rigid amyloid cores are associated with increased proliferation, is compatible with a scenario in which mutations that increase nucleation drive the formation of alternative fibrils, for example, with more exposed flanking regions that facilitate surface-catalyzed secondary nucleation processes.

Finally, it has been suggested that different types of oligomeric species, on- or off-pathway towards fibril formation, may be toxic to cells and potentially causative of disease. Importantly, our assay is not providing direct information on the specific properties of the toxic species formed during aggregation. However, the agreement of nucleation scores with clinical genetic data proves our assay is reporting on a mechanism relevant to human AD and that can be further used to predict the outcome of mutations of currently unknown significance.

Genetic interactions to infer structural conformations

We envision that more complex libraries may be useful to rationalize the specific conformations present at the transition state. For example, a library of double mutants, if sufficiently exhaustive, could shed light on the interactions required - within and between molecules - to generate the first nucleus that then seeds further aggregation.

The Arrhenius equation links the kinetic rate constant (k) to the highest free energy of Gibbs relative to the initial state (ΔG^\ddagger) as:

$$k = A \exp \left(-\frac{\Delta G^\ddagger}{RT} \right)$$

where A is a prefactor, R is the universal gas constant and T is temperature (Cohen et al. 2018). By this means, nucleation scores, which are indeed kinetic rate constants, can be fitted to a

thermodynamic model and values of ΔG for each mutant variant can be extracted. By using a double mutant cycle (DMC) analysis, we may then identify which residues and mutations have non-additive effects, meaning they are energetically coupled and forming required interactions at the transition state (Horowitz 1996). However, this type of analysis is not trivial as it requires coverage of the entire dynamic range of the effects of mutations and multiple measurements for each single substitution background. This means that many if not all possible double mutant combinations need to be tested, resulting in highly complex libraries, for example, of >310,000 double mutants in A β 42, difficult to tackle in our assay currently limited to ~150,000 variants by experimental scalability.

In a parallel study (see *Chapter III. The mutational landscape of a prion-like domain* in the *Results* section), we proved that genetic interactions are very convenient to illuminate sequence-to-structure relationships in disordered proteins. By quantifying interactions in the prion-like domain (PRD) of TAR DNA-binding protein 43 (TDP-43), we identified two secondary structural elements, an α -helix and a β -strand, that form *in vivo* and were previously determined *in vitro* for fragments of TDP-43. These results highlight how unstructured regions may be partially structured *in vivo* while the protein is under selection for a specific function.

One assay, multiple proteins; one protein multiple assays

It is worth noting that phenotypic outcomes may be caused by a combination of various altered mechanisms and functions (X. Li and Lehner 2020). For example, it was recently shown that biophysical ambiguities prevented from accurately predicting the outcome of mutations on protein allostery: binding to an interaction partner could be affected by changes in stability or binding affinity. This was solved by testing both phenotypes in two different DMS approaches with specific selections (Faure et al. 2022), illustrating the power of DMS in tackling phenotypic ambiguities. Thus, one library can be tested in multiple selection assays targeting a specific phenotype and in addition, each selection assay can be used for different libraries (Cagiada et al. 2021).

Our DMS nucleation approach, here piloted in A β 42, is easily and readily transferable to other amyloidogenic proteins, not only mammalian or human ones but also yeast and bacteria prion proteins. Furthermore, other disordered sequences such as PRDs are also potential candidates to test. For example, the PRD of TDP-43 - our other model protein - drives its self-assembly. It has been shown that *in vitro* and under specific conditions, TDP-43 can form liquid condensates as well as amyloid aggregates. We showed that at least inside the yeast cell, TDP-43 toxic variants adopt a liquid de-mixed state while insoluble and bigger assemblies prevent cell toxicity. Pairing TDP-43 with the DMS nucleation approach remains to be done, but we envision it may be very informative to further characterize the nature of the assemblies with different toxicity in yeast. In addition, the DMS nucleation approach in TDP-43 may reveal whether a completely different sequence to A β 42, or any of its mutant variants, are able to nucleate amyloids in a similar way to classic ones.

The A β 42 library can also be tested in the DMS toxicity approach, to verify that none of the classified as non-nucleating variants in our dataset are simply disappearing in the population due to toxicity. We proved in a pilot experiment with a handful of A β 42 variants that none of them has altered toxicity (data not shown), but further assessment of the complete library is missing.

In both nucleation and toxicity approaches, we cannot exclude the possibility that expression levels or degradation affect variants in distinct ways. This is a common caveat for many DMS approaches, which can be solved for example, with a fusion to a fluorescent tag that can be quantified by flow cytometry. However, this is currently not doable in our studies due to library design. In the TDP-43 work, we could verify by western blotting at small scale, that there is no bias in protein expression for toxic *versus* non-toxic variants, proving that toxicity is not directly related to protein overexpression. In the case of A β 42 we are tracking a gain of function, meaning that degraded or low expressed variants may be classified as non-nucleators, which in any case, is a better scenario than misclassifying them as nucleators.

One solution to account for degradation in the DMS nucleation approach is to run the experiment using a different yeast strain, with pre-formed aggregates in the background as previously shown in (Chandramowlishwaran et al. 2018). In this scenario, non-nucleating variants would survive in the population because SupN nucleates on top of pre-existing aggregates, in contrast to degraded variants, that would disappear in the population.

Other DMS approaches could be used for both A β 42 and TDP-43. For example, an assay that reports on condensation would be very informative for TDP-43, to more systematically assess the biophysical state of the toxic variants identified in our study, which we suggested have a more liquid-like state compared to the non-toxic ones. Another option would be to introduce other factors in the DMS experiments, such as molecular chaperones known to have a disaggregating role and that are potential therapeutic strategies.

One example of how different DMS selections report on different mechanisms, is the comparison between our A β 42 DMS approach with a Sup35N fusion tracking nucleation, and the DMS approach used in (Gray et al. 2019), in this case with A β 42 fused to dihydrofolate reductase (DHFR) and tracking solubility. Data from the two datasets do not correlate and only the nucleation assay correctly classifies all fAD mutations. Solubility scores from the DHFR assay, but not the nucleation scores from our assay, can be largely explained on the basis of hydrophobicity. Overall, this suggests that inside the yeast cell there are at least two different mechanisms of aggregation: one, that is driven by hydrophobicity and is not reporting on disease, and the other, that is instead exquisitely conformation-specific and extremely relevant to human disease.

Conclusions

- The impact of mutations on nucleation reveals a modular organization of A β 42: the C-terminus is very sensitive to mutations and forms the amyloid core required for nucleation, while the disordered N-terminus is enriched in mutations that increase aggregation and are causative of disease.
- *In vivo* nucleation scores accurately discriminate all familial Alzheimer's disease mutations in A β 42, proving that this DMS nucleation approach is tracking a mechanism that is extremely relevant to disease, making it a useful resource for clinical interpretation of genetic variation in A β 42.
- The atlas of amyloid aggregation discovers novel A β 42 mutations of all classes - including substitutions, insertions, deletions and truncations - that accelerate nucleation and are likely pathogenic.
- The atlas of amyloid aggregation also provides mechanistic insights into the process of nucleation: it reveals a central hotspot region between residues 17 to 27 at the N-terminus where mutations of all types increase nucleation.
- The mutational landscape of the prion-like domain of TDP-43 reveals that mutations promoting self-assembly into large insoluble foci reduce toxicity in yeast cells, most likely by titrating the protein away from toxic liquid condensates.
- Genetic interactions identify structural elements in TDP-43 that form *in vivo* and can be potentially used to determine critical contacts in the transition state of A β 42 nucleation.
- Deep mutagenesis is a powerful tool to study the effect of mutations on disease, function and conformation of intrinsically disordered proteins, which are otherwise difficult to characterize by means of existing biophysical approaches and computational predictors.

References

- 1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korb, et al. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74.
- Ackermann, Ang, Kanter, and Tsigelny. 1998. "Identification of Pairwise Interactions in the α -Neurotoxin-Nicotinic Acetylcholine Receptor Complex through Double Mutant Cycles." *Bollettino Della Societa Italiana Di Biologia Sperimentale*. [https://www.jbc.org/article/S0021-9258\(18\)38297-8/abstract](https://www.jbc.org/article/S0021-9258(18)38297-8/abstract).
- Adzhubei, Ivan, Daniel M. Jordan, and Shamil R. Sunyaev. 2013. "Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2." *Current Protocols in Human Genetics / Editorial Board, Jonathan L. Haines ... [et Al.]* Chapter 7 (January): Unit7.20.
- Ahmed, Mahiuddin, Judianne Davis, Darryl Aucoin, Takeshi Sato, Shivani Ahuja, Saburo Aimoto, James I. Elliott, William E. Van Nostrand, and Steven O. Smith. 2010. "Structural Conversion of Neurotoxic Amyloid-Beta(1-42) Oligomers to Fibrils." *Nature Structural & Molecular Biology* 17 (5): 561–67.
- Alberti, Simon, and Dorothee Dormann. 2019. "Liquid-Liquid Phase Separation in Disease." *Annual Review of Genetics* 53 (December): 171–94.
- Alberti, Simon, and Anthony A. Hyman. 2016. "Are Aberrant Phase Transitions a Driver of Cellular Aging?" *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 38 (10): 959–68.
- Aronin, N., K. Chase, C. Young, E. Sapp, C. Schwarz, N. Matta, R. Kornreich, B. Landwehrmeyer, E. Bird, and M. F. Beal. 1995. "CAG Expansion Affects the Expression of Mutant Huntingtin in the Huntington's Disease Brain." *Neuron* 15 (5): 1193–1201.
- Arosio, Paolo, Tuomas P. J. Knowles, and Sara Linse. 2015. "On the Lag Phase in Amyloid Fibril Formation." *Physical Chemistry Chemical Physics: PCCP* 17 (12): 7606–18.
- Arosio, Paolo, Thomas C. T. Michaels, Sara Linse, Cecilia Månsson, Cecilia Emanuelsson, Jenny Presto, Jan Johansson, Michele Vendruscolo, Christopher M. Dobson, and Tuomas P. J. Knowles. 2016. "Kinetic Analysis Reveals the Diversity of Microscopic Mechanisms through Which Molecular Chaperones Suppress Amyloid Formation." *Nature Communications* 7 (March): 10948.
- Arpino, James A. J., Samuel C. Reddington, Lisa M. Halliwell, Pierre J. Rizkallah, and D. Dafydd Jones. 2014. "Random Single Amino Acid Deletion Sampling Unveils Structural Tolerance and the Benefits of Helical Registry Shift on GFP Folding and Structure." *Structure* 22 (6): 889–98.
- Ayala, Youhna M., Laura De Conti, S. Eréndira Avendaño-Vázquez, Ashish Dhir, Maurizio Romano, Andrea D'Ambrogio, James Tollervy, et al. 2011. "TDP-43 Regulates Its mRNA Levels through a Negative Feedback Loop." *The EMBO Journal* 30 (2): 277–88.
- Bachas, Sharrol, Goran Rakocevic, David Spencer, Anand V. Sastry, Robel Haile, John M. Sutton, George Kasun, et al. 2022. "Antibody Optimization Enabled by Artificial Intelligence Predictions of Binding Affinity and Naturalness." *BioRxiv*. <https://doi.org/10.1101/2022.08.16.504181>.
- Baeza-Centurion, Pablo, Belén Miñana, Juan Valcárcel, and Ben Lehner. 2020. "Mutations Primarily Alter the Inclusion of Alternatively Spliced Exons." *ELife* 9 (October). <https://doi.org/10.7554/eLife.59959>.
- Baldwin, Andrew J., Tuomas P. J. Knowles, Gian Gaetano Tartaglia, Anthony W. Fitzpatrick, Glyn L. Devlin, Sarah Lucy Shammass, Christopher A. Waudby, et al. 2011. "Metastability of Native Proteins and the Phenomenon of Amyloid Formation." *Journal of the American Chemical Society* 133 (36): 14160–63.
- Benilova, Iryna, Rodrigo Gallardo, Andreea-Alexandra Ungureanu, Virginia Castillo Cano, An Snellinx, Meine Ramakers, Carmen Bartic, Frederic Rousseau, Joost Schymkowitz, and Bart De Strooper. 2014. "The Alzheimer Disease Protective Mutation A2T Modulates Kinetic and Thermodynamic Properties of Amyloid- β ($A\beta$) Aggregation." *The Journal of Biological Chemistry* 289 (45): 30977–89.
- Bolognesi, Benedetta, Samuel I. A. Cohen, Pablo Aran Terol, Elin K. Esbjörner, Sofia Giorgetti, Maria F. Mossuto, Antonino Natalello, et al. 2014. "Single Point Mutations Induce a Switch in the Molecular Mechanism of the Aggregation of the Alzheimer's Disease Associated $A\beta_{42}$ Peptide." *ACS Chemical Biology*. <https://doi.org/10.1021/cb400616y>.

