An N-glycosylation hotspot in immunoglobulin κ light chains is associated with AL amyloidosis

Alice Nevone[1,2], Maria Girelli[1,2], Silvia Mangiacavalli[3], Bruno Paiva[4], Paolo Milani[1,2], Pasquale Cascino[1,2], Maggie Piscitelli[1,2], Valentina Speranzini[5], Claudio S. Cartia[3], Pietro Benvenuti[1,3], Ibai Goicoechea[4], Francesca Fazio[6], Marco Basset[1,2], Andrea Foli[1,2], Martina Nanci[1,2], Giulia Mazzini[1,2], Serena Caminito[1,2], Melania Antonietta Sesta[1,2], Simona Casarini[1,2], Paola Rognoni[1,2], Francesca Lavatelli[1,2], Maria Teresa Petrucci[6], Pier Paolo Olimpieri[7], Stefano Ricagno[5,8], Luca Arcaini[2,3], Giampaolo Merlini[1,2], Giovanni Palladini[1,2], Mario Nuvolone[1,2]

1 Amyloidosis Research and Treatment Center, Foundation IRCCS Policlinico San Matteo
2 Department of Molecular Medicine, University of Pavia, Pavia, Italy
3 Division of Hematology, Foundation IRCCS Policlinico San Matteo, Pavia, Italy
4 Clinica Universidad de Navarra, Centro de Investigacion Medica Aplicada (CIMA), Instituto de Investigacion Sanitaria de Navarra (IDISNA), CIBER-ONC number CB16/12/00369, Navarra, Spain
5 Dipartimento di Bioscienze, Università Degli Studi di Milano, Milan, Italy.
6 Hematology, Department of Translational and Precision Medicine, Azienda Ospedaliera Policlinico Umberto I, Sapienza University of Rome, Rome, Italy
7 Department of Public Health and Infectious Diseases, Sapienza University of Rome, Rome, Italy.
8 Institute of Molecular and Translational Cardiology, IRCCS Policlinico San Donato, Milan, Italy.

Corresponding authors:
Mario Nuvolone, MD PhD
Amyloidosis Research and Treatment Center
Foundation IRCCS Policlinico San Matteo
Viale Golgi 19, 27100 Pavia, Italy
Phone: +39-0382-501297
Fax: +39-0382-502990
Email: mario.nuvolone@unipv.it

Giovanni Palladini, MD PhD
Amyloidosis Research and Treatment Center
Foundation IRCCS Policlinico San Matteo
Viale Golgi 19, 27100 Pavia, Italy
Phone: +39-0382-502994
Fax: +39-0382-502990
Email: giovanni.palladini@unipv.it

Competing interests statement:
PC, GP and MNu are inventors on a patent application related to immunoglobulin sequencing.

Keywords: N-glycosylation, monoclonal gammopathies, AL amyloidosis, multiple myeloma, prediction.

Abstract

Immunoglobulin light chain (AL) amyloidosis is caused by a small, minimally proliferating B cell/plasma cell clone secreting a patient-unique, aggregation-prone, toxic light chain (LC). The pathogenicity of LCs is encrypted in their sequence, yet molecular determinants of amyloidogenesis are poorly understood.

Higher rates of N-glycosylation among clonal κ LCs from patients with AL amyloidosis compared to other monoclonal gammopathies indicate that this post-translational modification is associated with a higher risk of developing AL amyloidosis.

Here, we exploited LC sequence information from previously published amyloidogenic and control clonal LCs and from a series of 220 patients with AL amyloidosis or multiple myeloma followed at our Institutions to define sequence and spatial features of N-glycosylation, combining bioinformatics, biochemical, proteomics, structural and genetic analyses. We found peculiar sequence and spatial pattern of N-glycosylation in amyloidogenic κ LCs, with most of the N-glycosylation sites laying in the framework region 3, particularly within the E strand, and consisting mainly of the NFT sequon, setting them apart with respect to non-amyloidogenic clonal LCs.

Our data further support a potential role of N-glycosylation in determining the pathogenic behavior of a subset of amyloidogenic LCs and may help refine current N-glycosylation-based prognostic assessments for patients with monoclonal gammopathies.

Main text

Introduction

Systemic immunoglobulin light chain (AL) amyloidosis, a prototypic monoclonal gammopathy of clinical significance, is caused by a typically small, minimally proliferating, bone-marrow residing B cell or plasma cell clone secreting a pathogenic light chain (LC), more commonly of the λ isotype[1]. Disease-associated, monoclonal immunoglobulin LCs are unstable, misfolds, and aggregate in the forms of amyloid fibrils, which deposit in the extracellular space of target organs, leading to cytotoxicity, progressive subversion of tissue architecture, and potentially fatal organ dysfunction[1]. The amyloidogenicity of immunoglobulin LCs is believed to be encrypted, at least in part, in their sequence, which is unique to each patient[2-7]. Yet, molecular determinants of LC misfolding and aggregation are poorly understood.

Early anecdotal observations on individual LC sequences from patients with AL amyloidosis displaying N-glycosylation suggested a potential role of this post-translational modification in determining amyloidogenicity, particularly for κ LCs[8-13]. Recently, using mass spectrometry (MS) on immunoprecipitated serum free LCs (MASS-FIX), investigators at the Mayo Clinic have demonstrated that 16.7-33% of clonal κ LCs among patients with AL amyloidosis are N-glycosylated, as opposed to 3.7% of clonal κ LCs of patients with other monoclonal gammopathies[14-17]. Also, N-glycosylation of clonal LCs could be detected at the stage of monoclonal gammopathy of undetermined significance (MGUS) or smoldering multiple myeloma (MM), years before the subsequent diagnosis of AL amyloidosis[18]. Moreover, a study based on bulk and single-cell RNA sequencing of primary, patient-derived plasma cells has shown that transcriptional programs related to protein N-linked glycosylation are selectively upregulated in

AL[19]. Collectively, these observations point towards a potential contribution of N-glycosylation in influencing the amyloidogenicity of a subset of patients with AL amyloidosis.

In the present study, we exploited LC sequence information from previously published amyloidogenic and control clonal LCs, as well as from a series of 220 newly-sequenced patients with AL amyloidosis or MM followed at our Institutions, to study sequence and spatial features of N-glycosylation, combining bioinformatics analyses with biochemical, proteomics, structural and genetic investigations. We found peculiar sequence and spatial features of N-glycosylation in amyloidogenic κ LCs, further supporting a potential role of N-glycosylation in determining the pathogenic behavior of a subset of these proteins and possibly refining current N-glycosylation-based prognostic assessments.

Materials and methods

Ethical statements

Clinical records and biological samples were from subjects referred to the Italian Amyloid Center or to the Department of Hematology, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy and to the University Hospital of Navarra, Pamplona, Spain for a diagnostic workout in the suspicion of systemic AL amyloidosis or MM. Per the Declaration of Helsinki, all patients gave their written informed consent for the use of their clinical data and biological samples for research purposes, and this study was approved by the local Institutional Review Board.

In silico prediction of N-glycosylation of immunoglobulin LCs

Nucleotide sequences from published literature or obtained in this study (as detailed in Supplementary Information) were translated *in silico* with the Sequence Manipulation Suite (SMS) tool (https://www.bioinformatics.org/sms2/). Aminoacidic sequences in FASTA format were then submitted to NetNGlyc (http://www.cbs.dtu.dk/services/NetNGlyc/), an online tool that employs artificial neural networks that examine the sequence context of NXS/NXT sequons to predict N-glycosylation sites in human proteins[20]. Default parameters were employed.

Deglycosylation and Western blotting analysis

Urine proteins were digested with the PNGase F enzyme (NEB) according to the manufacturer's instructions. The enzyme was replaced with water for undigested control samples. Proteins were then boiled in reducing sample buffer, separated by SDS-PAGE in 4-15% Mini-protean TGX precast gels and electrotransferred onto 0.2µm PVDF membranes (Bio-Rad). Immunoblotting was performed using polyclonal rabbit antibody anti-human kappa LCs (DAKO, 1:13000) and polyclonal

secondary swine antibody anti-rabbit immunoglobulins HRP (DAKO, 1:10000). Blots were then developed by Immobilon Western chemiluminescent HRP substrate (Millipore).

Mass spectrometry

Urine proteins were digested with PNGase F glycerol-free enzyme (New England Biolabs), in non-denaturant conditions. Undigested control samples were prepared as detailed above. After reduction with dithiothreitol and alkylation with iodoacetamide, samples were digested with trypsin, purified on tip C18, and analyzed using a Dionex Ultimate 3000 UHPLC system coupled with a Q Exactive mass spectrometer (Thermo Fisher Scientific)[21]. Protein identification was obtained using the Proteome Discoverer software, version 2.0 (Thermo Fisher Scientific). The search was performed against the human proteome database (Uniprot) plus the LC sequence of interest. Enzymatic cleavage was set as full tryptic. Carbamidomethylation of cysteines as static modification, cyclization of N-terminal Q residues to form pyroglutamate, and deamidation of N residues was included as a dynamic modification. The presence of N-linked glycosylation at specific amino acid positions was evaluated indirectly, by examining the urinary tryptic digest at baseline and after PNGase F digestion for the presence of N deamidation.

