UNIVERSITÀ DEGLI STUDI DI TRIESTE

XXXII CICLO DEL DOTTORATO DI RICERCA IN

Biomedicina Molecolare

TITOLO DELLA TESI

Characterization of anti-transglutaminase 2 (TG2) antibodies from Celiac patients by an integrated experimental and computational approach

Settore scientifico-disciplinare: **BIO/13**

DOTTORANDO
Emanuela Rizzo

COORDINATORE
Prof.ssa Germana Meroni

SUPERVISORE DI TESI
Prof. Daniele Sblattero

ANNO ACCADEMICO 2018/2019

Abstract

Celiac disease (CD) is a gluten sensitive enteropathy. It is characterized by small-intestinal mucosal injury and nutrient malabsorption produced by dietary exposure to wheat gluten and similar proteins contained in rye and barley. Approximately 1% of the population is affected by CD, even if most of the affected individuals remain undiagnosed. CD is characterized by the production, upon gluten ingestion, of high affinity IgA and IgG antibodies (Abs) directed against gliadin, a gluten component, and the autoantigen Tissue Transglutaminase 2 (TG2). Reactivity against TG2 is an important hallmark of CD, since nearly all CD patients develop autoantibodies against TG2. Accumulation of TG2-specific IgA plasma cells occurs in the small-intestinal mucosal lesion of untreated CD patients, and these B cell clones also contribute to the generation of serum IgA anti TG2.

Given the clinical-pathological meaning of anti-TG2 Abs in CD, their characteristics have been investigated. Anti-TG2 Abs are widely studied and well characterized:

- They show a limited number of somatic hypermutations.
- Naïve antibodies show high affinity.
- The gene usage is restricted to certain gene families and to a preferred VH-VL pairing, mostly belonging to IGHV5-51:IGKV1-5 gene fragments.
- They mostly recognize 3 TG2 epitopes and most of them, characterized by the presence of IGHV5-51 gene segment, bind TG2 on the so-called Epitope1.

Here, we propose to deeply investigate the structural determinants of anti-TG2 Epitope1 Abs, which have not been elucidated yet.

Phage display is a powerful tool to study Abs. Therefor phage display libraries of single-chain variable fragments (scFvs) obtained from the Ab repertoire derived from intestinal biopsy lymphocytes are used to study the immunological response in CD. Phage display libraries permit the analysis and screening of a large number of antigen-Ab interactions, and also the easy recovery of the scFv gene sequence thanks to bacteriophages characteristic.

Since the *IGHV5* gene is the most represented in anti-TG2 Abs, libraries characterized by this specific *IGHV5* have been generated. The libraries are composed by IGHV derived from 55 CD patient's lymphocyte paired with 4 VLs from anti TG2-Abs from celiac disease patients previously isolated.

ScFv libraries have been selected for TG2 binding and 102 reactive clones have been bioinformatically analysed.

Computational methods are considered to be a way to generate *in silico* hypotheses, helping to interpret and guide experiments. In the 1980s, Ab modelling started with the unexpected discovery that most of the complementarity determining regions (CDRs) loops adopt a limited number of conformations. Since a limited variability has been observed in the length and composition of loops L1-3 and H1-2, we focused on the length and composition of loop H3, usually playing a key role in the antigen recognition. The CDRH3s of the 102 TG2 reactive clones are characterized by a preferential length of 14 amino acids (aas) and 4 aas are conserved, hence called "*key residues*". The design of a "*consensus*" VH CDRH3 sequences was made based on Hidden Markov Models (HMMs) by HMMER. The "*consensus*" CDRH3 is 14aa long and has the 4 "*key residues*". The ScFv with the *in silico* designed "*consensus*" CDRH3 was generated using the backbone of a well characterized anti-TG2 Ab. Ab 3D models have been built with Modeller using the structure of the reference Ab as a template. The so generated scFv was tested *in vitro* as scFv displayed on phage surface, and also as scFv-Fc (fragment constant), IgG and IgA format, to confirm the reactivity in the physiological structure.

Moreover, mutants carrying random aas in the CDRH3 sequence (characterized by the *"key residues"*) were generated and analysed to deeply investigate the role of the CDRH3 aas in the interaction with TG2. To elucidate the role of single CDRH3 aas, CDRH3s with point mutations have been designed. Point mutants were generated, firstly as scFvs displayed on phages, and secondly as scFv-Fcs, to finally investigate the affinity of the different mutants.

Our results suggest that phage display technology is an excellent tool to study Abs specific for an immune disease. In fact, the scFvs displayed on phages surface have shown to preserve the characteristics of the Abs expressed in the patients and no bias has been identified, confirming the goodness of our samples and technology. Moreover, a "*consensus*" sequence has been generated *in silico*. Amazingly, the ideal sequence was not present neither in the panel of anti-TG2 Abs nor in TG2 positive clones identified by other research groups. The "*ideal"* anti-TG2 Abs preserve specificity for TG2 Epitope 1, and the reactivity was incredibly maintained even in more physiological formats, as IgG and IgA. *In silico* and *in vitro* technology allowed the rapid analyses and generation of recombinant Abs, which facilitates the deep investigation of the role of aas in CDRH3 and the generation of Abs with a very high affinity.

Table of contents

1.Introduction

1.1 Celiac disease (CD)

The first clear description of CD was given by Samuel Gee in 1888. He suggested that dietary might be the treatment for this disease. During the next century, various diets were tried, however without identifying the component that triggers the disease. In 1909, Herter suggested that fats were better tolerated than carbohydrates, and in 1921 Howland identified the intolerance to carbohydrates. In 1950 Wim Dicke, during his doctoral thesis, established that the exclusion of wheat, rye and oats from the diet led to dramatic improvement in the general condition of the child[1]. This observation was later confirmed by Dicke's colleagues[2]. Up to now, we can define the CD as a gluten sensitive enteropathy. It is characterized by small-intestinal mucosal injury and nutrient malabsorption produced by dietary exposure to wheat gluten storage proteins and similar protein contained in rye and barley[3]. In addition to gluten ingestion, patients need genetic predisposition to develop CD. Genes necessary to trigger the CD are the human leukocyte antigen (HLA) HLA/DQ2 and/or HLA/DQ8 haplotypes[4]. Following the ingestion of gluten by CD patients, anti-gliadin antibodies (AGAs) and anti-tissue transglutaminase 2 (anti-TG2) antibodies are generated and released in the serum. Those antibodies (Abs) become serological indicators of CD[5]. After the specific Abs production, CD patients display various degree of intestinal inflammation. In fact, gluten ingestion triggers a mucosal immune response. This results from an increased intraepithelial lymphocyte to a severe subepithelial mononuclear cell infiltration that leads to villous atrophy and crypt hyperplasia[6]. Moreover, CD patients may be affected by extra intestinal symptoms or, on the other hand, remain asymptomatic[7,8]. CD is widespread since it affects approximately 1% of the population, and it is considered to be common in Europe, the US, Australia, Mexico and South American countries [9]. Up to now, the only efficient treatment for CD is the strict adherence to a gluten-free diet[10].

1.1.1 Prevalence and incidence

Until the 1990s, CD was considered a gastrointestinal disorder mainly affecting children[11]. More precisely, CD may be diagnosed at any age, but presents itself typically in infancy (between 9 and

24 months) or between the third or fourth decade of life[12,13]. Moreover, CD is more common in females than in males, with a ratio of 3:1[14].

Affecting 1% of the world population, CD is a public health problem[9,15], and is therefore a highly studied disease. However, the information is not equally provided from all countries. Most of the screening studies have been performed in Europe and show variation from country to country. In Europe, high seroprevalence of CD is found in Sweden, Finland, Turkey, the United Kingdom, Italy, the Czech Republic and Portugal, whereas in Russia, Poland and Switzerland, CD is less common[16]. Similar studies were performed even in other countries, and data showed that the prevalence values for celiac disease were 0.4% in South America, 0,5% in Africa and North America, 0,6% in Asia and 0,8% in Oceania (this last one is comparable to Europe)[15]. Moreover, a high prevalence of celiac disease has also been reported from India and some countries in middle-eastern Asia and Africa[17,18].

The celiac society was founded in the United Kingdom in 1968, and other celiac organizations were set up in other counties progressively, such as in California and Canada. In 1979, the AIC (Associazione Italiana Celiachia) was founded, followed 9 years later by the Association of European Celiac Societies.

Although diagnoses of CD are increasing, the disorder remains heavily undiagnosed.

1.1.2 Factors that triggers the gluten sensitivity

There is not a specific factor which can trigger the raising of CD, but the observation of various circumstances lead to identification of some factors that trigger the development of CD. Indeed, the concurrence of CD in 75% of monozygotic twins suggested a tendency of familiarity in CD and led to study the genetic factors[1] (Fig. 1). In 1972, Falchuk and colleagues, and Stokes and colleagues studied a histocompatibility antigen associated with the pathology[19,20]. It is now known that the development of the disease requires both the ingestion of gluten and genetic predisposition (Fig. 1). The genetic predisposition has been showed not only between twins, but also in the first-degree relatives of CD patients. Indeed, relatives of patients exceed the predisposition of the general population, with the percentage growing to 8%[21]. Human leukocyte antigen (HLA)-DQ2 and HLA-DQ8 haplotypes were identified as genetic suspects for CD development[4]. These variants have been estimated to contribute around 25-40% of the genetic risk[22–24]. The presence of HLA-DQ2 or HLA-DQ8 haplotypes is necessary, but luckily not sufficient for the development of the pathology, as the presence of these HLA-DQ2 or HLA-DQ8 is a characteristic of around 40% of the European and North American populations, whereas CD affects 40 times less people[11].

HLA-DQ2 and HLA-DQ8 are dimeric class II major histocompatibility complex molecules expressed on the surface of the antigen-presenting cells (APCs). They consist of an α-chain and a β-chain encoded by specific variants of HLA-DQA1 and HLA-DQB1 genes respectively. More than 90% of CD patients are HLA-DQ2 positive and the rest carry HLA-DQ8[11]. Only 2%-5% of the population carrying the gene develop the CD. This implicates that other genetic as well as environmental factors contribute to the manifestation of the pathology[25,26] (Fig. 1).

Potentially, environmental factors, including consumption of gluten, infection in early years of life, lower economic status and the quality of the hygienic environment[27–29] could be related to the risk factors for the development of CD (Fig. 1).

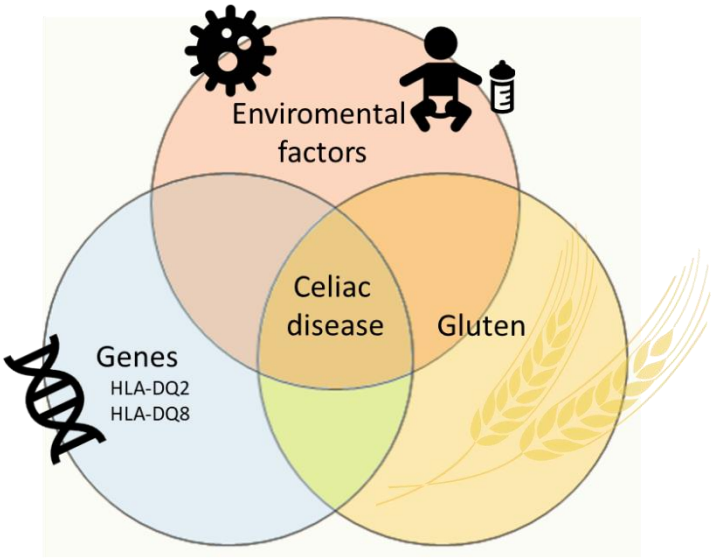


Figure 1. ***Factors that trigger the development of celiac disease.*** *There are different factors that trigger the development of the celiac disease pathology. Environmental factors include gluten ingestion, but also other environmental factors, like microorganisms and the timing of the introduction of gluten in the diet. The presence of HLA-DQ2 and HLA-DQ8 genes is necessary, but not enough for the insurgence of the pathology. Therefore, all these factors contribute for the development of the pathology.*

1.1.2.1 Gluten

The most important environmental factor that triggers the pathology is gluten (Fig. 1). Indeed, the gut lesions of most CD patients disappear when gluten is completely excluded from the diet, and reappear after gluten ingestion[30]. Gluten is one of the few digestion-resistant proteins and it is constituted by several non-digestible immunogenic peptides[31], characteristics which increase the role of gluten in triggering the development of the pathology.

Gluten is an external trigger factor for the development of CD. Gluten proteins are composed by glutenin polymers (subdivided into high and low molecular weight glutenins) and gliadin monomers.

Gluten is a complex storage protein found in different grains, such as wheat, rye and barley. Glutenins compose the alcohol-insoluble fraction, whereas gliadins make up the alcohol soluble part. Gliadins and glutenins are rich in proline and glutamine residues. This composition makes these proteins extremely resistant to proteolytic degradation by gastric and pancreatic enzymes, therefore, gliadins and glutenins are not easily degraded in the gastrointestinal tract[32,33].

1.1.2.2 Genetic

Once gluten is ingested, the gliadin peptides gain access to the intestinal lamina propria[34]. The gliadin peptides are deamidated and are recognized by APC through HLA class II molecules. Deamidated gliadins are taken up by APCs, which present gliadins to gluten-specific $CD4^+$ T cells, allowing the activation of an immune response mediated by $CD4^+$ T cell[9] (Fig. 2).

HLA is a gene complex encoding the human major histocompatibility complex (MHC), which role is to present peptide antigens to lymphocyte. HLA molecules are composed of two subunits, α-chain and β-chain encoded respectively by specific variants of the *HLA-DQA1* and *HLA-DQB1* genes[11,35]. HLA-DQ2, also called HLA-DQ2.5 heterodimer, is encoded by the *HLA-DQA1*05* and *HLA-DQB1*02*, whereas, HLA-DQ8 is encoded by heterodimer *HLA-DQA1*03* and *HLA-DQB1*03* (these genes are located on chromosome 6p21)[36–38]. The strength of the immunological response is given by the heterodimer HLA-DQ2.5 on the APC. Approximately 90% of patients display this HLA on APC, and most of the rest carry HLA-DQ8[9,34,37,39]. Moreover, it was shown that there is a higher risk of developing the disease if the HLA-DQ2.5 is in homozygosis[40]. Interestingly, HLA-DQ2 has been shown to be associated with early onset of the disease, whereas HLA-DQ8 has been shown to be associated with onset of the CD in adults[41]. As said before, the presence of couple of heterodimers is necessary but not sufficient for developing CD.

Indeed, even among the genetic factors, non-HLA regions have been associated with the development of the CD[23,36,37,39]. Genome-wide association studies have found 18 risk factors in the MHC region aside from the MHC class II genes that are involved in triggering CD. HLA class I genes are involved in the development of CD as well as in other immune-mediated diseases[37]. Interestingly, some non-HLA regions are shared with other autoimmune diseases, such as type 1 diabetes and rheumatoid arthritis[30].

1.1.2.3 Immune mechanism and the generation of autoantibodies

The circulation of antibodies against gluten has been reported in 1985, and three years later the antibodies against other food components were identified. Anti-endomysium IgAs were detected 15 years later, in 1983[1]. In 1997, tissue transglutaminase was identified as the major autoantigen

responsible for the development of CD[5]. Gluten peptides are characterized by amino acids glutamine at key positions which can be deamidated by TG2[42]. These deamidations lead to the modification of glutamine residues in glutamic acid, modifications that increase the binding affinity of gluten peptides to HLA-DQ2 or HLA-DQ8 molecules on the APCs[43]. The gliadin peptides bound to HLA can be presented to CD4+T helper cells[44,45] (Fig. 2).

Gliadin-specific $CD4^+$ T cells recognize the gliadin deamidated peptide presented by HLA through T cell receptors (TCRs). Indeed, TCRs specific for gliadin peptides have been identified in CD patients[46]. TCRs are generated in random process, and it could be that TCRs specific for gliadin are generated in only a minority of HLA-DQ2 and HLA-DQ8 positive individuals, which could be an explanation of why only a percentage of the population carrying HLA-DQ2 and HLA-DQ8 develop the pathology[46]. Gluten-specific $CD4^+$ T cells could secrete cytokines after activation, like IFNγ and IL-21[47], leading to an inflammation of the small intestinal lamina propria area which causes tissue damage (Fig. 2). This process is the link between adaptative and innate immunity, and triggers the increase of mucosal expression of IL-15,IL-18 and type I interferons[11].

Moreover, gluten-specific $CD4^+$ cells are also involved in the generation of antibody responses which is characteristic of CD. After the interaction of APCs with gliadin specific $CD4^+$ cells, $CD4^+$ cells help anti-TG2 and anti-gliadin B cells promoting their activation and differentiation into plasma cells secreting antibodies against both deamidated gliadin peptides and TG2[48]. The production of anti-TG2 antibodies in the presence of gluten suggests that gliadin specific T cells are involved in the activation of TG2-specific B cells, probably thanks to the TG2-gliadin complex[49,50] (Fig.2).

Both anti-TG2 and anti-gliadin Abs can be detected in the circulation of CD patients. Moreover, anti-TG2 antibodies are present in the small intestinal mucosa[51,52]. Previous data suggested that Anti-TG2 Abs both circulating and in the intestinal mucosa were probably produced by plasma cells in the small intestinal mucosa. However, more recent data suggest that anti-TG2 antibodies in the serum are secreted by plasma cells that are clonally related to intestinal TG2-specific plasma cells but reside outside the gut[53,54]. Both anti-gliadin and anti-TG2 antibodies probably are involved in the pathogeneses of CD, increasing the permeability of the epithelial barrier, allowing the gliadin peptides to cross the lamina propria[55].

Moreover, there are other autoantibodies against other transglutaminase members family involved in other autoimmune disease, like anti-TG3 and anti-TG6 Abs, involved respectively in dermatitis herpetiformis and gluten ataxia, probably involved in the generation of the extraintestinal manifestation of CD[55].

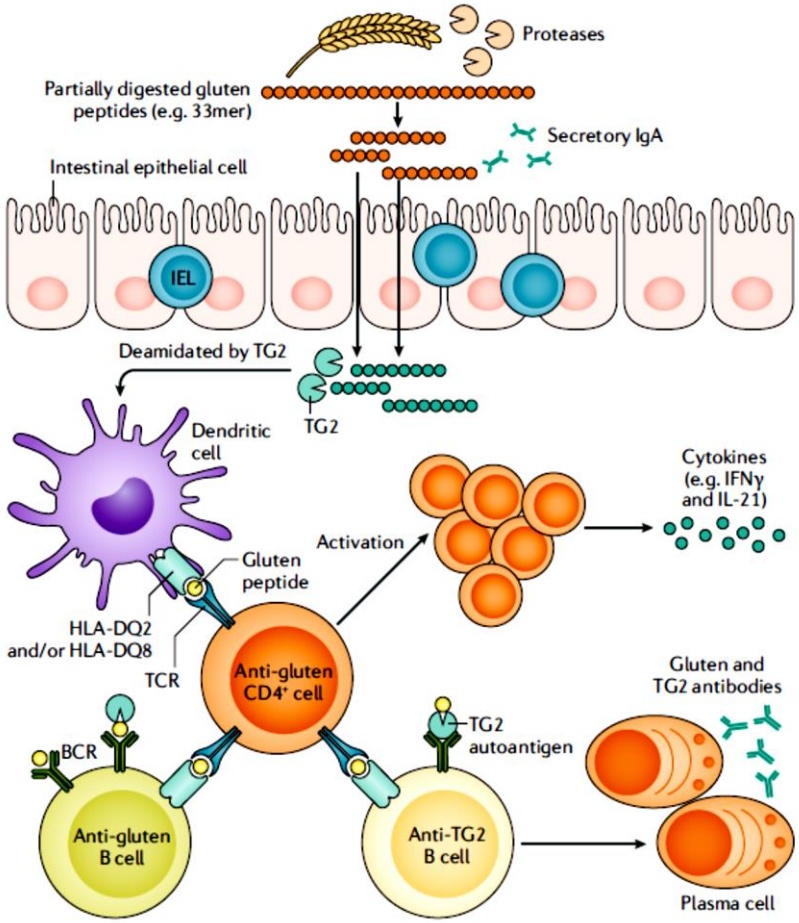


Figure 2. ***Adaptative immune response involved in celiac disease****. Gluten is rich in proline, a characteristic which confers it a high resistance to proteolytic degradation at the digestion level. In the lamina propria, the gliadin peptides are deamidated by TG2, which modifies glutamine residues into glutamic acid, increasing their affinity to HLA-DQ2 and HLA-DQ8. Deamidated gliadins are taken up by APCs, including dendritic cells, which present gliadins to gluten-specific CD4⁺ T cells.*

Both gluten-specific and TG2-specific B cells have been suggested to act as antigen-presenting cells in CD. B cells recognize their antigens (gliadin peptides and TG2-gliadin complexes) through surface B cell receptors (BCRs), internalize them and present gliadin to gluten specific CD4⁺ cells. After the interaction of HLA-DQ2 or HLA-DQ8 with gliadin peptides and TCRs, both T cells and B cells would be activated. Activated CD4⁺ T cells secrete inflammatory cytokines, including IFNγ and IL-21, generating an inflammatory environment in the small intestinal lamina propria. Once activated, the B cells can differentiate into plasma cells that secrete antibodies against gluten and TG2. (Edited from Lindfors et al., 2019)

1.1.2.4 Other factors

The gluten is the main triggering factor. However, like genetic factors, it is not sufficient for the development of CD. Indeed, other socio-economic conditions were suggested to be responsible for triggering the pathology[27]. Furthermore, microorganisms are suggested to play both a pathogenic role and protective role in the development of CD: several studies have shown an association between CD and change in the microbiome composition. In this regard, *Clostridium*, *Prevotella* and *Actynomyces* are found in CD patients whereas *Helicobacter pylori* might protect individuals from the development of the pathology, even if in an undefined manner[11].

Early life feeding practice was suggested to be involved in the development of CD. However there are opposite results in different studies, so there is no conclusive answer for the age of introduction of gluten in the diet, and other studies are needed to clarify this point[11].

1.1.3 Diagnosis

CD is an autoimmune disease associated with gastrointestinal symptoms. It was thought to be associated to childhood diseases, whereas, now it is known that CD can occur in the first two years of life or in the second or third decades[11,31]. In paediatric patients under three years old, CD can appear with diarrhea, loss of appetite, abdominal pain and distension and failure to grow up. In older children and adults the onset of CD is characterized by diarrhea, bloating, constipation, abdominal pain and weight loss[31]. Furthermore, CD is characterized by non-gastrointestinal findings such as dermatitis herpetiformis, chronic fatigue, joint pain or inflammations, iron deficiency anaemia, migraines, depression, attention-deficient disorder, epilepsy, osteoporosis, dental enamel defects, and other autoimmune disorders even including type 1 diabetes mellitus[9]. Nevertheless, CD may be asymptomatic and patients can be found only by screening in risk groups[21].

The diagnosis of CD is established after positive celiac serological tests to identify antibodies specific to the outbreak of CD, intestinal biopsy to evaluate the intestinal mucosa morphology, and genetic HLA predisposition. Serological tests include EmAs, antibodies specific to TG2 in the endomysium (a form of perivascular connective tissue), anti-TG2 and anti-deamidate gliadin antibodies. The EmA test is based on an indirect immunofluorescence, whereas both anti-TG2 antibodies and anti-deamidate gliadin antibodies are revealed by enzyme-linked immunosorbent assay ELISA. In all these tests IgAs are revealed. However, some autoimmune diseases, including CD, are characterized by a deficient IgAs production. Indeed, CD patients show IgAs deficiency ten times more commonly than the general population[1] when the total IgAs level is analysed [11,56]. Afterwards, if patients result

seronegative, the diagnoses are based on the detection of small intestinal mucosal damages. The diagnoses are carried out in a gluten containing diet. Adults diagnosis require both seropositivity and the identification of small-mucosa lesions in patient *HLA-DQA1* and *HLA-DQB1*[9]. For paediatric patients, the paediatric guidelines from Europe present a diagnosis without biopsy. Indeed if the paediatric patient shows anti-TG2 IgAs 10 times the cut off, detectable EmA in HLA-DQ2/HLA-DQ8 positivity, and symptoms suggestive of CD, duodenal biopsy cab be avoided[57].

Up to now, the only efficient treatment is a strictly gluten free diet, which patients must start the diet only after the diagnosis of CD. Any food with a gluten amount <20mg/kg is considered gluten free[58].

1.2 Transglutaminase 2

Transglutaminases are a family of enzymes which catalyze post-translation modification of proteins. There are 8 members of the family with enzymatic activity, TGs1-7 and blood coagulation factor XIII[59]. TG2 is a multifunctional protein, and in contrast with the other members of the family, it is ubiquitously expressed. Furthermore, it is present both in the intracellular and in the extracellular compartment, despite lucking of a secretory leader. It is also localized in the nucleus. TG2 is involved in different physiological pathways and its activity is regulated by cofactors. Moreover, TG2 has been found to be down or over regulated in some different diseases[55,60].

Human TG2 belongs to the transglutaminase family. It is a monomer composed by 686 amino acids (the molecular weight is 78KDa), and encoded by the *TGM2* gene[61]. Recombinant human TG2 can be expressed in different organisms; Sf9 insect cells were used at first, followed by *E. coli* cells for the production of the recombinant TG2. Analyses of hydrogen/deuterium exchange by mass spectrometry indicate that the expression system does not affect the folding of the enzyme[62] (see chapter 4.3 Epitope on TG2 for more details). As all the other members of the transglutaminase family, TG2 is composed by four domains; at the amino-terminal (N-term) there is a β-sandwich domain containing the fibronectin and integrin binding site, followed by Catalytic core domain which contains the regulatory site, and two β-barrel domains at the C-term (β-barrel1 andβ-barrel2)[60,63] (Fig. 3A). TG2 is involved in a great number of different pathways, so the activity of TG2 could be fine regulated; indeed the catalytic activity is regulated by Ca^{++}, guanine nucleotides and redox potential[63,64]. In its active site, TG2 binds up to six Ca^{++}[62]. In a physiological condition, where, the intracellular environment is characterized by low presence of Ca^{++} and the presence of guanosine

triphosphate (GTP) or guanosine diphosphate (GDP), the protein is in a closed conformation (Fig 3B left) and the active site is covered, and the protein is inactive. In conditions of stress, the level of Ca^{++} increases and the TG2 changes conformation adopting an aligned or open conformation (Fig)[60,65] (Fig. 3B right). In the open form, the catalytic site, made up of C227, H335 and D358, is exposed[55,60]. Most of the extracellular TG2s have been shown to be inactive in spite of conditions favoring activation. This inactive status is due to the oxidizing extracellular environment which promotes the formation of reversible disulfide bond between two vicinal cysteines[64]. Extracellular TG2 can be activated in response to stimuli[66].