- Bolognesi, Benedetta, Andre J. Faure, Mireia Seuma, Jörn M. Schmiedel, Gian Gaetano Tartaglia, and Ben Lehner. 2019. "The Mutational Landscape of a Prion-like Domain." *Nature Communications* 10 (1): 4162.
- Bolognesi, Benedetta, Janet R. Kumita, Teresa P. Barros, Elin K. Esbjorner, Leila M. Luheshi, Damian C. Crowther, Mark R. Wilson, Christopher M. Dobson, Giorgio Favrin, and Justin J. Yerbury. 2010. "ANS Binding Reveals Common Features of Cytotoxic Amyloid Species." *ACS Chemical Biology* 5 (8): 735–40.
- Buell, Alexander K. 2022. "Stability Matters, Too – the Thermodynamics of Amyloid Fibril Formation." *Chemical Science*, February. <https://doi.org/10.1039/D1SC06782F>.
- Cagiada, Matteo, Kristoffer E. Johansson, Audrone Valanciute, Sofie V. Nielsen, Rasmus Hartmann-Petersen, Jun J. Yang, Douglas M. Fowler, Amelie Stein, and Kresten Lindorff-Larsen. 2021. "Understanding the Origins of Loss of Protein Function by Analyzing the Effects of Thousands of Variants on Activity and Abundance." *Molecular Biology and Evolution* 38 (8): 3235–46.
- Calderon, Diego, Andria Ellis, Riza M. Daza, Beth Martin, Jacob M. Tome, Wei Chen, Florence M. Chardon, et al. 2020. "TransMPRA: A Framework for Assaying the Role of Many Trans-Acting Factors at Many Enhancers." *BioRxiv*. <https://doi.org/10.1101/2020.09.30.321323>.
- Carter, P. J., G. Winter, A. J. Wilkinson, and A. R. Fersht. 1984. "The Use of Double Mutants to Detect Structural Changes in the Active Site of the Tyrosyl-TRNA Synthetase (*Bacillus Stearothermophilus*)." *Cell* 38 (3): 835–40.
- Chandramowliswaran, Pavithra, Meng Sun, Kristin L. Casey, Andrey V. Romanyuk, Anastasiya V. Grizel, Julia V. Sopova, Aleksandr A. Rubel, Carmen Nussbaum-Krammer, Ina M. Vorberg, and Yury O. Chernoff. 2018. "Mammalian Amyloidogenic Proteins Promote Prion Nucleation in Yeast." *The Journal of Biological Chemistry* 293 (9): 3436–50.
- Chapman, Matthew R., Lloyd S. Robinson, Jerome S. Pinkner, Robyn Roth, John Heuser, Marten Hammar, Staffan Normark, and Scott J. Hultgren. 2002. "Role of *Escherichia Coli* Curli Operons in Directing Amyloid Fiber Formation." *Science* 295 (5556): 851–55.
- Chen, Dailu, Kenneth W. Drombosky, Zhiqiang Hou, Levent Sari, Omar M. Kashmer, Bryan D. Ryder, Valerie A. Perez, et al. 2019. "Tau Local Structure Shields an Amyloid-Forming Motif and Controls Aggregation Propensity." *Nature Communications* 10 (1): 2493.
- Chen, Wei-Ting, Chen-Jee Hong, Ya-Tzu Lin, Wen-Han Chang, He-Ting Huang, Jih-Ying Liao, Yu-Jen Chang, et al. 2012. "Amyloid-Beta (A β) D7H Mutation Increases Oligomeric A β 42 and Alters Properties of A β -Zinc/Copper Assemblies." *PloS One* 7 (4): e35807.
- Chesmore, Kevin, Jacqueline Bartlett, and Scott M. Williams. 2018. "The Ubiquity of Pleiotropy in Human Disease." *Human Genetics* 137 (1): 39–44.
- Chimon, Sandra, Medhat A. Shaibat, Christopher R. Jones, Diana C. Calero, Buzulagu Aizezi, and Yoshitaka Ishii. 2007. "Evidence of Fibril-like β -Sheet Structures in a Neurotoxic Amyloid Intermediate of Alzheimer's β -Amyloid." *Nature Structural & Molecular Biology* 14 (12): 1157–64.
- Chiti, Fabrizio, and Christopher M. Dobson. 2017. "Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade." *Annual Review of Biochemistry* 86 (June): 27–68.
- Cohen, Samuel I. A., Paolo Arosio, Jenny Presto, Firoz Roshan Kurudenkandy, Henrik Biverstål, Lisa Dolfe, Christopher Dunning, et al. 2015. "A Molecular Chaperone Breaks the Catalytic Cycle That Generates Toxic A β Oligomers." *Nature Structural & Molecular Biology*. <https://doi.org/10.1038/nsmb.2971>.
- Cohen, Samuel I. A., Risto Cukalevski, Thomas C. T. Michaels, Anđela Šarić, Mattias Törnquist, Michele Vendruscolo, Christopher M. Dobson, Alexander K. Buell, Tuomas P. J. Knowles, and Sara Linse. 2018. "Distinct Thermodynamic Signatures of Oligomer Generation in the Aggregation of the Amyloid- β Peptide." *Nature Chemistry* 10 (5): 523–31.
- Cohen, Samuel I. A., Sara Linse, Leila M. Luheshi, Erik Hellstrand, Duncan A. White, Luke Rajah, Daniel E. Otzen, Michele Vendruscolo, Christopher M. Dobson, and Tuomas P. J. Knowles. 2013. "Proliferation of Amyloid-B42 Aggregates Occurs through a Secondary Nucleation Mechanism." *Proceedings of the National Academy of Sciences of the United States of America* 110 (24): 9758–63.
- Cohen, Samuel I. A., Michele Vendruscolo, Christopher M. Dobson, and Tuomas P. J. Knowles. 2012. "From Macroscopic Measurements to Microscopic Mechanisms of Protein Aggregation." *Journal of Molecular Biology* 421 (2–3): 160–71.
- Colom-Cadena, Martí, Ellen Gelpi, Sara Charif, Olivia Belbin, Rafael Blesa, Maria J. Martí, Jordi Clarimón, and Alberto Lleó. 2013. "Confluence of α -Synuclein, Tau, and β -Amyloid Pathologies

- in Dementia with Lewy Bodies.” *Journal of Neuropathology and Experimental Neurology* 72 (12): 1203–12.
- Colon, and Kelly. 1992. “Partial Denaturation of Transthyretin Is Sufficient for Amyloid Fibril Formation in Vitro.” *Biochemistry*.
- Colvin, Michael T., Robert Silvers, Qing Zhe Ni, Thach V. Can, Ivan Sergeev, Melanie Rosay, Kevin J. Donovan, et al. 2016. “Atomic Resolution Structure of Monomorphic A β 42 Amyloid Fibrils.” *Journal of the American Chemical Society* 138 (30): 9663–74.
- Coyote-Maestas, Willow, Yungui He, Chad L. Myers, and Daniel Schmidt. 2019. “Domain Insertion Permissibility-Guided Engineering of Allosteric Ion Channels.” *Nature Communications* 10 (1): 290.
- Coyote-Maestas, Willow, David Nedrud, Yungui He, and Daniel Schmidt. 2022. “Determinants of Trafficking, Conduction, and Disease within a K⁺ Channel Revealed through Multiparametric Deep Mutational Scanning.” *ELife* 11 (May). <https://doi.org/10.7554/eLife.76903>.
- Cremades, Nunilo, Samuel I. A. Cohen, Emma Deas, Andrey Y. Abramov, Allen Y. Chen, Angel Orte, Massimo Sandal, et al. 2012. “Direct Observation of the Interconversion of Normal and Toxic Forms of α -Synuclein.” *Cell* 149 (5): 1048–59.
- Cukalevski, Risto, Xiaoting Yang, Georg Meisl, Ulrich Weininger, Katja Bernfur, Birgitta Frohm, Tuomas P. J. Knowles, and Sara Linse. 2015. “The A β 40 and A β 42 Peptides Self-Assemble into Separate Homomolecular Fibrils in Binary Mixtures but Cross-React during Primary Nucleation.” *Chemical Science*. <https://doi.org/10.1039/c4sc02517b>.
- Das, Madhurima, Christopher J. Wilson, Xiaohu Mei, Thomas E. Wales, John R. Engen, and Olga Gursky. 2016. “Structural Stability and Local Dynamics in Disease-Causing Mutants of Human Apolipoprotein A-I: What Makes the Protein Amyloidogenic?” *Journal of Molecular Biology* 428 (2 Pt B): 449–62.
- De Jonghe, C., C. Esselens, S. Kumar-Singh, K. Craessaerts, S. Serneels, F. Checler, W. Annaert, C. Van Broeckhoven, and B. De Strooper. 2001. “Pathogenic APP Mutations near the Gamma-Secretase Cleavage Site Differentially Affect Abeta Secretion and APP C-Terminal Fragment Stability.” *Human Molecular Genetics* 10 (16): 1665–71.
- Di Fede, Giuseppe, Marcella Catania, Michela Morbin, Giacomina Rossi, Silvia Suardi, Giulia Mazzoleni, Marco Merlin, et al. 2009. “A Recessive Mutation in the APP Gene with Dominant-Negative Effect on Amyloidogenesis.” *Science* 323 (5920): 1473–77.
- Diss, Guillaume, and Ben Lehner. 2018. “The Genetic Landscape of a Physical Interaction.” *ELife* 7 (April). <https://doi.org/10.7554/eLife.32472>.
- Dobson, Christopher M. 2003. “Protein Folding and Misfolding.” *Nature* 426 (6968): 884–90.
- Domingo, Júlia, Pablo Baeza-Centurion, and Ben Lehner. 2019. “The Causes and Consequences of Genetic Interactions (Epistasis).” *Annual Review of Genomics and Human Genetics* 20 (August): 433–60.
- Domingo, Júlia, Guillaume Diss, and Ben Lehner. 2018. “Pairwise and Higher-Order Genetic Interactions during the Evolution of a tRNA.” *Nature* 558 (7708): 117–21.
- Dyson, H. Jane, and Peter E. Wright. 2005. “Intrinsically Unstructured Proteins and Their Functions.” *Nature Reviews. Molecular Cell Biology* 6 (3): 197–208.
- Emond, Stéphane, Philippe Mondon, Sandra Pizzut-Serin, Laurent Douchy, Fabien Crozet, Khalil Bouayadi, Hakim Kharrat, Gabrielle Potocki-Véronèse, Pierre Monsan, and Magali Rемаud-Simeon. 2008. “A Novel Random Mutagenesis Approach Using Human Mutagenic DNA Polymerases to Generate Enzyme Variant Libraries.” *Protein Engineering, Design & Selection: PEDS* 21 (4): 267–74.
- Emond, Stéphane, Maya Petek, Emily J. Kay, Brennen Heames, Sean R. A. Devenish, Nobuhiko Tokuriki, and Florian Hollfelder. 2020. “Accessing Unexplored Regions of Sequence Space in Directed Enzyme Evolution via Insertion/Deletion Mutagenesis.” *Nature Communications* 11 (1): 3469.
- Erwood, Steven, Teija M. I. Bily, Jason Lequyer, Joyce Yan, Nitya Gulati, Reid A. Brewer, Liangchi Zhou, Laurence Pelletier, Evgueni A. Ivakine, and Ronald D. Cohn. 2022. “Saturation Variant Interpretation Using CRISPR Prime Editing.” *Nature Biotechnology* 40 (6): 885–95.
- Esposito, Daniel, Jochen Weile, Jay Shendure, Lea M. Starita, Anthony T. Papenfuss, Frederick P. Roth, Douglas M. Fowler, and Alan F. Rubin. 2019. “MaveDB: An Open-Source Platform to Distribute and Interpret Data from Multiplexed Assays of Variant Effect.” *Genome Biology* 20 (1): 223.
- Falcon, Benjamin, Wenjuan Zhang, Alexey G. Murzin, Garib Murshudov, Holly J. Garringer, Ruben Vidal, R. Anthony Crowther, Bernardino Ghetti, Sjors H. W. Scheres, and Michel Goedert. 2018. “Structures of Filaments from Pick’s Disease Reveal a Novel Tau Protein Fold.” *Nature* 561 (7721): 137–40.
- Falcon, Benjamin, Wenjuan Zhang, Manuel Schweighauser, Alexey G. Murzin, Ruben Vidal, Holly J. Garringer, Bernardino Ghetti, Sjors H. W. Scheres, and Michel Goedert. 2018. “Tau Filaments

- from Multiple Cases of Sporadic and Inherited Alzheimer's Disease Adopt a Common Fold." *Acta Neuropathologica* 136 (5): 699–708.
- Fang, Yu-Sheng, Kuen-Jer Tsai, Yu-Jen Chang, Patricia Kao, Rima Woods, Pan-Hsien Kuo, Cheng-Chun Wu, et al. 2014. "Full-Length TDP-43 Forms Toxic Amyloid Oligomers That Are Present in Frontotemporal Lobar Dementia-TDP Patients." *Nature Communications*. <https://doi.org/10.1038/ncomms5824>.
- Faure, Andre J., Júlia Domingo, Jörn M. Schmiedel, Cristina Hidalgo-Carcedo, Guillaume Diss, and Ben Lehner. 2022. "Mapping the Energetic and Allosteric Landscapes of Protein Binding Domains." *Nature* 604 (7904): 175–83.
- Faure, Andre J., Jörn M. Schmiedel, Pablo Baeza-Centurion, and Ben Lehner. 2020. "DiMSum: An Error Model and Pipeline for Analyzing Deep Mutational Scanning Data and Diagnosing Common Experimental Pathologies." *Genome Biology* 21 (1): 207.
- Fayer, Shawn, Carrie Horton, Jennifer N. Dines, Alan F. Rubin, Marcy E. Richardson, Kelly McGoldrick, Felicia Hernandez, et al. 2021. "Closing the Gap: Systematic Integration of Multiplexed Functional Data Resolves Variants of Uncertain Significance in BRCA1, TP53, and PTEN." *American Journal of Human Genetics* 108 (12): 2248–58.
- Fernandez-Escamilla, Ana-Maria, Frederic Rousseau, Joost Schymkowitz, and Luis Serrano. 2004. "Prediction of Sequence-Dependent and Mutational Effects on the Aggregation of Peptides and Proteins." *Nature Biotechnology* 22 (10): 1302–6.
- Ferrone, F. 1999. "Analysis of Protein Aggregation Kinetics." *Methods in Enzymology* 309: 256–74.
- Findlay, Gregory M. 2021. "Linking Genome Variants to Disease: Scalable Approaches to Test the Functional Impact of Human Mutations." *Human Molecular Genetics* 30 (R2): R187–97.
- Findlay, Gregory M., Evan A. Boyle, Ronald J. Hause, Jason C. Klein, and Jay Shendure. 2014. "Saturation Editing of Genomic Regions by Multiplex Homology-Directed Repair." *Nature* 513 (7516): 120–23.
- Findlay, Gregory M., Riza M. Daza, Beth Martin, Melissa D. Zhang, Anh P. Leith, Molly Gasperini, Joseph D. Janizek, Xingfan Huang, Lea M. Starita, and Jay Shendure. 2018. "Accurate Classification of BRCA1 Variants with Saturation Genome Editing." *Nature* 562 (7726): 217–22.
- Firnberg, Elad, and Marc Ostermeier. 2012. "PFunkel: Efficient, Expansive, User-Defined Mutagenesis." *PloS One* 7 (12): e52031.
- Fitzpatrick, Anthony W., Tuomas P. J. Knowles, Christopher A. Waudby, Michele Vendruscolo, and Christopher M. Dobson. 2011. "Inversion of the Balance between Hydrophobic and Hydrogen Bonding Interactions in Protein Folding and Aggregation." *PLoS Computational Biology* 7 (10): e1002169.
- Fitzpatrick, Anthony W. P., Benjamin Falcon, Shaoda He, Alexey G. Murzin, Garib Murshudov, Holly J. Garringer, R. Anthony Crowther, Bernardino Ghetti, Michel Goedert, and Sjors H. W. Scheres. 2017. "Cryo-EM Structures of Tau Filaments from Alzheimer's Disease." *Nature* 547 (7662): 185–90.
- Fowler, Douglas M., Carlos L. Araya, Sarel J. Fleishman, Elizabeth H. Kellogg, Jason J. Stephany, David Baker, and Stanley Fields. 2010. "High-Resolution Mapping of Protein Sequence-Function Relationships." *Nature Methods* 7 (9): 741–46.
- Fusco, Giuliana, Serene W. Chen, Philip T. F. Williamson, Roberta Cascella, Michele Perni, James A. Jarvis, Cristina Cecchi, et al. 2017. "Structural Basis of Membrane Disruption and Cellular Toxicity by α -Synuclein Oligomers." *Science* 358 (6369): 1440–43.
- Galvagnion, Céline, Alexander K. Buell, Georg Meisl, Thomas C. T. Michaels, Michele Vendruscolo, Tuomas P. J. Knowles, and Christopher M. Dobson. 2015. "Lipid Vesicles Trigger α -Synuclein Aggregation by Stimulating Primary Nucleation." *Nature Chemical Biology* 11 (3): 229–34.
- Gasperini, Molly, Lea Starita, and Jay Shendure. 2016. "The Power of Multiplexed Functional Analysis of Genetic Variants." *Nature Protocols* 11 (10): 1782–87.
- Gazit, Ehud. 2002. "A Possible Role for Pi-Stacking in the Self-Assembly of Amyloid Fibrils." *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology* 16 (1): 77–83.
- Gersing, Sarah, Matteo Cagiada, Marinella Gebbia, Anette P. Gjesing, Gireesh Seesankar, Amelie Stein, Anna L. Gloyn, et al. 2022. "A Multiplexed Assay of Human Glucokinase Reveals Thousands of Potential Disease Variants with Both Decreased and Increased Activity." *BioRxiv*. <https://doi.org/10.1101/2022.05.04.490571>.
- Gonzalez, Courtney E., Paul Roberts, and Marc Ostermeier. 2019. "Fitness Effects of Single Amino Acid Insertions and Deletions in TEM-1 β -Lactamase." *Journal of Molecular Biology* 431 (12): 2320–30.