Molecular Modeling

Multiple sequence alignments between the reference structure and the patient-derived sequences were produced using ClustalOmega with standard settings. The primary alignment was then used as input for secondary structure annotations (residues 1-107), which were produced using ESPript 3.0 in advanced mode. Mapping of the confirmed N-glycosylation sites onto the three-dimensional architecture of immunoglobulin κ LCs was carried out using the deposited structure of a

144 noncovalent Bence-Jones κ LC full dimer (PDB-ID:1B6D). Structural representations were produced
145 using Pymol.
146
147 Genomic analyses
148 Mutations resulting in a predicted N-glycosylation site and their corresponding genomic regions
149 were analyzed individually through ENSEMBL/GnomAD to verify the presence of candidate N-
150 glycosylation progenitor sites as previously defined[22]. The region of interest of germline *IGKV4-01*
151 gene from Pt. 73 was amplified from peripheral blood using the following primers (5'-3': Fwd
152 primer: GCCACCATCAACTGCAAGTC, Rev primer: ATTTCCACCTTGGTCCCTTGG) and cycling
153 conditions: 30 sec at 98°C; 35 cycles with 10 sec at 98°C, 30 sec at 68°C and 30 sec at 72°C; 1 min
154 at 72°C. The obtained amplicon was analyzed through Sanger sequencing according to standard
155 protocols.
156
157 Statistical analyses
158 Performance of predicted N-glycosylation site (any glycosylation versus FR3-DE glycosylation)
159 within the κ LCs as a prognostic factor for risk refinement has been evaluated by mean of the
160 Matthew Correlation Coefficient (MCC)[23] using R statistical software version 3.6.1 and the mltools
161 package.
162
163

Results

Sequence-based prediction of immunoglobulin LC N-glycosylation

N-glycosylation of proteins typically occurs on N-residues within NXS/T consensus sequences (termed sequon, where N is asparagine, X is any residue, more commonly not a proline, S is serine and T is threonine) in the presence of appropriate sequence context[24]. We used NetNGlyc to predict potential N-glycosylation sites within the variable region of LCs from previously published amyloidogenic and control LCs[7, 25]. A total of 835 variable region sequences were analyzed, including 597 amyloidogenic LCs (142 κ and 455 λ) and 238 control clonal LCs (127 κ and 111 λ) from patients with non-AL plasma cell disorders, mainly MM.

Among amyloidogenic κ LCs, NetNGlyc predicted the presence of at least one N-glycosylation site in 44 out of 142 patients (31%) (Fig. 1A, Suppl. Fig. 1). No variable region sequence presented 2 or more predicted sites.

We then investigated the spatial distribution of the predicted N-glycosylation sites and sequon usage. Interestingly, considering the 44 amyloidogenic κ LCs predicted to be N-glycosylated within their variable region, the putative N-glycosylation site lay within the framework region 3 (FR3) in 36 cases (82%), mostly within the E strand (57%) and, to a less extent, within the D strand (25%) (Fig. 1A). Also, the putative N-glycosylation site was NFT in 22 (50%) of sequences. Of note, a putative NFT site was invariably present within the E strand of FR3 (Fig. 1A).

Conversely, non-amyloidogenic, clonal κ LC sequences from patients with other, non-AL plasma cell disorders were predicted to be N-glycosylated only in 17 out of 127 cases (13%), with no preferential spatial distribution or sequon usage (Fig. 1A, Suppl. Fig. 1).

Conversely, a predicted N-glycosylation site was seen only in 28 out of 455 (6%) amyloidogenic λ LC sequences, mainly laying within the CDR3 region and in clones derived from the *IGLV2-14*

germline gene (Suppl. Fig 1, Suppl. Fig. 2A). A similar proportion of predicted N-glycosylation site was present in non-amyloidogenic, clonal λ LC sequences (Suppl. Fig 1, Suppl. Fig. 2A).

To validate these observations in an independent series of patients, we sequenced 220 consecutive patients with monoclonal gammopathies followed at our Institutions, including 119 patients with AL amyloidosis (n= 24 κ and n= 95 λ) and 101 patients with MM (n= 68 κ and n= 33 λ). Again, we used NetNGlyc to predict potential N-glycosylation sites within the variable region of clonal LCs. Isotype (κ versus λ) restriction and germline gene usage were in line with the expectations for AL amyloidosis and MM patients (Suppl. Fig. 3). Among patients with κ-expressing clones, NetNGlyc predicted the presence of one N-glycosylation site in 9 out of 24 patients with AL amyloidosis (38%), in 7 cases (78%) within the E strand of FR3, and consisting of an NFT sequon in 4 (44%) cases (Fig. 1B). Conversely, a predicted N-glycosylation site was identified only in 5 out of 68 patients with MM expressing a κ LC (7%), mostly outside of FR3. Among patients with λ-expressing clones, a predicted N-glycosylation site was seen only in 1 out of 95 patients with AL amyloidosis (1%) and in 2 out of 33 patients with MM (6%) (Fig. 1B).

Biochemical and proteomics confirmation of N-glycosylation prediction

To verify the accuracy of NetNGlyc prediction, we analyzed urine samples from 95 patients affected by AL amyloidosis or MM from our series (23 κ and 72 λ clones, of which 7 and 1 predicted to be N-glycosylated, respectively). Urinary proteins were digested with protein N-glycosidase F (PNGase F), an amidase, which hydrolyses the bond between the innermost GlcNAc and asparagine residues of glycoproteins with N-linked oligosaccharides, thus releasing the intact oligosaccharides and leaving deamidated asparagine residues within proteins/peptides[26]. Digested urinary proteins were then analyzed by Western blotting. In this assay, unglycosylated LCs are expected to migrate at the level corresponding to 25 kDa proteins, while N-glycosylated LCs

212 migrate as a >25-KDa band. Both clonal LCs (Bence Jones proteins) and polyclonal LCs can be
213 detected, the latter being more abundant in patients with glomerular damage and substantial
214 proteinuria, as in the case of amyloid renal involvement. Therefore, in the case of patients with
215 substantial proteinuria and detection of multiple bands with Western blotting on urine samples at
216 diagnosis, we also analyzed samples at the time of achievement of a hematologic response after
217 chemotherapy, whenever available, to better discriminate between clonal and polyclonal LCs,
218 considering the therapy-induced reduction of clonal LCs. In all cases, undigested urinary protein
219 samples were analyzed as controls. Interestingly, a shift in electrophoretic mobility upon PNGase F
220 digestion, indicative of the presence of an N-linked glycan, was seen in 8 out of 8 cases (100%)
221 with a predicted N-glycosylation site, and in none of the 87 cases (0%) lacking a predicted N-
222 glycosylation site (Fig. 2A). Thus, these biochemical analyses confirmed the accuracy of NetNGlyc
223 prediction.
224 To further verify that the N-glycosylation occurred at the identified sequon within the FR3 region,
225 among AL patients with an N-glycosylated clonal κ LC, we selected one case lacking amyloid renal
226 involvement and clinically significant proteinuria (patient 24), to minimize the amount of non-
227 clonal κ LCs (and other proteins) within the urine sample. The clonal κ LC sequence of this patient
228 displayed three asparagine residues (N34, N37, and N90, with N90 lying in the FR3 region, being
229 part of the NIS sequon and predicted to be glycosylated). To account for the possibility of
230 spontaneous deamidation of asparagine residues[27], we split the urine sample into two parts, one
231 undergoing PNGase F digestion, while the other being exposed to the same buffer and same
232 incubation conditions, yet lacking the enzyme. Urinary proteins from patient 24 with or without
233 enzyme were then analyzed with MS. Of note, in both samples the clonal κ LC sequence was the
234 protein with the highest sequence coverage (99% in both cases), with multiple peptides, including
235 peptides spanning all three CDRs (Fig. 2B). We then compared deamidation in all three available N

236 residues within the clonal κ LC sequence in undigested and PNGase F-digested urine samples.
237 Interestingly, only N90 of the NIS sequon showed a substantial increase in deamidation upon
238 PNGase F digestion (Fig. 2B).
239 Collectively, these results are in agreement with the N-glycosylation occurring at the expected
240 sequon of the clonal LC sequence, as predicted by NetNGlyc, and corroborate the observation of
241 an N-glycosylation hotspot in amyloidogenic κ LCs. For the subsequent analyses, we, therefore,
242 focused on κ LCs.
243
244 **Progenitor glycosylation sites, rather than genomic variants, explain the N-glycosylation hotspot**
245 **in amyloidogenic κ LCs**
246 As germline *IGKV* genes (and allelic variants thereof) do not encode for any sequon, it has been
247 suggested that the presence of a sequon in expressed κ LCs be the result of somatic
248 hypermutation[9, 22]. This process has been proposed to occur particularly at the level of specific
249 sites, termed progenitor glycosylation sites, laying mainly within the CDRs and within the DE loop,
250 where one single base pair substitution would suffice to introduce a putative N-glycosylation site[9,
251 22]. In this context, the potential contribution of rare, germline-encoded single-nucleotide variants
252 in the generation of sequons within immunoglobulin κ LCs has remained unaddressed.
253 We first compared the 51 amyloidogenic κ LC sequences predicted to be N-glycosylated with
254 available nucleotide sequence information (42 previously published sequences and 9 from our
255 series) with their corresponding germline gene/allele and verified that in 32 cases (63%) the
256 predicted N-glycosylation was indeed the actualization of a progenitor glycosylation site, occurring
257 as a result of a single nucleotide substitution[22]. Two, three, or four nucleotide substitutions were
258 seen in 15, 2, and 2 cases, respectively. Of note, in all the 51 sequences, the observed nucleotidic
259 mutation(s) resulted in the acquisition of an N residue in the context of an NXS/T sequon (Fig. 3A).

We then exploited the availability of genomic sequence information from the Genome Aggregation Database (gnomAD)[28] to verify the occurrence, within the general population, of nucleotide substitutions leading to the acquisition of an N residue in the context of an NXS/T sequon which we had identified in amyloidogenic κ LC sequences.