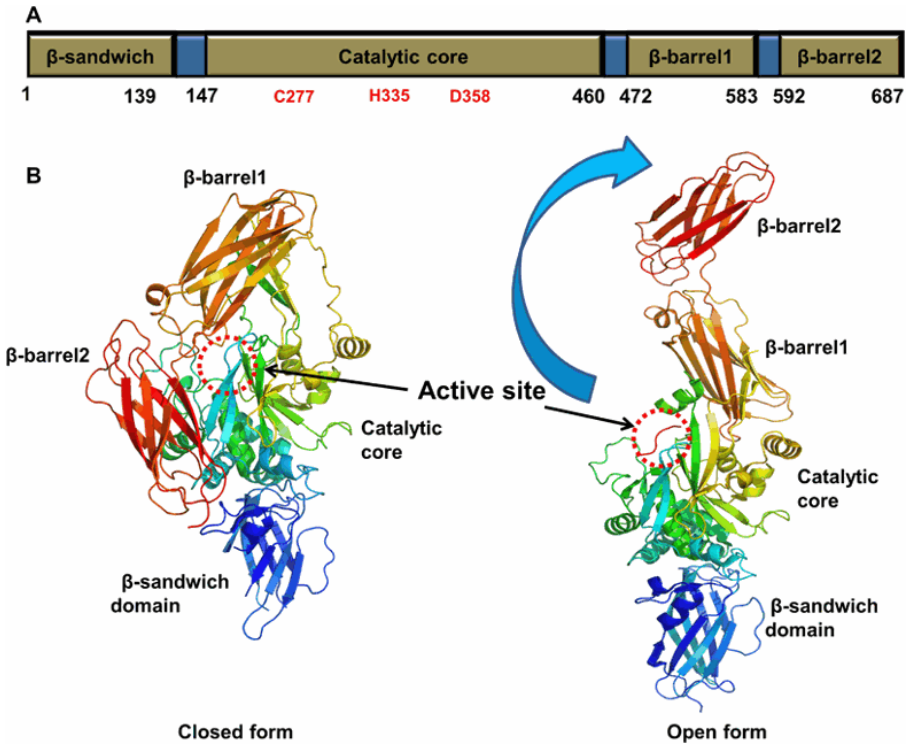


Figure *3*. *Structural domain boundary and dynamic three-dimensional structures of transglutaminase 2 (TG2). A)* *The four distinct domains are indicated by coloured boxes with amino acid positions (below). The catalytic triad comprising C277, H335, and D358, is shown in* red. ***B)****The ribbon structure of the compact, closed form (inactive form) and the expanded, opened form (active form) are shown. A large conformational change upon irreversible inhibitor binding leads to exposure of the active site (marked by the* red-dotted circle*). The chains from the N- to the C-termini are coloured from* blue to red. *(Edited from Lee and Park, 2017)*

1.2.1 Activity

TG2 is involved in the post-translational modification of target proteins by cross-linking, deamidation and amine incorporation, which are all Ca^{++} dependent activities.

TG2 can deamidate the γ-carboxamide group of a specific glutamine residue on a protein during the cross-linking process with an ε-amino group of a lysine residue[67] (Fig. 4). This activity is allowed by the exposition of the cysteine residue in the core domain of the protein. Otherwise, the γ-carboxamide group of a specific glutamine residue is replaced with water, and the glutamine residue is converted into a negatively charged glutamic acid in a deamidation reaction[67] (Fig.4). Moreover, TG2 catalyzes the cleaving of isopeptide bonds[68] (Fig. 4). Deamidation of the gliadin peptides results in their increased binding affinity with HLA complex, which triggers the activation of the immune system. Moreover, TG2 is also able to cross-link gliadin peptides to itself, which allows the cross presentation of both gliadin peptides and TG2s. This event might provide an explanation for the generation of the TG2 autoantibodies characteristic of CD[55].

TG2, as multifunctional protein, is characterized also by Ca^{++} independent activity. Therefore, TG2 acts as G-protein, thus participating in transmembrane signaling[69,70] (Fig. 4). The protein disulfide isomerase (PDI) activity of TG2 has been reported. The TG2 PDI activity is not affected by the level of Ca^{++} and GTP, so that cysteine in the active site for the transaminase activity is not important for this enzymatic activity[60]. On the other hand, the presence of Ca^{++} has been shown to inhibit the kinase activity of TG2[71] which is involved in really important pathways. Indeed, P53, histone H1, H2A, H2B, H3, H3, and Retinoblastoma proteins are those among the substrate which are phosphorylated by TG2[60].

Furthermore, TG2 can also act as a scaffold protein on the extra cellular matrix (ECM) and this activity acts in as a cross-linking activity in an independent manner. TG2 has been suggested to form a ternary complex with fibronectin (Fn) via direct integration with integrins on the cell membrane[72–74]. The cross-link activity of TG2 is well characterized in the presence of ECM turnover[74].

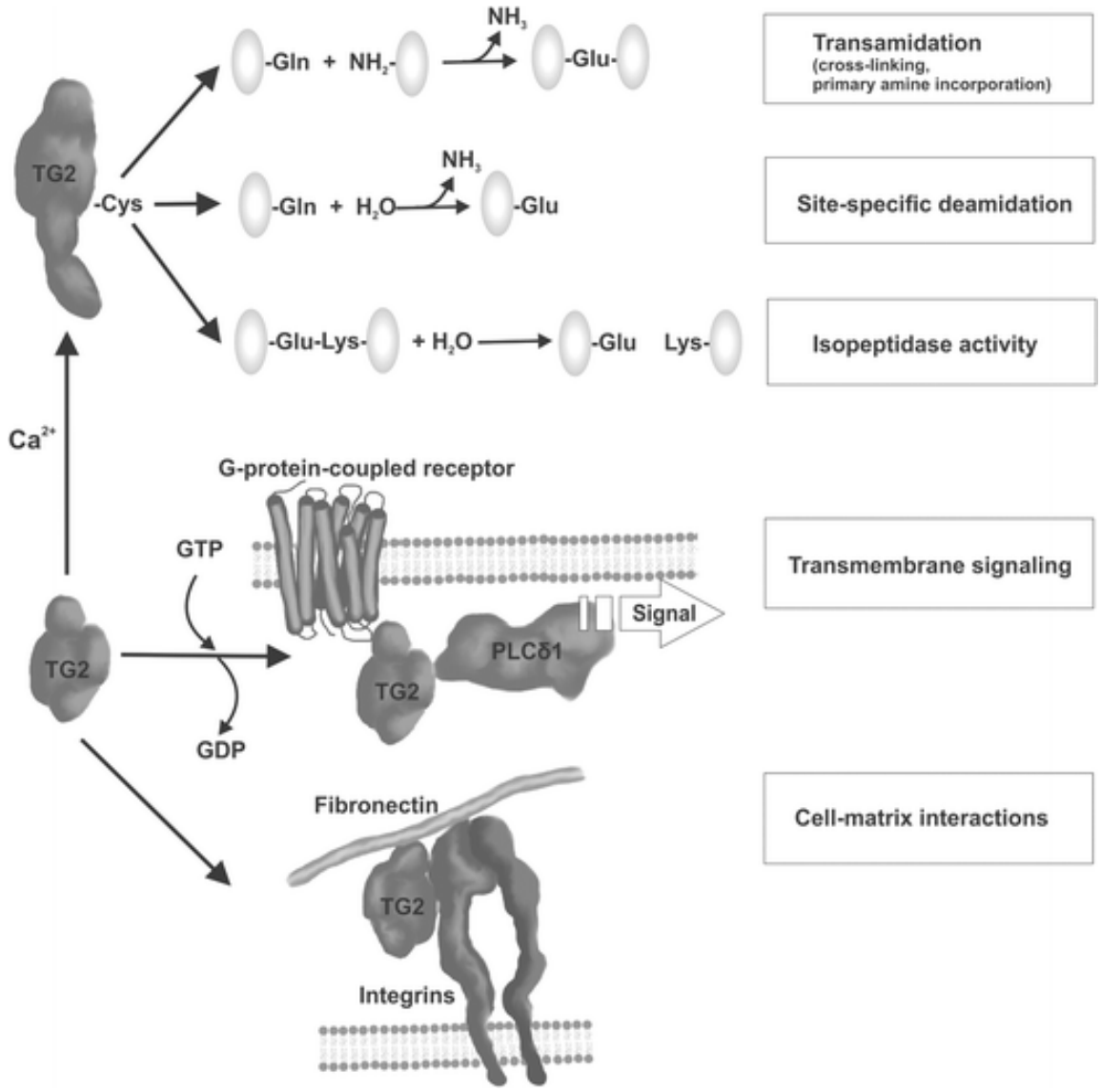


Figure 4. ***Main biochemical functions of transglutaminase 2 (TG2****). TG2 catalyzes Ca^{++}-dependent post-translational modification of proteins by transamination or deamidation of specific polypeptide-bound glutamines. Depending on the substrate, the transamination can lead either to crosslinking of proteins through the generation of ε-(γ-glutamyl) lysine isopeptide bonds or incorporation of small-molecule amines such as polyamines and histamine into proteins. Alternatively, the glutamine substrate can react with water in a TG2-catalyzed deamidation reaction resulting in the conversion of the reactive glutamine residue into a negatively charged glutamic acid. TG2 is also capable of cleaving isopeptide bonds by virtue of its isopeptidase activity. In addition, TG2 possesses GTPase activity allowing it to participate in signal transduction. Moreover, independent of its enzymatic activity in the extracellular environment, TG2 acts as an integrin-binding adhesion coreceptor for fibronectin and thus has a role in cell attachment and spreading. (Edited from Rauhavirta et al., 2019)*

1.2.2 TG2 enzymatic activity in the CD pathogenesis

TG2 has a crucial role in the CD pathogenesis. Up to now, the site of TG2 enzymatic activity remains uncertain. One designated site could be the small intestine. In fact, TG2 has been detected in the epithelial and endothelial cells as well as in the basement membrane[75]. In normal physiological condition, TG2 in the small-intestinal environment is its inactive conformation[66]. Despite the environment, TG2 in the intestinal mucosa can be activated by the activation of the toll-like receptor 3[66]. Moreover, the pH of the environment can affect the type of enzymatic activity of TG2. In the small basement membrane the pH supports the cross-linking activity instead of the deamidation activity[55]. In contrast the pH in the intestinal lumen decreases from pH 7.3 to pH 6.6 a pH at which the TG2 catalyzes the deamidation[76]. It is possible that the deamidation of gliadin peptides could occur in a permissive environment during the contact with intestinal epithelial cells[55,77]. The deamidation catalyzed by TG2 could also occur in the lymphoid tissue[78].

1.2.3 Other pathways involved

As previously described, TG2 is a multifunctional enzyme and the multiple activities of TG2 are involved in various cellular activities, such as apoptosis, proliferation, differentiation, cell adhesion, angiogenesis, stabilization of ECM and bone development. Moreover, beyond the role of TG2 in the development of CD, TG2 is linked to many other human diseases. In fact, TG2 is involved in neurodegeneration, inflammatory disease, diabetes, tissue fibrosis, and the formation of certain types of cancer[60].

1.3 Antibody (Ab)

The immune system is a defence against pathogens. The first answer in a host's defence is provided by the innate immune system, which is a rapid response. However, this rapid reaction is characterized by the lack of immune memory, and it results in a non-specific response. Vertebrates also have the more specific adaptative immune system. This immune response is specific for each antigen, and lymphocytes of the immune response evolved in order to recognize a vast variety of different antigens. Antigen is considered any molecules specifically recognized by high specific proteins on the lymphocytes. Antigens can be recognized by T and B lymphocyte receptors. T lymphocytes receptors (T cells receptors, TCRs) are displayed on the cellular surface. TCRs recognize peptide presented by MHC. In contrast to TCRs, the immunoglobulins (Igs) are proteins involved in

the direct recognition of antigens and Igs are displayed on the surface of B lymphocytes. Every lymphocyte codifies an Ig specific for an antigen. Once differentiated in plasma cells, the B lymphocytes can secrete Igs as Abs[79].

1.3.1 Structure

In 1890, von Behring and Kitasato reported the existence of an agent in the blood that could neutralized the diphtheria toxin[80]. Almost forty years later, in 1939, Tisalius and Kabat used electrophorese to separate immunized serum into his different fractions: albumin, α-globulin, β-globulin and γ-globulin. Moreover, it was shown that the serum is depleted of γ-globulin fraction after incubation of serum with the antigen [80]. Gerald and Edelman worked on the identification of the Ig structure. The heterogenicity was the major challenge for Edelman during Ig structure understanding[81], in an age, the 1950s, when immunology was considered only a sub-discipline. It was thought that the interaction of Ig with the antigen was due to an "*instructional*" process, where the antigen served as a template for the folding of the binding site of the Igs[82]. However, Edelman did not agree with the instructional theory, because this theory did not provide that Igs were not generated against self-antigens[81]. From '59 to '69 the "*instructional*" theory was abandoned in favor of "*selection*" theory[83], where the Igs were synthetized prior to antigen exposure and made up a repertoire of proteins[81]. The "*clonal selection*" theory proposed the concept that Abs naturally possess an affinity for the antigens and are selected from a large group of pre-existing globulins[81].

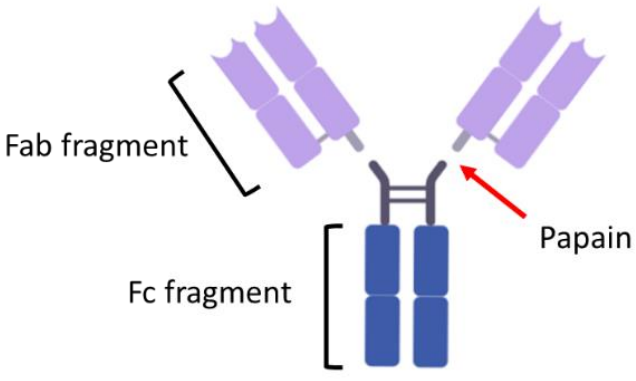


Figure 5. ***Structure of Ig after the cleavage of the Ab with papain enzyme****. These subunits correspond to the two Fab fragments (in purple) and Fc fragment (in blue). Biorender.*

Both Edelman and Porter focused on the understanding of the Igs structure. Porter's experiments had been performed on the rabbit's IgGs. In 1959, using the enzyme papain, Porter cleaved the IgGs in three pieces of about 50KDa each. These subunits correspond to the two Fabs (antigen-bindings) and constant Fc (crystallizable) fragments (Fig. 5)[84]. Papain hydrolyzes peptide bounds in IgGs in order to generate three fragments, with similar molecular weight but different charges. Moreover, the crystal of Fc fragments from Abs with different specificity are homogeneous. On the other hand, the lack of capacity of Fab fragments to form crystals correlates with structural heterogeneity of these antigen specific fractions and so the different amino acid sequence[81]. At the same time, Edelman showed that the reduction of disulfide bonds of Abs in the presence of denaturating agents led to the dissociation of the molecule into smaller pieces, light (L) and heavy (H) chains[85]. It was proposed that Igs must be composed of more than one polypeptides chain linked by disulphide bonds and so, the Igs consist of multiple subunit of polypeptide chains[86] (Fig. 6). In the 1969, Edelman and colleagues demonstrated that L chains are composed of variable (V) and constant (C) regions[87] (Fig. 6). In the same year, Press and Hogg demonstrated that the H chains are also composed by V and C regions[88].

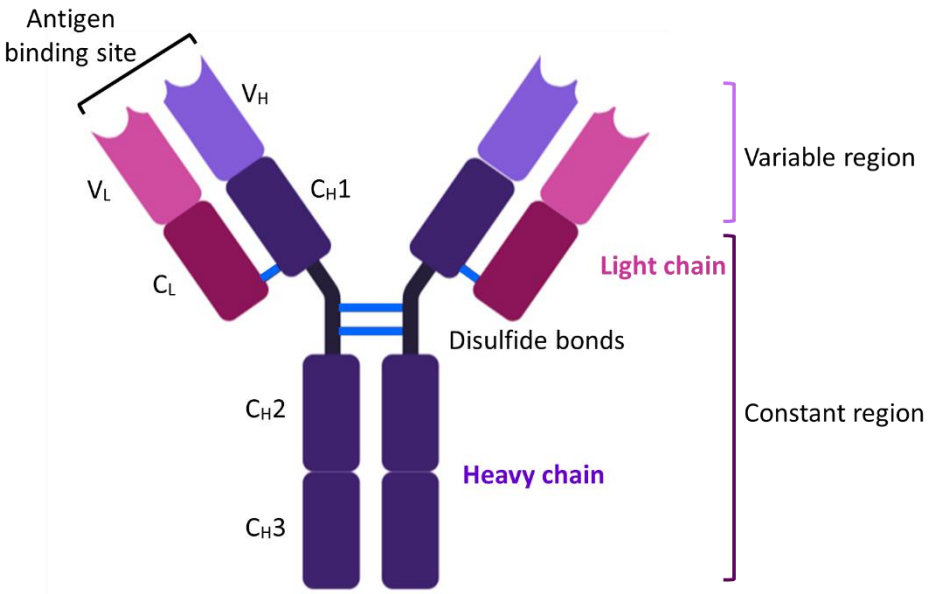


Figure 6. ***Structure of the Ig****. Ig is a heterodimeric protein composed of two H (purple) and two L (pink) chains. The two H chains are joined to each other thanks to disulphide bounds (blue) at the level of hinge region (line in purple). There is also a disulphide bound between the L and the H chains (blue). Both L and H chains are composed by a V and C regions. The union of VL and VH generate the antigen binding site.*
The L chain has a variable region and one constant domain, CL, whereas the H chain has a variable region and three constant domains (CH1, CH2 and CH3). Biorender

From those years till half of the previous century, the Igs have been deeply investigated. We know that Igs are a Y-shaped proteins composed by three portions with similar weight, as previously described (Fig. 5). Moreover, all the Abs are composed by two couple of L and H polypeptide chains paired. H chains are polypeptides from 55kDa to 77kDa, whereas L chains have a molecular weight around 25kDa. Disulfide bonds join the two H chains of an Ab and another disulfide bond joins the H chain with the L chain (Fig. 6). The two L chains and the two H chains that compose the Ab are identical[89].

Both the L and H chains have a C region and a V region (Fig. 6) (called CL, VL for the L chain and CH and VH for the H chain). Each tip of the "Y" is composed by the N-term, for both the L and H chain. These regions are the most variable regions and are, hence involved in the antigen binding. The force of the interaction between the Ab binding domain (paratope) is called affinity. In addition, the Ab has two equal Ab binding domains, so each Ab can bound two antigens on a surface. This structural characteristic drives up the total force of the interaction, which is called avidity[79].

The C-term of the L chain is the CL region. There are two type of L chains, which are called lambda (λ) or kappa (κ) in honor of their discoverers, Korngold and Lipari [90]. The λ chain has at least six functional segments which, however, have not been shown to have different functions, whereas the κ chain has only one constant segment[89]. The L chain, composed by one CL and VL domain, has a mass of approximately 25kDa[80].

The CH region is involved in the communication with the other components of the immune system. There are five types of CH fragments, 5 classes of isotypes. Some isotypes have sub isotypes, which confer its different functions to the Ab. The five main isotypes are: Ig M (IgM), Ig D (IgD), Ig G (IgG), Ig A (IgA) and Ig E (IgE). In contrast to CL chains, CH chains contain three or four C domains, depending on the isotype[80]. The CH1 domain is in the Fab, which is in fact composed by VH, CH1, VL and CL domains, whereas, CH2, CH3 and eventually CH4 generate the Fc region. Between CH1 and CH2 there is the hinge region, where there is the formation of disulfide bounds between the two C chains (Fig. 6). The base part of the "Y" is composed by these 2 (for IgD, IgG and IgA) or 3 (for IgM and IgE) subunits of the CH and contains a glycosylation site [91]. The H chain, composed of three CH domains and a VH domain, has a mass of approximately 55kDa[80].

1.3.2 Igs genes

The Igs repertoire is the result of all the specificity of the Igs in every individual. The total number in human is at least 10^{11}. There were two theories for the generation of the Igs repertoire. The first

was the germline theory, in which every Ig is codified by a different gene, implying that the Abs repertoire is inherited. On the other side, the somatic diversification theory proposed that the Abs repertoire is generated from a limited number of inherited V-region sequences in the B cell, sequences that in each individual can undergo modifications. The cloning of Ig genes revealed that the Abs repertoire is generated by DNA rearrangements during the B-cells development. The DNA sequence encoding for the V region is assembled at each locus by selection from a relatively small group of inherited gene segments[79].

After gene rearrangement, in mature B lymphocyte the V sequences are closer than in non-lymphoid cells. This rearrangement is called somatic recombination and it characterizes the B lymphocyte[79]. The Ig H and L chains are each encoded by a separate multigene family[92,93]. V and C domains are encoded by independent elements. V(D)J gene segments for the V domain, and an individual exon codifies for C segments. The *loci* for the H chain are on the chromosome 14q32.33, whereas loci for L chain are in different chromosomes: the Lκ is on the chromosome 2p11.2 while the Lλ is on the chromosome 22q11.2[89,94–96] (Fig. 7 A,B,C).

1.3.3 Variable regions

The V domains are functionally divided into three hypervariable domains, called complementary determining regions (CDRs), which are located between four regions of stable sequence called framework regions (FRs)[80] (Fig. 7D). The creation of the V domain is directed by recombination signal sequences(RSSs), that flank the rearranging gene segments. Each RSS is composed of 7 well conserved nucleotides, which are followed by a spacer of 12 or 23 non-conserved nucleotides, themselves followed by 9 less well conserved nucleotides. The spacer of 12 or 23 nucleotides reflects one or two DNA elix turns, so that the hepatmer and the nonamer are on the same side of the DNA molecule. A gene fragment with a spacer of 23 base pairs (bps) can join only with a gene segment with a 12 bps long and viceversa, avoiding wasteful rearrangements[79,80]. The number of functional genes are reported on Figure 7[80], but the pseudogenes are not taken into account[97].

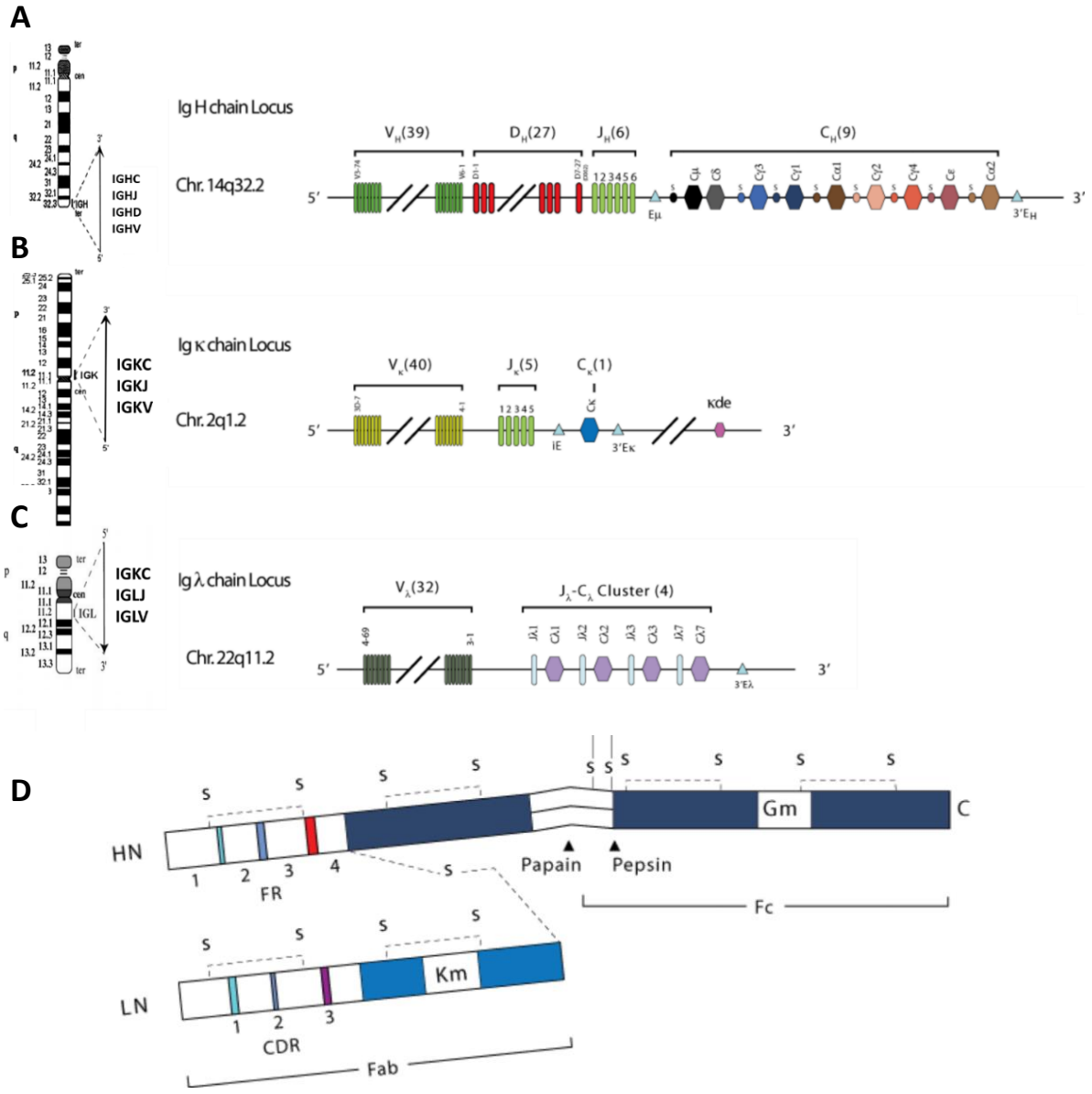


Figure 7. ***Organization of Ab***. ***A)*** *human IGH locus at 14q32.33. The vertical line indicates the localization. The number of genes in the locus is shown. It comprises 72 genes for the V region, among which 39 IGHV, 27 IGHD and 6 IGHJ, and 9 genes which codify for the C region* ***B)*** *human IGK locus is in 2p11.2. The vertical line indicates the localization. The number of genes defines the potential IGK repertoire. The V region comprises 45 genes, among which 40 IGHV and 5 IGHJ, whereas the constant region is codified by only one gene.* ***C)*** *human IGL locus is in 22q11.2. The vertical line indicates the localization. The number of genes defines the potential IGL repertoire, and it comprises 32 IGHV genes and 4 IGHJ which are alternated with the 4 IGLC genes. After somatic rearrangement of the DNA, where V(D)J genes are joined, the DNA is ready for transcription and transduction to generate the polypeptide chains* ***D)*** *H and L chains are reported. They are joined by the disulphide bounds and the disulphide bounds from the H chain indicate the bounds with the other H chain. Both H and L chains have the C region (blue and light blue respectively) and the V region. Each V region contains 3 CDRs and 4 FRs. The 4 Frs are indicated on the top on the H chain, whereas, on the bottom the L chain is reported with the 3 CDRs indicated (Edited from Marie-Paule Lefranc 2001 and Schroeder et al., 2010).*

The H chain locus is shematized on Figure 7A. Near the telomer there is the V region, which is composed by three gene fragments: variable (V), joining (J) and diversity (D). VH genes can be grouped in 7 families[98]. From telomer to centromer, after VH genes, 27 DH genes can be found[99] followed by 6 JH genes. These genes are rearranged in the VDJ order following the 12/23 rule. The VH gene segment contains FR1,2 and 3, CDRH1 and 2 and the amino terminal portion of CDRH3. The DH gene forms the middle part of CDRH3 and the JH gene generates the carboxy terminal part of CDRH3 and FR4 (Fig 7A and 7D)[80]. To increase the diversity of CDRH3s there are other mechanisms in addtion to VDJ rearrangjament. The first allows the DH gene segment to be in all the six potential frames. The second is the formation of a hirpin during the. A nick subsequently solves this hairpin, but it could generate a 3' overhang that creates a palindromic extension, named the P junction. Third, during the recombination process, the gene segment can undergo a loss of one of the seven nucleotides. Fourth, a polymerase can add nucleotides to replace or add to the original germline sequence. All these factors allowe the generation of the huge diversity that characterizes CDRH3s, both in length and structure, so that more than 10^{10} different CDRH3 can be generated[80]. The V domains of both κ and λ chains are composed by the junction of V and J segments. In fact, L chains lack the D gene segment[79]. Genes for VL are rearranged in V-J order following the 12/23 rule[79] as for H chains, and, as for H chains again, the CDRLs1-2 of the L chains are on the V segment while the CDRL3 is composed by a carbossi terminal of V fragment and the amino terminal of the J fragment[79]. The κ locus schematized on Figure 7B is on the chromosome 2p11.2[100]. There are 40 functional V genes, and 5 J genes. VJ fragments are joined with one Cκ fragment[80]. The λ locus schematized in Figure 7B is on chromosome 22q11.2. There are 32 Vλ functional genes, after the Vλ genes followed by Jλ genes, which are alternated with Cλ genes[80]. The nomenclature of the Igs is based on the IMGT®, which is the international ImMunoGeneTics information system® (http://www.imgt.org). The IMGT® was created in June 1989 in Montpellier, by Marie-Paule Lefranc (University of Montpellier and CNRS) to characterize the genes and alleles of the antigen receptors, immunoglobulins (IG) or antibodies[101,102]. The number into the name of the genes reflects the nomenclature but not the chromosomal position.