- Gray, Vanessa E., Ronald J. Hause, Jens Luebeck, Jay Shendure, and Douglas M. Fowler. 2018. "Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data." *Cell Systems* 6 (1): 116-124.e3.
- Gray, Vanessa E., Katherine Sitko, Floriane Z. Ngako Kameni, Miriam Williamson, Jason J. Stephany, Nicholas Hasle, and Douglas M. Fowler. 2019. "Elucidating the Molecular Determinants of A β Aggregation with Deep Mutational Scanning." *G3* 9 (11): 3683-89.
- Gremer, Lothar, Daniel Schölzel, Carla Schenk, Elke Reinartz, Jörg Labahn, Raimond B. G. Ravelli, Markus Tusche, et al. 2017. "Fibril Structure of Amyloid- β (1-42) by Cryo-Electron Microscopy." *Science* 358 (6359): 116-19.
- Griffith, Daniel, and Alex S. Holehouse. 2021. "PARROT Is a Flexible Recurrent Neural Network Framework for Analysis of Large Protein Datasets." *ELife* 10 (September). <https://doi.org/10.7554/eLife.70576>.
- Guenther, Elizabeth L., Qin Cao, Hamilton Trinh, Jiahui Lu, Michael R. Sawaya, Duilio Cascio, David R. Boyer, Jose A. Rodriguez, Michael P. Hughes, and David S. Eisenberg. 2018. "Atomic Structures of TDP-43 LCD Segments and Insights into Reversible or Pathogenic Aggregation." *Nature Structural & Molecular Biology* 25 (6): 463-71.
- Guerrero-Ferreira, Ricardo, Lubomir Kovacic, Dongchun Ni, and Henning Stahlberg. 2020. "New Insights on the Structure of Alpha-Synuclein Fibrils Using Cryo-Electron Microscopy." *Current Opinion in Neurobiology* 61 (April): 89-95.
- Haass, Christian, and Dennis J. Selkoe. 2007. "Soluble Protein Oligomers in Neurodegeneration: Lessons from the Alzheimer's Amyloid Beta-Peptide." *Nature Reviews. Molecular Cell Biology* 8 (2): 101-12.
- Haddock, Hugh K., Adam S. Dingens, Sarah K. Hilton, Julie Overbaugh, and Jesse D. Bloom. 2018. "Mapping Mutational Effects along the Evolutionary Landscape of HIV Envelope." *ELife* 7 (March). <https://doi.org/10.7554/eLife.34420>.
- Harrison, Alice Ford, and James Shorter. 2017. "RNA-Binding Proteins with Prion-like Domains in Health and Disease." *Biochemical Journal* 474 (8): 1417-38.
- Hartl, F. Ulrich, and Manajit Hayer-Hartl. 2009. "Converging Concepts of Protein Folding in Vitro and in Vivo." *Nature Structural & Molecular Biology* 16 (6): 574-81.
- Hasle, Nicholas, Anthony Cooke, Sanjay Srivatsan, Heather Huang, Jason J. Stephany, Zachary Krieger, Dana Jackson, et al. 2020. "High-Throughput, Microscope-Based Sorting to Dissect Cellular Heterogeneity." *Molecular Systems Biology* 16 (6): e9442.
- Hatami, Asa, Sanaz Monjazebe, Saskia Milton, and Charles G. Glabe. 2017. "Familial Alzheimer's Disease Mutations within the Amyloid Precursor Protein Alter the Aggregation and Conformation of the Amyloid- β Peptide." *The Journal of Biological Chemistry* 292 (8): 3172-85.
- Hellstrand, Erik, Barry Boland, Dominic M. Walsh, and Sara Linse. 2010. "Amyloid β -Protein Aggregation Produces Highly Reproducible Kinetic Data and Occurs by a Two-Phase Process." *ACS Chemical Neuroscience* 1 (1): 13-18.
- Herrera, Fernando E., Alessandra Chesi, Katerina E. Paleologou, Adrian Schmid, Adriana Munoz, Michele Vendruscolo, Stefano Gustincich, Hilal A. Lashuel, and Paolo Carloni. 2008. "Inhibition of Alpha-Synuclein Fibrillization by Dopamine Is Mediated by Interactions with Five C-Terminal Residues and with E83 in the NAC Region." *PLoS One* 3 (10): e3394.
- Hindorff, Lucia A., Praveen Sethupathy, Heather A. Junkins, Erin M. Ramos, Jayashri P. Mehta, Francis S. Collins, and Teri A. Manolio. 2009. "Potential Etiologic and Functional Implications of Genome-Wide Association Loci for Human Diseases and Traits." *Proceedings of the National Academy of Sciences of the United States of America* 106 (23): 9362-67.
- Høie, Magnus Haraldson, Matteo Cagiada, Anders Haagen Beck Frederiksen, Amelie Stein, and Kresten Lindorff-Larsen. 2022. "Predicting and Interpreting Large-Scale Mutagenesis Data Using Analyses of Protein Stability and Conservation." *Cell Reports* 38 (2): 110207.
- Hori, Yukiko, Tadafumi Hashimoto, Yosuke Wakutani, Katsuya Urakami, Kenji Nakashima, Margaret M. Condrón, Satoshi Tsubuki, Takaomi C. Saido, David B. Teplow, and Takeshi Iwatsubo. 2007. "The Tottori (D7N) and English (H6R) Familial Alzheimer Disease Mutations Accelerate Abeta Fibril Formation without Increasing Protofibril Formation." *The Journal of Biological Chemistry* 282 (7): 4916-23.
- Horovitz, A. 1996. "Double-Mutant Cycles: A Powerful Tool for Analyzing Protein Structure and Function." *Folding and Design* 1 (6): R121-6.
- Horovitz, A., and A. R. Fersht. 1990. "Strategy for Analysing the Co-Operativity of Intramolecular Interactions in Peptides and Proteins." *Journal of Molecular Biology* 214 (3): 613-17.

- Houben, Bert, Frederic Rousseau, and Joost Schymkowitz. 2022. "Protein Structure and Aggregation: A Marriage of Necessity Ruled by Aggregation Gatekeepers." *Trends in Biochemical Sciences* 47 (3): 194–205.
- Iadanza, Matthew G., Matthew P. Jackson, Eric W. Hewitt, Neil A. Ranson, and Sheena E. Radford. 2018. "A New Era for Understanding Amyloid Structures and Disease." *Nature Reviews. Molecular Cell Biology* 19 (12): 755–73.
- Jahn, Thomas R., and Sheena E. Radford. 2005. "The Yin and Yang of Protein Folding." *The FEBS Journal* 272 (23): 5962–70.
- Jain, Pankaj C., and Raghavan Varadarajan. 2014. "A Rapid, Efficient, and Economical Inverse Polymerase Chain Reaction-Based Method for Generating a Site Saturation Mutant Library." *Analytical Biochemistry* 449 (March): 90–98.
- Jiang, Wei, and Liang Chen. 2021. "Alternative Splicing: Human Disease and Quantitative Analysis from High-Throughput Sequencing." *Computational and Structural Biotechnology Journal* 19: 183–95.
- Julien, Philippe, Belén Miñana, Pablo Baeza-Centurion, Juan Valcárcel, and Ben Lehner. 2016. "The Complete Local Genotype–Phenotype Landscape for the Alternative Splicing of a Human Exon." *Nature Communications* 7 (1): 1–8.
- Kant, Rob van der, Nikolaos Louros, Joost Schymkowitz, and Frederic Rousseau. 2022. "Thermodynamic Analysis of Amyloid Fibril Structures Reveals a Common Framework for Stability in Amyloid Polymorphs." *Structure* 30 (8): 1178-1189.e3.
- Karthika, S., T. K. Radhakrishnan, and P. Kalaichelvi. 2016. "A Review of Classical and Nonclassical Nucleation Theories." *Crystal Growth & Design*. <https://doi.org/10.1021/acs.cgd.6b00794>.
- Kayed, Rakez, Elizabeth Head, Jennifer L. Thompson, Theresa M. McIntire, Saskia C. Milton, Carl W. Cotman, and Charles G. Glabe. 2003. "Common Structure of Soluble Amyloid Oligomers Implies Common Mechanism of Pathogenesis." *Science* 300 (5618): 486–89.
- Khan, Moien Abdul Basith, Muhammad Jawad Hashim, Jeffrey Kwan King, Romona Devi Govender, Halla Mustafa, and Juma Al Kaabi. 2020. "Epidemiology of Type 2 Diabetes - Global Burden of Disease and Forecasted Trends." *Journal of Epidemiology and Global Health* 10 (1): 107–11.
- Khan, Tarique, Tejbir S. Kandola, Jianzheng Wu, Shriram Venkatesan, Ellen Ketter, Jeffrey J. Lange, Alejandro Rodríguez Gama, et al. 2018. "Quantifying Nucleation In Vivo Reveals the Physical Basis of Prion-like Phase Behavior." *Molecular Cell* 71 (1): 155-168.e7.
- Kinney, and McCandlish. n.d. "Massively Parallel Assays and Quantitative Sequence–Function Relationships." *Annual Review of ... Antitrust Law Developments*. <http://kinneylab.labsites.cshl.edu/wp-content/uploads/sites/60/2018/08/KinneyMcCandlish2019.pdf>.
- Knowles, Tuomas P. J., Michele Vendruscolo, and Christopher M. Dobson. 2014. "The Amyloid State and Its Association with Protein Misfolding Diseases." *Nature Reviews. Molecular Cell Biology* 15 (6): 384–96.
- Knowles, Tuomas P. J., Christopher A. Waudby, Glyn L. Devlin, Samuel I. A. Cohen, Adriano Aguzzi, Michele Vendruscolo, Eugene M. Terentjev, Mark E. Welland, and Christopher M. Dobson. 2009. "An Analytical Solution to the Kinetics of Breakable Filament Assembly." *Science*. <https://doi.org/10.1126/science.1178250>.
- Kowalsky, Caitlin A., Justin R. Klesmith, James A. Stapleton, Vince Kelly, Nolan Reichkitzer, and Timothy A. Whitehead. 2015. "High-Resolution Sequence-Function Mapping of Full-Length Proteins." *PloS One* 10 (3): e0118193.
- Kummer, Markus P., and Michael T. Heneka. 2014. "Truncated and Modified Amyloid-Beta Species." *Alzheimer's Research & Therapy* 6 (3): 28.
- Kyte, J., and R. F. Doolittle. 1982. "A Simple Method for Displaying the Hydrophobic Character of a Protein." *Journal of Molecular Biology* 157 (1): 105–32.
- Landrum, Melissa J., Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church, and Donna R. Maglott. 2014. "ClinVar: Public Archive of Relationships among Sequence Variation and Human Phenotype." *Nucleic Acids Research* 42 (Database issue): D980-5.
- Li, Jixi, Thomas McQuade, Ansgar B. Siemer, Johanna Napetschnig, Kenta Moriwaki, Yu-Shan Hsiao, Ermelinda Damko, et al. 2012. "The RIP1/RIP3 Necrosome Forms a Functional Amyloid Signaling Complex Required for Programmed Necrosis." *Cell* 150 (2): 339–50.
- Li, Xianghua, Jasna Lalic, Pablo Baeza-Centurion, Riddhiman Dhar, and Ben Lehner. n.d. "Changes in Gene Expression Shift and Switch Genetic Interactions." <https://doi.org/10.1101/578419>.
- Li, Xianghua, and Ben Lehner. 2020. "Biophysical Ambiguities Prevent Accurate Genetic Prediction." *Nature Communications* 11 (1): 4923.