Out of 51 sequences, there were 32 unique nucleotide substitutions. Of these, 15 were annotated as potential single nucleotide polymorphisms (SNPs) in GnomAD, with sequence information available from a median of 244,818 alleles (interquartile range, IQR: 196,845 - 246,071 alleles) (Fig. 3A). Ten of these SNPs were identified in only 1-3 alleles out of a median of 245,301 alleles, thus qualifying as ultra-rare variants[29].

Considering the 9 amyloidogenic κ LC sequences from our series, for which we had access to genomic DNA, only the clonal κ LC sequence identified in patient 73 contained an SNP (rs748402676, within the *IGKV4-1* gene), which was observed in only 3 out of 244,651 alleles (2 from Latino/admixed American and 1 from European non-Finnish individuals) within the GnomAD dataset. To verify if this SNP associated with the formation of the N-glycosylation site in patient 73 was present already at the genomic level, we sequenced this region of the *IGKV4-1* gene in gDNA from the peripheral blood of patient 73. Of note, Sanger sequencing confirmed the lack of this SNP within gDNA from this patient (Fig. 3B). Overall, these data support the idea that somatic hypermutation in the context of progenitor glycosylation sites, rather than ultra-rare genomic variants, is responsible for the N-glycosylation hotspot in amyloidogenic κ LCs.

The structural context of N-glycosylation in amyloidogenic κ LCs

To gain insight into the structural context of the observed N-glycosylation of amyloidogenic κ LCs, we used the available structure of a full dimer of a Bence-Jones κ LC with accession number 1B6D in the Protein Data Bank[30]. Guided by multiple primary sequence alignment between 1B6D and

the patient-derived amyloidogenic κ LC sequences confirmed, in this work, to be N-glycosylated within the DE loop, we mapped the N-glycosylation sites onto the three-dimensional structure (Fig. 4). The exact position of the glycosylated N residue is variable in the seven sequences: four κ LCs are N-glycosylated in position 88 (according to IMGT numbering), two in 86, and one in position 90; in all cases, the glycosylated N is mapping on the beta-strand E. We then inspected the other patient-derived sequences, predicted to be N-glycosylated, and mapped representative ones onto the same structural model (Suppl. Fig. 4). In this larger set of sequences, the position of the N-glycosylation is more variable. Overall, we observed the involvement of three different immunoglobulin regions (FR1, CDR1, FR3), all located on the solvent-exposed surface of the variable domain (strands B, D, E, and hypervariable loop 1), thus not directly affecting the monomer-monomer interface in the context of the full dimeric protein.

Prognostic significance of N-glycosylation site mapping within κ LCs

In light of our observations, we asked whether incorporating the knowledge of the spatial distribution of the predicted N-glycosylation site may refine risk prediction. To this aim, we first merged LC sequence information from patients with AL amyloidosis and non-AL, clonal plasma cell disorders coming from published literature and the present study. We then used the presence of a putative N-glycosylation site at any position within the clonal κ LC or specifically within the FR3 region to classify the clonal κ LC as potentially amyloidogenic or not. We then applied this rule to our cohort of merged sequences, defined each sequence as true or false positive and true or false negative based on the presence or absence of the predictor and the corresponding clinical category, and we computed the Matthews Correlation Coefficient (MCC). Of note, incorporating the knowledge of the spatial distribution of the putative N-glycosylation site led to an increase in positive predictive rate from 0.71 to 0.83 and MCC from 0.25 to 0.30 (Fig. 5). Collectively, these

data indicate a distinctive sequence and regional distribution of N-glycosylation site among amyloidogenic κ LCs which can be exploited to identify clonal κ LCs more likely to be associated with AL amyloidosis.

Discussion

In this study, we identified a novel N-glycosylation hotspot in immunoglobulin κ LCs that are associated with AL amyloidosis. Our data suggest that somatic hypermutations in the context of progenitor glycosylation sites within the FR3 region are the main determinants of κ LC glycosylation. The knowledge of the spatial distribution of the putative N-glycosylation site leads to an increase in the positive predictive rate of the amyloidogenicity of the sequence of a clonal κ LC. This may improve the capability to detect dangerous κ LCs and achieve early diagnosis that still represents an unmet need.

The addition of N-glycans to proteins influences their folding, molecular trafficking, and binding properties and N-glycosylated proteins are crucially involved in multiple physiological functions, including cell adhesion, receptor activation, signal transduction, and endocytosis[31, 32].

The most abundant glycoproteins in human serum are represented by immunoglobulins[33]. Besides the well-characterized, germline gene-encoded N-glycosylation site within the constant region of immunoglobulin heavy chains, which imparts effector functions like binding to Fc receptors[34, 35], immunoglobulin heavy and LCs can occasionally display an N-glycosylation site within their variable region as a result of somatic hypermutation during affinity maturation[22, 36, 37]. N-glycans within the immunoglobulin variable region impact antibody stability and affinity, thus forming an additional layer of diversification within the antibody repertoire[22, 38].

The acquisition of potential N-glycosylation sites in the immunoglobulin variable region through somatic hypermutation is a distinctive feature of follicular lymphoma and has been reported also in subsets of Burkitt's lymphoma and diffuse large B cell lymphoma[39-42]. In this setting, it has been proposed that N-glycans of tumor-associated surface immunoglobulins form a bridge with microenvironmental lectins, thus providing persistent activating signals to tumor cells[43-45]. The

potential pathogenic role of N-glycosylation in the immunoglobulin variable region in other clinical settings is less well understood.

N-glycosylation has been long suspected to be a determinant of κ LC amyloidogenicity[9, 46]. It was suggested that N-glycosylation may support fibril formation by improving binding capacity at extracellular matrix sites (therefore creating favorable fibrillogenic conditions), by altering intrinsic protein characteristics (such as protein solvation), and posing steric hindrance or protection from protease degradation during clearance[47, 48]. The latest hypothesis is further supported by a recently reported structure of a λ LC *ex vivo* fibril[49], carrying different post-translational modifications, including N-glycosylation. In this study, it was observed that the glycosylation sites are exposed on the surface of both the fibrillar and native form of the protein, as for our κ models for LC. It is believed that such modification, in the case of the λ fibril, contributes to determining the overall fold and fibrillar morphology, at the same time making the amyloid more resistant to proteolytic degradation[49].

Recent clinical observations have renewed the interest in N-glycosylation of monoclonal LCs, particularly of the κ type: 1) The increased prevalence of N-glycosylation among κ monoclonal LCs from patients with AL amyloidosis compared to patients with other plasma cell disorders[14, 16]; 2) the notion that N-glycosylation of monoclonal LCs is an independent and potent risk factor of MGUS progression towards a symptomatic plasma cell disorder, particularly AL amyloidosis[15]; 3) the fact that N-glycosylation of monoclonal LCs predates the onset of symptomatic AL amyloidosis by years and is present already at the stage of MGUS or smoldering MM[18]. Indeed, the International Myeloma Working Group has recently warranted further investigations to clarify the relationship and implications of N-linked glycosylation and AL amyloidosis[50].

By combining sequence information in our patient population and from previous patients' series with bioinformatic analyses, we have found that a substantial proportion of amyloidogenic κ LC

sequences display a predicted N-glycosylation motif (NXS/T). Moreover, and somewhat unexpectedly, we found that such motif is mostly located within a specific region of the κ LC variable region (that is in strand E or D, within FR3) and displays an NFT motif in about half the cases. Structural analyses confirmed that the N-glycosylation sites experimentally verified in this work are exposed at the surface of the LC monomer and map on a region not involved in dimer formation. This holds also for the larger set of amyloidogenic κ LCs from public repositories which were predicted to be N-glycosylated. Thus, structural analysis suggests that N-glycosylation should not affect the formation of the dimer interface and more in general the dimer assembly.

Importantly, biochemical and proteomic analyses have confirmed the accuracy of N-glycosylation prediction, demonstrating altered electrophoretic mobility +/- the appearance of N deamidation upon PNGase F digestion of Bence Jones proteins from AL patients with a predicted N-glycosylated monoclonal LC. The recognition of a specific spatial and sequence preference of the N-glycosylation site in a significant proportion of amyloidogenic κ LCs suggests a potential pathogenic effect of this post-translational modification.

From a biochemical and biophysical point of view, a growing body of evidence indicates that LC amyloidogenicity correlates with several biophysical properties, including reduced thermodynamic/kinetic stability and protein dynamics. Moreover, beta strands D and E are one hotspot for those mutations and modifications that contribute to altered protein stability and subsequent variation of aggregation propensity[51-60].

The addition of a sugar moiety to the side chain of key surface residues such as asparagine could contribute to the destabilization of the native state of the LC, therefore promoting protein misfolding, while offering extra charged binding surfaces that would create a localized accumulation of the modified LCs, favoring aggregation.

382 On the other hand, glycans are bulky, highly flexible, and soluble molecules. Thus, compared to
383 the un-glycosylated counterpart, a glycosylated protein displays an increased steric hindrance and
384 solubility, properties that are regarded as protective against protein aggregation and/or
385 degradation. For example, in polymerogenic proteins such as serpins, N-glycosylation is protective
386 against the formation of pathologic polymers[61, 62]. Further studies investigating the possible effect
387 of the presence, sequence, and localization of an N-glycosylation site on LC pathogenicity are
388 warranted.
389 Different from amyloidogenic κ LCs, monoclonal κ LCs associated with other plasma cell disorders
390 display a lower prevalence of predicted N-glycosylation sites and lack a preferential spatial
391 distribution or sequon usage. Should this observation be further supported by the analysis of a
392 higher number of sequences, this would imply that incorporating the knowledge of the spatial
393 distribution of the predicted N-glycosylation site may refine risk prediction of an N-glycosylated
394 clonal κ LC sequence towards AL.
395
396

Acknowledgments

We thank Andrea Patrignani and staff of the Functional Genomic Center Zurich for PacBio library preparation and sequencing. This work was supported by grants from the Amyloidosis Foundation (MNu), the Italian Ministry of Health (Ricerca Finalizzata, grant #GR-2018-12368387) (MNu), the Italian Ministry of Research and Education (PRIN 20207XLJB2) (SR, GP), the CARIPLO Foundation (grant #2018-0257) (MNu), Fondazione ARISLA (project TDP- 43-STRUCT) (SR), Cancer Research UK [C355/A26819], FC AECC and AIRC under the Accelerator Award Program (BP, GM, GP, MNu).