1.3.4 Antigen binding domain

The interaction of the two polypeptide chains generate the antigen binding site[103]. It is known that the antigen binding site of Igs is formed by six hypervariable regions from both light and heavy variable domains[103,104] (Fig. 8). Five of six of the hypervariable regions usually show a limited number of chain conformations and a very similar local conformation. These are called canonical structures[104]. Moreover, the local conformation of the canonical structure is well conserved[105], suggesting that in most of the Igs it is possible to predict with great accuracy the local conformation of the hypervariable regions from the sequence, except for the loop H3[104]. In the antigen binding site, CDR loops out of the V region backbone formed by two sheet of β-pleated strands allow the formation of a stable binding site[106]. As previous described, CDRs, except CDR3, are encoded by the germline sequence, whereas CDR3s derive from recombination events, suggesting the variability of these loops[106]. Furthermore, the CDRH3 loop has a wide range of variation in both length and shape[105]. To further analyze the role of the VH chain, Abs with the same VL but different VHs have been identified to be specific for different antigen[107]. Amazingly, it was shown that changing only the CDRH3 changes the specificity of the initial Ab[108], suggesting that the CDRH3 plays a crucial role in the specificity for the antigen binding. On the other hand, it was shown that Abs with the same CDRH3 bind a target with different affinity, suggesting an importance of additional antibody structure beyond the CDR3 in modulating the affinity[109]. To support the importance of the V chains backbone, Abs carrying the same CDRH3 have been identified from both specific and non-specific Abs. Specific Abs have been investigated, and a high level of homology and the same rearrangement

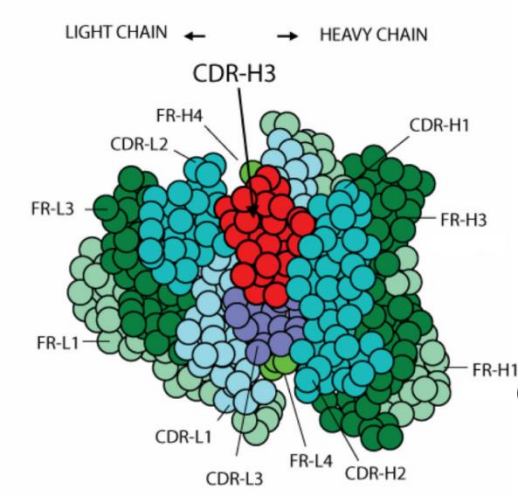


Figure 8. ***The antigen binding site****. The light chain is on the left, whereas the heavy chain is on the right side. Both chains have the four FRs and the three CDRs. The CDR-H3 has a central location (red) (Edited from Schroeder et al., 2010).*

for both H and L chains have been identified[109], whereas the analyses of non-specific Abs show very different VL genes. Surprising even CDRH1 and CDRH2 were largely diverse[109]. These data suggest that conformations of CDRH3s are modified not only by the sequence, but also by the structural environment, especially by the VH and VL pairing and that CDRH3 is necessary yet insufficient for specific binding[109].

To further increase the Igs diversity and affinity for the antigen another mechanism is involved. The variable domain genes undergo a somatic hypermutation (SHM), which changes one base pair every 1000 per cell cycle. These errors can be introduced through two different mechanisms, the first mechanism targets hot spot motif, whereas the second involves an error prone DNA synthesis that introduces a mismatch with the template. These mechanisms are activated after the interaction with the antigens[80].

1.3.5 Constant regions of the heavy chain

When human plasma was electrophoresed, certain protein groups were ascribed a Greek letters to designate where the protein falls in order of the mobility. In 1939, Tiselius and Kabat showed that the gamma fraction of electrophoresed serum contains the largest amount of Igs. The first isotype of Igs discovered was called γ-globulin since it was in the γ region of the mobility. After five years IgM was discovered, and subsequently IgA, IgD and as last IgE was described, in 1966[110]. The order in which the Abs develop is different than the order of their discovery. In fact, the VH chain in an early stage is expressed in association with the μ H chain to produce IgM, and after, by alternatively splicing, IgD. Subsequently, during the development, in response to different antigenic stimulation, the V domain may associate with other isotypes (IgG, IgA and IgE)[80].

Heavy chain genes are on chromosome 14 and CH genes follow the genes that codify for the VH region (Fig. 7). The C region is defined as CH1-CH2-CH3 for IgG, IgA and IgD, whereas the additional domain CH4 is present in IgM and IgE. The Fc fragment defines the isotype and the subclass of the Igs. Each CH region folds into a fairly constant structure consisting of a 3 strand-4 strand beta sheets pinned blocked together by an intrachain disulfide bond. Fc fragments mediate the effector function, binding Fc receptors on effector cells or activating other immune mediators. CH2s of the Igs are glycosylated, and the glycosylation changes in accordance with the isotype. Glycans interact with hydrophobic pocket on the Fc domain that stabilize the Ig structure[80].

IgM is the first Ig expressed during B cell development with the V region not yet undergone to affinity maturation. IgM can specifically bind the target with the high avidity. In fact, IgMs are usually

multimeric, meaning pentameric and rarely hexameric. IgMs are linked to each other by disulphide bonds at the CH4 subunit. Moreover, a peptide chain, the J-chain, binds of two monomers. IgMs coat the antigen for destruction[80].

IgDs have a very low expression and it is not clear why mature B cells need the expression of IgD in addition to IgM. It has been proposed that IgDs could have a role in the modulation of the humoral immune response[111].

IgG is the simplest, most abundant isotype found in the body and represent 75% of total Igs in human. There are 4 subclasses, from 1 to 4 that reflect the rank order of the serum level of these Abs in the blood. The IgG subclasses shown different functional activities: IgG1 and IgG3 are generally induced in response to protein antigen, whereas IgG2 and IgG4 are associated in response to polysaccharide antigens[80].

IgAs, like IgMs, can take advantage of avidity for the binding. IgAs can form dimers, especially at the mucosa level, with the J-chain helping this association. IgAs in the serum are less present than IgGs, but IgA levels are much higher than IgG on the mucosal surface and in secretions, like in saliva and breast milk. There are two subclasses of IgA, IgA1 and IgA2. IgAs protect the mucosal surface from toxins, virus and bacteria by binding to them at the mucosal surface. IgAs may have a role in the protection of intestinal tissue by taking up the antigen to dendritic cells.[80]

Last, there are IgEs, which are the lowest Igs in the human serum and also have the shortest half-life. IgEs are associated with hypersensitivity and allergic reactions as well as the response to parasitic worm infections[80].

1.4 Anti Transglutaminase2 antibodies from CD patients

The serum of CD patients includes Abs recognizing TG2, and the level of Abs increases after exposure to glutens and decreases in a gluten-free diet. Abs reactive to TG2 have been identified from repertoires derived from intestinal biopsy lymphocytes, but not from peripheral blood lymphocytes[52]. Moreover, accumulation of TG2-specific IgA plasma cells occurs in the small-intestinal mucosal lesion of untreated CD patients and these B cell clones also contribute to the generation of serum IgAs anti-TG2[52,112]. Interestingly, it has been reported that in CD patients approximately 10% of Ab-secreting plasma cells in the duodenal mucosa produce TG2-specific Abs[113]. In autoimmune diseases B cells produce antibodies, and they are also involved in the secretion of cytokines or presentation of the antigen to T cells. It was suggested that B cells can be the main APC (Fig. 2) for $CD4^+$ T cells in autoimmune diseases[114]. Plasma cells are the dominant type

presenting gluten in the gut of CD patients. It was hence proposed that B-cells are involved in stimulating pathogenic, gluten-specific T cells[115]. The production of TG2-specific IgAs and IgGs probably results from a collaboration between TG2-specific B cells and gluten-specific $CD4^+$ T cells, where BCR mediates the uptake of the complex TG2-gluten[49]. TG2 has a double action on gluten, deamidation and transamidation[76,116]. Recent data indicate that TG2-specific plasma cells in CD target the N-region on the TG2 and this epitope biases the presentation of deamidated gluten

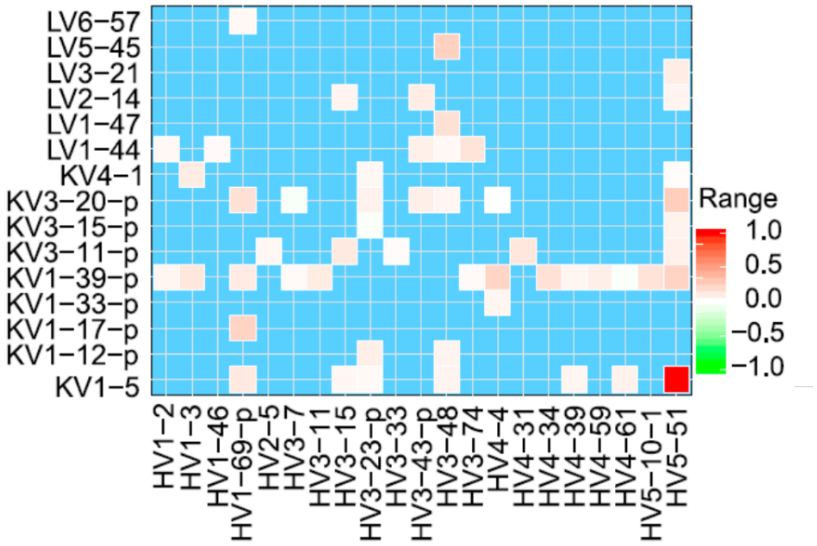


Figure 9. ***TG2-specific (TG2+) BCRs show strong VH:VL pairing preferences****. Heat map showing the relative VH:VL pairing frequencies for frequently observed pairs. The color intensity index for each pair was obtained by dividing the difference in frequency between TG2+ and non–TG2-specific. (TG2) PCs with the highest difference value. Averages of frequency values from 10 patients were used. (Edited from Roy et al., 2017).*

peptides to T cells by B cells binding enzymatically active TG2[117]. Moreover, Abs that target N-terminal of TG2 are suggested to be linked to the development of the patology[117]. In the next few pages other characteristics of anti-TG2 Abs are presented, however other studies are necessary to deeply investigate these antibodies.

Finally, Abs specific for other members of TGs family have been identify in CD or other CD-related conditions[50].

1.4.1 Preferential VH and VL usage

Analyses of anti-TG2 Abs have shown that there is a preferential use of some VH gene families. In 2001 it was demonstrated that the VH gene use was restricted to three, VH5, VH3 and VH1, of the

seven human Ab VH familes[52]. The binding of these anti-TG2 Abs has been examined and showed competition with the patient sera[52]. Due to their importance in the autoimmune response and the high level of anti-TG2 Abs the attention was focused on the characterization of these Abs. Since the VH plays the most important role in the antigen binding, a lot of studies focused on the characterization of VH chains of anti-TG2 Abs. VH5 gene family, especially the gene VH5-51 was identified as the most expressed in anti-TG2 antibodies from CD patients[113,118,119], whereas in databases of Igs repertoire VH3 is the most represented, followed by VH1 and VH4, and VH5 is only the fourth most expressed[120,121]. Interestingly, even the VL chains of anti-TG2 Abs from CD patients show a preferential usage of κ light chain[113]. The anti-TG2 Abs VH-VL pairing has been analyzed with single cell high-throughput sequencing on gut lesion plasma cells, and a preferred pairing has been identified[119]. The most preferred pairings in anti-TG2 Abs (TG2+) are VH5-51:KV1-5 and VH5-51:KV1-39 (Fig. 9), whereas these preferred pairings are not observed neither in the set of Abs non-specific for TG2 (TG2-)[119] nor in a naïve library[120,121]. Interestingly, these preferential pairings are not limited to a single CD patient, but they recur in different individuals[119].

1.4.2 Few somatic hypermutations and high affinity

Surprising, these Abs show high affinity to TG2 despite the few somatic hypermutations[113,119]. The affinity of anti-TG2 antibodies is in the nM range. A low number of somatic hypermutations is in VH5-51 of anti TG2+ Abs, whereas other VH genes show a higher level of mutations than VH5. Moreover, this low number of mutations is restricted to CD, whereas other Abs from influenza show a higher number of somatic hypermutations[113]. Despite the low number of VH and VL of anti-TG2 Abs mutations, phylogenetic analysis has demonstrated that clones with a particular mutation expanded more than others, probably due to a possible increase of affinity[119]. Furthermore, the reversion of the recurrent mutation leads to a reduction of affinity in TG2 binding[113,119]. Notably, limited VH-VL pairings and limited somatic hypermutations have been identified even in anti-gliadin IgAs of gut plasma cells lesion of CD patients[122].

1.4.3 Epitopes on TG2

It was demonstrated that Abs anti TG2 recognize the antigen in four different epitopes [62,65,118,119,123]. It was shown that epitope 1 is the most targeted, and that notably Abs that bind this epitope have VH5[118]. Anti-TG2 epitope 2 Abs use IGHV3 gene segments, whereas anti-TG2 epitope 3 Abs mainly use IGHV4 gene segments. The fourth epitope is partially overlapped so the three major epitopes

are considered[118]. Anti-TG2 epitope1 Abs are reactive to mouse TG2 (mTG2), which shares 84% of its sequence identity with human TG2 (hTG2), whereas Abs from others epitopes are not so cross-reactive[118]. Notably the Abs anti-TG2 from CD patients intestinal biopsy are specific for TG2 and cross-reactivity for other members of transglutaminase family was not identified[118].

Briefly, TG2 can assume two conformations, open and active in presence of a high level of Ca^{++}, and closed and inactive with low level of Ca^{++} and in presence of GTP (as previously described in chapter Transglutaminase 2) (Fig. 3). TG2 produced in *E. coli* exists mostly in the closed conformation, whereas Sf9-produced TG2 is in the open conformation[118]. The affinity for the open conformation was shown to be slightly higher than for the closed conformation. Eventually, addition of Ca^{++} leads to the same conformation independently of the producer organism[118]. Regardless, difference in affinity for the different conformations has been assessed, and the affinity was only weakly affected or not affected at all by the allosteric regulation, suggesting that anti-TG2 Abs from CD patients bind the antigen in a region that undergoes to only small structural variations upon ligand binding[118]. N-terminal of TG2 has been identified as the location of the three epitopes targeted by anti-TG2 Abs from CD patients[118] (Fig. 10B). Surprisingly, the affinity of Abs that target this region showed similar binding affinities[117]. Abs that target non N-terminal region of TG2 could bind either a region in the core domain near the N-terminal or C-terminal of TG2, but these Abs have only been identify at low

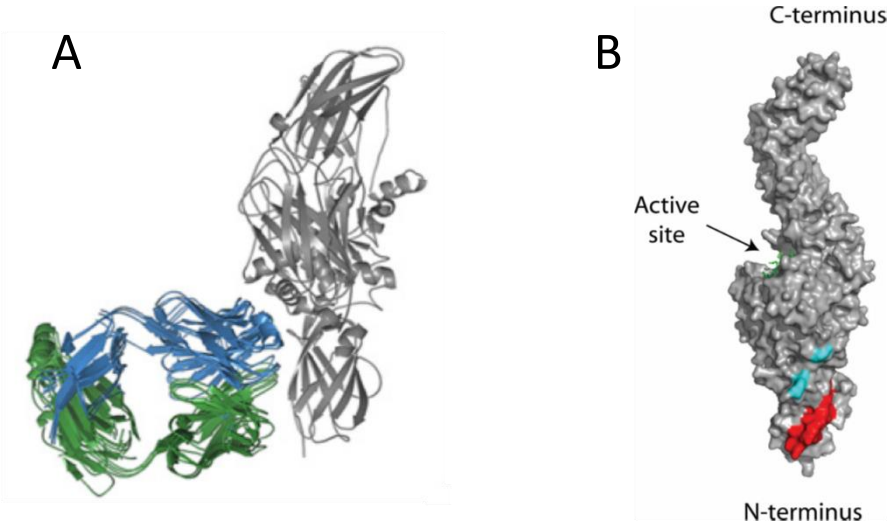


Figure 10. ***Models of the interaction of TG2-GTP with 679-14-E06 Fab fragment. A)*** *TG2-GTP is colored in gray (PDB code 4PYG), and the light and heavy chains of the Fab fragment 679-14-E06 (PDB code 4ZD3) are colored in green and blue, respectively.* ***B)*** *Regions targeted by TG2-specific autoantibodies. Surface representation of the open TG2 structure with a bound active site inhibitor shown in green (PDB code 2Q3Z). The region in the N-terminal domain that has been implicated in binding to fibronectin (aa 88–106) is shown in red and overlaps with epitope 1, whereas epitope 2/3 are shown in cyan. (**A)**Edited from Chen et al., 2015.* ***B)*** *Edited from Iversen et al., 2013).*

level[117]. Anti-TG2 epitope 1 Abs have been identified as the most abundant in CD patients, hence a lot of effort has been done on the characterization of these Abs. One referred Ab that binds epitope 1 is called 679-14-E06[65,113,118,124,125]. The crystal structure of the Fab Fragment of 679-14-E06 has been identified, and the binding with the antigen could be well characterized (Fig. 10 A)[65]. Both structures are available on PDB. The code for TG2-GTP is 4PYG (Fig. 10 A)[126] and the one for 679-14-E06 Fab is 4ZD3 (Fig. 10A)[65]. The open conformation has also been deposited on PDB with code 2Q3Z (Fig 10B)[127]. 4PYG is in the close conformation, which is the predominant conformation produced in *E. coli*, so this backbone has been considered in this study. Thanks to these 3D structures, it has been demonstrated that both the H and L chains of 679-14-E06 are involved in the epitope 1 binding[65].

In this work we focus on the characterization of anti-TG2 Abs targeting the epitope1 on TG2, since most of the IgAs and IgGs against TG2 from CD patients bind this epitope. For our purposes we take advantage from the backbone structure of the 679-14-E06 Ab deposited on PDB, as previously described.

As general knowledge about anti-TG2 Abs, additional information is necessary. It has been demonstrated that TG2 is exposed on the cellular surface thanks to the binding with integrins or heparane sulfate chains of syndecan-4, and it is involved in the adhesion and migration[72,73]. One additional target is collagen VI[125]. Moreover, TG2 predominantly binds fibronectin in the ECM and interestingly, epitope 1 on TG2 overlaps with the fibronectin-binding site of TG2 (Fig. 10B). Beyond this binding site, TG2 can be bound to ECM thanks to its second C-terminal β-barrel domain[125]. This hypothesis has been supported by the finding that very few plasma cells from CD patients recognize this region, suggesting that ECM-TG2 blocks the epitope, preventing the activation of C-terminal specific B-cells[117].

Both CD patients and potential CD patients with untreated disease showed an increase of TG2-specific IgA IGHV5 family, which binds the N-terminal epitope, suggesting that anti N-terminal Abs are produced on the onset of the pathology. In the early stages, the N-terminal is blocked by the TG2 N-terminal, allowing for an increased production of non-N-terminal TG2 epitopes in the progression of the disease.

In addition, auto-antibodies anti-TG2 have been identified even in non-CD individuals suffering from other diseases, such as inflammatory bowel disease or viral infections. However, these anti-TG2 Abs target other epitopes, different from the epitopes recognized by anti-TG2 Abs from CD patients[55,119].

1.5 Phage display technology

Methods for generating Abs have been developed in 1980 with the production of polyclonal Abs from animal immunization[128]. In 1975 the hybridoma technology was introduced by Koheler and Milstein[129]. Mice were immunized with the target antigens, the spleens harvested, and the B-cells immortalized by cell fusion. The so obtained hybridomas were then screened to identify those making monoclonal Abs against the target[130]. Hybridoma technology provides the capability to produce a homogeneous and purified Abs preparation that improves tracking, detection, and quantification. However, the Abs are produced in mouse organism, and this feature limits their usage for therapy, due to an immune reaction. Precisely, mouse monoclonal Abs were seen as foreign when injected into patients[128]. Moreover, the hybridoma technology is expensive and the isolation of the specific antibodies require a long time. However, the main limit of the hybridoma is due to the complication to generate Abs against toxic and highly conserved antigens, as well as these antigens not being stable in animal system[128].

At the beginning of the 1980s, solutions began to emerge through the application of protein engineering. The first focus of protein engineering was to turn mouse monoclonal Abs into their human counterparts[131]. Bacteria were at first used for the expression of Abs, but with poor yields. Lymphoid cells were later proposed as host cells, with a more suitable achievement[130].

In 1985, George Smith published a new technology: an expression vector where the peptides could be displayed on a filamentous bacteriophage (Fig. 11) by genetic fusion to a coat protein that mediates the bacterial infectivity of the phage[130,132]. Primers to amplify the Ig variable regions were identified in 1989[133], and in the same year a library of Fab from immunized mice were generated, and specific Fab isolated[134]. This technology has then been improved, and in 1990 McCafferty, Griffiths, Winter and Chiswell published an article entitled: “Phage antibodies: filamentous phage displaying antibody variable domains” in which they displayed VH linked with VL with a short flexible peptide, thus creating a single chain fragment variable (scFv) (Fig. 11) as fusion protein with the phage pIII coat protein, on the phage surface[135]. This paved the way for the era of phage display technology. This technology is a robust, easy to perform and inexpensive method to select specific antigen binders.

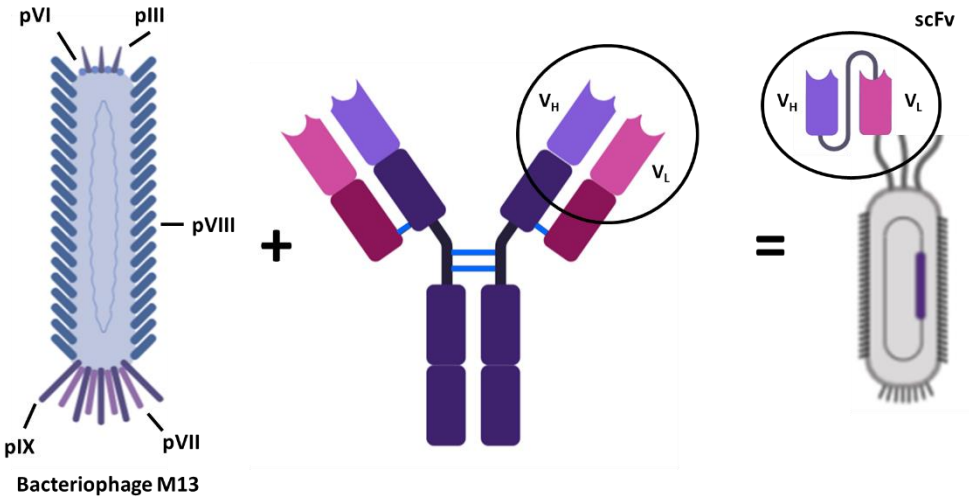


Figure 11. ***Schematic of the display of antibody fragment on phage surface***. *The left shows the bacteriophage M13 structure. The proteins of the phage are indicated and the circle inside the phage represent the ssDNA. In the middle the Ab structure is shown and VL and VH are rounded by a circle. On the right the VH joined by a linker to VL is shown. This structure is a single chain fragment variable (scFv) (in the circle). ScFv is displayed on the phage surface as fusion protein with pIII of the phage, and the phagemid DNA is enclosed in the phage coat protein. The genes in the phagemid codify for the scFv display on the phage surface. Biorender.*

1.5.1 Phage

Phages are viruses that infect bacteria. They are widely used in research, because they can be propagated to enormous numbers inexpensively in microbiological cultures, they are easily freed from impurities and are easy to manipulate in the laboratory. The virions, the phage particles, have a coat, called capsid, inside the capsid there is the phage genome, whose genes encode for phage proteins, including the coat proteins. The capsid mediates the infection on the host cells, and when the phage attaches the uninfected bacterium, transfers to the host the phage genome. Once inside the cell, the phage genes reprogram the cell's machinery to make progeny virions, which are release from the host and can go to infect other bacteria[136].

Different phages infect different bacteria, and filamentous phages infect *E. coli* cells. The filamentous phage M13 (here the focus is placed on M13, but other strains, as f1 and fd exist) infects *E. coli* via F pili. In fact, M13 can infect "male" *E. coli* cells (which are F^+)[137]. The M13 phage is neither temperate nor lytic. M13 phage establishes a chronic infection in its host, where it continuously releases new phages[137]. The genome of the M13 is a single strand DNA (ssDNA) 6407 nucleotides

long which is enclosed in coat proteins (Fig. 11). The genome consists of 9 genes that codify for 11 proteins. Five of these proteins are coat proteins, whereas the other six are involved in the replication and assembly of the phage. M13 is about 1µm long and 6nm across, so, the capsid is long and thin and consist of a long tubular array of 2700 major coat proteins (pVIII) (Fig. 11)[137]. The phage is chapped at each end by minor coat proteins, which are present in approximately five copies. The head of the virion is composed by the minor coat proteins pIII and pVI, whereas the tail is composed by the coat proteins pVII and pIX[137] (Fig. 11).

1.5.2 Life cycle

As previously said, the phage M13 is a non-lytic filamentous phage. F pilus of the "male" *E. coli* cells can be attached by pIII of the M13. The natural disassembling of the pilus automatically brings the phage closer to the surface of the cell. Here, the circular ssDNA genome of the phage enters in the bacterium, where, thanks to the host machinery, ssDNA is converted in double strand DNA (dsDNA) like replicative form (RF). The RF undergoes rolling circle replication and serves as template for expression of the phage proteins. Phage proteins are involved, beyond in the coat, in the replication, assembly and extrusion. Phage progeny are assembled by packaging the ssDNA into protein coats and extruded through the membrane into the medium[137,138].

1.5.3 Phagemid libraries

The M13 bacteriophage became the most favorite candidate for phage display thanks to its easy manipulation and the coupling of genotype and phenotype. As previously indicated, George Smith displayed a peptide fragment of EcoRI as fusion protein with the pIII on the heat of the filamentous phage. Smith checked first the infectivity of the virions, and later the effective display of the peptides[132]. Initially the protein of interest had been fused in the genome of the phage. Later, as an alternative, phagemid vectors based on smaller minimal plasmid have been used[137]. Phagemid vector contains bacterial and phage origins of replication, a leader sequence, multiple cloning sites, an antibiotic resistance gene, a coat protein (pIII or pVIII) and a weak promoter (*lacZ*)[139]. Notably, a phagemid vector alone is not able to produce infective phage particles: an helper phage, such as M13KO7, is indispensable to provide the necessary genes encoding the proteins for the phage replication, packaging and assembly[139]. The helper phage has been modified to have a defective packaging signal, therefore, the replication and packaging is less efficient than the phagemid vector

that carries the wild-type M13 intergenic region. This leads to a production of a greater number of phages carrying the phagemid vector than the genome of the helper phage.

Peptides could be displayed on the surface of both the pIII and the pVIII. However, due to the high rate of pVIII, the display of large peptides is not well tolerated, and pIII display is more efficient, even if there is a low-level of display[128].

Since 1990, different Abs formats have been employed in the construction of antibody-displaying phage libraries. Already in 1988 the format scFv had been proposed[140]. Combining these knowledges, a scFv phage libraries could be generated. To generate a scFv library, genes of VH and VL chains of antibodies are prepared by reverse transcription from the mRNA from B-lymphocyte. The VH and VL chains genes are amplified and assembled to form single genes VL-linker-VH (Fig. 11). The so assembled scFvs are cloned in the phagemid vectors. *E. coli* cells are transformed with the phagemid vectors previously obtained. Cells carrying phagemid can grow and they can be infected with the helper phage to produce the recombinant phages that display scFv antibodies fragments as fusion protein to the pIII coat protein (Fig. 11)[138].