- Ling, Shuo-Chien, Magdalini Polymenidou, and Don W. Cleveland. 2013. “Converging Mechanisms in ALS and FTD: Disrupted RNA and Protein Homeostasis.” *Neuron* 79 (3): 416–38.
- Lovell, Simon C., Ian W. Davis, W. Bryan Arendall 3rd, Paul I. W. de Bakker, J. Michael Word, Michael G. Prisant, Jane S. Richardson, and David C. Richardson. 2003. “Structure Validation by Calpha Geometry: Phi, Psi and Cbeta Deviation.” *Proteins* 50 (3): 437–50.
- Lövestam, Sofia, Fujiet Adrian Koh, Bart van Knippenberg, Abhay Kotecha, Alexey G. Murzin, Michel Goedert, and Sjors H. W. Scheres. 2022. “Assembly of Recombinant Tau into Filaments Identical to Those of Alzheimer’s Disease and Chronic Traumatic Encephalopathy.” *ELife* 11 (March). <https://doi.org/10.7554/eLife.76494>.
- Lührs, Thorsten, Christiane Ritter, Marc Adrian, Dominique Riek-Loher, Bernd Bohrmann, Heinz Döbeli, David Schubert, and Roland Riek. 2005. “3D Structure of Alzheimer’s Amyloid-Beta(1-42) Fibrils.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (48): 17342–47.
- Macdonald, Christian B., David Nedrud, Patrick Rockefeller Grimes, Donovan Trinidad, James S. Fraser, and Willow Coyote-Maestas. 2022. “Deep Insertion, Deletion, and Missense Mutation Libraries for Exploring Protein Variation in Evolution, Disease, and Biology.” *BioRxiv*. <https://doi.org/10.1101/2022.07.26.501589>.
- Maji, Samir K., Marilyn H. Perrin, Michael R. Sawaya, Sebastian Jessberger, Krishna Vadodaria, Robert A. Rissman, Praful S. Singru, et al. 2009. “Functional Amyloids as Natural Storage of Peptide Hormones in Pituitary Secretory Granules.” *Science* 325 (5938): 328–32.
- Mannini, Benedetta, Roberta Cascella, Mariagiorgia Zampagni, Maria van Waarde-Verhagen, Sarah Meehan, Cintia Roodveldt, Silvia Campioni, et al. 2012. “Molecular Mechanisms Used by Chaperones to Reduce the Toxicity of Aberrant Protein Oligomers.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (31): 12479–84.
- Maricque, Brett B., Hemangi G. Chaudhari, and Barak A. Cohen. 2018. “A Massively Parallel Reporter Assay Dissects the Influence of Chromatin Structure on Cis-Regulatory Activity.” *Nature Biotechnology*, November. <https://doi.org/10.1038/nbt.4285>.
- Martin, Erik W., Alex S. Holehouse, Ivan Peran, Mina Farag, J. Jeremias Incicco, Anne Bremer, Christy R. Grace, Andrea Soranno, Rohit V. Pappu, and Tanja Mittag. 2020. “Valence and Patterning of Aromatic Residues Determine the Phase Behavior of Prion-like Domains.” *Science* 367 (6478): 694–99.
- Matreyek, Kenneth A., Lea M. Starita, Jason J. Stephany, Beth Martin, Melissa A. Chiasson, Vanessa E. Gray, Martin Kircher, et al. 2018. “Multiplex Assessment of Protein Variant Abundance by Massively Parallel Sequencing.” *Nature Genetics* 50 (6): 874–82.
- McCormick, James W., Marielle Ax Russo, Samuel Thompson, Aubrie Blevins, and Kimberly A. Reynolds. 2021. “Structurally Distributed Surface Sites Tune Allosteric Regulation.” *ELife* 10 (June). <https://doi.org/10.7554/eLife.68346>.
- Meinema, Anne C., Justyna K. Laba, Rizqiya A. Hapsari, Renee Otten, Frans A. A. Mulder, Annemarie Kralt, Geert van den Bogaart, C. Patrick Lusk, Bert Poolman, and Liesbeth M. Veenhoff. 2011. “Long Unfolded Linkers Facilitate Membrane Protein Import through the Nuclear Pore Complex.” *Science* 333 (6038): 90–93.
- Meisl, Georg, Luke Rajah, Samuel A. I. Cohen, Manuela Pfammatter, Anđela Šarić, Erik Hellstrand, Alexander K. Buell, et al. 2017. “Scaling Behaviour and Rate-Determining Steps in Filamentous Self-Assembly.” *Chemical Science*. <https://doi.org/10.1039/c7sc01965c>.
- Meisl, Georg, Catherine K. Xu, Jonathan D. Taylor, Thomas C. T. Michaels, Aviad Levin, Daniel Otzen, David Klenerman, et al. 2022. “Uncovering the Universality of Self-Replication in Protein Aggregation and Its Link to Disease.” *Science Advances* 8 (32): eabn6831.
- Meisl, Georg, Xiaoting Yang, Birgitta Frohm, Tuomas P. J. Knowles, and Sara Linse. 2016. “Quantitative Analysis of Intrinsic and Extrinsic Factors in the Aggregation Mechanism of Alzheimer-Associated A β -Peptide.” *Scientific Reports* 6 (January): 18728.
- Meisl, Georg, Xiaoting Yang, Erik Hellstrand, Birgitta Frohm, Julius B. Kirkegaard, Samuel I. A. Cohen, Christopher M. Dobson, Sara Linse, and Tuomas P. J. Knowles. 2014. “Differences in Nucleation Behavior Underlie the Contrasting Aggregation Kinetics of the A β 40 and A β 42 Peptides.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (26): 9384–89.
- Melamed, Daniel, David L. Young, Caitlin E. Gamble, Christina R. Miller, and Stanley Fields. 2013. “Deep Mutational Scanning of an RRM Domain of the *Saccharomyces Cerevisiae* Poly(A)-Binding Protein.” *RNA* 19 (11): 1537–51.
- Michaels, Thomas C. T., Anđela Šarić, Johnny Habchi, Sean Chia, Georg Meisl, Michele Vendruscolo, Christopher M. Dobson, and Tuomas P. J. Knowles. 2018. “Chemical Kinetics for Bridging

- Molecular Mechanisms and Macroscopic Measurements of Amyloid Fibril Formation.” *Annual Review of Physical Chemistry* 69 (April): 273–98.
- Mok, Sue-Ann, Carlo Condello, Rebecca Freilich, Anne Gillies, Taylor Arhar, Javier Oroz, Harindranath Kadavath, et al. 2018. “Mapping Interactions with the Chaperone Network Reveals Factors That Protect against Tau Aggregation.” *Nature Structural & Molecular Biology* 25 (5): 384–93.
- Monsellier, Elodie, Matteo Ramazzotti, Niccolò Taddei, and Fabrizio Chiti. 2008. “Aggregation Propensity of the Human Proteome.” *PLoS Computational Biology* 4 (10): e1000199.
- Morimoto, Akira, Kazuhiro Irie, Kazuma Murakami, Yuichi Masuda, Hajime Ohigashi, Masaya Nagao, Hiroyuki Fukuda, Takahiko Shimizu, and Takuji Shirasawa. 2004. “Analysis of the Secondary Structure of Beta-Amyloid (Abeta42) Fibrils by Systematic Proline Replacement.” *The Journal of Biological Chemistry* 279 (50): 52781–88.
- Mullaney, Julianne M., Ryan E. Mills, W. Stephen Pittard, and Scott E. Devine. 2010. “Small Insertions and Deletions (INDELs) in Human Genomes.” *Human Molecular Genetics* 19 (R2): R131–6.
- Murray, Brian, Mirco Sorci, Joseph Rosenthal, Jennifer Lippens, David Isaacson, Payel Das, Daniele Fabris, Shaomin Li, and Georges Belfort. 2016. “A2T and A2V A β Peptides Exhibit Different Aggregation Kinetics, Primary Nucleation, Morphology, Structure, and LTP Inhibition.” *Proteins* 84 (4): 488–500.
- Niblock, Michael, and Jean-Marc Gallo. 2012. “Tau Alternative Splicing in Familial and Sporadic Tauopathies.” *Biochemical Society Transactions* 40 (4): 677–80.
- Ogden, Pierce J., Eric D. Kelsic, Sam Sinai, and George M. Church. 2019. “Comprehensive AAV Capsid Fitness Landscape Reveals a Viral Gene and Enables Machine-Guided Design.” *Science* 366 (6469): 1139–43.
- Oliveberg, Mikael. 2010. “Waltz, an Exciting New Move in Amyloid Prediction.” *Nature Methods* 7 (3): 187–88.
- Olson, C. Anders, Nicholas C. Wu, and Ren Sun. 2014. “A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain.” *Current Biology: CB* 24 (22): 2643–51.
- Otzen, Daniel, and Roland Riek. 2019. “Functional Amyloids.” *Cold Spring Harbor Perspectives in Biology* 11 (12). <https://doi.org/10.1101/cshperspect.a033860>.
- Patwardhan, Rupali P., Choli Lee, Oren Litvin, David L. Young, Dana Pe’er, and Jay Shendure. 2009. “High-Resolution Analysis of DNA Regulatory Elements by Synthetic Saturation Mutagenesis.” *Nature Biotechnology* 27 (12): 1173–75.
- Pellarin, Riccardo, Philipp Schuetz, Enrico Guarnera, and Amedeo Caflisch. 2010. “Amyloid Fibril Polymorphism Is under Kinetic Control.” *Journal of the American Chemical Society* 132 (42): 14960–70.
- Perutz, M. F., T. Johnson, M. Suzuki, and J. T. Finch. 1994. “Glutamine Repeats as Polar Zippers: Their Possible Role in Inherited Neurodegenerative Diseases.” *Proceedings of the National Academy of Sciences of the United States of America* 91 (12): 5355–58.
- Petkova, A. T., G. Buntkowsky, F. Dyda, R. D. Leapman, W-M Yau, and R. Tycko. 2004. “Solid State NMR Reveals a PH-Dependent Antiparallel Beta-Sheet Registry in Fibrils Formed by a Beta-Amyloid Peptide.” *Journal of Molecular Biology* 335 (1): 247–60.
- Portelius, Erik, Nenad Bogdanovic, Mikael K. Gustavsson, Inga Volkmann, Gunnar Brinkmalm, Henrik Zetterberg, Bengt Winblad, and Kaj Blennow. 2010. “Mass Spectrometric Characterization of Brain Amyloid Beta Isoform Signatures in Familial and Sporadic Alzheimer’s Disease.” *Acta Neuropathologica* 120 (2): 185–93.
- Qiang, Wei, Wai-Ming Yau, Jun-Xia Lu, John Collinge, and Robert Tycko. 2017. “Structural Variation in Amyloid- β Fibrils from Alzheimer’s Disease Clinical Subtypes.” *Nature* 541 (7636): 217–21.
- Qiang, Wei, Wai-Ming Yau, Yongquan Luo, Mark P. Mattson, and Robert Tycko. 2012. “Antiparallel β -Sheet Architecture in Iowa-Mutant β -Amyloid Fibrils.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (12): 4443–48.
- Raimondi, Sara, Fulvio Guglielmi, Sofia Giorgetti, Sonia Di Gaetano, Angela Arciello, Daria M. Monti, Annalisa Relini, et al. 2011. “Effects of the Known Pathogenic Mutations on the Aggregation Pathway of the Amyloidogenic Peptide of Apolipoprotein A-I.” *Journal of Molecular Biology* 407 (3): 465–76.
- Rentzsch, Philipp, Daniela Witten, Gregory M. Cooper, Jay Shendure, and Martin Kircher. 2019. “CADD: Predicting the Deleteriousness of Variants throughout the Human Genome.” *Nucleic Acids Research* 47 (D1): D886–94.
- Rollins, Nathan J., Kelly P. Brock, Frank J. Poelwijk, Michael A. Stiffler, Nicholas P. Gauthier, Chris Sander, and Debora S. Marks. 2019. “Inferring Protein 3D Structure from Deep Mutation Scans.” *Nature Genetics* 51 (7): 1170–76.

- Rubin, Alan F., Hannah Gelman, Nathan Lucas, Sandra M. Bajjalieh, Anthony T. Papenfuss, Terence P. Speed, and Douglas M. Fowler. 2018. "Correction to: A Statistical Framework for Analyzing Deep Mutational Scanning Data." *Genome Biology* 19 (1): 17.
- Rumbley, J., L. Hoang, L. Mayne, and S. W. Englander. 2001. "An Amino Acid Code for Protein Folding." *Proceedings of the National Academy of Sciences of the United States of America* 98 (1): 105–12.
- Sanborn, Adrian L., Benjamin T. Yeh, Jordan T. Feigerle, Cynthia V. Hao, Raphael JI Townshend, Erez Lieberman Aiden, Ron O. Dror, and Roger D. Kornberg. 2021. "Simple Biochemical Features Underlie Transcriptional Activation Domain Diversity and Dynamic, Fuzzy Binding to Mediator." *ELife* 10 (April). <https://doi.org/10.7554/eLife.68068>.
- Sant'Anna, Ricardo, Carolina Braga, Nathalia Varejão, Karinne M. Pimenta, Ricardo Graña-Montes, Aline Alves, Juliana Cortines, Yraima Cordeiro, Salvador Ventura, and Debora Foguel. 2014. "The Importance of a Gatekeeper Residue on the Aggregation of Transthyretin: Implications for Transthyretin-Related Amyloidoses." *The Journal of Biological Chemistry* 289 (41): 28324–37.
- Šarić, Anđela, Alexander K. Buell, Georg Meisl, Thomas C. T. Michaels, Christopher M. Dobson, Sara Linse, Tuomas P. J. Knowles, and Daan Frenkel. 2016. "Physical Determinants of the Self-Replication of Protein Fibrils." *Nature Physics* 12 (9): 874–80.
- Šarić, Anđela, Yasmine C. Chebaro, Tuomas P. J. Knowles, and Daan Frenkel. 2014. "Crucial Role of Nonspecific Interactions in Amyloid Nucleation." *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.1410159111>.
- Šarić, Anđela, Thomas C. T. Michaels, Alessio Zaccone, Tuomas P. J. Knowles, and Daan Frenkel. 2016. "Kinetics of Spontaneous Filament Nucleation via Oligomers: Insights from Theory and Simulation." *The Journal of Chemical Physics* 145 (21): 211926.
- Sasaguri, Hiroki, Per Nilsson, Shoko Hashimoto, Kenichi Nagata, Takashi Saito, Bart De Strooper, John Hardy, Robert Vassar, Bengt Winblad, and Takaomi C. Saido. 2017. "APP Mouse Models for Alzheimer's Disease Preclinical Studies." *The EMBO Journal* 36 (17): 2473–87.
- Sawaya, Michael R., Michael P. Hughes, Jose A. Rodriguez, Roland Riek, and David S. Eisenberg. 2021. "The Expanding Amyloid Family: Structure, Stability, Function, and Pathogenesis." *Cell* 184 (19): 4857–73.
- Scheibel, Thomas, Jesse Bloom, and Susan L. Lindquist. 2004. "The Elongation of Yeast Prion Fibers Involves Separable Steps of Association and Conversion." *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.0308754101>.
- Scheidt, Tom, Urszula Łapińska, Janet R. Kumita, Daniel R. Whiten, David Klenerman, Mark R. Wilson, Samuel I. A. Cohen, et al. 2019. "Secondary Nucleation and Elongation Occur at Different Sites on Alzheimer's Amyloid- β Aggregates." *Science Advances* 5 (4): eaau3112.
- Schmidt, Matthias, Alexis Rohou, Keren Lasker, Jay K. Yadav, Cordelia Schiene-Fischer, Marcus Fändrich, and Nikolaus Grigorieff. 2015. "Peptide Dimer Structure in an A β (1-42) Fibril Visualized with Cryo-EM." *Proceedings of the National Academy of Sciences of the United States of America* 112 (38): 11858–63.
- Schmidt, Matthias, Sebastian Wiese, Volkan Adak, Jonas Engler, Shubhangi Agarwal, Günter Fritz, Per Westermark, Martin Zacharias, and Marcus Fändrich. 2019. "Cryo-EM Structure of a Transthyretin-Derived Amyloid Fibril from a Patient with Hereditary ATTR Amyloidosis." *Nature Communications*. <https://doi.org/10.1038/s41467-019-13038-z>.
- Schmiedel, Jörn M., and Ben Lehner. 2019. "Determining Protein Structures Using Deep Mutagenesis." *Nature Genetics* 51 (7): 1177–86.
- Schraivogel, Daniel, Terra M. Kuhn, Benedikt Rauscher, Marta Rodríguez-Martínez, Malte Paulsen, Keegan Owsley, Aaron Middlebrook, et al. 2022. "High-Speed Fluorescence Image-Enabled Cell Sorting." *Science* 375 (6578): 315–20.
- Sekijima, Yoshiki, R. Luke Wiseman, Jeanne Matteson, Per Hammarström, Sean R. Miller, Anu R. Sawkar, William E. Balch, and Jeffery W. Kelly. 2005. "The Biological and Chemical Basis for Tissue-Selective Amyloid Disease." *Cell* 121 (1): 73–85.
- Seuma, Mireia, and Benedetta Bolognesi. 2022. "Understanding and Evolving Prions by Yeast Multiplexed Assays." *Current Opinion in Genetics & Development* 75 (June): 101941.
- Seuma, Mireia, Andre J. Faure, Marta Badia, Ben Lehner, and Benedetta Bolognesi. 2021. "The Genetic Landscape for Amyloid Beta Fibril Nucleation Accurately Discriminates Familial Alzheimer's Disease Mutations." *ELife* 10 (February). <https://doi.org/10.7554/eLife.63364>.
- Seuma, Mireia, Ben Lehner, and Benedetta Bolognesi. 2022. "An Atlas of Amyloid Aggregation: The Impact of Substitutions, Insertions, Deletions and Truncations on Amyloid Beta Fibril Nucleation." *BioRxiv*. <https://doi.org/10.1101/2022.01.18.476804>.
- Shi, Yang, Wenjuan Zhang, Yang Yang, Alexey G. Murzin, Benjamin Falcon, Abhay Kotecha, Mike van Beers, et al. 2021. "Structure-Based Classification of Tauopathies." *Nature* 598 (7880): 359–63.