Author contribution:

Conceived, designed and supervised the project: MNu.

Provided financial support to the project: SR, BP, GMe, GP, MNu.

Performed clinical evaluations, provided patients' samples and/or collected clinical data: SM, BP, PM, CCS, PB, IG, FF, MB, AF, MNa, MTP, LA, GMe, GP, MNu.

Maintained biorepository: AN, MG, PC, SCam, MAS, SCas.

Curated clinical database: AN, PB, FF, MNu.

Retrieved published sequences and performed *in silico* analyses: AN, MG, PC, MP, PR, MNu.

Processed bone marrow samples, prepared sequencing libraries and analyzed LC sequences: AN, MG, PC, MP, MNu.

Performed PNGase F digestion and Western blotting: MG, PC, MP.

Performed mass spectrometry analyses: GMa, SCam, FL.

Performed genomic analyses and Sanger sequencing: AN, MG, PC, MP, SCas.

Performed molecular modelling and structural analyses: VS, SR.

Performed statistical analyses: PPO.

Prepared figures: AN, PC, MP, VS, SR, MNu.

Wrote the manuscript: AN, MNu.

Read, edited and approved the manuscript: All authors

Competing interests:

PC, GP and MNu are inventors on a patent application related to immunoglobulin sequencing.

Data availability statement

The data generated in this study are available within the article and its supplementary data files. LC sequences have been deposited to GenBank (MZ595009-MZ595094).

References

1. Merlini G, Dispenzieri A, Sanchorawala V, Schonland SO, Palladini G, Hawkins PN*, et al.* Systemic immunoglobulin light chain amyloidosis. *Nat Rev Dis Primers* 2018 Oct 25; **4**(1)**:** 38.

2. Comenzo RL, Zhang Y, Martinez C, Osman K, Herrera GA. The tropism of organ involvement in primary systemic amyloidosis: contributions of Ig V(L) germ line gene use and clonal plasma cell burden. *Blood* 2001 Aug 1; **98**(3)**:** 714-720.

3. Abraham RS, Geyer SM, Price-Troska TL, Allmer C, Kyle RA, Gertz MA*, et al.* Immunoglobulin light chain variable (V) region genes influence clinical presentation and outcome in light chain-associated amyloidosis (AL). *Blood* 2003 May 15; **101**(10)**:** 3801-3808.

4. Perfetti V, Casarini S, Palladini G, Vignarelli MC, Klersy C, Diegoli M*, et al.* Analysis of V(lambda)-J(lambda) expression in plasma cells from primary (AL) amyloidosis and normal bone marrow identifies 3r (lambdaIII) as a new amyloid-associated germline gene segment. *Blood* 2002 Aug 1; **100**(3)**:** 948-953.

5. Prokaeva T, Spencer B, Kaut M, Ozonoff A, Doros G, Connors LH*, et al.* Soft tissue, joint, and bone manifestations of AL amyloidosis: clinical presentation, molecular features, and survival. *Arthritis Rheum* 2007 Nov; **56**(11)**:** 3858-3868.

6. Perfetti V, Palladini G, Casarini S, Navazza V, Rognoni P, Obici L*, et al.* The repertoire of lambda light chains causing predominant amyloid heart involvement and identification of a preferentially involved germline gene, IGLV1-44. *Blood* 2012 Jan 5; **119**(1)**:** 144-150.

7. Bodi K, Prokaeva T, Spencer B, Eberhard M, Connors LH, Seldin DC. AL-Base: a visual platform analysis tool for the study of amyloidogenic immunoglobulin light chain sequences. *Amyloid* 2009 Mar; **16**(1)**:** 1-8.

8. Dwulet FE, O'Connor TP, Benson MD. Polymorphism in a kappa I primary (AL) amyloid protein (BAN). *Mol Immunol* 1986 Jan; **23**(1)**:** 73-78.

9. Stevens FJ. Four structural risk factors identify most fibril-forming kappa light chains. *Amyloid* 2000 Sep; **7**(3)**:** 200-211.

10. Omtvedt LA, Bailey D, Renouf DV, Davies MJ, Paramonov NA, Haavik S*, et al.* Glycosylation of immunoglobulin light chains associated with amyloidosis. *Amyloid* 2000 Dec; **7**(4)**:** 227-244.

11. Karimi M, Sletten K, Westermark P. Biclonal systemic AL-amyloidosis with one glycosylated and one nonglycosylated AL-protein. *Scand J Immunol* 2003 Apr; **57**(4)**:** 319-323.

12. Myran T, Husby G, Kyle RA, Sletten K. The amino acid sequence of a glycosylated AL-chain from a patient with primary amyloidosis. *Amyloid* 2004 Jun; **11**(2)**:** 109-112.

13. Connors LH, Jiang Y, Budnik M, Theberge R, Prokaeva T, Bodi KL*, et al.* Heterogeneity in primary structure, post-translational modifications, and germline gene usage of nine full-length amyloidogenic kappa1 immunoglobulin light chains. *Biochemistry* 2007 Dec 11; **46**(49)**:** 14259-14271.

14. Kumar S, Murray D, Dasari S, Milani P, Barnidge D, Madden B*, et al.* Assay to rapidly screen for immunoglobulin light chain glycosylation: a potential path to earlier AL diagnosis for a subset of patients. *Leukemia* 2019 Jan; **33**(1)**:** 254-257.

15. Dispenzieri A, Larson DR, Rajkumar SV, Kyle RA, Kumar SK, Kourelis T*, et al.* N-glycosylation of monoclonal light chains on routine MASS-FIX testing is a risk factor for MGUS progression. *Leukemia* 2020 Oct; **34**(10)**:** 2749-2753.

16. Mellors PW, Dasari S, Kohlhagen MC, Kourelis T, Go RS, Muchtar E*, et al.* MASS-FIX for the detection of monoclonal proteins and light chain N-glycosylation in routine clinical practice: a cross-sectional study of 6315 patients. *Blood Cancer J* 2021 Mar 4; **11**(3)**:** 50.

17. Milani P, Murray DL, Barnidge DR, Kohlhagen MC, Mills JR, Merlini G*, et al.* The utility of MASS-FIX to detect and monitor monoclonal proteins in the clinic. *Am J Hematol* 2017 Aug; **92**(8)**:** 772-779.

18. Kourelis T, Murray DL, Dasari S, Kumar S, Barnidge D, Madden B*, et al.* MASS-FIX may allow identification of patients at risk for light chain amyloidosis before the onset of symptoms. *Am J Hematol* 2018 Nov; **93**(11)**:** E368-E370.

19. Alameda D, Goicoechea I, Vicari M, Arriazu E, Nevone A, Rodriguez S*, et al.* Tumor cells in light-chain amyloidosis and myeloma show different transcriptional rewiring of normal plasma cell development. *Blood* 2021 Jun 16.

20. Gupta R, Brunak S. Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac Symp Biocomput* 2002**:** 310-322.

21. Lavatelli F, Mazzini G, Ricagno S, Iavarone F, Rognoni P, Milani P*, et al.* Mass spectrometry characterization of light chain fragmentation sites in cardiac AL amyloidosis: insights into the timing of proteolysis. *J Biol Chem* 2020 Dec 4; **295**(49)**:** 16572-16584.

22. van de Bovenkamp FS, Derksen NIL, Ooijevaar-de Heer P, van Schie KA, Kruithof S, Berkowska MA*, et al.* Adaptive antibody diversification through N-linked glycosylation of the immunoglobulin variable region. *Proc Natl Acad Sci U S A* 2018 Feb 20; **115**(8)**:** 1901-1906.

23. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 2020 Jan 2; **21**(1)**:** 6.

24. Imperiali B, Hendrickson TL. Asparagine-linked glycosylation: specificity and function of oligosaccharyl transferase. *Bioorg Med Chem* 1995 Dec; **3**(12)**:** 1565-1578.

25. Malecka A, Troen G, Tierens A, Ostlie I, Malecki J, Randen U*, et al.* Immunoglobulin heavy and light chain gene features are correlated with primary cold agglutinin disease onset and activity. *Haematologica* 2016 Sep; **101**(9)**:** e361-364.

26. Maley F, Trimble RB, Tarentino AL, Plummer TH, Jr. Characterization of glycoproteins and their associated oligosaccharides through the use of endoglycosidases. *Anal Biochem* 1989 Aug 1; **180**(2)**:** 195-204.

27. Palmisano G, Melo-Braga MN, Engholm-Keller K, Parker BL, Larsen MR. Chemical deamidation: a common pitfall in large-scale N-linked glycoproteomic mass spectrometry-based analyses. *J Proteome Res* 2012 Mar 2; **11**(3)**:** 1949-1957.

28. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q*, et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020 May; **581**(7809)**:** 434-443.

29. Hernandez RD, Uricchio LH, Hartman K, Ye C, Dahl A, Zaitlen N. Ultrarare variants drive substantial cis heritability of human gene expression. *Nat Genet* 2019 Sep; **51**(9)**:** 1349-1355.