There are three types of antibody repertoire:

- Naïve antibody libraries

  A naïve library is a collection of Igs from circulating B-cells. The rearranged V-genes were amplified from IgM isotype of healthy, or non-immune, donors. A single pot library is generated from different donors, in order to increase the number of possible random VL-VH pairing. The naïve libraries have the extraordinary ability to be used to screen for multiple antigens. Over peptides, this kind of libraries could be used to select Abs fragment that recognize toxins and self-antigens and target multiple antigens with no prior exposure. The disadvantage of this library is that the naïve repertoire is polyreactive and the antibodies obtained are often lower in affinity compared to immune libraries[139].

- Immune antibody libraries

  To generate an immune antibody library, IgG (or, in particular case, IgA) mRNA is obtained from immune donors, such as diseased or infected patients. The choice of sample depends of the action of the disease/infections. Differently from IgMs used for naïve library, here the B-cells are activated and undergo the affinity maturation process. This library allows the isolation of high-affinity binders for the specific target. Immune libraries are very useful to study the response against disease and infections.

One limit of this library is that is not possible to generate an immune library for self-antigen, except for immune libraries derived from autoimmune diseases[139]. In presence of an autoimmune disease, immune libraries are the best tool to deeply understand and investigate the immune response. Other limits of the immune library is the not so easy availability of human donor samples and the lack of both human and animals donor for deadly antigens[139].

- Semi-synthetic and synthetic antibody libraries

  Semi-synthetic libraries consist of partially natural sequence mixed with chemically synthetized sequences. This union allows to maintain a natural framework and to design the diversity of the antibodies. CDRs are designed, to generate an artificial diversity, whereas the framework sequences are pre-determinate, and the choice of the frameworks allows an improved stability, higher expression, and reduced aggregation. This would allow a successful antibody panning and application.[139]

  Synthetic libraries are done with chemically derived antibodies. They are artificially created repertoires. It could be done by *in vitro* reconstruction of CDRs randomization. However, the construction is based on *de novo* synthesis and the Abs diversity is *in silico* designed. Synthetic libraries are optimized for the framework choice, and both the type of framework and the diversity of the CDRs can be decided. These libraries allow the generation of Abs repertoire against even toxic and self-antigen. Bioinformatics analyses based on pre-existing data play a crucial role in the *in silico* design of the Abs repertoire. Bioinformatics analyses will study specific epitopes, antigen-antibody interactions and affinity maturation design, variable gene segment recombination and structural prediction of the variable regions. Analyses of CDR regions provide valuable information on CDRs predominance[139]. This approach allows the generation of Abs affinity that is in the order of nM-pM.

In vitro affinity maturation strategies can be used to further increase the specificity of the Abs. Some of these strategies mimic the *in vivo* affinity maturation.

1.5.4 Affinity selection

After the phage display idea, George Smith focused on the development of an effective procedure for affinity selection. The term "panning" or "bio panning" has been used[136]. Since those years, the bio panning key steps have been conserved and are schematized in Figure 12. Moreover, different typology of phage libraries can be used. The antigen of interest is immobilized on a solid surface, and the phage display library is successively joined. The phages which display Ab fragments, which bind immobilized antigen, are captured on a substrate surface, whereas the unbinding phages are removed in the washing step. Specific phages that bind the antigen are eluted and used to infect *E. coli* cells. *E. coli* cells carrying the phagemid vector are recovered. This bacteria library can be infected with the helper phage in order to generate a new phage library, that now is enriched in phages that display antigen-specific Ab fragments. This enrich library is than ready for a new bio panning circle. The washing step is normally varied between rounds of bio panning to increase the level of stringency. After several rounds of panning, generally three or four, an enriched population

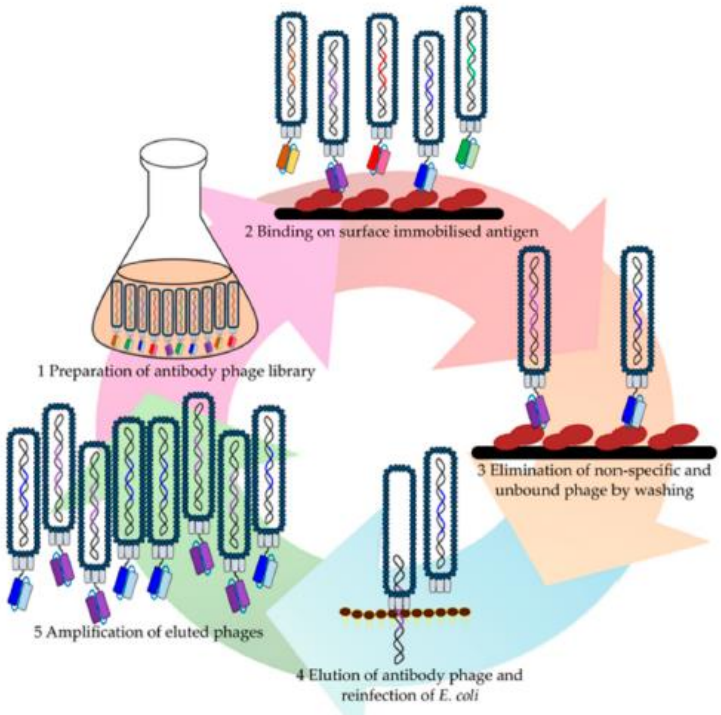


Figure 12. ***Bio panning process***. *A few steps are repeated to select Ab segments specific for the antigen. 1)Phage display library can be generated. 2) The antigen of interest is binding on surface and the phage library is joined to allow the interaction between the Ab fragment and the antigen. 3) A washing step allows the elimination of non-specific and unbound phages. 4) Phages bound to the antigen can now be eluted and used to infect E. coli cells. 5) The infected cells can be infected with the helper phage to amplify the phages that were selected for the binding to the antigen. This phage library is ready for a new cycle of bio panning. (Edited from Lim et al., 2019).*

is obtained and can be identified. The phages of the last round of bio panning will be tested as monoclonal Ab fragments. The specific monoclonal Abs can be identified by DNA sequencing[139].

1.5.5 Recombinant antibodies

After the introduction of the hybridoma technology and protein engineering, the first focus of the protein engineers was how to turn mouse monoclonal Abs into their human counterparts. This allows the use of mAbs in clinics, avoiding the immune response against the mouse Abs. Chimeric mAbs were generated at first, in which only the constant part is human, but the variable region is mouse. Then, humanized mAbs have been introduced, in which only the antigen-binding loops were mouse[130]. With the arising of phage display technology, a generation of fully human Abs became possible.

Once the specific scFv for the interest target has been identified by phage display, it is easy to combine the scFv with the human Fc region. This fusion generates the scFv-Fc. Abs are produced in mammalian cell culture, then the recombinant Abs are tested for affinity and specificity for the target[141].

Moreover, after identifying the VH and VL gene sequences of the scFv isolated by phage display, the full immunoglobulin can be produced. The full length immunoglobulins have the desired effector functions[142]. In 2002, Adalimumab (Humira) was the first human Ab that was approved for therapy by the Food and Drug Administration (FDA)[143]. This mAb opened the way to isolation of recombinant Abs from phage displayed derived Abs, and up to now, there are 7 mAbs approved by the FDA[143].

1.5.6 Antibody analysis

In addition to *in vitro* analysis, after identifying the nucleotide sequence of the VH and VL of the specific scFv for the antigen of interest, interesting *in silico* analyses can be performed. As previously mentioned for the nomenclature of the Igs, IMGT® is a powerful website to characterize the genes and alleles of the Abs[101]. The creation of IMGT® marked the birth of immunoinformatic, since it is at the interface between immunogenetics and bioinformatics[101]. The IMGT® database and tools, which built on the IMGT-ONTOLOGY axiom and concepts, bridge the gap between genes, sequences and three-dimensional structure[101]. IMGT® allows to standardize the delimitation of FRs and CDRs. For a two-dimensional graphical representation, IMGT® supply to the users the IMGT/Collier-de-Perles tool. Results from IMGT Colliers-de-Perles allow to quickly visualize amino acids that are important for the 3D structural configuration[101].

After the analysis of the Igs sequences, the identification of the genes and the FRs and CDRs, an additional instrument to label the interest residue is the Hidden Markov Model (HMM). HMMs are a formal foundation for making probabilistic models of linear sequence labeling problems. HMMs can identify profiles of multiple sequences and identify sites of interest[144].

To deeply understand the interaction of an antibody with the target antigen, the 3D structure can be investigated. In addition to experimental methods capable of generating structural models in atomic detail, like X-ray crystallography, nuclear magnetic resonance, neutron diffraction and cryo-electron microscopy, computational structure prediction methods are available[145]. To deeply investigate at the molecular level the binding Ab-Ag, a union of the *in vitro* study with the computational protein-protein interface prediction (docking) provides an excellent tool[145].

Computational antibody engineering has focused on improving binding affinities or biophysical characteristics[146], since, CDR backbones cluster into distinct groups of canonical structure[147] with preferential amino acid sequence. Computational antibody engineering has focused on improving characteristics of existing Abs, rather than designing *de novo*. In this work we focus on the affinity maturation, and analyses of the amino acids involved in the binding. Rather than studies of the antibody-antigen interaction and affinity maturation, *in silico* studies can be applied to identify mutations that confer thermo-resistance, improving association rates and identifying aggregation prone regions[146]. In the *de novo* design of CDRs, canonical CDR backbone loop structures need to be chosen first, then followed by the selection of amino acid chains using energy minimization, which corresponds to maximizing interactions[146]. This strategy, with the design of all the six CDRs, has been reported to generate antibodies with high sensitivity, in a range of 4-50nM[146]. The study of Entzminger and colleagues has here been reported, where they generated an Optimal CDR *de novo* design[146]. The choice of CDRs was based on a database of canonical structure backbones for CDRs derived from known antibody structures. CDRH3 has been reported to have ten-times more structures than other CDRs. Given the position of the antigen, their computational method calculates a score where CDR backbone atoms close to the antigen are rewarded as having the potential to contribute to binding. FRs regions were from known scFv structures, and libraries have been designed, and the energy of the interaction has been calculated (low energies indicating favourable interactions). The PyMOL Molecular Graphic System calculates the polar contacts. Once the VL that interacts with the antigen has been identified, like in the *in vitro* strategy, it is possible to modify only the CDRHs[146]. And only at the end of the study, validate *in silico* analyses *in vitro*.

Finally, RosettaAntybody is a server to predict the antibody structures.

2 Aim of the thesis

Celiac disease is a gluten-sensitive enteropathy characterized by an autoimmune response with mainly the production of IgA antibodies against gliadin and self-antigen tissue transglutaminase (TG2). For this purpose, studying the immune-response against TG2 enzyme could provide a major understanding about the development of celiac disease. In this work our purpose is to investigate the structural features of anti-TG2 antibodies with the IGHV5-51 chain, since it has been demonstrated that IGHV5-51:IGVK1-5 is the VH-VL preferred pairing found in anti-TG2 antibodies of celiac patients. The characteristics of anti-TG2 antibodies have been investigated, given their clinical-pathological meaning in CD, especially for antibodies characterized by those chains. Furthermore, it has been demonstrated that anti-TG2 antibodies mostly recognize three epitopes on TG2, and most of them, characterised by the presence of the IGHV5-51 gene segment, bind TG2 on the so-called Epitope 1 [62,65,118]. Moreover, it has been demonstrated, that anti-epitope 1 antibodies are the first anti-TG2 produced upon the onset of the pathology[117]. Here, we propose to investigate the sequence and the structural determinants of anti-TG2 Epitope 1 antibodies.

To this purpose, first, we generated phage display libraries of scFvs from CD patients' intestinal biopsy lymphocytes. We chose the phage display technology because it allows a rapid screening and selection of scFvs specific for the antigen. Due to a previous knowhow, generally, scFvs from the phage libraries reflect very precise characteristics of the antibodies anti TG2. The characteristics of these libraries were deeply analysed, both to validate the libraries and to investigate new features. The IGHV5 chain, which is the most identified in anti-TG2 antibodies was characterized in detail to mark both common features and differences between IGHV5 chains that compose anti-TG2 antibodies and non-specific antibodies.

ScFvs selected for TG2 binding and reactive clones were bioinformatically analysed. Since CDRH3 is the most variable region, we focused our attention on the VH CDRH3 loop, which is also the most important for antigen-binding.

Computational methods are universally accepted as important tools for the study of small molecules drugs. Computational methods are generally considered a way to generate *in silico* hypotheses, helping to interpret and guide experiments [148]. In the 1980s, antibody modelling started with the unexpected discovery that most of the CDR loops adopt a limited number of conformations called canonical structures [105]. CDRH3s from anti-TG2 scFvs were analysed with bioinformatics analyses, and a "*consensus*" sequence was designed based on Hidden Markov Models (HMMs).

ScFv with the "*consensus*" CDRH3 designed *in silico* was generated using an isolated backbone of a previous anti-TG2, and scFv carrying the "*consensus*" CDRH3 was assessed for the binding to TG2, as well as for the specific interaction with the epitope 1, to validate the hypotheses of a sequence designed *in silico*.

To further validate these data, the *consensus* clone has been produced in more physiological formats: scFv-Fc, IgG and IgA. The affinity of the consensus and the parental antibodies have been calculated in order to investigate if the strategy of designing CDRH3 *in silico* could be a tool to increase the affinity of antibodies, even if the parental antibodies show very high affinity. These data could support the combine strategy of *in silico* and *in vitro* technologies to generate high affinity and specific antibodies.

We also investigated the role of single amino acid position in the CDRH3 loop of anti-TG2 antibodies, to deeply investigate their role in antigen-binding. For this purpose, *in silico* random libraries have been projected and specific point mutants were proposed to be generated and investigated. Analyses *in vitro* have been performed to assess the clones proposed by computational analyses.

3 Materials and methods

3.1 Abbreviations

AP: Alkaline phosphatase

APS: Ammonium persulfate

CHO: Chinese hamster ovary

DTT: Dithiothreitol

EDTA: Ethylene diamine tetra acetic acid

ELISA: Enzyme-linked immunosorbent assay

HRP: Horseradish Peroxidase

IPTG: Isopropyl-β-D-1-tiogalattopiranoside

SDS PAGE: Sodium dodecyl sulphate polyacrylamide gel electrophoresis

scFv: single chain fragment variable

TMB: 3,3',5,5'-Tetramethylbenzidine

3.2 Materials

Stock solutions of antibiotics (Sigma) are prepared by dissolving kanamycin at 50mg/mL in water and ampicillin at 100mg/mL in water. Kanamycin and ampicillin stocks are filtered with 0.22µm filter device and stored at –20°C.

2xTY liquid broth: 16g bacto-tryptone, 10g bacto-yeast and 5g NaCl to 1L of ddH2O. Final pH 7.0. Agar plates are prepared by adding 1.5% bacto-agar to 2xTY broth.

2xTYA: 2xTY liquid broth supplemented with 100µg/mL ampicillin.

2xTYAG: 2xTY liquid broth supplemented with 100µg/mL ampicillin and 1% of glucose.

2xTYAK: 2xTY liquid broth supplemented with 100µg/mL ampicillin and 50µg/mL kanamycin.

Restriction endonucleases, T4 DNA ligase and buffers are purchased from New England Biolabs and used according to the manufacturer suggestions and standard molecular biology procedures.

GenElute™ Plasmid Miniprep Kit (Sigma) for plasmid DNA preparation, following the instructions of the manufacturer.

GenElute™ Gel Extraction Kit and **GenElute PCR Clean-Up Kit** (Sigma) for DNA purification from agarose gel and restriction reactions, respectively, following the instructions of the manufacturer.

The DNA Clean and Concentrator™-5 kit (Zymo Research) for purification and concentration of the ligation reaction, following the instructions of the manufacturer.

TG1 Electrocompetent Cells (Lucigen) and 1mm gap cuvette (BTX) for ligase transformation, following the instructions of the manufacturer.

Solution for phages precipitation: 20% (w/v) polyethylene glycol (PEG) 6000 in 2.5M NaCl. The solution is filtered through a 0.22µm filter before use.

PBS: 8g NaCl, 0.2g KCl, 1.44g Na2HPO4 and 0.24g KH2PO4 in 1L H2O, final pH 7.4.

PBS/Tween: PBS supplemented with 0.1% (v/v) of Tween-20.

2% MPBS: 2g non-fat milk powder /100mL PBS.

4% MPBS: 4g non-fat milk powder /100mL PBS.

5% MPBS: 5g non-fat milk powder /100mL PBS.

HRP-conjugated antibodies: purchased from Jackson ImmunoResearch.

AP-conjugated antibodies purchased from Sigma.

CCMB80 for preparation of competent *E. coli* cells

11.8 g $CaCl_2$ (dihydrate), 4 g $MnCl_2$ (tetrahydrate), 2 g $MgCl_2$ (hexahydrate), 10 mM K-acetate (pH 7), 10% Glycerol, H_2O to 1 L. Adjust pH to 6.4. Filtration with 0.2 µm filter.

TAE buffer for DNA electrophoresis on agarose gels

Tris Base 4.84 g, EDTA (pH 8 0.5 M) 2mL, Acetic acid 1.14mL, H_2O to 1L.

Transfer buffer

Glycine 2.9g, Tris Base 5.8g, SDS (SDS 10% 3.75mL) 0.37g, Methanol 200mL, H_2O to 1L.

SDS Running Buffer for protein electrophoresis on acrylamide gels

Glycine 14.4g, Tris Base 3g, SDS 1g, H_2O to 1L.

Loading buffer 6x for DNA samples (agarose gels)

40% glycerol, 60% H_2O, 0,1% (w/v) bromophenol blue.

Running gel for SDS-polyacrylamide gel

12% polyacrylamide mix (29% acrylamide, 1% bisacrylamide) 2mL, Tris 1.5M pH 8.8 1.3mL, SDS 10% 50µL, APS 50µL, TEMED 5µL, H_2O 1.6mL.

Stacking gel for SDS-polyacrylamide gel

5% polyacrylamide mix (29% acrylamide, 1% bisacrylamide) 330µL, Tris 1M pH 6.8 250µL, SDS 10% 20µL, APS 20µL, TEMED 2µL, H_2O 1.4mL.

Loading buffer 2x for protein samples (acrylamide gels)

100mM Tris pH 6.8, 4% SDS, 0.2% bromophenol blue, 2% β-mercaptoethanol, 20% glycerol.

Lysis Buffer

20mM Tris pH 8, 500mM NaCl, 0.1% Triton X100, 5mM imidazole.

Solution A

20mM Tris pH 8, 500mM NaCl, 5mM imidazole.

Elution Buffer

20mM Tris pH 8, 500mM NaCl, 300mM imidazole.

Destaining solution

30% Methanol, 10% Acetic acid, 70% H_2O.

Coomassie solution

225mL of Methanol, 25mL of Acetic acid, 0.625g Brilliant Blue, H_2O to 500mL.

3.2.1 Oligonucleotides:

All the following primers are in the direction 5'->3'

FULL INV2 VH SENSE	AGGTGGCGCCCATGCCGAGGTGCAGCTGGTGCAGTC
FULL INV2 VH ANTI	TGGTGCTAGCTGAAGAGACGGTGACCATTGTCCC
FULL INV2 VL SENSE	ATCGGCGCGCATGCCGACATCCAGATGACCCAGTCTCC
FULL INV2 VL ANTI	CCACCGTACGGTTTGATTTCCACCTTGGTCC
SENSESHUFFLED/RANDOM	P-TTTGATATCTGGGGCCAAGGG
ANTI CONSENSUS	AGCATCGGTACTATCATAGCTGCGGGGTCTCGCACAGTAATACATG
ANTI RANDOM	P-CGSAYCMNNMNNATCMNNMNNMNNCGGTCTCGCACAGTAATACATGG
ANTI SHUFFLED	GCTCGGGCTATAGCGATCGGTCGCATCTCTCGCACAGTAATACATGG
SENSE NEG	TTTGATTATTGGGGCCAAGGGACAATGG
ANTI NEG	ATAATAGCCGCCGCCGCTGCTGCCATGTCTCGCACAGTAATACATGG
ANTI_Ran_A/L/S/T/M/V	P-CGSAYCCRHMNNATCMNNMNNMNNCGGTCTCGCACAGTAATACATGG
PRO_INV2_ANTI	AGCATCTAGACTATCATAGTACGGCGGTCTCGCACAGTAATACATGG
MUT2	AGCATCTAGACTATCATAGTAGCGGGGTCTCGCACAGTAATACA
MUT2	AGCATCTAGACTATCATAGCTATGGGGTCTCGCACAGTAAT
MUT3	TGTATTACTGTGCGAGACCCCGCAGCTATGATAGTCTAGATGCT
MUT4	AGCATCGGTACTATCATAGTAATGGGGTC
MUT5	AGCATCTAGACTATCCGCGTAATGGGGTCTCGCACAGT
MUT6	AGCATCTAGGTTATCATAGTAATGGGGTCTCG
MUT7	AGCATCAATACTATCATAGTAATGGGGTC

Other primers used in this study are published in Sblattero and Bradbury,1998[149] and Sblattero et al., 2004[150].

All primers were purchased from Eurofins Scientific.

3.3 Bacterial strains

The bacterial strains used in this study were:

- *Escherichia coli* DH5αF' (F'*/endA1 hsdR17* (rK2 mK1) *supE44 thi-1 recA1 gyrA* (Nalr) *relA1 D (lacZYA-argF) U169 deoR (*F80*dlacD(lacZ)M15))* was used for phage propagation.
- *Escherichia coli BL21-CodonPlus(DE3)-RIPL strain* B F– *ompT hsdS* (rB– mB–) *dcm+* Tetr *gal λ*(DE3) *endA* Hte *[argU proL/*Camr*] [argU ileY leuW* Strep/Specr*]*

The phage used in this study is: Helper phage M13KO7

3.4 Method

- RNA from intestinal biopsy lymphocytes was provided by Fabiana Ziberna from *IRCCS materno infantile Burlo Garofolo, Trieste.*
- The phage display libraries were generated in collaboration with *IRCCS materno infantile Burlo Garofolo, Trieste.*
- The sequencing by Illumina technology was performed by Fabiana Ziberna from *IRCCS materno infantile Burlo Garofolo, Trieste.*
- Bioinformatic analyses, like HMM and Molecular docking were performed by Anna Vangone and Romina Oliva, respectively from *KAUST Catalysis center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia and Department of Sciences and Technologies, University "Parthenope" of Naples, Italy.*

The experimental procedures reported in the following chapters, both in method and common procedure, about phage display are detailed in Dal Ferro et al., 2019[142]*.*

3.4.1 Intestinal Biopsy Lymphocyte RNA preparation and library construction

The total RNA from intestinal biopsy lymphocyte from 55 previously untreated CD adult patients with high tiers of anti-human TG2 was prepared at IRCCS materno Burlo Garofolo. All the patients had HLA-DQ2 histocompatibility antigens. cDNA was synthesized by using random hexamers and SuperScript II reverse transcriptase (Life Technologies). Ig V regions were amplified by using specific

V region primers (as described by Sblattero and Bradbury, 1998) [149] and assembled into scFv[151] before cloning into pDAN5 vector[152].

3.4.2 Ligation and electroporation of the scFv library

Both phagemid cloning vector pDAN5[152] and purified scFv fragments library were sequentially digested with BssHII and NheI enzymes for 2h at 50°C and for 4h at 37°C, respectively. The ligation reaction was prepared in 70µl volume with 1 µg of double-digested and purified vector and 0.5 µg of double digested and purified scFv (phagemid: insert molar ratio of 1:3), incubating at 16°C overnight. The reaction was purified and concentrated in 7µl using a commercial kit (Zymo research). The ligation mix was electroporated into TG1 electrocompetent cells. Transformations were pooled, plated on three 2xTY agar plates with ampicillin (20 cm diameter each) and grown O/N at 28°C. Two dilutions were also plated to estimate library diversity. The next day colonies were scraped up in 2xTY 20% glycerol and frozen in 1mL.

3.4.3 Recombinant TG2 production

Recombinant hTG2 was expressed in pET-28b vector by *E. coli* BL21-RIPL strain, while mTG2 was expressed in pTrcHis-B (Invitrogen) vector by DH5αF' cells. We followed the following procedure. Inoculate a single colony from a fresh agar plate added with 50µg/mL kanamycin or 100µg/mL ampicillin (kanamycin resistance marker for hTG2, whereas ampicillin was used for mTG2) and grow cells O/N at 30°C in 2xTY liquid broth added with the specific resistance drug and 1% glucose. Dilute cells 1:100 (in 200-800mL) in 2xTY liquid broth with the specific resistance drug and grow at 28°C until $O.D._{600}$ 0.5, induced with 0.2mM IPTG and incubated at 25°C for about 5 hours. Centrifugate 7000 rpm for 15' at 10°C and resuspend the pellet with lysis buffer (10mL/g/bacteria) and freeze at -80°C O/N. The next day, thaw the resuspended pellet on ice and add lysozyme (1mg/g/bacteria) and incubate shaking on ice for 30' until solution become very viscous. Add DNAse 50µg/mL and stir 20'on ice. Sonicate the solution on ice, 1' on, 15" off, 30" on, 15" off, and 30" on at 30% amplitude on ice. Centrifugate at 11000rpm 30' at 10°C. Collect supernatant and filter it with 0,45µm filter.

3.4.4 Recombinant TG2 purification

The supernatant was collected, and the protein was purified by affinity chromatography of the bacterial extract on Ni-NTA Superflow (Qiagen) resin. Ni-NTA resin binds specifically His-tag in the N-term of recombinant TG2 protein. The purification was done as follow. Prepare a purification

column with Ni-NTA resin (we used around 400µL of resin, for 400mL of bacteria culture). Wash the column with 10mL of solution A and discard it. The column is enclosed by ice. Then, apply sample and store the flow-throughput. Wash the column with solution A (20mL for 400µL of resin). Elute the protein with 5mL of Elution buffer and collect about 12 elution fractions of 200µL each. Remaining buffer is used to wash. Wash the column with 20mL of solution A.

3.4.5 Protein dialysis

After checking protein concentration and degradation by SDS-PAGE, more concentrated fractions are chosen and joined. Centrifugate at 14000 rpm 15' at 10°C to eliminate micro aggregates. Charge the supernatant in a dialysis membrane (Spectra/Por dialysis membrane, Spectrumlabs), then close the membrane with clamps and dialyze the supernatant O/N in 1000 more volumes than the dialyzed fraction of PBS at 10°C. Recover the dialyzed sample and transfer it in a 1.5mL tube. Centrifuge at 14000 rpm for 15' at 10°C and collect sample. Protein concentration and degradation are checked by SDS-PAGE.

3.4.6 Rescuing phagemid particles from library

One aliquot of the library was diluted in 11 mL (1 mL for $O.D._{600nm}$ calculation) of 2xTYAG medium and grown with shaking (200rpm) at 37°C. After reaching $O.D._{600nm}$ 0.5, bacteria, concentrated as $5x10^8$/mL, were infected with a 20-fold excess of helper phage and left at 37°C for 45 min, standing. Bacteria were centrifuged for 15 min at 2800 x g, supernatant discarded, and cells pellet dissolved in 40 mL of 2xTYAK medium. Bacteria were grown with shaking (200rpm) at 28°C overnight. The next day, bacteria were centrifuged for 20 min at 6300 x g, 10 °C. The supernatant containing the phage particles were mixed with 1/5 volume of PEG/NaCl solution and left on ice for 45 min.