- Sormanni, Pietro, Francesco A. Aprile, and Michele Vendruscolo. 2015. "The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility." *Journal of Molecular Biology* 427 (2): 478–90.
- Spillantini, Maria Grazia, and Michel Goedert. 2013. "Tau Pathology and Neurodegeneration." *Lancet Neurology* 12 (6): 609–22.
- Staller, Max V., Alex S. Holehouse, Devjane Swain-Lenz, Rahul K. Das, Rohit V. Pappu, and Barak A. Cohen. 2018. "A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain." *Cell Systems* 6 (4): 444–455.e6.
- Staller, Max V., Eddie Ramirez, Sanjana R. Kotha, Alex S. Holehouse, Rohit V. Pappu, and Barak A. Cohen. 2022. "Directed Mutational Scanning Reveals a Balance between Acidic and Hydrophobic Residues in Strong Human Activation Domains." *Cell Systems* 13 (4): 334–345.e5.
- Starita, Lea M., Nadav Ahituv, Maitreya J. Dunham, Jacob O. Kitzman, Frederick P. Roth, Georg Seelig, Jay Shendure, and Douglas M. Fowler. 2017. "Variant Interpretation: Functional Assays to the Rescue." *American Journal of Human Genetics* 101 (3): 315–25.
- Starr, Tyler N., Allison J. Greaney, William W. Hannon, Andrea N. Loes, Kevin Hauser, Josh R. Dillen, Elena Ferri, et al. n.d. "Shifting Mutational Constraints in the SARS-CoV-2 Receptor-Binding Domain during Viral Evolution." <https://doi.org/10.1101/2022.02.24.481899>.
- Starr, Tyler N., Allison J. Greaney, Sarah K. Hilton, Daniel Ellis, Katharine H. D. Crawford, Adam S. Dingens, Mary Jane Navarro, et al. 2020. "Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding." *Cell* 182 (5): 1295–1310.e20.
- Stenson, Peter D., Matthew Mort, Edward V. Ball, Katy Evans, Matthew Hayden, Sally Heywood, Michelle Hussain, Andrew D. Phillips, and David N. Cooper. 2017. "The Human Gene Mutation Database: Towards a Comprehensive Repository of Inherited Mutation Data for Medical Research, Genetic Diagnosis and next-Generation Sequencing Studies." *Human Genetics* 136 (6): 665–77.
- Stephens, Amberley D., Maria Zacharopoulou, and Gabriele S. Kaminski Schierle. 2019. "The Cellular Environment Affects Monomeric α -Synuclein Structure." *Trends in Biochemical Sciences* 44 (5): 453–66.
- Sun, Song, Jochen Weile, Marta Verby, Yingzhou Wu, Yang Wang, Atina G. Cote, Iosifina Fotiadou, et al. 2020. "A Proactive Genotype-to-Patient-Phenotype Map for Cystathionine Beta-Synthase." *Genome Medicine* 12 (1): 13.
- Sun, Song, Fan Yang, Guihong Tan, Michael Costanzo, Rose Oughtred, Jodi Hirschman, Chandra L. Theesfeld, et al. 2016. "An Extended Set of Yeast-Based Functional Assays Accurately Identifies Human Disease Mutations." *Genome Research*. <https://doi.org/10.1101/gr.192526.115>.
- Sunde, M., L. C. Serpell, M. Bartlam, P. E. Fraser, M. B. Pepys, and C. C. Blake. 1997. "Common Core Structure of Amyloid Fibrils by Synchrotron X-Ray Diffraction." *Journal of Molecular Biology* 273 (3): 729–39.
- Tabet, Daniel, Victoria Parikh, Prashant Mali, Frederick P. Roth, and Melina Claussnitzer. 2022. "Scalable Functional Assays for the Interpretation of Human Genetic Variation." *Annual Review of Genetics*, September. <https://doi.org/10.1146/annurev-genet-072920-032107>.
- Tareen, Ammar, Mahdi Kooshkbaghi, Anna Posfai, William T. Ireland, David M. McCandlish, and Justin B. Kinney. 2022. "MAVE-NN: Learning Genotype-Phenotype Maps from Multiplex Assays of Variant Effect." *Genome Biology* 23 (1): 98.
- Tartaglia, Gian Gaetano, Sebastian Pechmann, Christopher M. Dobson, and Michele Vendruscolo. 2007. "Life on the Edge: A Link between Gene Expression Levels and Aggregation Rates of Human Proteins." *Trends in Biochemical Sciences* 32 (5): 204–6.
- Tartaglia, Gian Gaetano, and Michele Vendruscolo. 2008. "The Zyggregator Method for Predicting Protein Aggregation Propensities." *Chemical Society Reviews* 37 (7): 1395–1401.
- Taylor, Alexander I. P., and Rosemary A. Staniforth. 2022. "General Principles Underpinning Amyloid Structure." *Frontiers in Neuroscience* 16 (June): 878869.
- Taylor, J. Paul, Robert H. Brown Jr, and Don W. Cleveland. 2016. "Decoding ALS: From Genes to Mechanism." *Nature* 539 (7628): 197–206.
- Thacker, Dev, Kalyani Sanagavarapu, Birgitta Frohm, Georg Meisl, Tuomas P. J. Knowles, and Sara Linse. 2020. "The Role of Fibril Structure and Surface Hydrophobicity in Secondary Nucleation of Amyloid Fibrils." *Proceedings of the National Academy of Sciences of the United States of America* 117 (41): 25272–83.
- Thu, Tran Thi Minh, Nguyen Truong Co, Ly Anh Tu, and Mai Suan Li. 2019. "Aggregation Rate of Amyloid Beta Peptide Is Controlled by Beta-Content in Monomeric State." *The Journal of Chemical Physics* 150 (22): 225101.

- Tiessen, Axel, Paulino Pérez-Rodríguez, and Luis José Delaye-Arredondo. 2012. "Mathematical Modeling and Comparison of Protein Size Distribution in Different Plant, Animal, Fungal and Microbial Species Reveals a Negative Correlation between Protein Size and Protein Number, Thus Providing Insight into the Evolution of Proteomes." *BMC Research Notes* 5 (February): 85.
- Tomiyama, Takami, Tetsu Nagata, Hiroyuki Shimada, Rie Teraoka, Akiko Fukushima, Hyoue Kanemitsu, Hiroshi Takuma, et al. 2008. "A New Amyloid Beta Variant Favoring Oligomerization in Alzheimer's-Type Dementia." *Annals of Neurology* 63 (3): 377–87.
- Tompa, Peter, and Monika Fuxreiter. 2008. "Fuzzy Complexes: Polymorphism and Structural Disorder in Protein-Protein Interactions." *Trends in Biochemical Sciences* 33 (1): 2–8.
- Tonner, Peter D., Abe Pressman, and David Ross. 2022. "Interpretable Modeling of Genotype–Phenotype Landscapes with State-of-the-Art Predictive Power." *Proceedings of the National Academy of Sciences* 119 (26): e2114021119.
- Törnquist, Mattias, Thomas C. T. Michaels, Kalyani Sanagavarapu, Xiaoting Yang, Georg Meisl, Samuel I. A. Cohen, Tuomas P. J. Knowles, and Sara Linse. 2018. "Secondary Nucleation in Amyloid Formation." *Chemical Communications* 54 (63): 8667–84.
- Trovato, Antonio, Fabrizio Chiti, Amos Maritan, and Flavio Seno. 2006. "Insight into the Structure of Amyloid Fibrils from the Analysis of Globular Proteins." *PLoS Computational Biology* 2 (12): e170.
- Ulamiec, Sabine M., David J. Brockwell, and Sheena E. Radford. 2020. "Looking Beyond the Core: The Role of Flanking Regions in the Aggregation of Amyloidogenic Peptides and Proteins." *Frontiers in Neuroscience*. <https://doi.org/10.3389/fnins.2020.611285>.
- Uptain, S. M., G. J. Sawicki, B. Caughey, and S. Lindquist. 2001. "Strains of [PSI(+)] Are Distinguished by Their Efficiencies of Prion-Mediated Conformational Conversion." *The EMBO Journal* 20 (22): 6236–45.
- Uversky, Vladimir N. 2015. "Intrinsically Disordered Proteins and Their (Disordered) Proteomes in Neurodegenerative Disorders." *Frontiers in Aging Neuroscience* 7 (March): 18.
- Van Cauwenbergh, Caroline, Christine Van Broeckhoven, and Kristel Slegers. 2016. "The Genetic Landscape of Alzheimer Disease: Clinical Implications and Perspectives." *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 18 (5): 421–30.
- Vetter, I. R., W. A. Baase, D. W. Heinz, J. P. Xiong, S. Snow, and B. W. Matthews. 1996. "Protein Structural Plasticity Exemplified by Insertion and Deletion Mutants in T4 Lysozyme." *Protein Science: A Publication of the Protein Society* 5 (12): 2399–2415.
- Wälti, Marielle Aulikki, Francesco Ravotti, Hiromi Arai, Charles G. Glabe, Joseph S. Wall, Anja Böckmann, Peter Güntert, Beat H. Meier, and Roland Riek. 2016. "Atomic-Resolution Structure of a Disease-Relevant A β (1-42) Amyloid Fibril." *Proceedings of the National Academy of Sciences of the United States of America* 113 (34): E4976-84.
- Wasmer, Christian, Adam Lange, Hélène Van Melckebeke, Ansgar B. Siemer, Roland Riek, and Beat H. Meier. 2008. "Amyloid Fibrils of the HET-s(218-289) Prion Form a Beta Solenoid with a Triangular Hydrophobic Core." *Science* 319 (5869): 1523–26.
- Weile, Jochen, and Frederick P. Roth. 2018. "Multiplexed Assays of Variant Effects Contribute to a Growing Genotype-Phenotype Atlas." *Human Genetics* 137 (9): 665–78.
- White, Helen E., Julie L. Hodgkinson, Thomas R. Jahn, Sara Cohen-Krausz, Walraj S. Gosal, Shirley Müller, Elena V. Orlova, Sheena E. Radford, and Helen R. Saibil. 2009. "Globular Tetramers of B2-Microglobulin Assemble into Elaborate Amyloid Fibrils." *Journal of Molecular Biology* 389 (1): 48–57.
- Wilson, D. S., and A. D. Keefe. 2001. "Random Mutagenesis by PCR." *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* Chapter 8 (May): Unit8.3.
- Wong, Winston. 2020. "Economic Burden of Alzheimer Disease and Managed Care Considerations." *The American Journal of Managed Care* 26 (8 Suppl): S177–83.
- Wrenbeck, Emily E., Justin R. Klesmith, James A. Stapleton, Adebola Adeniran, Keith E. J. Tyo, and Timothy A. Whitehead. 2016. "Plasmid-Based One-Pot Saturation Mutagenesis." *Nature Methods*. <https://doi.org/10.1038/nmeth.4029>.
- Wright, Peter E., and H. Jane Dyson. 2015. "Intrinsically Disordered Proteins in Cellular Signalling and Regulation." *Nature Reviews. Molecular Cell Biology* 16 (1): 18–29.
- Xiao, Yiling, Buyong Ma, Dan McElheny, Sudhakar Parthasarathy, Fei Long, Minako Hoshi, Ruth Nussinov, and Yoshitaka Ishii. 2015. "A β (1-42) Fibril Structure Illuminates Self-Recognition and Replication of Amyloid in Alzheimer's Disease." *Nature Structural & Molecular Biology* 22 (6): 499–505.