30. Roussel A, Spinelli S, Deret S, Navaza J, Aucouturier P, Cambillau C. The structure of an entire noncovalent immunoglobulin kappa light-chain dimer (Bence-Jones protein) reveals a weak and unusual constant domains association. *Eur J Biochem* 1999 Feb; **260**(1)**:** 192-199.

31. Ohtsubo K, Marth JD. Glycosylation in cellular mechanisms of health and disease. *Cell* 2006 Sep 8; **126**(5)**:** 855-867.

32. Helenius A, Aebi M. Intracellular functions of N-linked glycans. *Science* 2001 Mar 23; **291**(5512)**:** 2364-2369.

33. Gudelj I, Lauc G, Pezer M. Immunoglobulin G glycosylation in aging and diseases. *Cell Immunol* 2018 Nov; **333:** 65-79.

34. Quast I, Keller CW, Maurer MA, Giddens JP, Tackenberg B, Wang LX*, et al.* Sialylation of IgG Fc domain impairs complement-dependent cytotoxicity. *J Clin Invest* 2015 Nov 2; **125**(11)**:** 4160-4170.

35. Arnold JN, Wormald MR, Sim RB, Rudd PM, Dwek RA. The impact of glycosylation on the biological function and structure of human immunoglobulins. *Annu Rev Immunol* 2007; **25:** 21-50.

36. Dunn-Walters D, Boursier L, Spencer J. Effect of somatic hypermutation on potential N-glycosylation sites in human immunoglobulin heavy chain variable regions. *Mol Immunol* 2000 Feb-Mar; **37**(3-4)**:** 107-113.

37. Anumula KR. Quantitative glycan profiling of normal human plasma derived immunoglobulin and its fragments Fab and Fc. *J Immunol Methods* 2012 Aug 31; **382**(1-2)**:** 167-176.

38. Sabouri Z, Schofield P, Horikawa K, Spierings E, Kipling D, Randall KL*, et al.* Redemption of autoantibodies on anergic B cells by variable-region glycosylation and mutation away from self-reactivity. *Proc Natl Acad Sci U S A* 2014 Jun 24; **111**(25)**:** E2567-2575.

39. Zhu D, McCarthy H, Ottensmeier CH, Johnson P, Hamblin TJ, Stevenson FK. Acquisition of potential N-glycosylation sites in the immunoglobulin variable region by somatic mutation is a distinctive feature of follicular lymphoma. *Blood* 2002 Apr 1; **99**(7)**:** 2562-2568.

40. Zhu D, Ottensmeier CH, Du MQ, McCarthy H, Stevenson FK. Incidence of potential glycosylation sites in immunoglobulin variable regions distinguishes between subsets of Burkitt's lymphoma and mucosa-associated lymphoid tissue lymphoma. *Br J Haematol* 2003 Jan; **120**(2)**:** 217-222.

41. Forconi F, Capello D, Berra E, Rossi D, Gloghini A, Cerri M*, et al.* Incidence of novel N-glycosylation sites in the B-cell receptor of lymphomas associated with immunodeficiency. *Br J Haematol* 2004 Mar; **124**(5)**:** 604-609.

42. Zabalegui N, de Cerio AL, Inoges S, Rodriguez-Calvillo M, Perez-Calvo J, Hernandez M*, et al.* Acquired potential N-glycosylation sites within the tumor-specific immunoglobulin heavy chains of B-cell malignancies. *Haematologica* 2004 May; **89**(5)**:** 541-546.

43. Coelho V, Krysov S, Ghaemmaghami AM, Emara M, Potter KN, Johnson P*, et al.* Glycosylation of surface Ig creates a functional bridge between human follicular lymphoma and microenvironmental lectins. *Proc Natl Acad Sci U S A* 2010 Oct 26; **107**(43)**:** 18587-18592.

44. Linley A, Krysov S, Ponzoni M, Johnson PW, Packham G, Stevenson FK. Lectin binding to surface Ig variable regions provides a universal persistent activating signal for follicular lymphoma cells. *Blood* 2015 Oct 15; **126**(16)**:** 1902-1910.

45. Amin R, Mourcin F, Uhel F, Pangault C, Ruminy P, Dupre L*, et al.* DC-SIGN-expressing macrophages trigger activation of mannosylated IgM B-cell receptor in follicular lymphoma. *Blood* 2015 Oct 15; **126**(16)**:** 1911-1920.

46. Bellotti V, Mangione P, Merlini G. Review: immunoglobulin light chain amyloidosis--the archetype of structural and pathogenic variability. *J Struct Biol* 2000 Jun; **130**(2-3)**:** 280-289.

47. Zhang C, Huang X, Li J. Light chain amyloidosis: Where are the light chains from and how they play their pathogenic role? *Blood Rev* 2017 Jul; **31**(4)**:** 261-270.

48. Blancas-Mejia LM, Misra P, Dick CJ, Cooper SA, Redhage KR, Bergman MR*, et al.* Immunoglobulin light chain amyloid aggregation. *Chem Commun (Camb)* 2018 Sep 20; **54**(76)**:** 10664-10674.

49. Radamaker L, Karimi-Farsijani S, Andreotti G, Baur J, Neumann M, Schreiner S*, et al.* Role of mutations and post-translational modifications in systemic AL amyloidosis studied by cryo-EM. *Nat Commun* 2021 Nov 5; **12**(1)**:** 6434.

50. Murray DL, Puig N, Kristinsson S, Usmani SZ, Dispenzieri A, Bianchi G*, et al.* Mass spectrometry for the evaluation of monoclonal proteins in multiple myeloma and related disorders: an International Myeloma Working Group Mass Spectrometry Committee Report. *Blood Cancer J* 2021 Feb 1; **11**(2)**:** 24.

51. Blancas-Mejia LM, Horn TJ, Marin-Argany M, Auton M, Tischer A, Ramirez-Alvarado M. Thermodynamic and fibril formation studies of full length immunoglobulin light chain AL-09 and its germline protein using scan rate dependent thermal unfolding. *Biophys Chem* 2015 Dec; **207:** 13-20.

52. Klimtchuk ES, Gursky O, Patel RS, Laporte KL, Connors LH, Skinner M*, et al.* The critical role of the constant region in thermal stability and aggregation of amyloidogenic immunoglobulin light chain. *Biochemistry* 2010 Nov 16; **49**(45)**:** 9848-9857.

53. Gonzalez-Andrade M, Becerril-Lujan B, Sanchez-Lopez R, Cecena-Alvarez H, Perez-Carreon JI, Ortiz E*, et al.* Mutational and genetic determinants of lambda6 light chain amyloidogenesis. *FEBS J* 2013 Dec; **280**(23)**:** 6173-6183.

54. Maritan M, Ambrosetti A, Oberti L, Barbiroli A, Diomede L, Romeo M*, et al.* Modulating the cardiotoxic behaviour of immunoglobulin light chain dimers through point mutations. *Amyloid* 2019; **26**(sup1)**:** 105-106.

55. Rottenaicher GJ, Weber B, Ruhrnossl F, Kazman P, Absmeier RM, Hitzenberger M*, et al.* Molecular mechanism of amyloidogenic mutations in hypervariable regions of antibody light chains. *J Biol Chem* 2021 Jan-Jun; **296:** 100334.

56. Kazman P, Vielberg MT, Pulido Cendales MD, Hunziger L, Weber B, Hegenbart U*, et al.* Fatal amyloid formation in a patient's antibody light chain is caused by a single point mutation. *Elife* 2020 Mar 10; **9**.

57. Weber B, Hora M, Kazman P, Gobl C, Camilloni C, Reif B*, et al.* The Antibody Light-Chain Linker Regulates Domain Orientation and Amyloidogenicity. *J Mol Biol* 2018 Dec 7; **430**(24)**:** 4925-4940.

58. Morgan GJ, Kelly JW. The Kinetic Stability of a Full-Length Antibody Light Chain Dimer Determines whether Endoproteolysis Can Release Amyloidogenic Variable Domains. *J Mol Biol* 2016 Oct 23; **428**(21)**:** 4280-4297.

59. Morgan GJ, Usher GA, Kelly JW. Incomplete Refolding of Antibody Light Chains to Non-Native, Protease-Sensitive Conformations Leads to Aggregation: A Mechanism of Amyloidogenesis in Patients? *Biochemistry* 2017 Dec 19; **56**(50)**:** 6597-6614.

60. Rennella E, Morgan GJ, Kelly JW, Kay LE. Role of domain interactions in the aggregation of full-length immunoglobulin light chains. *Proc Natl Acad Sci U S A* 2019 Jan 15; **116**(3)**:** 854-863.

655 61. Moriconi C, Ordonez A, Lupo G, Gooptu B, Irving JA, Noto R*, et al.* Interactions between N-linked
656 glycosylation and polymerisation of neuroserpin within the endoplasmic reticulum. *FEBS J* 2015
657 Dec; **282**(23)**:** 4565-4579.

658

659 62. Visentin C, Broggini L, Sala BM, Russo R, Barbiroli A, Santambrogio C*, et al.* Glycosylation Tunes
660 Neuroserpin Physiological and Pathological Properties. *Int J Mol Sci* 2020 May 3; **21**(9).

661

662

663

664 **Figure legends**

665

666 **Figure 1. Sequence and spatial features of predicted N-glycosylation of amyloidogenic and non-**
667 **amyloidogenic clonal κ light chains.**

668 The pie chart shows the percentage of known amyloidogenic (AL) and non-amyloidogenic (Non-AL)
669 clonal κ light chains predicted to be N-glycosylated by NetNGlyc (in **A** from published literature
670 and in **B** from this study). Numbers (N) of unglycosylated (Unglyc) and N-glycosylated (N-glyc)
671 sequences are indicated. The exact location of the N residue predicted to be glycosylated and the
672 sequon type (NXS/T) are displayed in the heatmaps (each row denotes one sequence/patient). The
673 corresponding germline gene is indicated with a color code. FR: framework region; CDR:
674 complementarity determining region; J: J region.