3.4.7 Panning of the library on TG2

The day before selection, 10µg of TG2 antigen in 1mL of PBS was coated in an immunotube (Nunc, Maxi Sorp immunotube) and incubated overnight at 4°C. The day of selection, the antigen solution was discarded, and unspecific sites blocked adding 4 mL of 2% MPBS to the immunotube for 1h. At the same time, phages particles were saturated in 1 mL of 2% MPBS. After blocking, the immunotube was washed once with PBS and phages in milk added to the tube. Phages were incubated on the antigen for 30 min on rotation and 1.5hrs standing, followed by 10 washes of the immunotube with PBS/Tween and 10 with PBS to remove unbound phages. The antigen-linked phages were eluted by adding 1 mL of DH5αF' cells at $O.D._{600\ nm}$ 0.5 to the immunotube and incubating at 37°C for 45min, standing. At the same time, 0.5mL

of cells were infected with the starting phages particles. After incubation at 37°C, the cells infected with the starting phages were plated in serial dilutions on a 2xTY agar plate with ampicillin to estimate the number of infecting particles (input). The cells infected with the selected phages (output of the selection) were plated on a 30cm 2xTY agar plate with ampicillin. A small volume of the output was used to plate serial dilutions and estimate the number of selected phages. Plates were incubated O/N at 28°C and the next day the output colonies were scraped up in 2xTY 20% glycerol and frozen in 1mL.

3.4.8 Screening of the phages on the antigen by phage-ELISA

96 random colonies coming from either bio panning of the starting library from IBL from CD patients or random libraries, were grown in microtiter plate in 120µl/well of 2xTYAG medium O/N at 37°C, shaking (200rpm). The next day, a copy plate was prepared from the first by transferring a small volume (5-10µl) of culture from each well into the corresponding wells of a second microtiter plate containing 2xTYAG medium. The copy plate was grown at 37°C for 3hrs, shaking (200rpm). After growing, 50µl of 2xTYAG medium containing the helper phage (helper phage to bacteria ratio 20:1) was added to each well and incubated 45min at 37°C, standing. The bacteria were spun down by centrifugation for 30min at 1000 x g, the supernatant discarded, and each pellet resuspended with 120µl of 2xTYAK medium. The copy plate was grown O/N at 28°C, shaking. A 96 plate (Nunc, MaxiSorp) was coated with 70µl/well of the hTG2 antigen in PBS (final concentration 5µg/mL) and incubated O/N at 4°C. The next day, the coating solution was discared and unspecific sites blocked with 130µl/well of 2% MPBS for 1h. At the same time, the copy plate was centrifuged for 30min at 1000 x g. After blocking, 35µl of the copy plate supernatants (containing the phages) were transferred from each well on the corresponding wells of the antigen plate and mixed with 35µl of 4% MPBS. The phages were incubated for 1.5hrs and the unbound particles removed by three washes with PBS/Tween and three washes with PBS. HRP-conjugated anti-M13 mAb (GE Healthcare or Progen) was used as indicated by manufacturer's instructions for antigen-linked phages detection. After washing (three times with PBS/Tween and three times with PBS), 70 µl of TMB were added to each well. The reaction was developed for 5-20min and blocked with 30µl/well of 2N H_2SO_4. Absorbance was measured at 450nm.

3.4.9 Analysis of the positive clones by PCR

The clones which resulted positive from phage-ELISA screening were firstly analyzed by PCR to confirm the presence of the scFv. The amplification of each clone was performed with VLPT2 and VHPT2 primers in 20µl of reaction volume, adding 1 µl of bacteria from master plate as template (the denaturation step allows the release of DNA from bacteria) and PCR BIO HiFi polymerase (PCR BIOSYSTEM) according to

manufacturer's instructions. The PCR products were loaded on a 1.5% agarose gel and those presenting the correct length were sequenced. Finger printing was assessed only for the starting library, but for random primer, finger printing is not available to screen different clones.

3.4.10 ScFv with different CDRH3

The oligo designed consider the codon usage of *E. coli*. Moreover, the random positions are codified by NNK, which allow the expression of all the amino acids, whereas the option between 2 or more amino acids was generated to allow the specific presence of those amino acids. ScFv with different CDRH3 was generated by inverse PCR using the oligo listed before. For the PCR with degenerated oligo, phosphorylated oligos were designed. Amplicons were purified from gel and ligated. *E. coli* cells were transforming with the ligated vector. Bacteria were placed on 2xTYA O/N and the next morning single colony was collected in a multiwall. Phage ELISA was performed as described in 4.4.8.

3.4.11 Production of the positive clones in different antibody formats

Sub-cloning of the scFv in eukaryotic expression vector. The pMB-SV5 vector[153] and pDAN vector, with the scFv of interest, were digested with BssHII and NheI. After digestion, the fragment of interest has been purified by agarose gel. The ligation reaction was prepared in 10-20 µl of reaction volume with 100 ng of double-digested and purified vector and 50 ng of double digested and purified scFv (vector: insert molar ratio of 1:3), incubating O/N at 16°C. Each ligation reaction was transformed into 50 µl of chemically competent bacteria. After O/N growing on agar plate with ampicillin, the obtained colonies were checked by PCR, using scFv external primers.

IgG and IgA have been generated using the 1 vector strategy described by Fang at al., 2005[154]. The vector had been modified from pcDNA 3.1 vector (Invitrogen) in the host laboratory (data not shown) and primers specific for amplify the VH and VL have been designed and listed before. Primers carry the restriction site for the cloning in the vector. First the VL has been cloned and after the VH had. Once the vector obtained, all the expression sequences were sequenced.

3.4.12 antibodies analyses

IMGT had been used for the analyses of the variable region of the antibodies[101].

The highlighter tool [155] allowed the analyses of the mutation and phylogenetic analyses.

Weblogo.barkley[156] has been used for the generation of Logos.

3.5 Common procedure

3.5.1 PCR

PCRBIO HiFi Polymerase (PCRBIOSYSTEMS) was used. See manufacture instruction.

3.5.2 DNA electrophoresis on agarose gels

Agarose (Sigma) gels with a concentration of 1.5% in TAE buffer were used to separate PCR products with amplicons of about 1kb or below; 0.8% agarose gels were used to separate plasmid DNA preparations, before and after digestions. 1.5µL of EuroSafe nucleic acid staining solution (20000X, EuroClone) was added to 30ml of agarose gel. Agarose gels with a concentration of 3% in TAE buffer were used for DNA-fingerprinting and 3µL of Green gel safe nucleic acid staining solution (10000X, Fisher Molecular Biology) was added to 30ml of agarose gel. 100 base-pairs plus and 1KB molecular weight markers were purchased from Fermentas.

3.5.3 SDS-PAGE and western blot

Acrylamide gels were prepared at 4% stacking and 12% resolving acrylamide concentration by Mini-PROTEAN, Tetra Vertical Electrophoresis Cell (Bio-rad) with a thickness of 0.75mm. In particular, the stacking was prepared with 30% acrylamide/bis-acrylamide solution (Sigma), 130 mM Tris/HCl pH 6.8, 0.1% SDS, 0.1% APS and 0.01% TEMED (N,N,N',N'tetrametiletilendiamina) (Sigma). The resolving was prepared with 30% acrylamide/bis-acrylamide solution (acrylamide 29.2 g/ bis-acrylamide 0.8 g in 100mL of H2O), 400 mM Tris/HCl pH 8.8, 0.1% SDS, 0.1% APS, 0.01% TEMED 0,01%. The protein samples were mixed with denaturing buffer (tris-HCl 61.5mM pH 6.8, SDS 2.5%, Glycerol 10%, bromophenol blue 0,0025%, β-mercaptoethanol, ratio sample to buffer 6:1), boiled for 5 minutes and loaded on gel. The run was conducted in running buffer (50 mM Tris, 384 mM glycine, 0.1% SDS) at 15mA until proteins reached the running gel, then the amperage was increased to 20 mA. The proteins separation on gel was monitored through pre-stained molecular standard (Smobio).

For western blot analysis, the proteins were transferred to nitrocellulose membrane (GE Healthcare, Amersham UK) using a Mini Trans-Blot module (Bio-rad). The transfer was performed at 250 V for 45 min in transfer buffer (25 mM Tris, 192 mM Glycine, 20% Methanol, pH 8.3). The membrane was blocked in 2% MPBS for 1h at RT, washed in PBS and incubated with primary antibody diluted in 2% MPBS (according to manufacturer's instruction, for commercial Abs) for 1h at RT. After three washes in PBS/Tween and three washes in PBS, the membrane was incubated with alkaline phosphatase-

conjugated antibody diluted in 2% MPBS (according to manufacturer's instruction) for 1h at RT. After washes (as described above), the membrane was developed with 0,3 mg/ml of BCIP (5-Bromo-4-Chloro-3-Indolylphosphate, Sigma) and 0,6mg/ml of NBT (Nitro Blue Tetrazolium, Sigma) in 10 mL of alkaline phosphatase buffer (100 mM Tris, 0.1 M NaCl, 5 mM $MgCl_2$).

3.5.4 ELISA

Costar ELISA strips were coated with hTG2, mTG2 or BSA as a control protein at 5µg/ml O/N at 4°C. One wash in PBS was performed. Wells were blocked with 2% MPBS at RT for 45'. One wash in PBS was performed. ScFv-Fc or IgA or IgG antibodies diluted 1:500 in 2% MPBS were incubated for 1 hour at RT. 3 washes in PBST and 3 washes in PBS were performed. Anti-human IgG HRP conjugated (Jackson ImmunoResearch Laboratories) diluted 1:5000 in 2% MPBS or anti-human IgA HRP (Sigma-Aldrich) diluted 1:2000 were incubated for 1 hour at RT, followed by 3 washes in PBST and 3 washes in PBS. Immunocomplexes were revealed with the chromogenic substrate Tetrametilbenzidine (TMB, Sigma), the reaction was stopped with H2SO4 and the plate was read at OD450 in a microplate reader (NanoQuant infiniteM200proTecan).

3.5.6 Phage ELISA

Single clones were grown, in agitation, in 1mL of 2xTY added with ampicillin and 1% glucose at 37°C to $O.D._{600}$ 0.5 when they were infected with helper phage without agitation at 37°C for 45'. Bacteria were then centrifuged at 4000 rpm for 15', the supernatant was discarded and the pellet was resuspended in 1ml of 2xTY added with ampicillin and kanamycin and grown O/N at 28°C. As previously described, O/N coating, blocking and washing were performed. Bacteria were centrifuged for 20' at 5000 rpm at 10°C and supernatant with phage-particles of individual clones were added to the wells in 1:2 ratio with 4% MPBS and incubated for 1 hour and 30' at RT, followed by 3 washes in PBST and 3 washes in PBS. Wells were added with horseradish peroxidase conjugated anti-M13 monoclonal antibody (GE Healthcare) diluted 1:5000 in 2% MPBS and incubated for 1 hour at RT, followed by 3 washes in PBST

3.5.7 Phage Competition ELISA

Preparation of phage-particles as mentioned in "Phage ELISA". Costar ELISA strips were coated with hTG2, mTG2 or BSA as a control protein at 5µg/ml and incubated O/N at 4°C. Wells were blocked and washed. Reference antibodies in scFv-Fc format were added, diluted in 1:1000 in 2% MPBS for the antibody recognizing Epitope 1, whereas the others were diluted 1:250 and incubated for 1h at

RT. 3 washes in PBST and 3 washes in PBS were performed. 35µL of solution from every well was discarded. Bacteria were centrifuged at 5000 rpm for 20' at 10°C and 35µL of supernatant of individual clones were added in each well and the supernatants were incubated for 1 hour and 30' at RT. After 3 washes in PBST and 3 washes in PBS, wells were added with horseradish peroxidase conjugated anti-M13 monoclonal antibody as described previously. Immunocomplexes were revealed as described previously.

3.5.8 Transfection of CHO cells

Chinese Hamster Ovary (CHO) Expi cells were transfected in 24-well plate following the instructions of the manufacturer (Thermo Fisher Scientific).

3.5.9 Antibodies purification from supernatant

ScFv-Fcs and IgGs were purified with POROSE Protein A Affinity Resin: MabCapture A Select whereas IgAs were purified with CaptureSelect IgA Affinity Resin (both resins from Thermo Fisher). Resin preparation: completely resuspend the resin by inversions and transfer the desired volume of resin (POROSE Protein A Affinity Resin: MabCapture A Select has a dynamic binding capacity of 37mg/mL, whereas CaptureSelect IgA Affinity Resin Select has a dynamic binding capacity of 16mg/mL) in a 15mL tube with 10mL of Tris/HCl 100mM pH8 and centrifuge at 800 rcf for 5' at 10°C, discard the supernatant and repeat the wash. Sample preparation: adjust the medium pH with Tris/HCl 1M pH 8 or with PBS 10X for IgA purification (ratio 1:10 with the final volume of supernatant) and add Tris/HCl 100mM pH8 (or PBS 1X for IgA purification) to 10mL final volume and put on ice on a gently agitation for 30'. Resin and sample incubation: filtrate sample with 0.2µm filter and transfer the resin into the medium containing tube and mix gently. Antibody elution: prepare 1.5mL tubes (according to the number of elution fractions to collect) with 20µL of Tris/HCl 1M pH8 (60µL for the IgAs) (added to neutralize the glycine acid pH) and put on ice. Transfer the sample into the column, collect the flow-through and wash the resin loading 10mL of Tris/HCl 100mM pH8 (or PBS 1X for IgA purification) into the column and collect the flow-through. Load 5mL of glycine 100mM pH3 (pH2 for IgAs) into the column. Collect multiple elution fractions of 200µL into the previously prepared tubes with 20 or 60µL of Tris/HCl 1M pH8, gently mix and put on ice. Leave the excess of glycine buffer to wash the resin. Wash the column with 10mL of Tris/HCl 100mM pH8 (or PBS 1X for IgAs). Notes: avoid the resin to dry during all the steps.

3.5.10 Preparation of chemical competent *E. coli*

DH5αF' bacteria strain plated and grown O/N at 37°C. One colony inoculated in 2mL of liquid broth without resistance drug. 1mL of bacteria culture inoculated in 100mL of 2xTY liquid broth at 37°C without resistance drug to $O.D._{600}$ 0.4. Bacteria were chilled in ice for 10' to stop the growth, centrifuged at 10°C for 15' at 2500 rpm and the supernatant was discarded. The bacterial pellet resuspended in 8mL of CCMB80 solution and put in ice for 20'. After centrifuging for 15' at 10°C at 2500 rpm, supernatant was discarded, bacterial pellet resuspended in 4mL of CCMB80 and dispensed in 80µL aliquots. Competent cells were immediately used or stocked at -80°C.

3.5.11 Bacterial transformation

10µL of ligation reaction mixture or 100ng of plasmid preparation were transferred into a tube containing 80µL of competent cells. The mixture was incubated in ice for 20'. Heat shock was applied at 42°C for 1 minute; bacteria were then resuspended in 500µL of 2xTY liquid broth and allowed to grow at 37°C in absence of selective antibiotic for 1 hour. Bacteria were then plated on antibiotic-containing agar plates and grown O/N at 30°C.

4 Results and discussion

The phage display libraries described in the following chapter were generated in collaboration with *IRCCS materno infantile Burlo Garofolo, Trieste.*

4.1 Isolation and analysis of TG2-specific IGHV5 antibodies from Celiac gut biopsy by phage display

As previously described, Phage display technology is a powerful tool to study antibodies, whereas, immune libraries are the best choice to study autoimmune diseases[52,150]. Since CD is an autoimmune enteropathy, immune libraries were generated. ScFvs have been amplified from CD patients' intestinal biopsy lymphocytes. These antibodies repertoires were then used to study the immunological response in CD patients[52]. Specifically, it was previously shown in the host laboratory that anti-TG2 specific antibodies (in the scFv format) can be isolated from phage antibody libraries constructed from lymphocytes present in CD gut biopsy specimens[52]. Therefore, libraries from 55 untreated individual patients (Table 1) have been constructed, thanks to the collaboration with *IRCCS materno infantile Burlo Garofolo, Trieste,* and in particular of Fabiana Ziberna. CD patients are 21 males and 34 females, aged between 1 and 56, and 2 and 59 respectively (Table 1). The RNAs from 55 CD patients' intestinal biopsy lymphocytes (Fig. 13) have been extracted and retrotrascribed, and the cDNA have been syntetized[52]. Since, *IGHV5* gene family has been reported to be the most gene present in anti-TG2 antibodies from CD patients[113,118,119], primers specific for this gene family were used[149] to limit the antibodies diversity of VH chains to the members of the VH5 family (Fig. 13). These primers amplify both *IGHV5* and *IGHV1* gene families, therefore the VHs amplified belong both to VH5 and to VH1. The amplification strategy was formerly described in (Sblattero et al., 2004)[150]. To deeply investigate the role of VH chains in TG2 binding, the immune libraries have been constructed with 4 VL chains, since it was demonstrated that VH5 of anti-TG2 scFv chains require specific VL chains to bind the antigen[150]. These specialised VL chains used for the

Sex	N. sample	Age (years)
MALE	21	13 (1-56)
FEMALE	34	19 (2-59)
TOTAL	55	17

Table 1. ***CD patients' characteristics****. Intestinal biopsy from 55 CD patients have been collected. The sex and average of the age of patients are listed.*

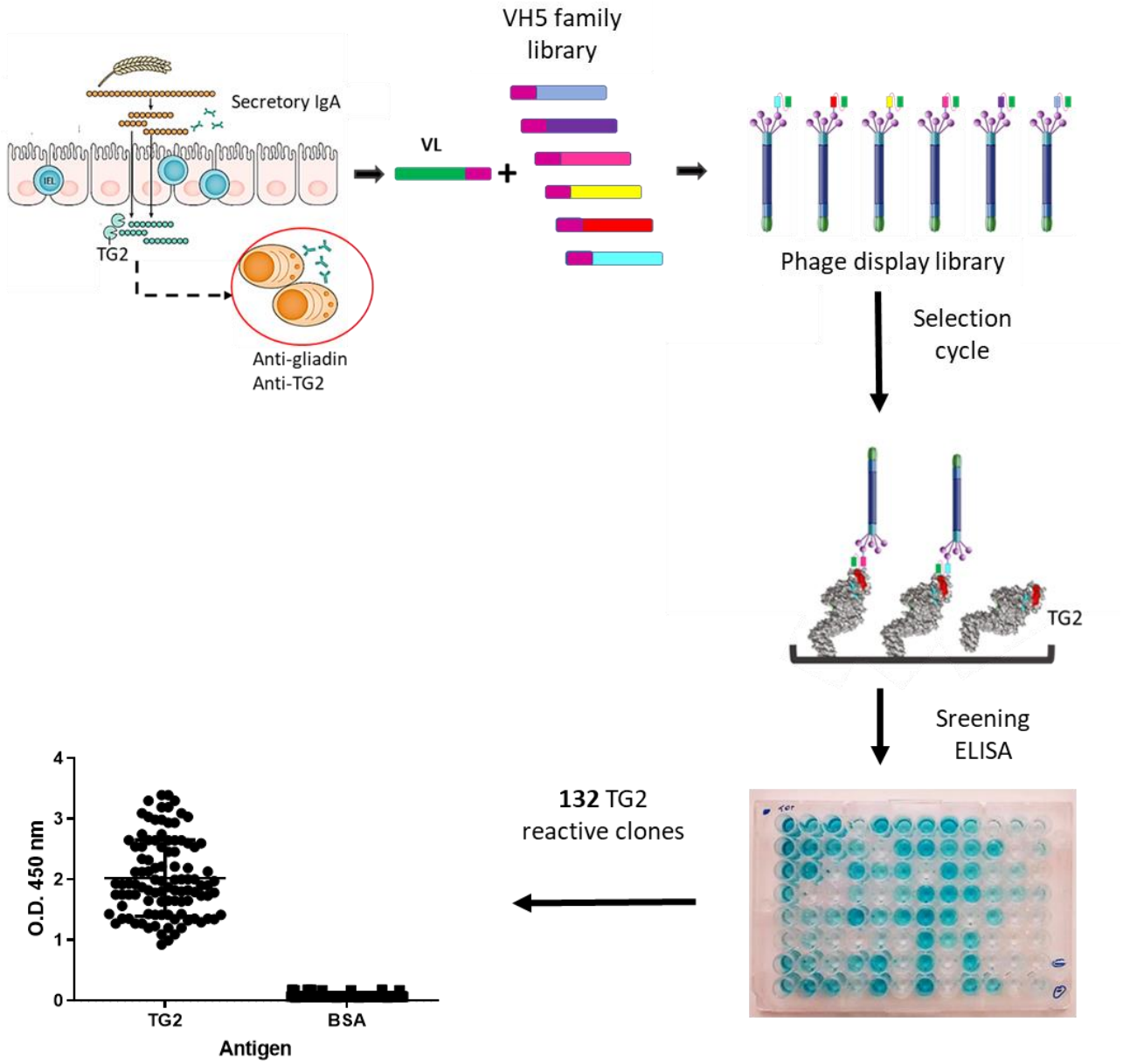


Figure 13. ***Phage display libraries generation and selection****. Intestinal biopsy lymphocytes are collected from CD patients. The VH5 family library is generated and VH5s are assembled with a fix VL in order to generate a phage display library. This library is characterized by a single VL and a library of VHs (only one VL is reported on the figure but in our study 4 VLs have been used). Libraries were pooled and a selection cycle on TG2 was performed. Selected clones were assessed. Phage ELISAs in 96 multi were performed to screen positive clones (in blue). Positive clones were tested to confirm the specificity for TG2. BSA was used as control to check the polyreactivity.*
(Intestinal reproduction with anti-TG2 is edited from Lindfors et al., 2019. TG2 is edited from Iversen et al., 2013).

libraries generation belong to anti-TG2 antibodies previously identified in the host laboratory[52,150]. Names of the previously identified clones are reported in Table 2, and VLs gene families are also listed (Table 2). Since not only a preferential VH usage but also a preferential VL gene families usage has been observed (Fig. 9)[119] has been observed, the libraries have been generated with preferring VH:VL pairings. The *IGKV1-5* gene family has been demonstrated to be the VL most paired with

IGHV5, therefore, libraries with two IGKV1-5 VLs were generated. The *IGKV1-39* gene family has been identified to pair with IGHV5, hence other libraries were generated using an *IGKV1-39* gene family. Only one VL belonging to the λ family (*IGLV1-14)* was used. The VH5 chains were PCR assembled[151] in a scFv format with the 4 different VLs, as described in Sblattero et al., 2004[146]. The phagemid vector used for the cloning procedures is pDAN, formerly described in Sblattero and Bradbury, 2000[157]. Phage display libraries were constructed (Fig. 13). Each library was subjected to a selection cycle with human TG2 as the target antigen (Fig.13) and specific clones were identified by ELISAs in a 96 multi wells plates (Fig. 13). After screening over 5000 single clones a set of 132 TG2-positive antibodies were isolated. Consecutively, positive clones were tested to verify the specificity for TG2. The ELISAs of specificity of these clones are reported in Figure 13. These clones were finally sequenced and analysed using IMGT[101]. 102 clones are characterized by *IGHV5* gene family usage, whereas 30 positive clones bear the *IGHV1*. We focus on the scFvs bearing IGHV5. Although the VH5 repertoire was initially equally paired to the 4 different VLs, the sequencing revealed that after selection on TG2, the 102 isolated antibodies showed a clear bias in VL gene usage. 80% of the TG2 reactive antibodies use one of the two VLs containing *IGKV1-5* gene and 17% of the clones were paired with the VL containing *IGKV1-39* and only 3 clones out of 102 contains the *IGLV2-14* (Table 2). These results are in agreement with previous reports[119,158–160], confirming a preferred κ chain usage in anti-TG2 antibodies from CD patients and preferential VH:VL pairing.

Clone	VL gene	# seqs
***2.8**	***IGKV1-5***	59
***4.1**	***IGKV1-5***	22
***2.6**	***IGKV1-39***	18
***2.12**	***IGLV2-14***	3

Table 2. ***VL characteristics***. *Name of the TG2 specific scFvs identified previously in the host laboratory. VL genes identified in the clones are reported, as well as the number of TG2 specific clone isolated from the libraries characterized by the usage of IGHV5 identified with the listed VL. (VLs from Marzari et al., 2001)*

4.2 Analyses of the mutations in TG2-specific *IGHV5* gene

We next analysed the VH5 repertoire of the 102 TG2-positive clones and in order to compare these clones to antibodies obtained by phage display we collected different sets of VHs deriving from the same 55 libraries. 30 clones are the TG2 positive bearing VH1, which is called TG2 positive (+) non VH5. Moreover, 98 negative clones have been sequenced, 53 clones bearing VH5 and 45 clones bearing VH1 (called non VH5).

Recapitulating, the clones that we analyse are:

(i) TG2-positive clones VH5 (102)

(ii) TG2-negative clones VH5 (53 clones);

(iii) TG2-positive clones non VH5 (30 clones);

(iv) TG2-negative clones non VH5 (45 clones).

Furthermore, we collected 9242 VHs belonging to the VH5 family derived from the sequencing of a phage display library build in the host laboratory using IgM from circulating peripheral blood lymphocytes (unpublished data). The total number of mutations presenting the VH5 obtained from unselected naïve IgM library as expected showed a median of 4 mutations (Fig. 14). The mutation level was reduced also in the in the selected populations of VH5 TG2-positive with a comparable number (5) of mutations per antibody (Fig. 14). These data are confirming the coherence of data

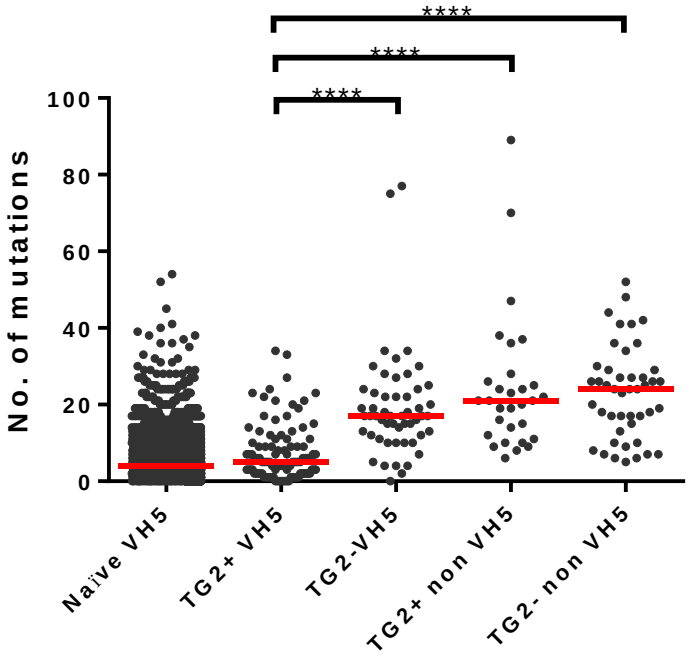


Figure 14. ***Characterization of VH from anti TG2 antibodies isolated with phage display technology.*** *Number of mutations of (9242) IGHV5 from a naïve phage display library, TG2+ (102) IGHV5 genes, TG2- (53) IGHV5, TG2 positive (+) (30) non IGHV5 and TG2 negative (-) (45) non IGHV5. TG2 positive and negative clones were from CD patients' intestinal biopsy library, which was generate and selected as phage display library. Median values were determined using unparametric t test. ****p<0,0001.*

from phage derived antibodies with previous results obtained with single cells sequencing[113,119]. We then confirmed that this limited number of somatic mutations is a feature of naïve clones and TG2-specific clones by showing that the levels of somatic hypermutations were significantly higher for all the other populations of gut derived antibodies. TG2-negative VH5 clones from the immune library have a median of 17 mutations, and TG2 positive or negative not containing VH5 genes have

respectively a median of 21 and 24 mutations (Fig. 14). The low number of mutations could be ascribed to a competitive advantage of B cells that target the epitope binding by *IGHV5-51* gene family IGs during B cell activation[117]. Moreover a low mutation rate of the variable regions of the Abs has been reported even for anti-gliadin Abs from CD patients[122], suggesting that it could be a characteristics of some Abs, with a determined gene family usage, produced by CD patients. These few mutations are characteristic for antibodies from CD patients. In fact, for example, Abs specific for rotavirus are characterized by a higher number of mutations[161]. Furthermore, others viral disease, such as influenza and HIV, which share with CD the chronic exposure to the antigen, are charachterized by a high degree of somatic hypermutations[162,163]. It could be possible that the Abs have a so high affinity, even without a high number of mutations, that it drives an expansion leading to the production of anti-TG2 antibodies with low mutation and with the usage of the same *IGHV5* gene.