- Xue, Wei-Feng, Steve W. Homans, and Sheena E. Radford. 2009. "Amyloid Fibril Length Distribution Quantified by Atomic Force Microscopy Single-Particle Image Analysis." *Protein Engineering, Design & Selection: PEDS* 22 (8): 489–96.
- Yang, Jie, Alexander J. Dear, Thomas C. T. Michaels, Christopher M. Dobson, Tuomas P. J. Knowles, Si Wu, and Sarah Perrett. 2018. "Direct Observation of Oligomerization by Single Molecule Fluorescence Reveals a Multistep Aggregation Mechanism for the Yeast Prion Protein Ure2." *Journal of the American Chemical Society* 140 (7): 2493–2503.
- Yang, Xiaoting, Georg Meisl, Birgitta Frohm, Eva Thulin, Tuomas P. J. Knowles, and Sara Linse. 2018. "On the Role of Sidechain Size and Charge in the Aggregation of A β 42 with Familial Mutations." *Proceedings of the National Academy of Sciences of the United States of America* 115 (26): E5849–58.
- Yang, Yang, Diana Arseni, Wenjuan Zhang, Melissa Huang, Sofia Lövestam, Manuel Schweighauser, Abhay Kotecha, et al. 2022. "Cryo-EM Structures of Amyloid- β 42 Filaments from Human Brains." *Science* 375 (6577): 167–72.
- Yoo, Haneul, Catherine Triandafillou, and D. Allan Drummond. 2019. "Cellular Sensing by Phase Separation: Using the Process, Not Just the Products." *The Journal of Biological Chemistry* 294 (18): 7151–59.
- Zbinden, Aurélie, Manuela Pérez-Berlanga, Pierre De Rossi, and Magdalini Polymenidou. 2020. "Phase Separation and Neurodegenerative Diseases: A Disturbance in the Force." *Developmental Cell* 55 (1): 45–68.
- Zhou, Juannan, and David M. McCandlish. 2020. "Minimum Epistasis Interpolation for Sequence-Function Relationships." *Nature Communications* 11 (1): 1782.
- Zhou, Lujia, Nathalie Brouwers, Iryna Benilova, Annelies Vandersteen, Marc Mercken, Koen Van Laere, Philip Van Damme, et al. 2011. "Amyloid Precursor Protein Mutation E682K at the Alternative β -Secretase Cleavage β' -Site Increases A β Generation." *EMBO Molecular Medicine* 3 (5): 291–302.

Annex I

Table with the impact on aggregation rates for A β 42 variants from the literature. The table reports on familial AD class (D: dominant, R: recessive), nucleation scores (NS), NS class at FDR10 (NS+: significantly increasing NS, NS-: significantly decreasing NS or WT-like NS) reported in our study, and the qualitative agreement with the published data from the literature (X. Yang et al. 2018; Bolognesi et al. 2014; Thacker et al. 2020; Hori et al. 2007; Meisl et al. 2016; W.-T. Chen et al. 2012; L. Zhou et al. 2011; Benilova et al. 2014; Morimoto et al. 2004; Murray et al. 2016; Thu et al. 2019).

A β variant	fAD class	NS	category FDR10	Yang et al., 2018 (https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1526152918/-/suppl.pdf)	Thacker et al., 2020 (https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1717411117/-/suppl.pdf)	Bolognesi et al., 2014 (https://pubs.acs.org/suppl/doi/10.1021/cb400616y)	Hori et al., 2006 (https://www.jbc.org/content/282/7/16161.long)	Meist et al., 2016 (https://www.nature.com/articles/srep18728)	Chen et al., 2012 (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0035807)	Zhou et al., 2011 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3377078/)	Benilova et al., 2014 (https://pubmed.ncbi.nlm.nih.gov/25253695/)	Morimoto et al., 2004 (https://www.jbc.org/content/279/5/052781.full)	Murray et al., 2016 (https://onlinelibrary.wiley.com/doi/10.1002/prot.24995)	Thu et al., 2019 (https://ai-p.sciatio.n.org/doi/full/10.1063/1.5096379)	Agreement	
A2T		-0.086	WT-like								Similar agg rate		Decreased/similar agg rate		V	
A2V	R	-1.189	NS-					Decreased agg rate			Similar agg rate		Decreased agg rate		V	
H6R	D	0.203	NS+				Increased agg rate								V	
D7H	D	0.485	NS+						Decreased agg rate						X	
D7N	D	0.312	NS+				Increased agg rate								V	
E11K	D	0.206	NS+							Increased agg rate					V	
A21G	D	0.223	NS+	Similar agg rate										-0.671	X	
A21S		-0.601	NS-	Decreased agg rate											V	
E22G	D	2.464	NS+		Increased agg rate									0.209	V	
E22K	D	0.619	NS+											0.545	V	
E22Q	D	1.810	NS+												V	
D23N	D	1.112	NS+											0.238	V	
V40S		-2.870	NS-		Decreased agg rate										V	
I41K		-2.973	NS-											-0.518	V	
A42S		-0.106	WT-like												V	
A42T	D	0.592	NS+												Increased agg rate	V

Annex II

Understanding and evolving prions by yeast
multiplexed assays



Understanding and evolving prions by yeast multiplexed assays

Mireia Seuma* and Benedetta Bolognesi*

Yeast genetics made it possible to derive the first fundamental insights into prion composition, conformation, and propagation. Fast-forward 30 years and the same model organism is now proving an extremely powerful tool to comprehensively explore the impact of mutations in prion sequences on their function, toxicity, and physical properties. Here, we provide an overview of novel multiplexed strategies where deep mutagenesis is combined to a range of tailored selection assays in yeast, which are particularly amenable for investigating prions and prion-like sequences. By mimicking evolution in a flask, these multiplexed approaches are revealing mechanistic insights on the consequences of prion self-assembly, while also reporting on the structure prion sequences adopt *in vivo*.

Address

Institute for Bioengineering of Catalonia (IBEC), The Barcelona Institute of Science and Technology, Baldiri Reixac 10-12, 08028 Barcelona, Spain

Corresponding author:

Benedetta Bolognesi (bbolognesi@ibecbarcelona.eu)

* Twitter account: [@mseumaar](#), [@Bennibolo](#)

Current Opinion in Genetics & Development 2022, 75:101941

This review comes from a themed issue on **Evolutionary Genetics**

Edited by **Christian Landry** and **Gianni Liti**

For complete overview of the section, please refer to the article collection, "[Evolutionary Genetics](#)"

Available online 28th June 2022

<https://doi.org/10.1016/j.gde.2022.101941>

0959-437X/© 2022 Elsevier Ltd. All rights reserved.

From classic prions to a wide set of prion-like proteins

Prions are infectious proteins that can self-template and self-propagate, acting as protein-based epigenetic elements [1]. Prion-forming sequences have challenged for decades our understanding of sequence–structure–function relationships. Not only have these proteins proved that the same primary sequence can adopt several stable folds, but also that their ability to self-assemble can provide phenotypic advantage, at least under certain circumstances. If in the last five years we all warmed up to the idea of functional self-assembly, thanks to extensive efforts aiming at deciphering the

role of condensation inside the cell [2], the hypothesis that the aggregation of specific proteins could be a common means to heritable phenotypic variability was far from trivial to formulate 30 years ago when the first yeast prions [*URE3*] and [*PSI+*] were characterized [3]. It is only thanks to pioneering work in *S. cerevisiae* that we now have a better idea of how prions arise and of the range of phenotypic outcomes they can result in. [4]. While paving the way for our current understanding of these sequences, the clever genetic manipulation of yeast in these early studies led to landmark mechanistic and structural insights, such as inferring the parallel in-register β -sheet arrangement of [*URE3*] and [*PSI+*] [5,6] (Figure 3a).

These studies also showed that prion phenotypes can be induced by increasing the frequency of conformational change by protein overproduction [7] and that the resulting phenotype is often similar to that of deleting or mutating the causal protein. Over time, other properties have been attributed to classic prions: they form amyloids, they are rich in Q/N with overall low sequence complexity, depend on Hsp104 for propagation, and are inheritable [8–10].

There is now growing evidence that a wider set of intrinsically disordered proteins in yeast can give rise to specific phenotypes upon self-assembly and that these can be inherited over several generations [11,12]. Although these sequences do not share all properties of classic prions (e.g. they do not necessarily form amyloids) [11], they are commonly also classified as prions or prion-like proteins. Even a subset of the human proteome is nowadays considered to be prion-like. These sequences are found in 1% of protein-coding genes [13], they are intrinsically disordered, and their low-complexity composition resembles that of classic yeast prions: rich in Q, N, but also S and Y. These sequences encode the information required for self-templated aggregation. Surprisingly, yeast proved to be an excellent model to explore many of these sequences, as they can be easily swapped for the prion domain of yeast proteins and assessed for their ability to support specific inheritable phenotypes [14,15].

Classic alignment approaches are not useful when looking at prions, as conservation is mostly evident at the level of composition rather than down to their exact primary sequence [16]. Composition is conserved, for

example, across 21 fungi that diverged over one billion years ago and homologs of prion proteins in different species have retained the ability to aggregate [17]. Moreover, the self-assembly properties of some more recently discovered prions show patterns of conservation all the way to humans [18•,19]. Nonetheless, many single amino acid changes in prion sequences are enough to prevent or enhance prion formation [20]. These polymorphisms are a key element in preventing prion propagation between species, a phenomenon commonly known as species barrier [21]. One single change in the amino acid sequence is sometimes sufficient to affect the ability of a ‘mutated’ monomer to template the aggregation of a wild-type molecule [22].

Fitness trade-offs

In yeast, the ability of prion sequences to switch to an aggregated state provides fitness advantages in a wide range of scenarios (bet hedging), with adaptation arising from the possibility of quickly acquiring complex traits that would be less likely to appear upon sequential selection of mutations contributing to them [23]. A typical example is the aggregation and sequestration of the yeast translation-termination factor Sup35p into aggregates, leading to the read-through of premature stop codons and the onset of the *[PSI+]* phenotype [4]. On the other hand, several recently discovered prions result instead in fitness advantage [24] by potentiating the action of their causal protein [18•]. The discovery of prions in at least a third of the wild yeast strains tested supports their role in the adaptation of yeast through a changing environment and it was even suggested that stress may promote prion switching [21,23,25–27].

Prion and prion-like domains (PRDs and PRLDs) were also shown to drive adaptive reversible protein condensation through liquid-liquid phase separation. Proteins containing these sequences can sense and rapidly respond to cellular stress — pH, starvation, and temperature — by condensing and temporarily sequestering or releasing proteins and transcripts [28••–30]. Condensed Ded1p promotes translation of stress mRNA upon heat-induced stress, an adaptive mechanism that has been fine-tuned to the growth temperature of each species [28••]. Another example consists of the protein Whi3 whose aggregation controls cell cycle in multinucleate cells, via sequestration of cyclin mRNAs, inducing phenotypic heterogeneity even in the absence of stress [31,32].

While providing interesting examples of the adaptive role of prions, these observations do not exclude that the states these sequences adopt can be toxic under certain circumstances. Indeed, the loss of function caused by aggregation can be detrimental and even lethal. Even when loss of function is not detrimental, such as for

Ure2p, the prion state can still drastically affect fitness, suggesting that the prion itself or intermediate assemblies in its formation are toxic for the cells [33]. On this line, a few versions of *[PSI+]* and *[URE3]* obtained in the lab have even been named ‘suicidal’ due to their high toxicity [34]. Although prions are found in wild yeast strains, their frequency is lower than certain viruses that are notoriously detrimental [35], suggesting they are actually overall harmful for the cells. This assumption has however been challenged by i) modeling the frequency of other reversible epigenetic elements [36,37] and ii) the finding that specific prions such as *[RNQ+]* and *[Het-S]* are instead detected in a vast majority of natural isolates [23,38].

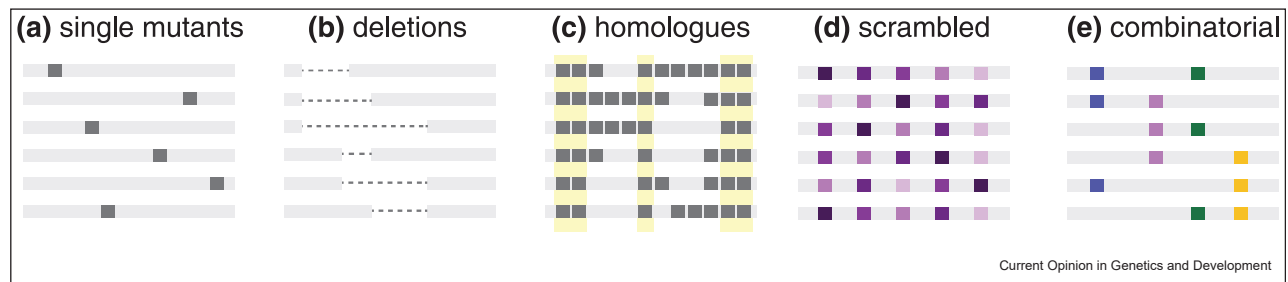
In the human proteome, PRLDs are particularly enriched in nucleic acid binding proteins [39,40], suggesting a role for these sequences in promoting functional condensation and temporal sequestration of RNAs also in more complex organisms. Many PRLDs exist in essential genes and aggregation of prion-like proteins plays a role in signal transduction from fungi to humans. This is the case of *Het-S* polymerization at the basis of heterokaryon incompatibility [41] and of the antiviral signaling cascade activated by MAVS prion switching [42]. However, PRDs and PRLDs can also drive pathological aggregation in devastating diseases, such as Fatal Familial Insomnia and Amyotrophic Lateral Sclerosis, and mutations in these protein regions cause dominant forms of disease [43–45]. The phenotypic consequences of these mutations are diverse and how different cell and tissue specificities depend on the identity and function of the aggregating sequence has not been fully elucidated yet. Altogether, the role and impact of these sequences in the human proteome is just another reminder of how the interplay between beneficial and detrimental effects of self-assembly and their sequence determinants is not trivial to decipher.

The multiplexed era

Altogether, prion conformation, function, toxicity, and environment all concur to the final phenotypic outcome of each cell. This interplay is particularly challenging to decipher, calling for well-defined assays to report — if possible — on just one of these biological mechanisms at once.

We suggest that high-throughput approaches such as multiplexed assays of variant effects (MAVEs) can be a useful tool to decouple the different layers of these complex systems. The basic principle behind MAVEs, also known as deep mutational scanning (DMS), is the construction of a library of thousands of variants that can be selected for a specific phenotype in a cell-based assay [46]. The performance of each variant is quantified by sequencing the population before and after selection [47]. Using mutations to disrupt or enhance a

Figure 1



Mutational library design. Examples of library design include (a) all possible single mutants in a given sequence to systematically explore the consequences of polymorphism [57•], (b) a range of deleted variants to identify key regions driving protein aggregation [58], (c) sequences mapping entire evolutionary trajectories [28••], (d) scrambled versions of the same sequence to gather insights about structural arrangements [5], and (e) different combinations of double mutants to infer specific residue–residue contacts in prion proteins [50].

process in order to understand it is the main power of MAVEs where variant libraries can be rationally designed to address a range of questions, from prion polymorphism to explore evolutionary paths or prion sequence space in a hypothesis-free manner [48–50] (Figure 1). Libraries can be selected in parallel assays that report on different phenotypes in variable experimental conditions mimicking fluctuating environments. Overall, the versatility of MAVEs makes them very suitable for studying the different layers of prion biology and the scale required for these approaches makes yeast an excellent system to employ: simple manipulation, large population, and fast generation time.