675

676 **Figure 2. Biochemical and proteomics confirmation of N-glycosylation prediction in**
677 **amyloidogenic κ light chains.**

678 **A** Western blot analysis of urinary kappa light chains from AL patients (Pt.) without (top) or
679 without renal involvement (bottom) at diagnosis or at the time of very good partial response to
680 therapy (VGPR), before (-) or after (+) digestion with protein N-glycosidase F (PNGase F). Red,
681 yellow and blue lines indicate molecular weight markers of 25 kDa, 20 kDa, and 15 kDa,
682 respectively. Arrow indicates >25 kDa band. Predicted N-glycosylation status (YES/NO) according
683 to NetNGlyc is indicated. The heatmap below blot images indicates (in shades of blue) the
684 differential light chain (dFLC) concentration (in milligrams, mg, per liter, L) at the time of urine
685 collection. **B** Liquid chromatography and tandem mass spectrometry in urine from patient 24, with
686 a clonal amyloidogenic κ light chain of the *IGKV2-28* gene, before (-) or after (+) digestion with
687 protein PNGase F. Physical distribution and absolute number (in shades of grey) of peptide ions

mapping against the clonal light chain sequence (protein sequence predicted based on the nucleotide sequence). Δ indicates a deamidated asparagine residue identified within a given peptide at the indicated position.

Figure 3. Genomic versus somatic origin of the predicted N-glycosylation site in amyloidogenic κ light chains.

A, left: For clonal amyloidogenic κ light chain sequences predicted to be N-glycosylated, the nucleotide sequence of the three codons resulting in the NXS/T site and the resulting amino acid residues (one-letter code) (Mutated sequence) are compared with the corresponding germline gene (Germline sequence). Nucleotides in orange denote mutations in the clonal light chain gene with respect to the corresponding germline gene which led to the generation of an NXS/T site (also in orange). Nucleotides in bold or in italic denote additional non-synonymous or synonymous mutations, respectively. The bar graph denotes the number of AL patients/sequences sharing an identical sequence at the same location (occurrence). The location of the N-glycosylation site within the variable region and the germline genes are indicated with a color code. **A, right:** For unique nucleotide substitutions leading to the acquisition of an N residue in the context of an NXS/T sequon within clonal amyloidogenic κ light chains, the SNP ID (according to dbSNP/GnomAD, NA if not annotated), the nucleotide change (denoted as Germ > Mut) and the position on the chromosome (Chr.2) are indicated. The bar graph denotes the number of AL patients with a nucleotide substitution corresponding to a specific SNP (occurrence). Germline genes are indicated with a color code. For SNPs annotated in GnomAD, figures on the right of the bar graph denote the number of mutant alleles out of the total number of sequenced alleles in GnomAD. **B** Sequence alignment of part of *IGKV4-1* gene from patient (Pt.) 73 as assessed with Sanger sequencing (Sanger seq.) on genomic DNA (gDNA) from peripheral white blood cells (top

712 sequence), and from the clonal κ light chain (bottom sequence). Letters in bold indicate the region
713 encompassing the NFT site. The nucleotide in orange denotes mutation with respect to the
714 germline genes which led to the generation of the NFT site (also in orange).

715

716 **Figure 4. Secondary structure analysis and structural mapping of N-glycosylation sites in**
717 **amyloidogenic κ light chains.**

718 Three-dimensional structure of model κ light chain 1B6D. Single variable domain (left) and full light
719 chain dimer (right) are displayed in cartoon representation. The portion of FR3 interested by N-
720 glycosylation is highlighted in orange, and the most frequent N-glycosylation site is marked by the
721 asterisk. N-term and C-term indicate the N- and C- termini of the variable domain. V_L and C_L
722 denote the variable and the constant domain of each monomer of the light chain, respectively.
723 Only beta strands D, E, and B are labelled on the left for simplicity. The arrow refers to the 3-axis
724 rotation (x=-15°, y=50°, z=15°) applied to go from one view to the other. Structural representation
725 was produced using Pymol (Molecular Graphics System, Version 2.4.1, Schrödinger, LLC).

726

727 **Figure 5. Prognostic significance of N-glycosylation site mapping within κ light chains**

728 Matthew Correlation Coefficient (MCC) considering the presence of N-glycosylation at any position
729 (Any N-glycosylation) or the presence of an N-glycosylation site-specifically at the D or E strand of
730 FR3 (FR3-DE N-Glycosylation) as a risk factor for amyloidogenicity of clonal κ light chains. N-
731 glycosylation prediction is performed on known amyloidogenic (AL) and non-amyloidogenic (non-
732 AL) clonal κ light chains based on NetNGlyc. True Positives (TP) were defined as AL κ sequences for
733 which a N-glycosylation site was predicted (in any region or specifically in the FR3-DE region), False
734 Positives (FP) as non-AL sequences with a predicted N-glycosylation site, True Negatives (TN) and

735 False negatives (FN) as non-AL and AL sequences without predicted N-glycosylation site
736 respectively.
737

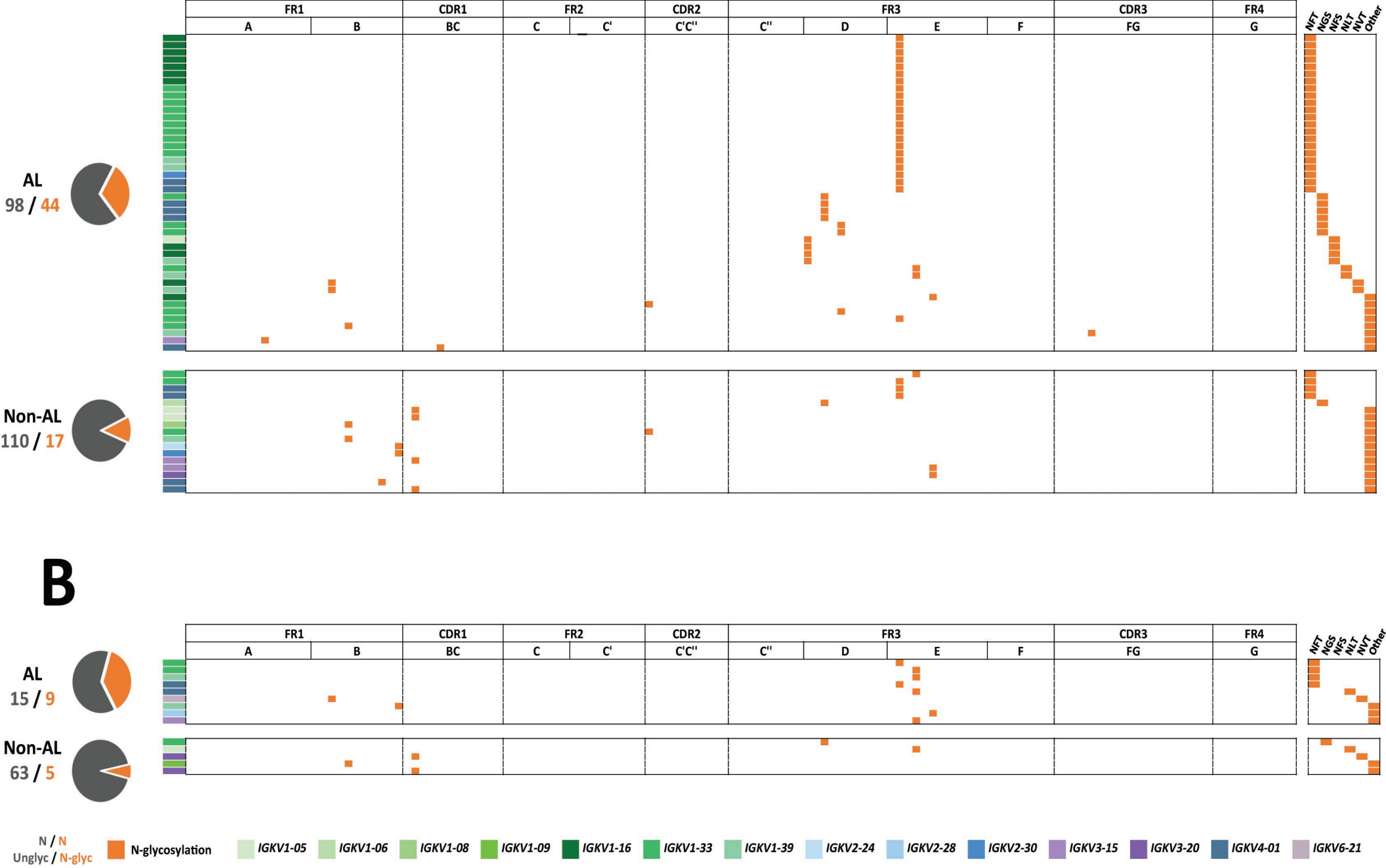

FR1
CDR1
FR2
CDR2
FR3
CDR3
FR4
A
B
BC
C
C'
C'C''
C''
D
E
F
FG
G
NFT
NGS
NFS
NLT
NVT
Other
AL
98 / 44
Non-AL
110 / 17
B
AL
15 / 9
Non-AL
63 / 5
N / N
Unglyc / N-glyc
N-glycosylation
IGKV1-05
IGKV1-06
IGKV1-08
IGKV1-09
IGKV1-16
IGKV1-33
IGKV1-39
IGKV2-24
IGKV2-28
IGKV2-30
IGKV3-15
IGKV3-20
IGKV4-01
IGKV6-21