After the identification of the number of mutations we focus on the analyses of the FRH1-3 and CDRH1-2, which is the region codified by *IGHV5* gene, before the CDRH3. IMGT analyses show the number of mutations by comparing each sequence with the VH region of the closest germline V-gene[101]. To identify the position of the mutations an alignment was performed with a bioinformatics tool (the highlighter tool)[155]. The naïve sequence of *IGHV5* gene is chosen as reference sequence. The sequences are sorted based on the number of silent/non silent mutations. The most similar sequence is placed at the top of the resulting graph, and the least similar at the bottom. The alignments are shown on Figure 15A (TG2 positive sequence) and B (TG2 negative sequences). Both sequences are characterized by the presence of IGHV5 segment. The accumulation graphic, below each alignment (Fig. 15C and D) summarizes the number of mutations of each sequence (they summarize the accumulation of the alignment of the sequences reported respectively on Figure 15A and B). These accumulation graphs show the number of mutations from the least to the most mutated VH sequences, to the sequences more mutated. The panel of the TG2 positive sequence shows limited non-silent mutations (in red), and some of them are recurrent.

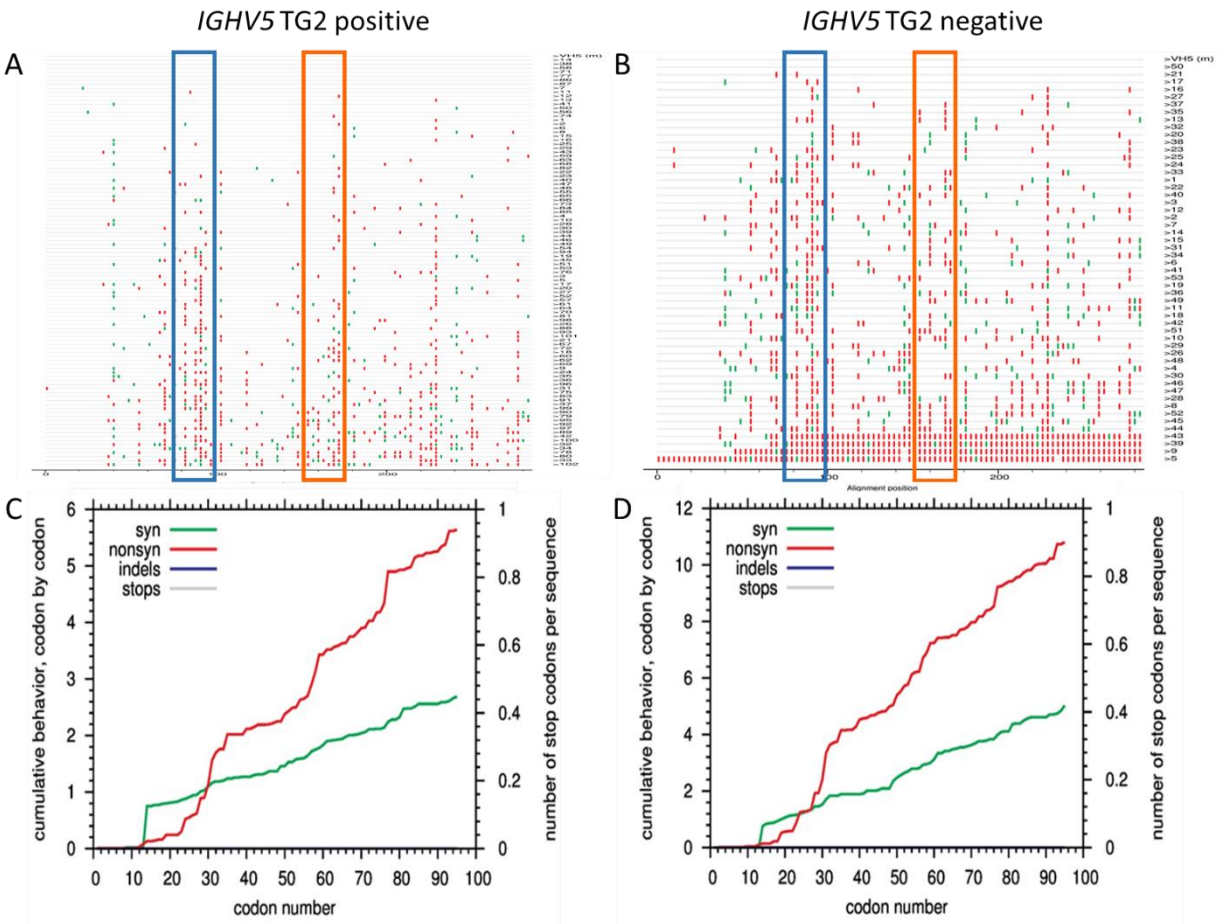


Figure 15. ***Mutation in VH5 genes. A and B)****The two columns represent the sequences, the blue boxed enclose CDRH1 and orange boxes CDRH2* ***A)*** *the VH5 TG2 positive and* ***B)*** *the VH5 TG2 negative. The silent mutations are reported in green, whereas the non-silent ones are in red. For both groups of sequences, a naïve sequence is used as reference. The link to the tool is reported.* ***C and D)*** *The mutations are summarized in the graphics below each alignment* ***C)*** *the VH5 TG2 positive and* ***D)*** *VH5 TG2 negative. The colours are the same as the top panels: silent mutations are in green, whereas non-silent ones are in red.*
https://www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter_top.html

This is in agreement with previous observations, where phylogenetic analysis demonstrated that clones with a particular mutation expanded more than others, possibly for the increased affinity for TG2[119]. For TG2 negative sequences (Fig. 15D), a higher number of mutations are shown than VH5 TG2 positive sequences (Fig. 15C). These data suggest that the level of mutations reported in Figure 14 are not localized only in the CDRH3 region, but also in the FR1-3 and CDRH1-2 (Figure 15 A B). CDRH1 are in the blue boxes, whereas CDRH2 are in the orange boxes on Figure 15A-B. However, comparing the total amount of mutations in the IGHV region and the mutations localized before CDRH3, these data suggest that for TG2+ the few mutations are present both in the CDRH3 and in the VH, whereas negative sequences show a higher number of mutations in the VH, but a major variability is suggested even in the CDRH3.

Since these data suggest that the variability is present in both FR1-3 and CDRH1-2 regions, we first focus on the analysis of these regions.

4.3 Anti TG2 positive CDR-H1 and H2 sequence analysis

We have analysed the length, sequences and conformations of the three hypervariable loops representing the antibody binding site. The hypervariable loops are indeed known to exhibit a finite repertoire of conformations, known as canonical structures [105,164–166]. First, we focused on the analyses of CDRH1 and CDRH2. To characterize the amino acids sequences of IGHV5 clones, we generate the "sequence logos" of CDRH1 and CDRH2 using Weblogo.barkley[156]. Sequence logos are graphical representations of the patterns within a multiple sequences alignment. Sequence logos provide a rich and precise description of sequences similarities and can rapidly reveal significant features of the alignment otherwise difficult to perceive. Each logo consists of stacks of letters, one stack for each position in the sequence. The overall height of each stack indicates the sequence conservation at that position, which is measured in bits, whereas the height of symbols reflects the relative frequency of the corresponding amino (or nucleic) acid at that position[156]. As can be seen on Figure 16, we found that, for TG2-positive clones, although some mutations are present, all H1 and H2 are highly conserved. The TG2 negative sequence shows more variability for some positions. For comparison, the logo of a panel of naïve sequences is reported. These data suggest a low variability of CDRH1 and CDRH2 between the *IGHV5* gene family. This low frequency in the amino acid changes is in agreement with previous data[119]. However, the reported molecular dynamic simulations have suggested an interaction of 679-14-E06, (an anti TG2 bearing IGHV5, previously

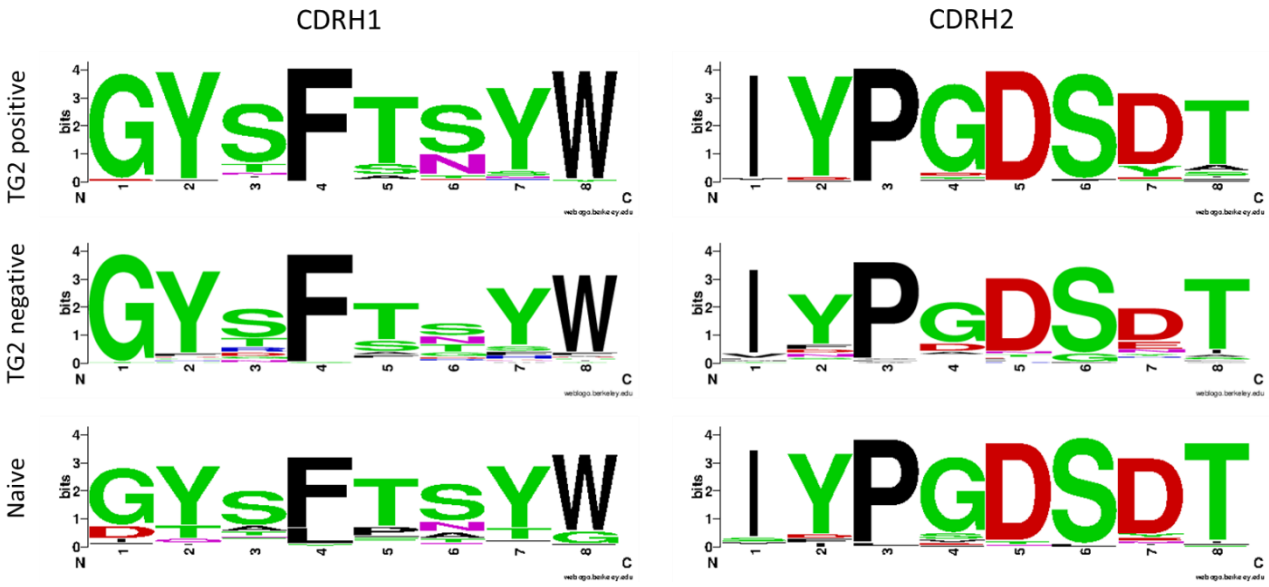


Figure 16. ***Characteristics of both VH CDR1 and VH CDR2****. Logos represent CDRs identified in TG2 positive clones, TG2 negatives clones and naïve clones. Logos are generated with WebLogo Berkley (Crooks et al., 2004) (https://weblogo.berkeley.edu/logo.cgi)*

described in chapter 1.4.3 Epitope on TG2) with TG2 via both CDRH2 and CDRH3 for the H chain[65]. These data suggest that CDRH2 could be involved in the TG2 binding. In fact, it is well conserved but the CDRH3 could play a predominant role due to the finding of a conserved sequence even in a TG2 negative clone. (Here the VL are not considered, because the VL for both positive and negative TG2 clones are the same 4, derived from TG2 positive clones). These data suggested us to investigate the length and composition of the CDRH3 loops.

4.4 TG2-specific IGHV5 Abs display CDRH3 length and *IGHJ* gene usage bias

We focused on the composition of the H3 loop, usually playing a key role in the antigen recognition. Moreover, for a specific antibody anti-TG2 characterized by an *IGHV5* gene, the role of H3 in the interaction with TG2 has been identified by Chen and colleagues[65]. First, we focused on the analyses of D and J genes that, joined to V gene, generate the CDRH3, by analysing D and J genes families pairing. Analysing the JH usage there is a clear difference between VH5 TG2-positive (TG2+) clones and VH5 TG2-negative (TG2-) and naïve clones (Fig 17A). 56% of the TG2+ clones use IGHJ3 gene while both the naïve library and the control clones were using preferentially the IGHJ4 (54%, Fig.

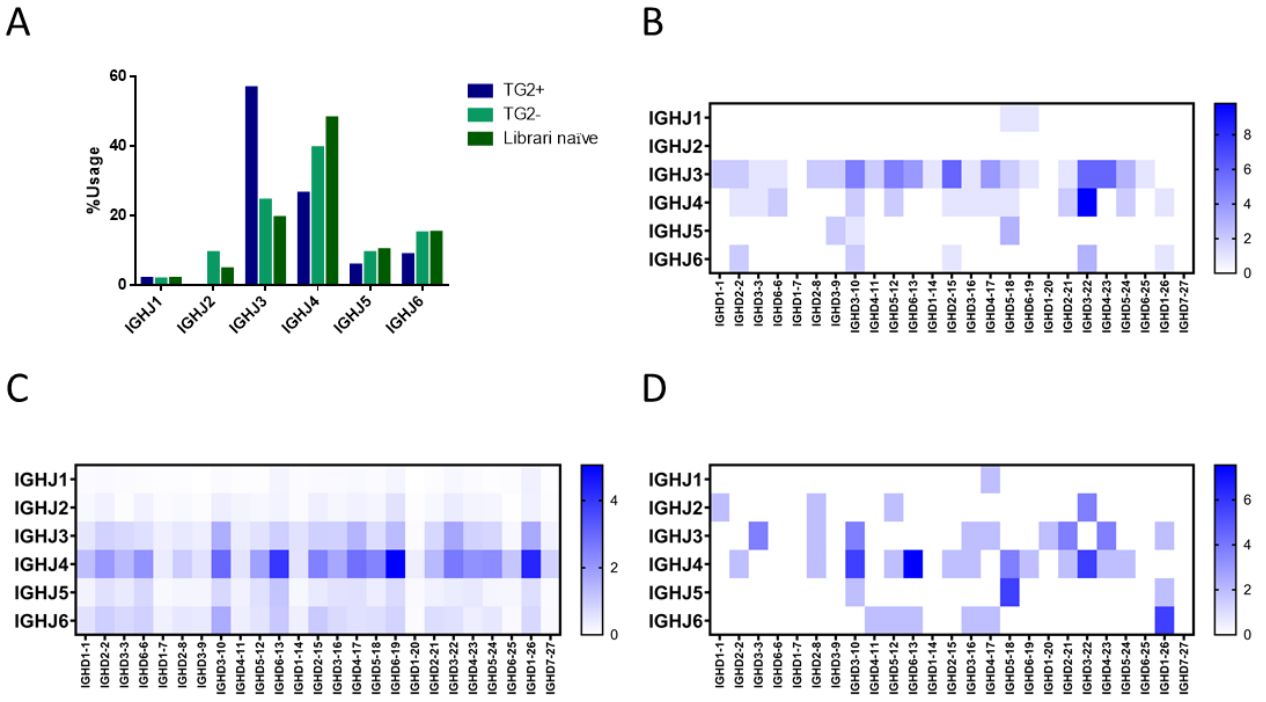


Figure 17. ***Characterization of VH from anti TG2 antibodies isolated with phage display technology***. ***A)*** *IGHJ frequency usage in IGHV5 antibodies. IGHD and IGHJ pairing with IGHV5 from* ***B)*** *TG2+ clones* ***C)*** *naïve library and* ***D)*** *TG2- from phage library.*

17A). The preferred IGHJ4 usage is reported also in other naïve repertoires[121], where the IGHV5 pairing with the IGHJ4 was 53%, whereas the pairing with IGHJ3 was reported to be around 22%[121]. These data suggest that the usage of phage display technology for the characterization of the Ab repertoire does not introduce bias, at least for the IGHV chain fragments pairing.

When the IGHD and IGHJ gene association was analysed the clones from the naïve library (Fig. 17C) as well as TG2-negative clones (Fig. 17D) showed a randomly distributed combination of D and J gene segments usage, while TG2-positive showed as well a broad usage of D genes with IGHD3-22:IGHJ4 pairing being the most frequently used (Fig. 17B).

We were wondering if the preferential IGHD and IGHJ gene usage pairing with IGHV5 could be due to a preferential position of these gene segments on the chromosome. Figure 7A shows the region of the chromosome where the HV, HD and HJ gene segments are located, and the numbers of the gene families as reported reflect the position on the chromosome. However, no preferred chromosomal position stood out from our analysis.

First, we focused on the gene composition of the CDRH3. Now, our goal is to characterize the CDRH3 involved in the TG2 binding. Analysing CDRH3s length (Fig. 18) we showed that IGHV5 naïve clones have an even distribution with a major frequency of the clones having 12 aa long loops. On the other hand, the TG2-positive clones show a clear bias in the distribution of the sequence length with more than 44% (45 out of 102) of the antibodies having a CDRH3 14 aa long. This observed preferred H3 length is in agreement with data reported by Roy and colleagues [119] obtained on single cell clones sequencing.

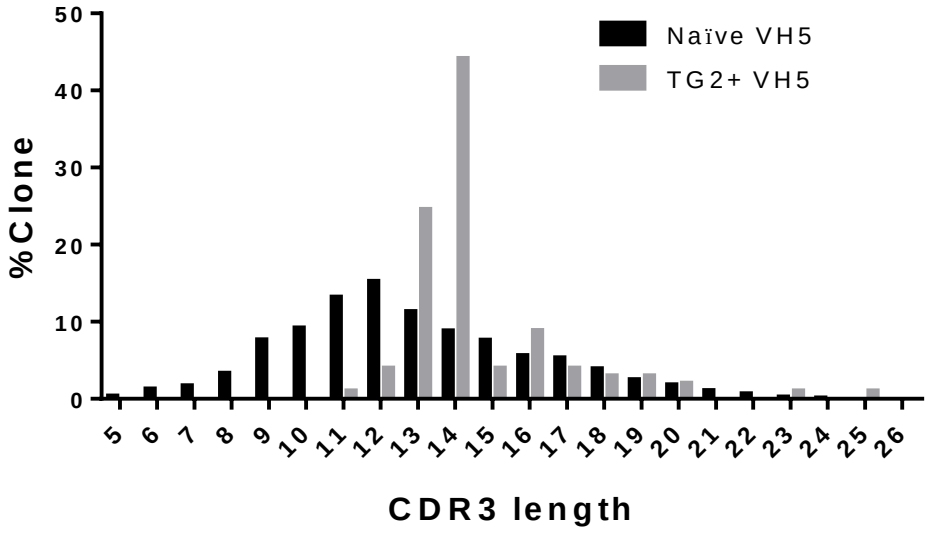


Figure 18. ***Characteristics of VH5 CDR3 length.*** *Aminoacidic length of CDR3 of Naïve and TG2+ generated with IGHV5 genes.*

4.5 Identification of TG2-positive CDRH3 key amino acid residues

The H3 sequences were further analyzed in order to highlight any bias in amino acid composition to rule out a role in antibody structure and function. TG2-positive clones with a CDRH3 of 14 aa were selected for further analysis, a total of 45 out of 102 sequences were retained (44% of the total Fig. 18). As control 810 VH5 sequences from the naive library having a CDRH3 of 14 aa were also used (9% of the total 9242 sequences VH5). We focus on the central region of the variable region, which is the most variable one, and is 9 amino acids long. Therefore, the variation of these CDRH3 have been analyzed performing sequence Logo analysis[156]. Logos of both the TG2-positive and naïve CDRH3 are reported in Figure 19. On the naïve sequences it is quite clear that the H3 loop sequences have a rather heterogeneous composition (Fig 19A).

After analyzing the TG2-positive sequences we could identify four "*key positions*", corresponding to specific amino acids preferentially present at specific loop positions of analysed CDRH3s.

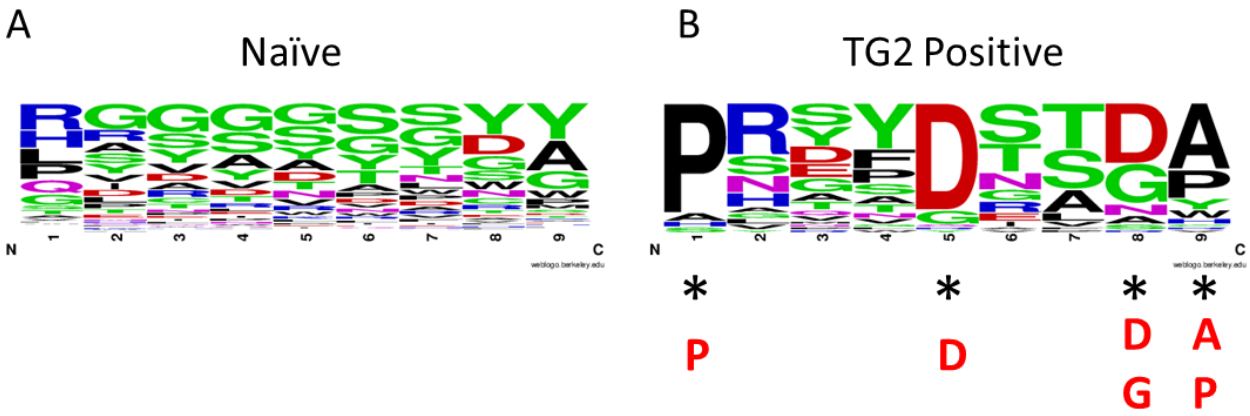


Figure 19. ***Characteristics of VH5 14aa long CDRH3***. ***A)*** *Logo from naïve phage library (810 clones).* ***B)*** *The Logo represents CDRH3 identified in TG2 positive clones (45 clones), "Key positions" are marked with an asterisk and specific amino acids for each of them are indicated in red. Logos were generated with WebLogo Berkeley. (Crooks et al., 2004)*

For the sake of simplicity, the amino acids position of the CDRH3 are going to be numbered as reported in Figure 19 and the position of IMGT nomenclature[102] is reported only in the subsequent list:

- position 1 in the Logo (4 IMGT) occupied by a Pro,
- position 5 in the Logo (8 IMGT) occupied by an Asp,
- position 8 in the Logo (11 IMGT) occupied by a Gly or an Asp,
- position 9 in the Logo (12 IMGT) occupied by Ala or a Pro,

(all "*key positions*" are indicated with an asterisk in Figure 19B)

We focused on the characterization of the "*key position"* in the VDJ rearrangement, where CDR3s are generated from the junction of IGHV-N1-IGHD-N2-IGHJ (Chapter 1.3.3). Analyzing the frequency of the Proline in position 1, we found that 38 out of 45 TG2+ clones with CDR3s 14aa long carry this amino acid. Amazingly in all the CDR3s the Prolines are codified by the N1 nucleotides (Table 3). This founding suggests a positive selection of TG2+ clones in CD patients. Up to now no bias allowing this junction in CD patients has been identified, although this could be investigated with future studies. Notably, one of the "*key positions*" identified by *in silico* analysis, Asp in position 5 on the Logo sequence, was previously shown to be an important residue for antigen binding being involved in the interaction with Lys-30 of TG2 [65]. 37 out of the 45 CDR3 analysed carry the aspartic acids in this position, and 33 Asp are codified by the *IGHD* gene (Table 3), while the other ones are codified by the N nucleotides. All the *IGHD3* genes (53% of the total CDR3s) and 50% of the *IGHD-non3* (35% of the total CDR3s) codify for the Asp. As previously shown (Fig. 17B) the most preferential IGHD gene family usage of the TG2+ clones is IGHD3, but even the presence of the others IGHD genes allows the presence of aspartic acid in this position.

Positions 8 and 9 are codified mostly by the *IGHJ3* and *IGHJ4* genes and only some nucleotides are generated from the N2, deriving by the D-J rearrangement (Table 3).

Amino acid	Total clone	N1	IGHD	N2	IGHJ
P position 1	**38**	38			
D position 5	**37**	2	33	2	
D or G position 8	**38**		5	8	25
A or P position 9	**33**			10	23

Table 3. ***Key positions****. The key position satisfied by the 45 clones are summarized. For each position the numbers of sequences that satisfy that criteria are reported (Total clone). Moreover, the region of the IGHV-N1-IGHD-N2-IGHJ rearrangement occupied by the amino acids are also reported.*

To quantify and give a statistical basis to these observations, we carried out other analyses. TG2-positive and naïve antibodies were stratified on the basis of the number of identified amino acids present in *key positions* in the H3 sequence. We can show that approximately 77% of CD derived TG2+ H3 loop present at least 3 key amino acid residues, with more than 55% satisfying the 4 "*key positions"* and 22% three of them (Fig. 20). On the opposite, the great majority (>53%) of naïve IGHV5 antibodies have an H3 loop with no *key position* occupied by identified residue and with only 4.9% of them presenting the *key* amino acids in 3 positions and only one out of 810 presenting all the 4 "*key"* amino acid residues (Fig 20).

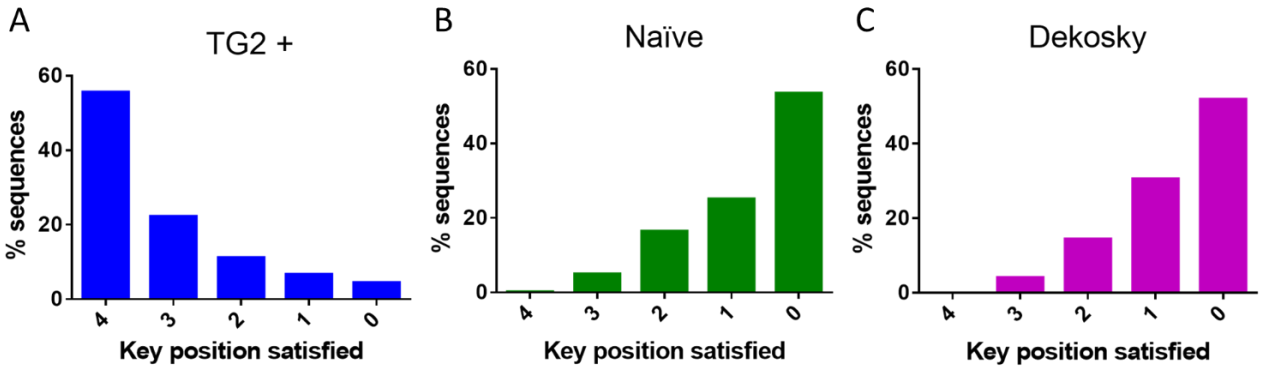


Figure 20. ***Key positions satisfied***. *Sequences 14 aa long that satisfy the key position* ***A)*** *TG2+ (Blue) and Naïve* ***B)*** *library generated in our laboratory (Green)* ***C)*** *published data (Purple, DeKosky et al., 2015) .*

Data from a published naïve repertoire are in agreement with the analyses of our naïve library[121]. To further confirm of our data, we analyzed the sequences published by DeKosky and colleagues[121], which allowed us to identify 851 sequences characterized by *IGHV5* gene usage with a CDRH3 14aa long. Among these 851 sequences, 51% show no *key position* in the H3 loop, and the 30% only 1 *key position* satisfied, and none all the 4 criteria. Data from a naïve repertoire agree with the data of a naïve phage display repertoire suggesting, another time, that the phage library well represents the Ab repertoire. Moreover, the *key position* identified in TG2 positive sequence suggest that well defined amino acids on the CDRH3 play a role in the specific binding with TG2.

4.6 Design *in silico* of the TG2-positive HCDR3 consensus sequence

In order to understand the importance of the "*key residues*" identified in the H3 loop we used a bioinformatics approach, in collaboration with Anna Vangone and Romina Oliva, respectively from *KAUST Catalysis center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, and Department of Sciences and Technologies, University "Parthenope" of Naples, Italy.* The Hidden Markov models (HMMs) (HMMER http://hmmer.org/), a statistical approach very used in the analysis of sequence profiles, has been tried when performing bioinformatic analyses. By applying the HMM approach they build a probabilistic model, called "HMM profile", starting from a multiple sequence alignment and later searching a database of sequences with a given HMM profile within a database of sequences looking for the meaning of these matches. Initially two HMM profiles were built: a positive profile (POS.hmm) was obtained starting from the set of the 45 TG-2 positive sequences

that outline the consensus, and a negative profile (NEG.hmm) was obtained starting from the set of the 810 naïve H3 sequences. Subsequently they created a database of sequences where we put together the 45 positive and the 810 negative sequences, for a total of 855 sequences and they calculated how significant the match of the sequences in the database was with both the positive and negative profiles.