Engineering selection

Next, we report on a series of selection assays that have already been employed or have great potential to be transferred to MAVEs to quantify prion properties at scale (Figure 2).

Toxicity

Prion toxicity can be assessed in a MAVE simply on the basis of cell viability (Figure 2a): over generations, cells carrying a toxic variant of the prion sequence will be depleted in the population, while those carrying a non-toxic variant will be enriched. This is quantified by deep sequencing before and after expression of the protein, an approach that was employed to quantitatively map toxicity for thousands of variants of the PRLD of the human protein TDP-43 [50].

Gain and loss of function

Cell viability can also be used to select for gain or loss of function (Figure 2b). For example, switching to [SMAUG+] or [GAR+] represents an adaptive advantage upon glucose depletion: [SMAUG+] hyperactivates the function of its causal protein Vts1 [18•], and [GAR+] supports yeast growth on mixed carbon sources [27]. In contrast, [SWI+] causes a loss of function of its causal protein and slows growth in nonfermentable carbon

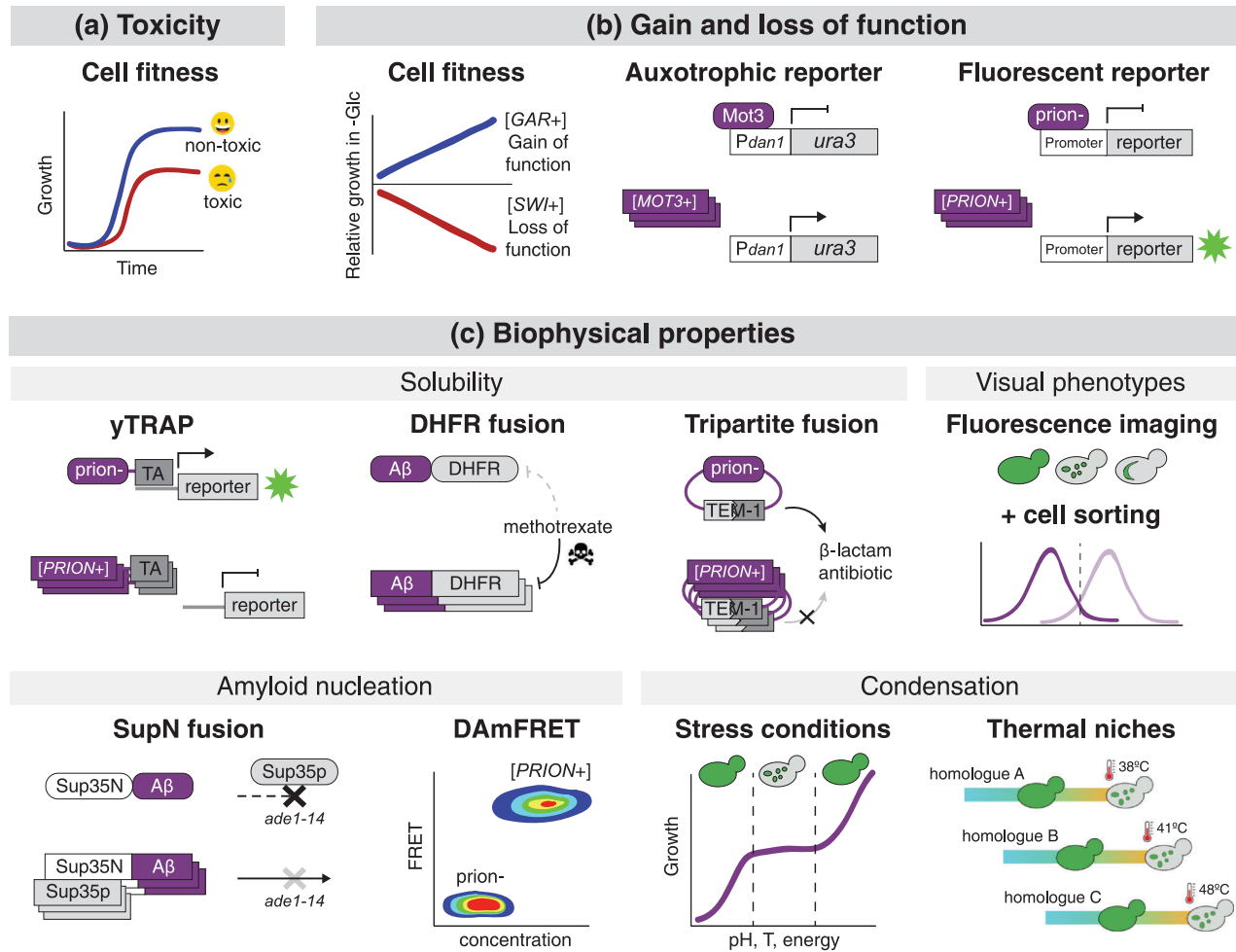
sources [51]. Thousands of variants of these proteins can be scored for their impact on prion formation by sequencing before and after selection in a medium lacking glucose. Mutational libraries can also be coupled to downstream auxotrophic reporters — such as *ura3* — or fluorescence reporters [52].

Biophysical state

Other genetically engineered yeast systems are suitable to select for specific protein physical states (Figure 2c). The yTRAP system couples the aggregation state of a protein of interest to the activity of a transcription activator acting on a fluorescence reporter, allowing variant discrimination using fluorescence-activated cell sorting [53]. A drug-resistance selection assay has been used in a MAVE to test aggregation of the amyloid beta (A β) peptide fused to dihydrofolate reductase (DHFR) [54]. Only when the DHFR is fused to a soluble A β variant, but not to an aggregating one, cells will be able to grow in the presence of methotrexate. Similarly, fusion to TEM-1 β -lactamase has been used to identify aggregation-prone sequences both in yeast and in the periplasmic space of *E. coli* [55]. The correct folding of the protein fusions brings the two halves of the enzyme in proximity, restoring function and allowing selection against β -lactam antibiotics. These assays share one common limitation: poorly expressed or degraded sequences would lead to the same phenotypic readout as those that aggregate.

There are two systems that instead track nucleation of protein assemblies, that is, the very first step in the formation of self-templating aggregates. One consists in fusing a sequence to the nucleation domain of Sup35p [56,57•]. Nucleation of Sup35p and induction of the [PSI+] phenotype is a readout of the ability of the fused sequence to nucleate amyloids. This approach was used to map > 17 000 A β variants [57•,58]. The other system, DAMFRET, is also particularly suited to run MAVEs. In this case, nucleation barriers and prion switching are observed by means of amphifluoric FRET and the

Figure 2



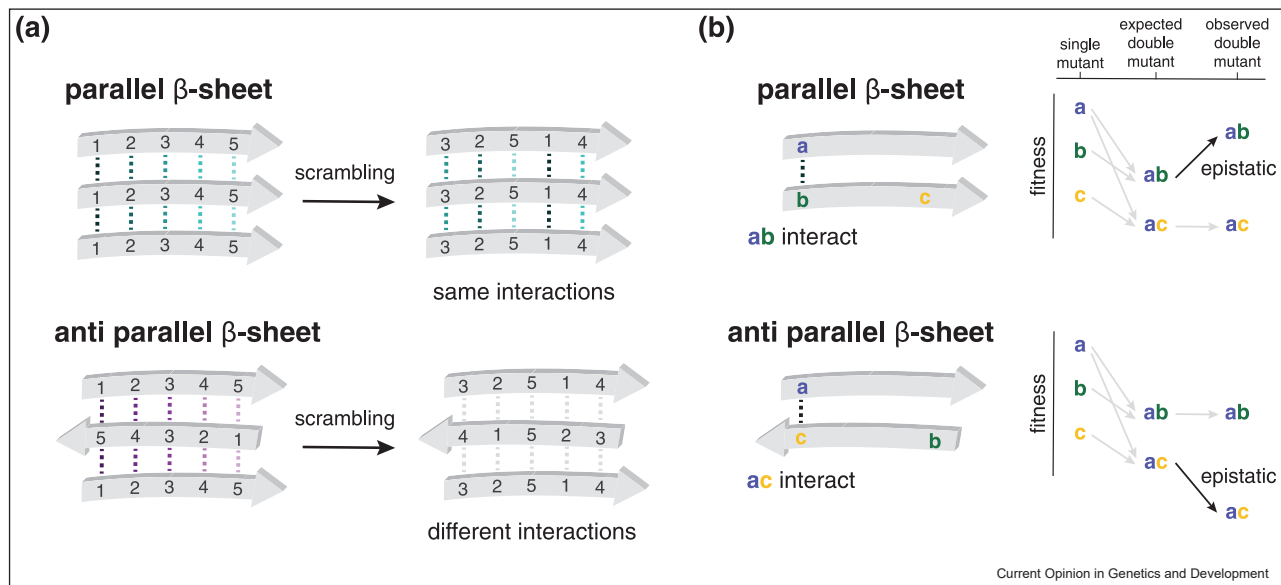
Experimental selection assays to quantify prion properties at scale. **(a)** Tracking cell growth over time reports on toxic and nontoxic prion variants [50]. **(b)** Gain and loss of function can be assessed with cell fitness, or an auxotrophic or a fluorescent reporter. Replacing *dan1* — transcriptionally repressed by Mot3p in normal conditions — with *ura3* and sequencing of variants growing in the absence of uracil can report on the loss of function induced by Mot3p prion switching to [MOT3+] [12]. Similarly, replacing *flo* family genes with *ura3* has been used to report on switching to [SWI+] and identify anti-prion chemical compounds in a high-throughput screening [52]. **(c)** Solubility can be quantified by the yTRAP system, where the prion is fused to a synthetic transcription-activation domain that recognizes a binding site upstream of a fluorescent reporter gene [53] by means of a DHFR fusion, where soluble protein variants allow the enzyme to remain soluble and functional in the presence of its competitive inhibitor methotrexate, and so to reduce DHF to THF [54,65], thanks to a tripartite fusion, where two domains of TEM-1 β -lactamase are fused to a prion protein. The enzyme can only be reconstituted and hence functional if the prion remains soluble, providing antibiotic resistance in bacteria or yeast cells [55]. Amyloid nucleation can be tracked with a supN fusion, in a yeast strain with a premature stop codon in the adenine gene. Endogenous full-length Sup35p, a translation-termination factor, recognizes the stop codon when soluble. SupN nucleation, induced by nucleation of the protein of interest, recruits Sup35p, causing a read-through of the stop codon and allowing growth in a medium lacking adenine [56,57]. The DAMFRET system tracks prion nucleation by fusing the protein of interest to a photoconvertible fluorescent protein. Emission of FRET signal quantifies concentration-dependent protein self-assembly in thousands of single cells in one single experiment [59,66]. Protein condensation can be assessed by cell growth in changing environments. For example, changes in pH, temperature, or nutrient availability can induce protein-phase separation, which ensures cellular fitness during recovery, a mechanism that has been shown to be adaptive and fine-tuned to a specific range of growth temperatures in different species [11,28]. Finally, prion phenotypes can also be screened and selected with a fluorescent tag and by means of imaging coupled to cell sorting [61].

frequency of nucleation is measured as a function of protein concentration in yeast cells [59].

The ability of prions to form condensates is also selectable, at least for those sequences that were shown to

promote cell viability in stress conditions [29,30]. These selection experiments can also be performed at different temperatures to report on the condensation of protein homologs from different species, which have adapted to their thermal niche [28].

Figure 3



Protein structure from yeast genetics. **(a)** In 2005, swapping of *sup35* with five scrambled versions of its sequence resulted in a $[PSI^+]$ phenotype, reporting on the specific arrangement of interactions required for prion switching to $[PSI^+]$ [5]. **(b)** In 2019, the interactions between thousands of double mutants in MAVEs were used to predict protein structure on the basis of the principle for which residues in close structural proximity are more epistatic [62,63], a method that proved powerful also to infer the *in vivo* structural signatures of a human PRLD [50].

Finally, visualizing prions with traditional microscopy showed that they can adopt multiple shapes and have different subcellular localizations [12,50]. Recently, fluorescence imaging has been coupled to cell-sorting, enabling the selection of variants of a library by a multiple set of morphological and spatial traits [60,61••].

Inferring *in vivo* protein conformation

Beyond illuminating genotype-to-phenotype relationships, MAVEs of combinatorial libraries can be used to infer structural elements since the genetic interactions between mutations in structurally proximal residues are likely to have nonindependent effects (i.e. be epistatic, Figure 3b) [62,63]. This approach provides a great opportunity to explore *in vivo* conformations of disordered proteins and is particularly appealing to study prions, which, due to their aggregation propensity, are otherwise very difficult to approach by traditional biophysical techniques. In this line, Wickner's vision was absolutely right and ahead of his time: yeast genetics can be extremely informative on protein structure (Figure 3).

Disclaimer: not just DNA

There is one element of prion biology that cannot be mimicked well by carefully tailored multiplexed assays. The very same DNA sequence often gives rise to different prion strains that differ in their amyloid structure, stability, and propagation [56,64]. Although this one-to-many relationship between genotype and phenotype cannot be captured by assays relying on DNA variation,

we believe that the power of MAVEs to massively assess phenotypes by scanning thousands of genotypes will provide the mechanistic insights required to guide also our understanding of those prion phenotypes not written in the coding sequence.

Conflict of interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

M.S. is supported by a fellowship from Agencia de Gestio d'Ajuts Universitaris i de Recerca (2019FI_B 01311). Work in the lab of B.B. is supported by the la Caixa Research Foundation project "DeepAmyloids" (LCF/PR/HR21/52410004), the Spanish Ministry of Science, Innovation and Universities (RTI2018-101491-A-I00 (MICIU/FEDER)), and the CERCA Program/Generalitat de Catalunya and the Centro de Excelencia Severo Ochoa.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest

1. Chien P, Weissman JS, DePace AH: **Emerging principles of conformation-based prion inheritance.** *Annu Rev Biochem* 2004, **73**:617-656.
2. Boeynaems S, Alberti S, Fawzi NL, Mittag T, Polymenidou M, Rousseau F, et al.: **Protein phase separation: a new phase in cell biology.** *Trends Cell Biol* 2018, **28**:420-435.