Figure 1

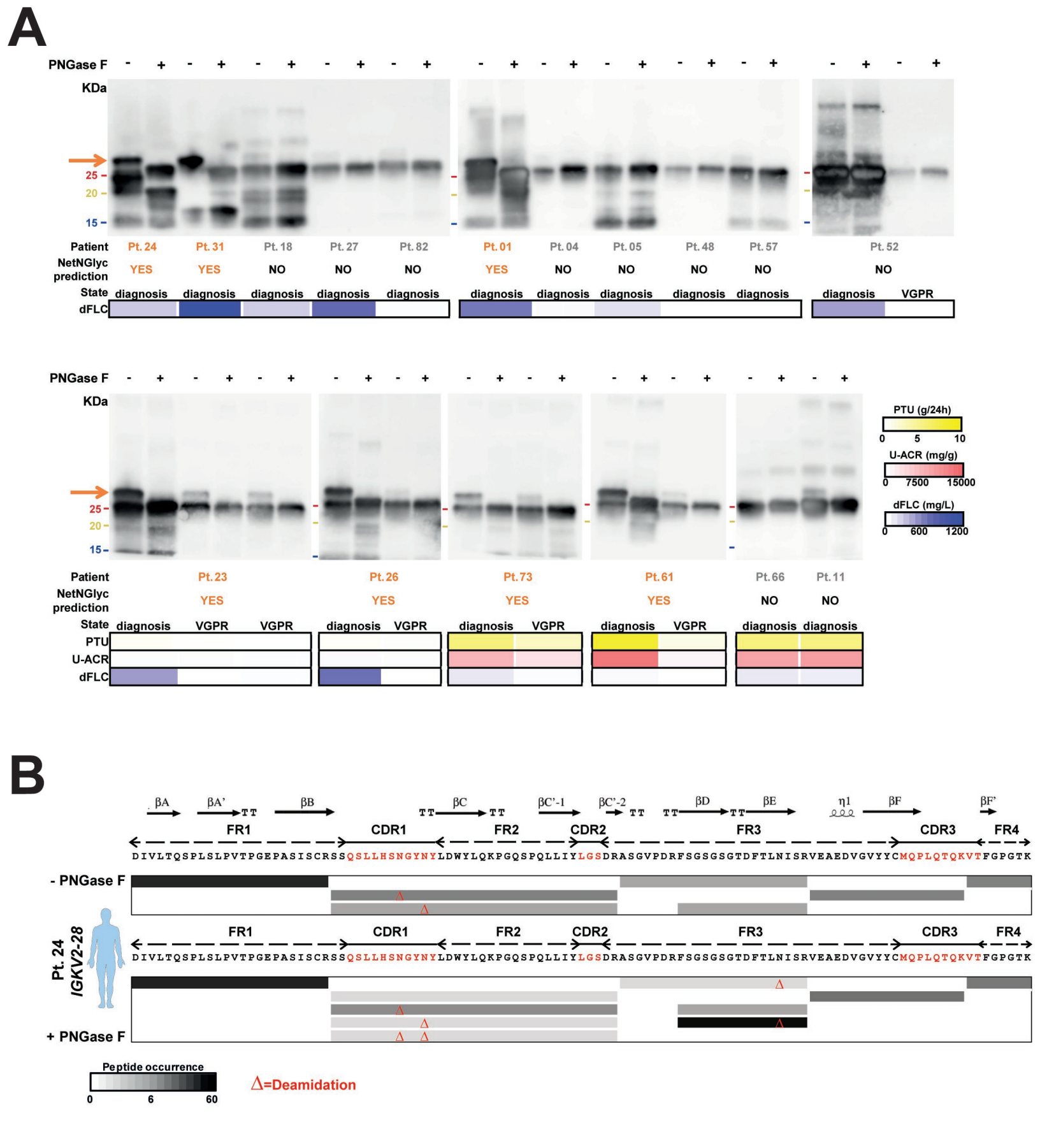

A
PNGase F
KDa
25
20
15
Patient
NetNGlyc prediction
State
dFLC
Pt. 24
Pt. 31
Pt. 18
Pt. 27
Pt. 82
Pt. 01
Pt. 04
Pt. 05
Pt. 48
Pt. 57
Pt. 52
YES
NO
diagnosis
VGPR
Pt. 23
Pt. 26
Pt. 73
Pt. 61
Pt. 66
Pt. 11
PTU
U-ACR
PTU (g/24h)
0 5 10
U-ACR (mg/g)
0 7500 15000
dFLC (mg/L)
0 600 1200
B
βA
βA'
βB
βC
βC'-1
βC'-2
βD
βE
η1
βF
βF'
FR1
CDR1
FR2
CDR2
FR3
CDR3
FR4
DIVLTQSPLSLPVTPGEPASISCRSSQSLLHSNGYNYLDWYLQKPGQSPQLLIYLGSDRASGVPDRFSGSGSGTDFTLNISRVEAEDVGVYYCMQPLQTQKVTFGPGTK
- PNGase F
+ PNGase F
Pt. 24
IGKV2-28
Peptide occurrence
0 6 60
Δ=Deamidation