For the negative profile, the results did not show a significant match with any positive sequence. The heterogeneity of the negative sequences is also reflected in the fact that only 34 negative sequences have an E-value <= 3 x 10^{-3} with the profile obtained from the 810 sequences.

Turning to the positive profile, all 45 positive sequences match the positive profile (36 sequences with E-value <1 x 10^{-4}). Even 16 negative sequences (1.9% of the total) show a significant match with the positive profile. Most of them have borderline E-value values, between 2 x 10^{-4} and 1 x 10^{-4}. Only 3 negative sequences correspond to the positive profile with E-values of the order of 10^{-5}.

Finally, by using a "plurality-rule consensus sequence" they have generated an ideal sequence from the probability distribution of the HMM profile. The sequence thus obtained is the following: CARPRSYDSTDAFDI (Fig. 21).

This sequence satisfies all the four *key positions*. Interestingly, the *"consensus*" CDR3 14aa long was not present either in the analysed panel of patients or in the panel of patients analysed by Roy and colleagues [119] nor in the VH control dataset even from other published library [120,121].

4.7 The CDRH3 *consensus* sequence allows the generation of TG2-positive scFv

In order to validate the data from *in silico* analyses, we generated an antibody carrying the "*consensus*" CDRH3 sequence and tested for binding to TG2 by ELISA. The *consensus* sequence is only the CDRH3 sequence, so we decided to insert this sequence in an antibody anti-TG2 isolated from CD patients and well characterized (Fig. 21)[113]. The choice of the backbone is fundamental, since the in the Ab binding CDRH3 plays a key role in the antigen recognition, but it is not enough[109]. An antibody that reflects the following feature has been chosen: i) the use as backbone of the *IGHV5-51* gene with the germline sequence with no mutation and paired with the *IGHJ3*; ii) the pairing of VH with the VL from *IGKV1-5*. The anti TG2 mAb 679-14-E06 (Fig.21)[113] IGHV5-51:IGKV1-5 was hence chosen. Moreover, this antibody is the prototype of CD derived antibodies reactive for the Epitope 1 [113,118] and its crystal structure is available [65]. Due to the importance of the VL in the

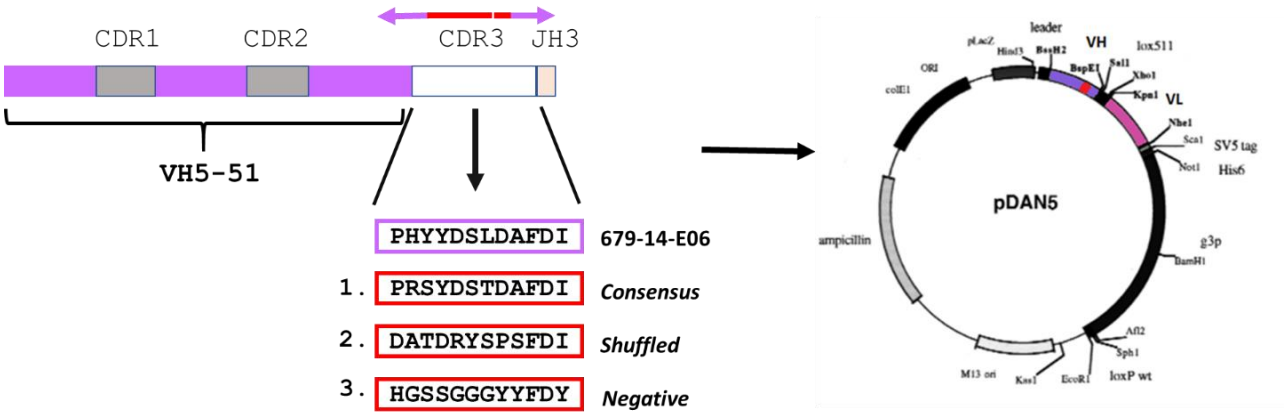


Figure 21. ***Schematic representation of the VH and how to change CDRH3****. VH is composed of VH5-51 and JH3 segments, the sequence of 679-14-E06 is naïve and the sequence of the CDRH3 is reported in the purple box. Only this sequence has been changed during the generation of the clones carrying different CDRH3s (in the red boxes). The sequence 1. is the consensus sequence. The sequences 2. and 3. are negative sequences. Sequence 2. is the shuffled sequence, i.e. the amino acids of the consensus sequence are present, but in different positions. Sequence 3. is a negative sequence generated with neg HMM. The primers used for inverse PCR are the narrows on the CDRH3, the red parts indicate the part of the CDRH3 that is different from the 679-14-E06 sequence, whereas the purple regions are the regions of 679-14-E06 that anneal on the pDAN vector carrying the 679-14-E06 backbone. After inverse PCR, gel purification of the PCR product and its ligation, a pDAN with the CDRH3 of interest (in red) is available. (pDAN is edited from Sblattero and Bradbury, 2000).*

TG2 binding[167], a VL from a specific TG2 antibody was necessary, and in addition its VL belongs to IGKV1-5. The 679-14-E06 monoclonal antibody was converted in the format scFv. VH and VL were cloned in the pDAN vector[152] (Fig. 22). *E. coli F'* cells were transformed with the pDAN vector, and the *E. coli* cells carrying the pDAN so constructed could be infected with the helper phage M13. After that the infection phages particles were produced and released in the culture media. These phages are displayed on the phage surface as fusion proteins with the pIII the scFv (Fig. 22). The phages are incubated in ELISA to assess the specificity of the scFv.

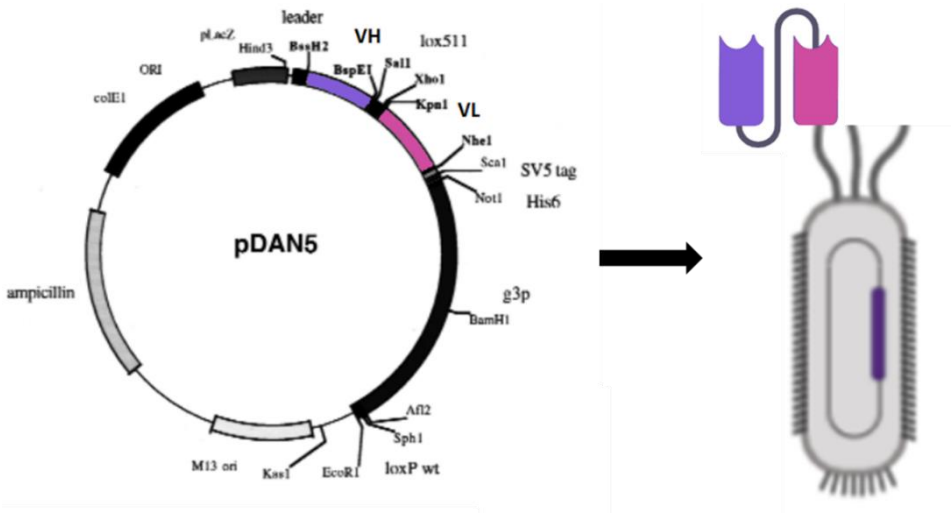


Figure 22. ***Map of the display vector pDAN5 with scFv inserted, and the phage displaying the scFv as fusion protein with pIII****. pDAN vector, with the genetic information for the VH and VL chains. Once phages were generated, thanks to the correlation genotype and phenotype, phages that display the scFv encode for that scFv. (Map of pDAN is edited from Sblattero and Bradbury 2000, the phage is designed with Biorender).*

We initially decided to test the antibody in the scFv format, thanks to a rapid generation of the scFv displayed on phage surface. Moreover, mAb 679-14-E06 in the scFv format was used as reference for the binding and the reactivity. ScFv carrying the consensus sequence was generated. Primers to perform an inverse PCR have been designed (see Oligonucleotides). The primers (represented as two arrows on Figure 21) have two essential characteristics: the sequence in 5' carry the sequence that we want to change in the vector (sequences in red on Figure 21), and the purple region (3') are designed to anneal on the vector with the region flanking the sequence that we want to change. The primers are designed to generate a determined CDR3. After ligation, vectors carry the mutations designed in CDR3 (Figure 21). For the generation of the scFv with the consensus CDR3, pDAN with 678-14-E06 has been used as template (Figure 21). The primers have been designed (named sense shuffled/random and anti-consensus, see 3.2.1 Oligonucleotides) to change the CDR3 of the 679-14-E06 with the *consensus* CDR3 (Figure 21). Upon purification of the PCR product and its ligation, the *E. coli* competent cells can be transformed. The clone has been sequenced to confirm the presence of the pDAN carrying the *consensus* CDR3 (Fig. 21).

As negative control we generated two H3 sequences: a "*shuffled*" sequence (Fig. 21), containing the same amino acids as the "*consensus*" sequence but in scrambled positions (this will allow to exclude the possibility of positivity solely due to charges of amino acids), and a "*negative*" sequence that was designed based on the HMM profile of the 810 VH5 14 aa long H3 negative sequences from the naïve library (Fig .21). As for the consensus sequence, pDANs carrying negatives sequences have been generated and checked. For the generation of shuffled clones, the primers named sense

shuffled/random and anti-shuffled have been used, whereas for negative clones primers called sense neg and anti-neg have been used (for the sequence see the chapter 3.2.1 Oligonucleotides). Phages containing the 4 different antibodies (listed in figure 21) were produced. The specificity of the phages were dissected by ELISA on TG2 and a control protein (BSA) (Fig. 23). Phage displaying the scFv containing the "*consensus*" CDR3 14aa long as well as control 679-14-E06 were highly positive on both human and mouse TG2 (Fig. 23). These data are in agreement with previous study in which it was demonstrated that the majority of antibodies carrying *IGHV5-51* gene family, especially 679-14-E06, were able to bind to mouse TG2 [118] with the scFv format allowing the specific binding. The antibodies carrying the two control VH CDR3s resulted negative in ELISA with TG2 antigens (Fig. 23), suggesting a critical and crucial role of the CDR3 sequence in the antigen recognition.

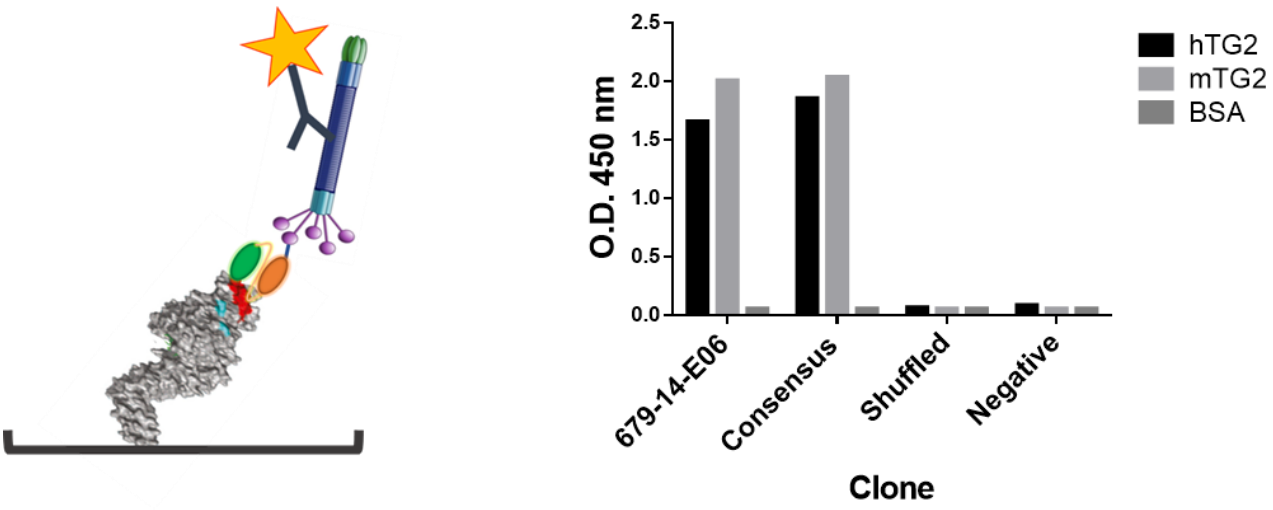


Figure 23. ***Phage ELISA****. Schematic representation of phage ELISA. The ELISA microplates were incubated with TG2 protein or BSA. After a blocking step and a wash step, phage particles carrying the scFv were incubated and, in a second time, anti-M13 HRP secondary antibody was incubated. On the right the ELISA result is reported. The signal that indicates a specific binding is appreciable when the phages displaying scFv from 678-14-E06 and consensus are displayed, whereas with both the negative control, shuffled and negative, a signal at the same level as the background is revealed. ELISA is performed both on human and mouse TG2 and BSA as negative control.*

We then carried out competition binding assay[141] to investigate if the "*consensus*" CDR3 antibody was binding to TG2 in the so-called epitope 1. A schematic procedure for competition ELISA is reported on Figure 24. Results (Fig. 24) suggest that the reference antibody 679-14-E06 and the *consensus* antibody compete for the same epitope 1 [118], that is located in TG2 N-terminal [62,65]. The competition for the epitope binding, as described in the schematic figure, could be identified with a reduction of ELISA signal. First a specific scFv-Fc was incubated, and if the scFv displayed on the phage surface binds the same epitope or a region nearby, the scFv-phage binding is obstructed by

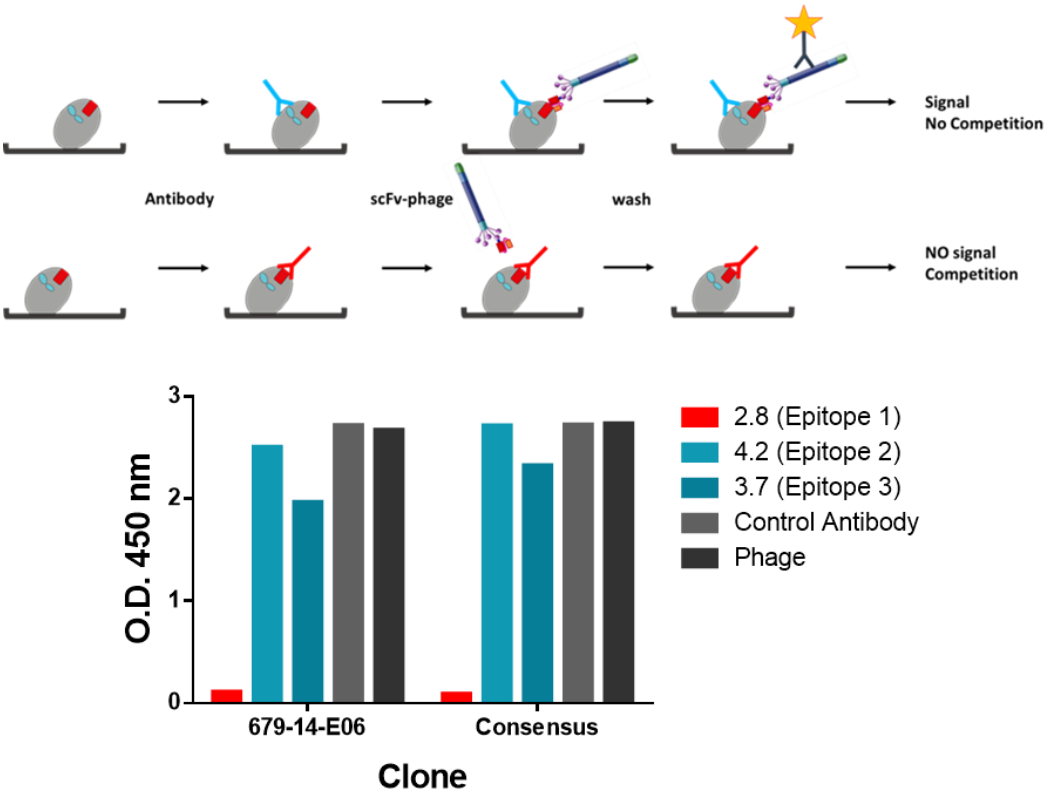


Figure 24. ***Competition phage ELISA****. On the top the schematic representation of competition ELISA is reported. The TG2 protein is represented as grey oval with the epitope 1 shown in red, while epitope 2 and 3 are shown in light blue. In the first case, antibodies recognizing non-epitope 1 were incubated with TG2. After the scFv-phage is incubated, if it recognizes a different epitope than the previously incubated antibody, scFv on phage surface can bind TG2 and a signal is detectable with the incubation with a secondary anti-M13 HRP conjugated antibody. In the second case, the scFv-Fc is specific for epitope 1 and no signal is obtained after the incubation of the scFv-phage that binds to epitope 1 and anti-M13 HRP antibody, suggesting that the investigated scFv antibody competes for the binding to epitope 1 of TG2 protein. On the bottom the result of phage display competition ELISA is reported. The first coating was performed with hTG2. scFv-Fcs against different epitopes (1, 2 and 3) were incubated as well as a scFv-Fc anti a non-related antigen (Control Antibody), and, as further control, a line without the antibody (Phage). Afterwards the scFv-phages were added to the well, and after a wash step, anti-M13 HRP secondary antibody was added. A chemiluminescent reaction was revealed after the addition of TMB substrate.*

the presence of scFv-Fc. Hence a competition is revealed as a reduction of ELISA signal. If phages were incubated with scFv-Fc that recognize a different epitope, they can bind the antigen and the signal is comparable to the signal from the incubation with a scFv-Fc non-specific for the antigen (Control Antibody on Figure 24) or the signal with phages without a scFv-Fc (labelled as Phage on F 24). Our data suggest that the scFv 679-14-E06 and the consensus sequence compete for the binding to the epitope 1.

4.8 CDRH3 *consensus* sequence in scFv-Fc and full size Ig formats

Although the VH and VL regions expressed in a scFv format in the phage display context are usually equally reactive when subcloned in a full-length antibody, we decided to check whether the *"consensus"* CDRH3 maintains reactivity when expressed in a more physiological condition. The scFv has been easily subcloned in a pHygro vector, which was modified as reported in Di Niro and colleagues[141]. As described, after digestion with BssHII and NheI from the pDAN vector (Fig. 21), the digested and purified scFv could be subcloned in the modified pHygro carrying the Fc (CH2 and CH3 from human IgG) (Fig. 25). The scFvs both of 678-14-E06 and carrying the *consensus* H3 loop have been cloned in the pHygro vector. To control the cloning, the region of the vector with the scFv has

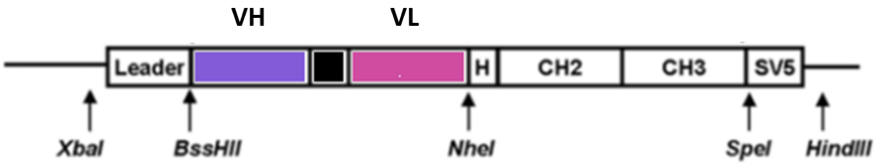


Figure 25. ***Schematic representation of the cloning vector to generate scFv-Fc.*** *Different scFvs can be exchanged by using restriction sites BssHII and NheI. ScFv is in frame with the Human IgG1 CH2-CH3 domains gene. This construct is in the modified vector pCDNA3.1/Hygro(+) as described by Di Niro and colleagues (Edited from Di Niro et al., 2007).*

been sequenced. The checked vector has been transfected in ExpiCHO cells for the expression of the scFv-Fc (Figure 26). ExpiCHO expression system is a high-yield transient expression system based on suspension-adapted Chinese Hamster Ovary cells. After the scFv-Fc construction, the full length antibody as IgG1 and IgA1[168] have been generated. The different antibody structures are schematized on Figure 26. For the design of the vector for the expression of IgG and IgA two approaches could be chosen: H and L chains could be cloned in two different expression vectors and

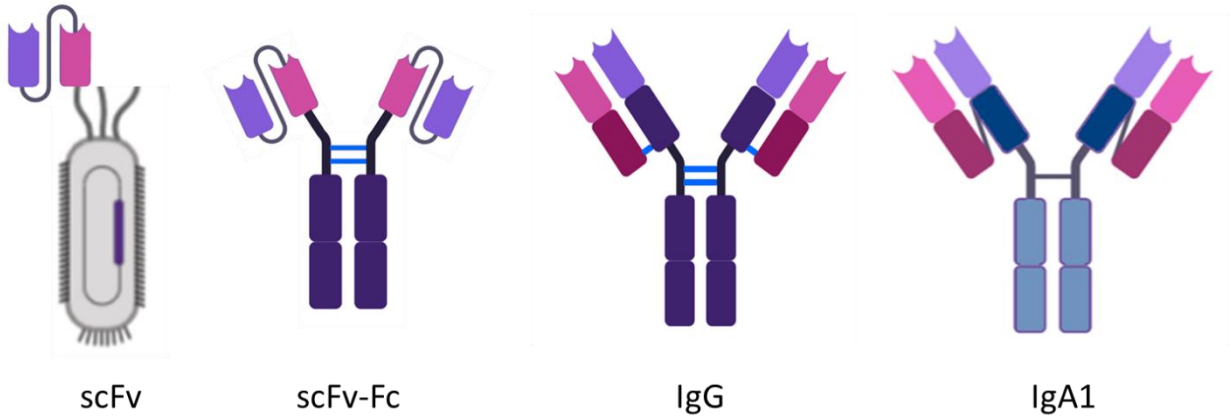


Figure 26. ***Schematic representation of antibody formats.*** *From left to right: scFv displayed on phage surface, scFv-Fc with human CH2-CH3, human IgG and human IgA1. (Biorender)*

cell co-transfected with both the two vectors, however it is not so easy to reach a balance of H chains and L chains. Therefore, we adopted the antibodies expression strategy described by Fang and colleagues[154]. Both H and L chains are cloned in the same vector and the use of the foot-and-mouth-disease virus (FMDV)-derived 2A self-processing sequence allows to express full-length antibodies from a single open reading frame (ORF). A 2A peptide sequence (about 24 amino-acids) is followed by a furin cleavage site[154]. This construct was used to link the antibody heavy and light chain sequences, and allowed a high level of full-length and functional monoclonal antibodies expression. The vectors were constructed in the host laboratory following the strategy published by Fang and colleagues, 2005[154]. Both 679-14-E06 and *consensus* antibodies were expressed in all the previous schematized (Fig. 26) antibody formats. ExpiCHO cells were used for a high expression level. The CHO medium culture contains antibodies. Production ELISA has been performed to check the presence of the antibodies (data not shown) and SDS-PAGEs under reducing condition of supernatant of the cell culture has been performed. Proteins could be visible after Coomassie staining (Fig. 27 A). To finely characterize the antibodies in the culture media, western blots under reducing conditions have been performed (Fig. 27B). Anti-IgG-AP (Alkaline Phosphatase) was used to detect scFv-Fc, IgG and L chain of IgA, whereas, anti-IgA-AP was used to detect the H chain of IgA. ScFv-Fc has a molecular weight (MW) of approximately 55kDa and the lower band is probably CH2 and CH3 without the scFv. IgG has a VH band at 55kDa and VL band at 25kDa and a band at 75kDa, due to the presence of VH plus VL. IgA VH was immunodetected with anti-IgA Ab (VH migrates in the gel at a MW of 55/60 kDa). On the other hand VL was immunodetected with anti-IgG Ab. The Abs have been purified as detailed. ScFv-Fcs and IgGs have been purified with Protein A agarose, whereas IgAs have been purified with IgA Affinity Matrix. After purification, the quality of the purified antibodies has been checked (Fig. 27 C), and they have been quantified (data not shown).

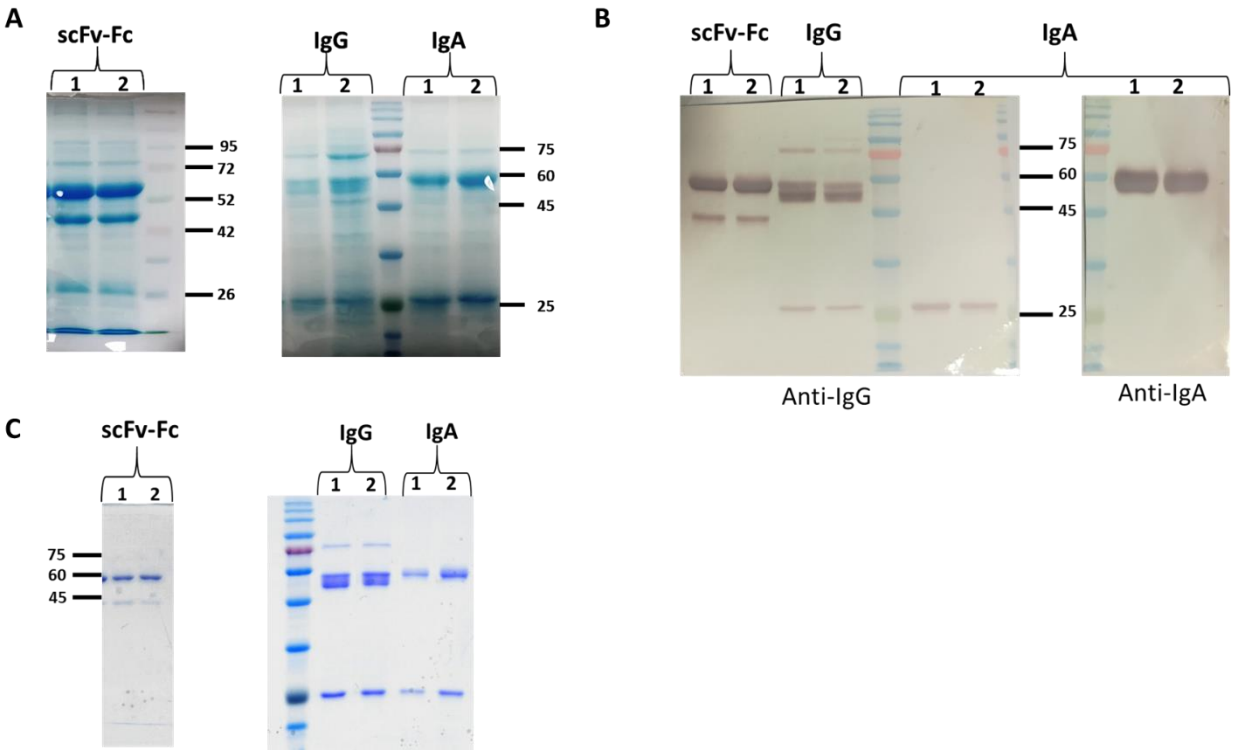


Figure 27. ***Antibodies.*** *Antibodies production in Expi CHO cells. Number 1 refers to 679-14E06 and number 2 to consensus.* ***A)*** *SDS-PAGE under denaturing conditions of Expi supernatants. Supernatants were loaded for each antibody in all the format, and gels have been stained with Coomassie blue* ***B)*** *Western blot of the 679-14-E06 sequence and the "consensus" sequence in IgA, IgG and scFv-Fc formats. ScFv-Fc, IgG, and VL chain of IgA are immunodetected with anti-IgG, whereas VH of IgA is immunodetected with anti-IgA.* ***C)*** *SDS-PAGE under denaturing conditions of scFv-FC, IgG and IgA after purification. Gels have been stained with Coomassie blue.*

For the scFvs only the 55kDa band is considered and for the IgG and IgA an average between the value of the concentration of the L chain and the H chain is considered.

The so produced, purified and quantified antibodies have been assessed for TG2 binding by ELISA. Affinity curves have been performed with hTG2 in coating and different concentration of antibody for each format. The analyses of the data were performed with GraphPad Prism, non-linear regression have been used and one site specific binding is chosen for analyses. (Given the experimental setting, one site specific binding has been chosen. In fact, we speculate that only one of the two antigen binding domains can interact with the hTG2 on the well surface).