3. Wickner RB: **[URE3] as an altered URE2 protein: evidence for a prion analog in *Saccharomyces cerevisiae***. *Science* (5158) 1994, **264**:566-569, <https://doi.org/10.1126/science.7909170>
4. Sparrer HE, Santoso A, Szoka FC, Weissman JS: **Evidence for the prion hypothesis: induction of the yeast [PSI⁺] factor by in vitro-converted Sup35 protein**. *Science* 2000, **289**:595-599.
5. Ross ED, Baxa U, Wickner RB: **Scrambled prion domains form prions and amyloid**. *Mol Cell Biol* 2004, **24**:7206-7213.
6. Ross ED, Minton A, Wickner RB: **Prion domains: sequences, structures and interactions**. *Nat Cell Biol* 2005, **7**:1039-1044.
7. Chernoff YO, Derkach IL, Inge-Vechtomov SG: **Multicopy SUP35 gene induces de-novo appearance of psi-like factors in the yeast *Saccharomyces cerevisiae***. *Curr Genet* 1993, **24**:268-270.
8. DePace AH, Santoso A, Hillner P, Weissman JS: **A critical role for amino-terminal glutamine/asparagine repeats in the formation and propagation of a yeast prion**. *Cell* 1998, **93**:1241-1252.
9. Chernoff YO, Lindquist SL, Ono B, Inge-Vechtomov SG, Liebman SW: **Role of the chaperone protein Hsp104 in propagation of the yeast prion-like factor [psi⁺]**. *Science* 1995, **268**:880-884.
10. Kawai-Noma S, Pack C-G, Kojidani T, Asakawa H, Hiraoka Y, Kinjo M, et al.: **In vivo evidence for the fibrillar structures of Sup35 prions in yeast cells**. *J Cell Biol* 2010, **190**:223-231.
11. Chakrabortee S, Byers JS, Jones S, Garcia DM, Bhullar B, Chang A, et al.: **Intrinsically disordered proteins drive emergence and inheritance of biological traits**. *Cell* 2016, **167**:369-381.e12.
12. Alberti S, Halfmann R, King O, Kapila A, Lindquist S: **A systematic survey identifies prions and illuminates sequence features of prionogenic proteins**. *Cell* 2009, **137**:146-158.
13. King OD, Gitler AD, Shorter J: **The tip of the iceberg: RNA-binding proteins with prion-like domains in neurodegenerative disease**. *Brain Res* 2012, **1462**:61-80.
14. Johnson BS, Snead D, Lee JJ, McCaffery JM, Shorter J, Gitler AD: **TDP-43 is intrinsically aggregation-prone, and amyotrophic lateral sclerosis-linked mutations accelerate aggregation and increase toxicity**. *J Biol Chem* 2009, **284**:20329-20339.
15. Ju S, Tardiff DF, Han H, Divya K, Zhong Q, Maquat LE, et al.: **A yeast model of FUS/TLS-dependent cytotoxicity**. *PLoS Biol* 2011, **9**:e1001052.
16. An L, Fitzpatrick D, Harrison PM: **Emergence and evolution of yeast prion and prion-like proteins**. *BMC Evol Biol* 2016, **16**:24.
17. Su T-Y, Harrison PM: **Conservation of prion-Like composition and sequence in prion-formers and prion-like proteins of *Saccharomyces cerevisiae***. *Front Mol Biosci* 2019, **6**:54.
18. Chakravarty AK, Smejkal T, Itakura AK, Garcia DM, Jarosz DF: **A non-amyloid prion particle that activates a heritable gene expression program**. *Mol Cell* 2020, **77**:251-265.e9.
- Example of a nonclassical prion which potentiates protein function upon self-assembly.
19. Harvey ZH, Chakravarty AK, Futia RA, Jarosz DF: **A prion epigenetic switch establishes an active chromatin state**. *Cell* 2020, **180**:928-940.e14.
20. Paul KR, Hendrich CG, Waechter A, Harman MR, Ross ED: **Generating new prions by targeted mutation or segment duplication**. *Proc Natl Acad Sci USA* 2015, **112**:8584-8589.
21. Tanaka M, Collins SR, Toyama BH, Weissman JS: **The physical basis of how prion conformations determine strain phenotypes**. *Nature* 2006, **442**:585-589.
22. Shida T, Kamatari YO, Yoda T, Yamaguchi Y, Feig M, Ohhashi Y, et al.: **Short disordered protein segment regulates cross-species transmission of a yeast prion**. *Nat Chem Biol* 2020, **16**:756-765.
23. Halfmann R, Jarosz DF, Jones SK, Chang A, Lancaster AK, Lindquist S: **Prions are a common mechanism for phenotypic inheritance in wild yeasts**. *Nature* 2012, **482**:363-368.
24. Garcia DM, Campbell EA, Jakobson CM, Tsuchiya M, Shaw EA, DiNardo AL, et al., Kaeberlein M, Jarosz DF: **A prion accelerates proliferation at the expense of lifespan**. *Elife* 2021, **10**:e60917, <https://doi.org/10.7554/eLife.60917> PMID: 34545808; PMCID: PMC8455135.
25. Tyedmers J, Madariaga ML, Lindquist S: **Prion switching in response to environmental stress**. *PLoS Biol* 2008, **6**:e294.
26. Holmes DL, Lancaster AK, Lindquist S, Halfmann R: **Heritable remodeling of yeast multicellularity by an environmentally responsive prion**. *Cell* 2013, **153**:153-165.
27. Jarosz DF, Brown JCS, Walker GA, Datta MS, Ung WL, Lancaster AK, et al.: **Cross-kingdom chemical communication drives a heritable, mutually beneficial prion-based transformation of metabolism**. *Cell* 2014, **158**:1083-1093.
28. Iserman C, Desroches Altamirano C, Jegers C, Friedrich U, Zarin T, Fritsch AW, et al.: **Condensation of Ded1p promotes a translational switch from housekeeping to stress protein production**. *Cell* 2020, **181**:818-831.e19.
- Sequences that evolved to condense at different temperatures.
29. Franzmann TM, Jahnel M, Pozniakovskiy A, Mahamid J, Holehouse AS, Nüske E, Richter D, Baumeister W, Grill W, Pappu RV, Hyman AA, Alberti S: **Phase separation of a yeast prion protein promotes cellular fitness**. *Science* (6371) 2018, **359**:eaa05654, <https://doi.org/10.1126/science.aao5654> PMID: 29301985.
30. Riback JA, Katanski CD, Kear-Scott JL, Pilipenko EV, Rojek AE, Sosnick TR, et al.: **Stress-triggered phase separation is an adaptive, evolutionarily tuned response**. *Cell* 2017, **168**:1028-1040.e19.
31. Zhang H, Elbaum-Garfinkle S, Langdon EM, Taylor N, Occhipinti P, Bridges AA, et al.: **RNA controls PolyQ protein phase transitions**. *Mol Cell* 2015, **60**:220-230.
32. Caudron F, Barral Y: **A super-assembly of Whi3 encodes memory of deceptive encounters by single cells during yeast courtship**. *Cell* 2013, **155**:1244-1257.
33. Nakayashiki T, Kurtzman CP, Edskes HK, Wickner RB: **Yeast prions [URE3] and [PSI⁺] are diseases**. *Proc Natl Acad Sci USA* 2005, **102**:10575-10580.
34. McGlinchey RP, Kryndushkin D, Wickner RB: **Suicidal [PSI⁺] is a lethal yeast prion**. *Proc Natl Acad Sci USA* 2011, **108**:5337-5341.
35. Wickner RB: **Prions and RNA viruses of *Saccharomyces cerevisiae***. *Annu Rev Genet* 1996, **30**:109-139.
36. Lancaster AK, Masel J: **The evolution of reversible switches in the presence of irreversible mimics**. *Evolution* 2009, **63**:2350-2362.
37. Lancaster AK, Bardill JP, True HL, Masel J: **The spontaneous appearance rate of the yeast prion [PSI⁺] and its implications for the evolution of the evolvability properties of the [PSI⁺] system**. *Genetics* 2010, **184**:393-400.
38. Dalstra HJP, Swart K, Debets AJM, Saupe SJ, Hoekstra RF: **Sexual transmission of the [Het-S] prion leads to meiotic drive in *Podospira anserina***. *Proc Natl Acad Sci USA* 2003, **100**:6616-6621.
39. Kraemer BC, Schuck T, Wheeler JM, Robinson LC, Trojanowski JQ, Lee VMY, et al.: **Loss of murine TDP-43 disrupts motor function and plays an essential role in embryogenesis**. *Acta Neuropathol* 2010, **119**:409-419.
40. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, et al.: **Identification and characterization of essential genes in the human genome**. *Science* 2015, **350**:1096-1101.
41. Saupe SJ, Daskalov A: **The [Het-s] prion, an amyloid fold as a cell death activation trigger**. *PLoS Pathog* 2012, **8**:e1002687.
42. Hou F, Sun L, Zheng H, Skaug B, Jiang Q-X, Chen ZJ: **MAVS forms functional prion-like aggregates to activate and propagate antiviral innate immune response**. *Cell* 2011, **146**:448-461.
43. Kim HJ, Kim NC, Wang Y-D, Scarborough EA, Moore J, Diaz Z, et al.: **Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS**. *Nature* 2013, **495**:467-473.

44. Patel A, Lee HO, Jawerth L, Maharana S, Jahnel M, Hein MY, *et al.*: **A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation.** *Cell* 2015, **162**:1066-1077.
45. Montagna P, Gambetti P, Cortelli P, Lugaresi E: **Familial and sporadic fatal insomnia.** *Lancet Neurol* 2003, **2**:167-176.
46. Fowler DM, Fields S: **Deep mutational scanning: a new style of protein science.** *Nat Methods* 2014, **11**:801.
47. Starita LM, Fields S: **Deep mutational scanning: a highly parallel method to measure the effects of mutation on protein function.** *Cold Spring Harb Protoc* 2015, **2015**:711-714.
48. Sanborn AL, Yeh BT, Feigerle JT, Hao CV, Townshend RJL, Lieberman Aiden E, *et al.*: **Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to Mediator.** *Elife* 2021, **10**:e68068.
49. Domingo J, Diss G, Lehner B: **Pairwise and higher-order genetic interactions during the evolution of a tRNA.** *Nature* 2018, **558**:117-121.
50. Bolognesi B, Faure AJ, Seuma M, Schmiedel JM, Tartaglia GG, Lehner B: **The mutational landscape of a prion-like domain.** *Nat Commun* 2019, **10**:4162.
51. Du Z, Park K-W, Yu H, Fan Q, Li L: **Newly identified prion linked to the chromatin-remodeling factor Swi1 in *Saccharomyces cerevisiae*.** *Nat Genet* 2008, **40**:460-465.
52. Du Z, Valtierra S, Cardona LR, Dunne SF, Luan C-H, Li L: **Identifying anti-prion chemical compounds using a newly established yeast high-throughput screening system.** *Cell Chem Biol* 2019, **26**:1664-1680.e4.
53. Newby GA, Kiriakov S, Hallacli E, Kayatekin C, Tsvetkov P, Mancuso CP, *et al.*: **A genetic tool to track protein aggregates and control prion inheritance.** *Cell* 2017, **171**:966-979.e18.
54. Gray VE, Sitko K, Kameni FZN, Williamson M, Stephany JJ, Hasle N, *et al.*: **Elucidating the molecular determinants of A β aggregation with deep mutational scanning.** *G3* 2019, **9**:3683-3689.
55. Ebo JS, Saunders JC, Devine PWA, Gordon AM, Warwick AS, Schiffrin B, *et al.*: **An in vivo platform to select and evolve aggregation-resistant proteins.** *Nat Commun* 2020, **11**:1816.
56. Chandramowlishwaran P, Sun M, Casey KL, Romanyuk AV, Grizel AV, Sopova JV, *et al.*: **Mammalian amyloidogenic proteins promote prion nucleation in yeast.** *J Biol Chem* 2018, **293**:3436-3450.
57. Seuma M, Faure A, Badia M, Lehner B, Bolognesi B: **The genetic landscape for amyloid beta fibril nucleation accurately discriminates familial Alzheimer's disease mutations.** *Elife* 2021, **10**:e63364.
- The first DMS approach to report on amyloid nucleation.
58. Seuma M, Lehner B, Bolognesi B: **An atlas of amyloid aggregation: the impact of substitutions, insertions, deletions and truncations on amyloid beta fibril nucleation.** *bioRxiv* 2022,, <https://doi.org/10.1101/2022.01.18.476804>
59. Khan T, Kandola TS, Wu J, Venkatesan S, Ketter E, Lange JJ, *et al.*: **Quantifying nucleation in vivo reveals the physical basis of prion-like phase behavior.** *Mol Cell* 2018, **71**:155-168.e7.
- A fluorescence tool to measure *in vivo* nucleation events via imaging flow cytometry.
60. Hasle N, Cooke A, Srivatsan S, Huang H, Stephany JJ, Krieger Z, *et al.*: **High-throughput, microscope-based sorting to dissect cellular heterogeneity.** *Mol Syst Biol* 2020, **16**:e9442.
61. Schraivogel D, Kuhn TM, Rauscher B, Rodríguez-Martínez M, Paulsen M, Owsley K, *et al.*: **High-speed fluorescence image-enabled cell sorting.** *Science* 2022, **375**:315-320.
- Unlocking the possibility of sorting and selecting cells on the basis of specific intracellular phenotypes.
62. Rollins NJ, Brock KP, Poelwijk FJ, Stiffler MA, Gauthier NP, Sander C, *et al.*: **Inferring protein 3D structure from deep mutation scans.** *Nat Genet* 2019, **51**:1170-1176.
63. Schmiedel JM, Lehner B: **Determining protein structures using deep mutagenesis.** *Nat Genet* 2019, **51**:1177-1186.
64. Itakura AK, Chakravarty AK, Jakobson CM, Jarosz DF: **Widespread prion-based control of growth and differentiation strategies in *Saccharomyces cerevisiae*.** *Mol Cell* 2020, **77**:266-278.e6.
65. Morell M, de Groot NS, Vendrell J, Avilés FX, Ventura S: **Linking amyloid protein aggregation and yeast survival.** *Mol Biosyst* 2011, **7**:1121-1128.
66. Posey AE, Ruff KM, Lalmansingh JM, Kandola TS, Lange JJ, Halfmann R, *et al.*: **Mechanistic inferences from analysis of measurements of protein phase transitions in live cells.** *J Mol Biol* 2021, **433**:166848.