Figure 2

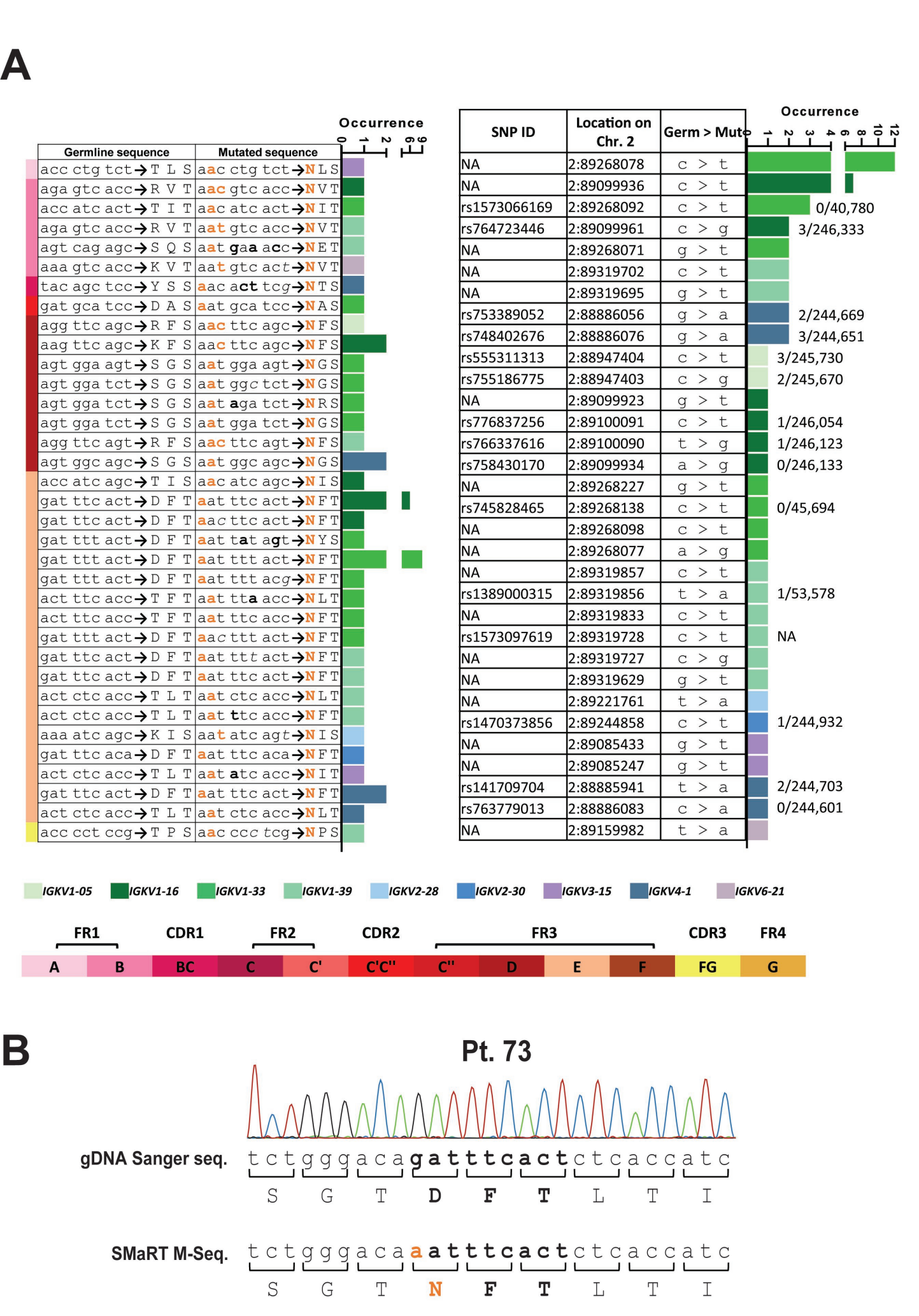

A
Occurrence
0 1 2 6 9
Germline sequence
Mutated sequence
acc ctg tct→T L S
aac ctg tct→N L S
aga gtc acc→R V T
aac gtc acc→N V T
acc atc act→T I T
aac atc act→N I T
aga gtc acc→R V T
aat gtc acc→N V T
agt cag agc→S Q S
aat gaa acc→N E T
aaa gtc acc→K V T
aat gtc act→N V T
tac agc tcc→Y S S
aac act tcg→N T S
gat gca tcc→D A S
aat gca tcc→N A S
agg ttc agc→R F S
aac ttc agc→N F S
aag ttc agc→K F S
aac ttc agc→N F S
agt gga agt→S G S
aat gga agt→N G S
agt gga tct→S G S
aat ggc tct→N G S
agt gga tct→S G S
aat aga tct→N R S
agt gga tct→S G S
aat gga tct→N G S
agg ttc agt→R F S
aac ttc agt→N F S
agt ggc agc→S G S
aat ggc agc→N G S
acc atc agc→T I S
aac atc agc→N I S
gat ttc act→D F T
aat ttc act→N F T
gat ttc act→D F T
aac ttc act→N F T
gat ttt act→D F T
aat tat agt→N Y S
gat ttt act→D F T
aat ttt act→N F T
gat ttt act→D F T
aat ttt acg→N F T
act ttc acc→T F T
aat tta acc→N L T
act ttc acc→T F T
aat ttc acc→N F T
gat ttt act→D F T
aac ttt act→N F T
gat ttc act→D F T
aat ttt act→N F T
gat ttc act→D F T
aat ttc act→N F T
act ctc acc→T L T
aat ctc acc→N L T
act ctc acc→T L T
aat ttc acc→N F T
aaa atc agc→K I S
aat atc agt→N I S
gat ttc aca→D F T
aat ttc aca→N F T
act ctc acc→T L T
aat atc acc→N I T
gat ttc act→D F T
aat ttc act→N F T
act ctc acc→T L T
aat ctc acc→N L T
acc cct ccg→T P S
aac ccc tcg→N P S
SNP ID
Location on Chr. 2
Germ > Mut
Occurrence
0 1 2 3 4 6 8 10 12
NA 2:89268078 c > t
NA 2:89099936 c > t
rs1573066169 2:89268092 c > t 0/40,780
rs764723446 2:89099961 c > g 3/246,333
NA 2:89268071 g > t
NA 2:89319702 c > t
NA 2:89319695 g > t
rs753389052 2:88886056 g > a 2/244,669
rs748402676 2:88886076 g > a 3/244,651
rs555311313 2:88947404 c > t 3/245,730
rs755186775 2:88947403 c > g 2/245,670
NA 2:89099923 g > t
rs776837256 2:89100091 c > t 1/246,054
rs766337616 2:89100090 t > g 1/246,123
rs758430170 2:89099934 a > g 0/246,133
NA 2:89268227 g > t
rs745828465 2:89268138 c > t 0/45,694
NA 2:89268098 c > t
NA 2:89268077 a > g
NA 2:89319857 c > t
rs1389000315 2:89319856 t > a 1/53,578
NA 2:89319833 c > t
rs1573097619 2:89319728 c > t NA
NA 2:89319727 c > g
NA 2:89319629 g > t
NA 2:89221761 t > a
rs1470373856 2:89244858 c > t 1/244,932
NA 2:89085433 g > t
NA 2:89085247 g > t
rs141709704 2:88885941 t > a 2/244,703
rs763779013 2:88886083 c > a 0/244,601
NA 2:89159982 t > a
IGKV1-05
IGKV1-16
IGKV1-33
IGKV1-39
IGKV2-28
IGKV2-30
IGKV3-15
IGKV4-1
IGKV6-21
FR1
CDR1
FR2
CDR2
FR3
CDR3
FR4
A
B
BC
C
C'
C'C''
C''
D
E
F
FG
G
B
Pt. 73
gDNA Sanger seq.
t c t g g g a c a g a t t t c a c t c t c a c c a t c
S G T D F T L T I
SMaRT M-Seq.
t c t g g g a c a a t t t c a c t c t c a c c a t c
S G T N F T L T I

Figure 3

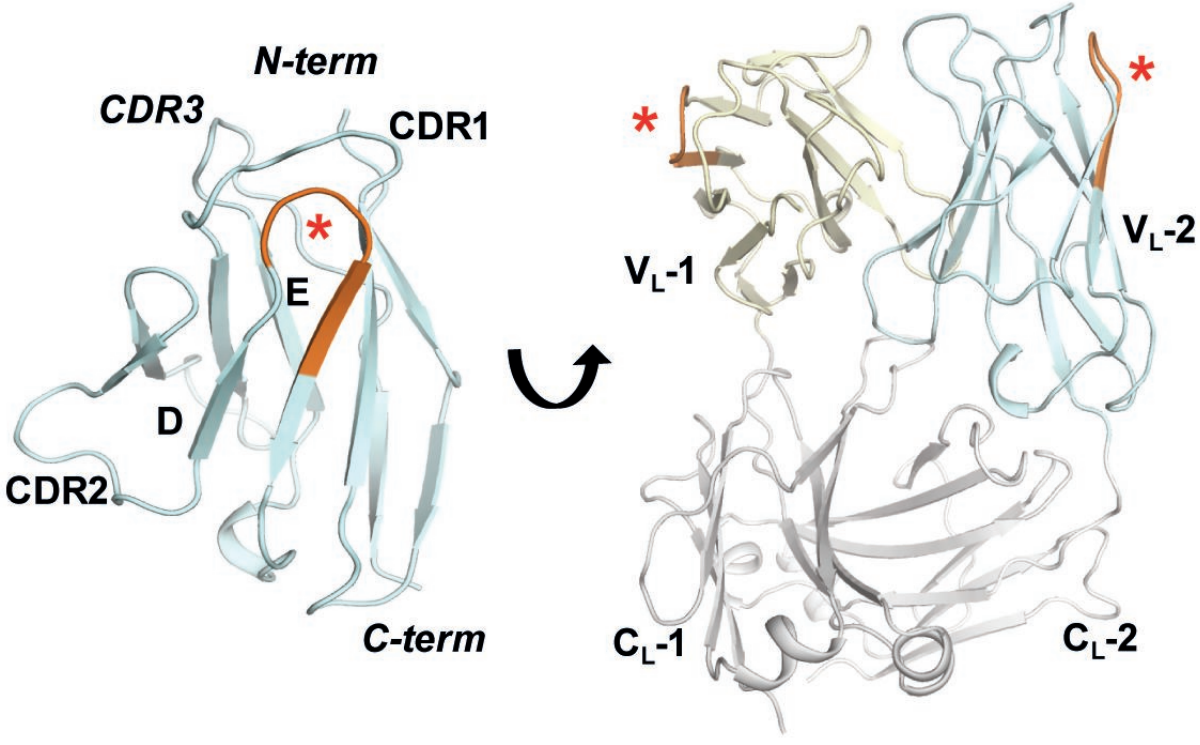
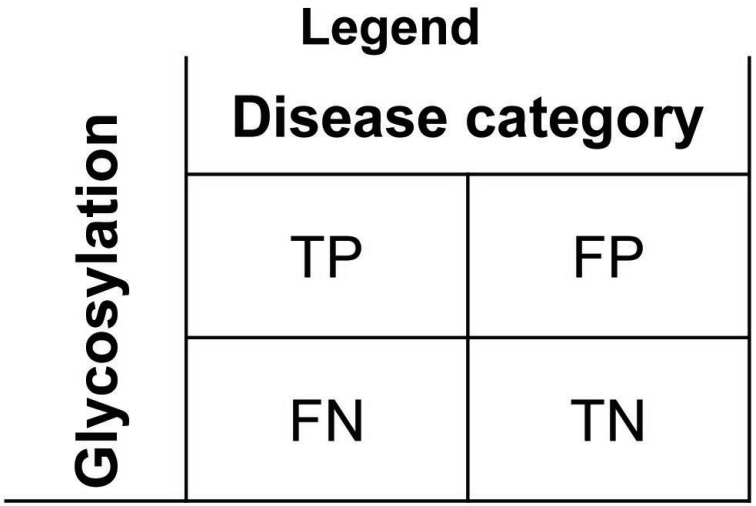
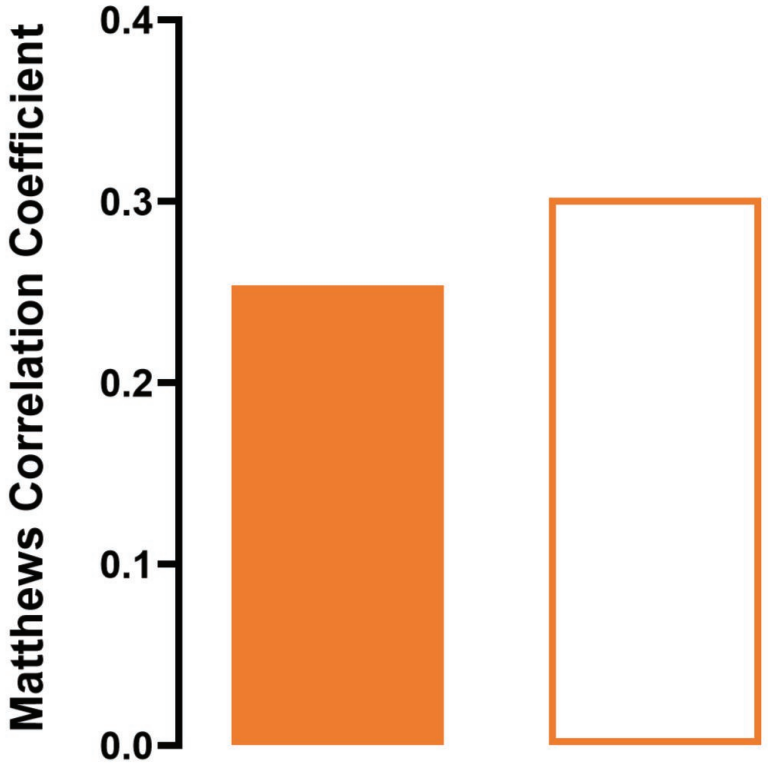


Figure 4





Any N-glycosylation

	AL	Non-AL
Yes	53	22
No	113	173

FR3-DE N-glycosylation

	AL	Non-AL
Yes	43	9
No	123	186

Figure 5