The affinity of the scFv-Fc is reported on Figure 28A, which shows that the *in silico* designed consensus binds to the TG2 with high affinity. The affinity has been calculated and the values are reported as Kd, expressed in ng/ml. Kd is the equilibrium binding constant, in the same units as the X axis (ng/ml). 17ng/ml and 34ng/ml are needed for the *consensus* and 679-14-E06 in scFv-Fc

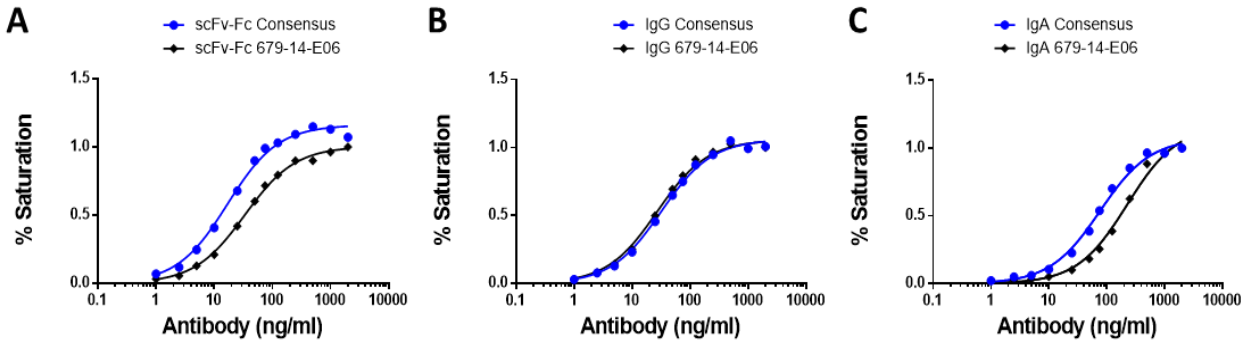


Figure 28. ***Affinity of antibodies in different formats****. Reactivity of different antibody formats for hTG2 binding, determined by ELISA, whereas BSA was used as negative control:* ***A)*** *scFv-Fc* ***B)*** *IgG and* ***C)*** *IgA1 (data on BSA were not shown).*

respectively to reach half of the saturation. All the analyses are performed with the Graph Pad software. The affinity of these antibodies is in the 10^{-10}M range, indicating a very high specificity. However, scFv-Fc is not present in nature, so, to better recreate the physiological condition, the IgG was first investigated. IgG is the most used Ig recombinant format, thanks to the high capability of binding of the protein A, which allowed an easy purification of both scFv-Fcs and IgGs. Nevertheless, to recreate the CD patient's antibody anti-TG2, the IgA format has also been generated and purified. The specificity of purified IgGs and IgAs were dissected. As previously reported for scFv-Fc, ELISAs to determine the binding were performed with different amount of Igs. Figures 28 B-C show that H3 consensus antibodies were fully functional, being able to bind human TG2 in all antibody formats. Moreover, the affinity of the antibodies was at a comparable level with that of 679-14-E06. The affinity of 679-14-E06 is comparable with previously reported data[119]. Unexpectedly, the *in silico* designed CDR3 not only binds TG2 in the phage display format, which is the format used for the preliminary screening that lead to the consensus CDRH3 construction, but preserves the specificity in a more physiological conditions. Moreover, the Ab 679-14-E05 has a naïve sequence, nevertheless its showing a high affinity, whereas other Abs before affinity maturation are undetectable with in vitro assays[169]. Surprisingly, *in silico* technology combined with *in vitro* technology allowed the generation of an antibody with very high affinity, which is not present in the panel of Ig repertoire of CD patients.

4.9 Dissecting fine specificity of VH CDRH3 analysis by mutant libraries

In order to gain insights into the paratope of the consensus antibody and to understand if the 4 "*key positions*" identified are the only requirement for producing an anti TG2 positive antibody we constructed libraries of mutants of the H3 consensus loop. 2 libraries were constructed. In the first one, the 4 fixed "*key positions*" were maintained , whereas 5 other amino acids of CDR3 were randomly mutated (Fig. 29 (4.)). Proline is allowed in position 1 of the CDR3, aspartic acid in 5, Glycine or aspartic acid and alanine or proline in position 9. All amino acids are allowed in the positions 2-4 and 6-7. The investigation of the structure of 679-14-E06 (PDB code 4ZD3)[65] suggests a preferred usage of small amino acid in position 7. We generated a second library, in which the only difference was a reduction of the possible diversity at position 7 that was limited to 6 possible amino acids, comprising A/L/S/T/M/V (Fig. 29 (5.)). The H3 loops with random amino acids were inserted into a backbone of the *IGHV5-51* gene with the germline sequence. As previously, for the choice of the backbone for the consensus sequence, the well characterized and with naïve VH5-51

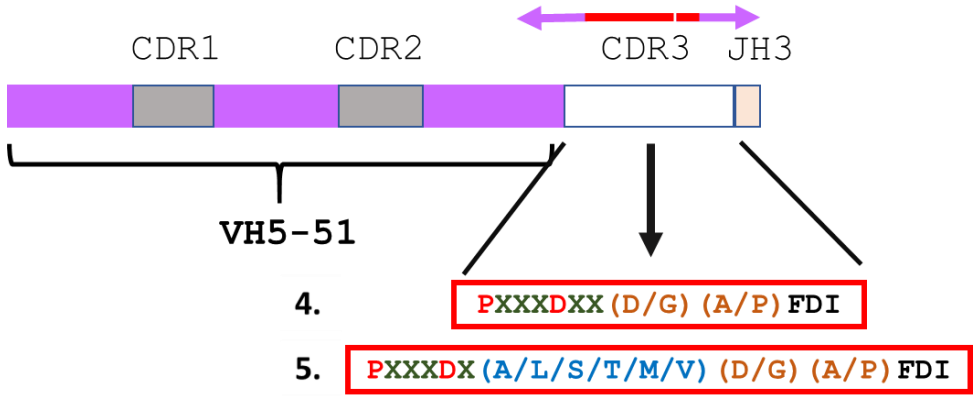


Figure 29. ***Random CDRH3 14aa long.*** *Schematic representation of the VH segment. Only CDRH3 was changed, whereas the backbone comes from 679-14-E06. CDR3 was designed as indicated. Schematic representation of VH CDRH3 with 4 fix positions (4.) and a fifth semi-fix position (5.). The P and D "key position" are reported in red, whereas D/G and A/P "key positions" are reported in orange, whereas the other amino acids were random (random aa positions showed in green). The "fifth semi-fix position" (sequence 5.) could admit one of the six amino acids indicated in blue.*

678-14-E06[113] was chosen and paired with the 679-14-E06 VL[113]. As previously for the generation of the consensus sequence, the libraries are generated with inverse PCR, the primers are reported in the chapter 3.2.1 Oligonucleotides. The primer sense shuffled/random was used for both libraries.

As primers anti, anti-random was used for library 4., whereas for library 5. anti-random A/L/S/T/M/V have been used. The PCR products were purified, ligated and *E. coli* cells were electroporated. For each library, random clones were selected and tested for TG2 reactivity by phage ELISA. Positive TG2 clones could be isolated from both libraries, and approximately 10 to 15% of antibodies tested

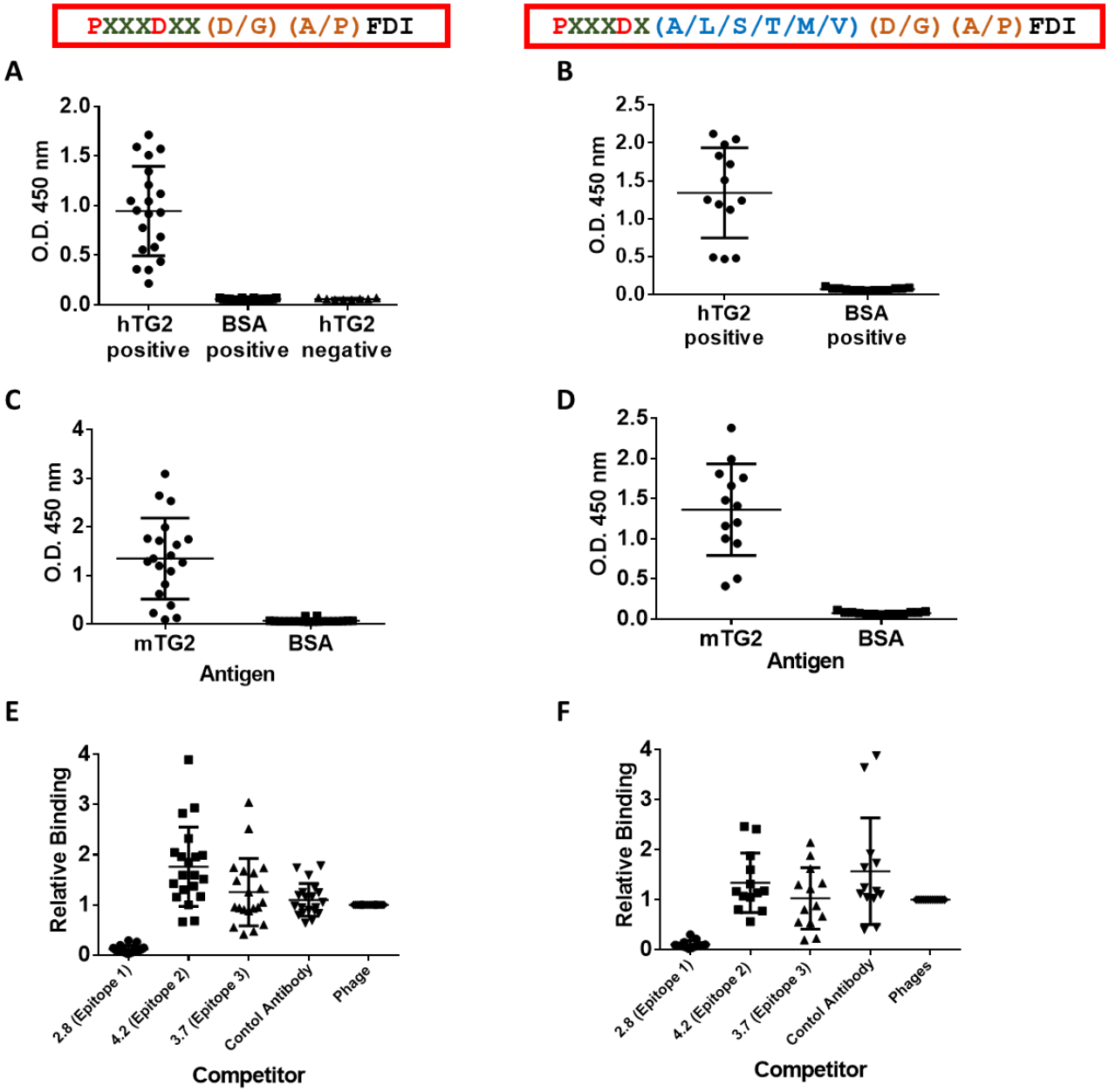


Figure 30. ***ELISA of clones with random VH CDRH3 14aa long***. ***A)*** *Phage ELISA of positive and negative clones on hTG2. Positive clones are tested even on BSA as negative control.* ***B)*** *Phage ELISA of positive clones on hTG2 and BSA as negative control.* ***C)*** *and* ***D)*** *positive clones from the respectively library reported on the top of the figure were assessed for the binding to mTG2 in ELISA, and all are specific even for mTG2. The same clones were tested in competition ELISA* ***E)*** *and* ***F)*** *and it is possible to see that the signal decreases when the competitor is an antibody that binds Epitope 1, suggesting that scFvs compete for the binding of Epitope 1 or a site nearby.*

positive with hTG2 (Fig. 30). TG2 reactive clones are specific and result negative with the unrelated BSA protein. Up to now only a few random negative clones (Fig. 30 A) have been sequenced and a higher number is under processing in the host laboratory. To better characterize positive clones from both libraries, the binding to mTG2 was dissected and the specificity for mTG2 has been confirmed (Fig. 30 B and C). Hence, the TG2 epitope specificity was investigated. As shown on Figure 30 E and F, all clones of both libraries show a low signal when incubated with the antibody that binds Epitope 1 (more details of competition ELISA are reported on Figure 24). These results suggest that all the scFv with random CDRH3 from both libraries 4. and 5. compete for the binding to Epitope 1. This further suggests that the presence of different amino acids level does not modify the epitope recognised.

TG2 positive and negative clones were sequenced and the CDRH3 composition was analysed for the frequency of amino acids present. The readout is presented in the logo on Figure 31 A and B fort positive clones and on Figure 31 C for TG2-negative clones.

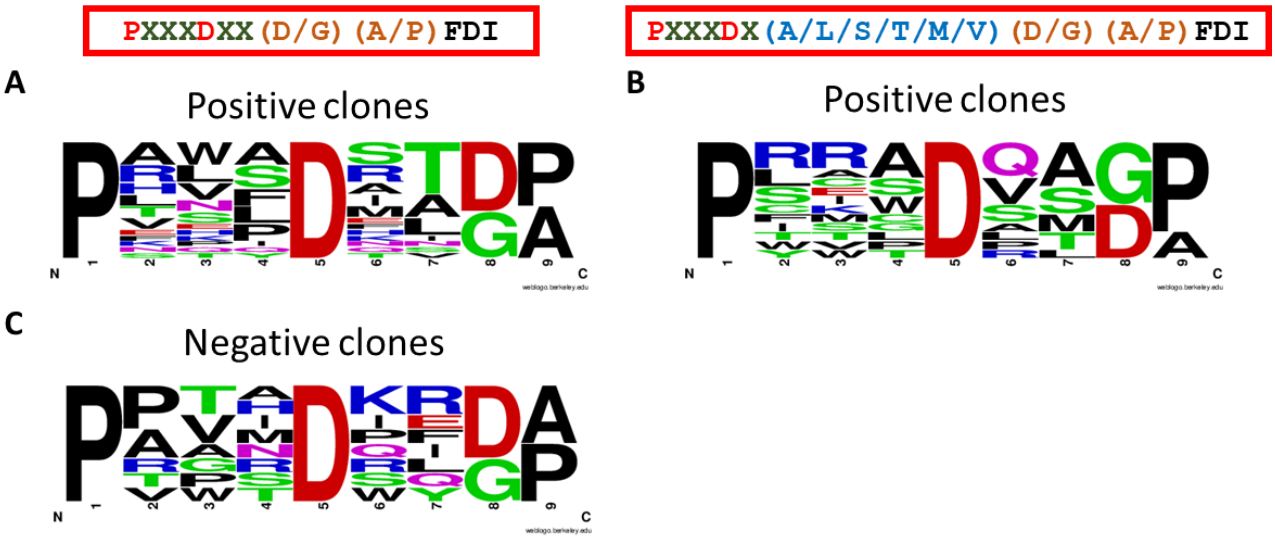


Figure 31. **Random VH CDRH3 14aa long. A)** Logos represent CDRH3 identified in TG2 positive clones and **C)** negative clones carrying "Random" CDRH3 reported on the top of the column. **B)** Logo represents CDRH3 identified in TG2 positive clones carrying random A/L/S/T/M/V. Logos were generated with WebLogo Berkeley. *(Crooks et al., 2004)*
(https://weblogo.berkeley.edu/logo.cgi)

4.10 Dissecting the fine specificity of binding to Epitope 1

In the attempt to finely dissect anti Epitope 1 antibodies-TG2 interactions, based on the CDR3s of TG2 reactive clones identified from the analyses of the two libraries, Anna Vangone and Romina Oliva, respectively from *KAUST Catalysis center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, and Department of Sciences and Technologies, University "Parthenope" of Naples, Italy*, generated, 8 mutants thanks to *in silico* analyses: 4 designed to increase and 4 to decrease the affinity (Table 4). Here we focus on the role of a single amino acid in the TG2 Epitope 1 binding. So, we investigate if the presence of a single or a couple of different amino acids can condition the binding. Since the 4 "*key positions*" are recurrent in TG2 positive clones (Fig. 19) here we focus on the role of the amino acids that were random in the previous chapter. As listed in Table 4 the Mut 1-4 could increase the affinity, whereas the Mut 5-7 and PP could decrease the affinity of the scFv for TG2. Moreover, to finely investigate the high affinity, the mutants 1-4 show mutations that are a mix between the 678-14-E06 and the *consensus* sequence CDR3. The mutants 5-7 carry amino acids that are either present in the panel of random negative sequences, or that in a docking between Ab and TG2 show a steric effect. Mut PP with the double Prolines change conformation, leads to an open conformation of the binding domain of the CDRH3.

Consensus	**CARPRSYDSTDAFDIW**
679-14-E06	**CARPHYYDSLDAFDIW**
Mut1	CARP**R**YYDSLDAFDIW
Mut2	CARPH**S**YDSLDAFDIW
Mut3	CARP**RS**YDSLDAFDIW
Mut4	CARPHYYDS**T**DAFDIW
Mut5	CARPHY**A**DSLDAFDIW
Mut6	CARPHYYD**N**LDAFDIW
Mut7	CARPHYYDS**I**DAFDIW
MutPP	CARP**P**YYDSLDAFDIW

Table 4. ***Point mutations. CDRH3s*** *of "consensus" and 679 are reported in bold. CDRH3 of the mutant clones, carrying point mutations, are listed below and mutations in relation to "consensus" or 679-14-E06 are highlighted in bold. Mutations that increase affinity are indicated in blue, whereas mutations for affinity decrease are indicated in red.*

To assess these mutants scFvs fuse in frame with pIII have been generated. As previously, the backbone is the sequence of 679-14-E06, and only the CDRH3 carries mutations. The mutants have been obtained with inverse PCR, on pDAN carrying scFv of 679-14-E06, with the primers listed in the paragraph Oligonucleotides 3.2.1. For all the inverse PCR the primer sense is the one named sense-shuffled/random, whereas the anti-primers report the name of the mutant clone. PCR products

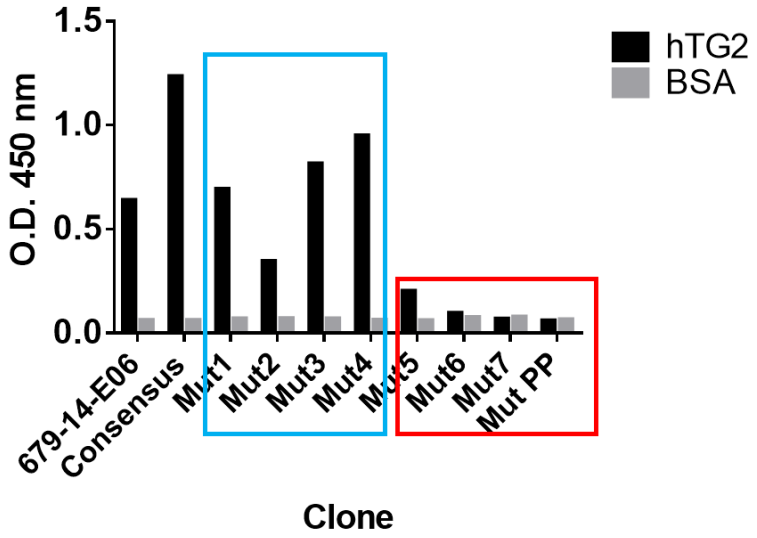


Figure 32. ***Phage ELISA of mutants in the scFv-phage format.*** *Phage ELISA of the mutants on TG2. BSA was used as negative control, whereas 679-14-E06 and consensus antibody were used as positive controls.*

were purified ligated and the *E. coli* competent cells were transformed. All the clones have been controlled by scFv sequencing. Phages that display the listed scFv were produced and assessed for specificity by phage ELISA made on hTG2 and an unrelated protein as negative control (Fig. 32). The ELISA results indicate that the mutants confirm the speculations. However, phage ELISA results depend both on the specificity of the scFv and on the stability of the scFv. Not all the scFvs have the same life-time, and it could happen that there is a proteolysis between the scFv and the pIII. Moreover, different scFvs have different expression levels, even if cloned in the same vector. Only

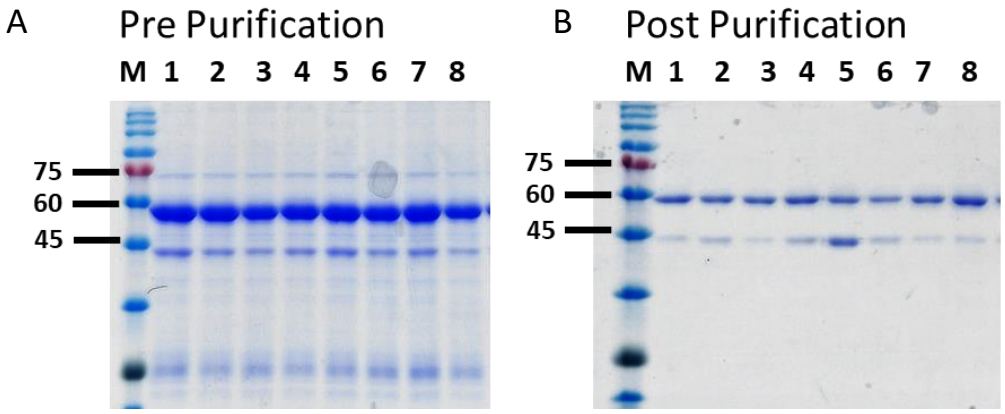
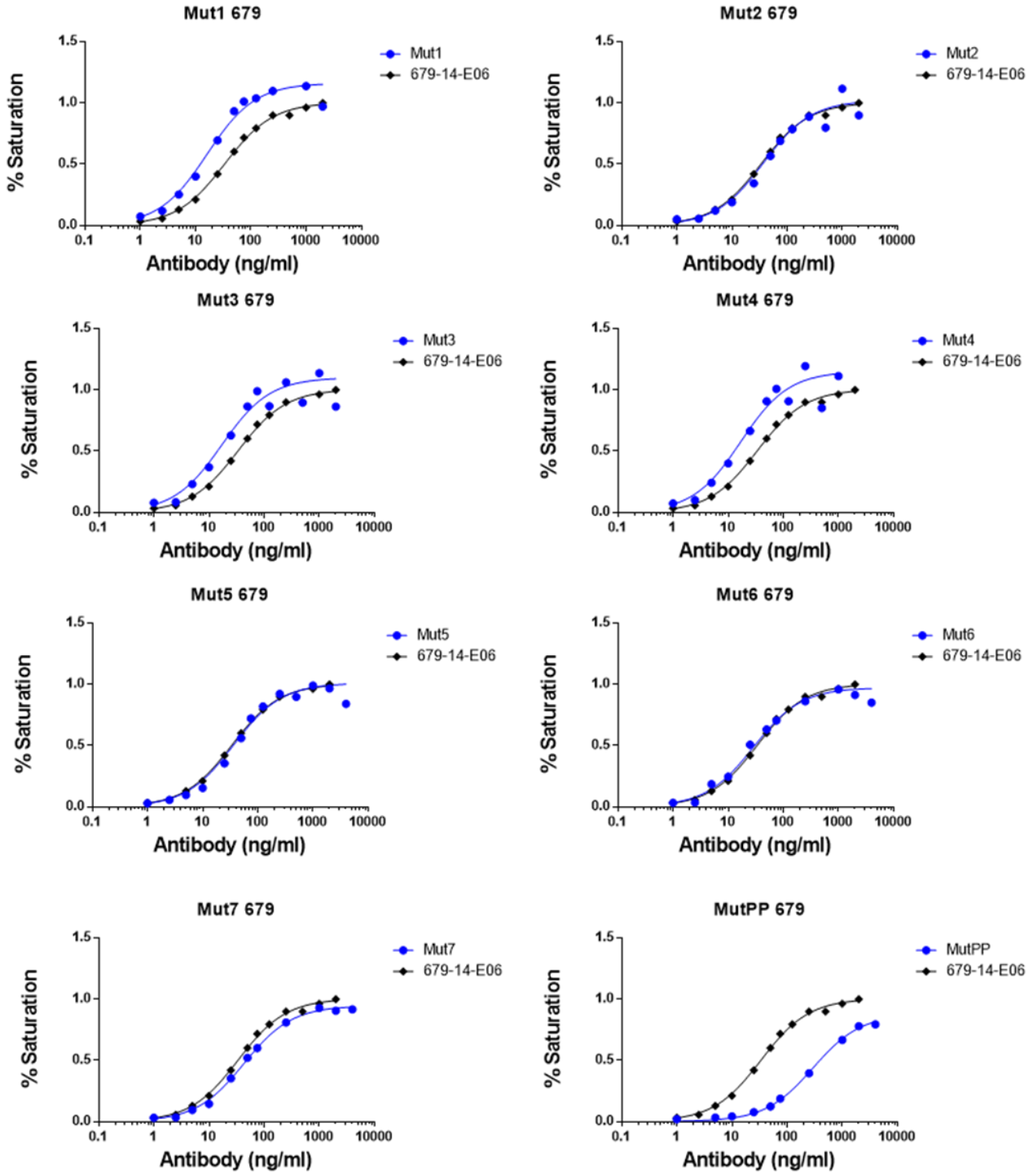


Figure 33. ***Antibodies.*** *Antibodies production in Expi CHO cells. Numbers refer to the number of mutant and number 8 is the clone Mut PP. SDS-PAGE under denaturing conditions of* ***A)*** *Expi supernatants and* ***B)*** *scFv-Fc after purification. Gels have been stained with Coomassie blue.*

10% of the phages display the scFv and most of them have only one scFv displayed on the surface[139]. Therefore, to finely investigate the affinity of the mutants, the scFvs have been subcloned in the



Mut1	Mut2	Mut3	Mut4	Mut5	Mut6	Mut7	MutPP	679-14-E06	Consensus
15,39	45,39	17,12	16,93	39,52	23,95	42,96	336,8	33,82	16,94

Figure 34. ***Affinity ELISA of mutants****. Binding of the mutants in the scFv-Fc format to TG2 was assessed by ELISA. In the table below the ng/ml to reach half of the saturation of each clone is reported. (The value of Consensus is referred to the Figure 28A)*

The analyses are performed with Graph Pad software https://www.graphpad.com/guides/prism/6/curve-fitting/index.htm?reg_comparing_models.htm)

modified pHygro[141]. Vectors have been used to transfected ExpiCHO, which could produce the scFv-Fc. After 10 days from the transfection of ExpiCHO in a 24 multi-well block, the supernatants have been collected and loaded on SDS-PAGE (Fig. 33 Pre purification) and the production of the scFv-Fcs was confirmed by production ELISA (data not shown). The so produced scFv-Fcs were purified and quantified. ScFv-Fc after purification are shown on Figure 33.

ScFv-Fcs have been assessed in different dilutions in ELISA on hTG2. All values are reported to 679-14-E06 as control and it was used as normalizer. These data indicate that clones Mut1-7 have an affinity that is comparable to 679-14-E06, and probably the low signal of Mut5 and Mut6 and no signal of Mut7 on phage-ELISA could be ascribed to a low level of scFv expression or stability. In fact, they show only a low reduction of affinity (Fig. 34), whereas Mut1,3 and 4 confirm their high affinity (Fig. 34) shown in phage ELISA. As expected, the mutant with the double Prolines (MutPP) shows a decrease in affinity compared to the control scFv-Fc (Fig. 34).

4.11 Recombinant TG2 protein expression

For all the experiments previously described, a wild-type both human and mouse TG2 proteins have been used. The protein expression is under the control of an inducible *lacZ* promotor (inducible with IPTG). Both recombinant hTG2 and mTG2 were produced in *E. coli* cells, the protein was then purified through Ni-NTA affinity resin, specific for histidine binding. After a wash step, the protein was eluted with imidazole, a competitor of histidine for the binding to Ni-NTA resin.

The elution fractions were collected and loaded in SDS-PAGE under denaturing conditions (Fig. 35). TG2 protein migrates at the expected MW of about 78 kDa, whereas the other bands are degradation forms, due to the thermolability of TG2 protein. The most concentrated fractions were collected in a single volume and it was dialyzed with PBS O/N at 10°C and quantified on SDS-PAGE through BSA calibration curve (data not shown). Only human TG2 protein is reported, but mouse TG2 shows a similar profile (data not shown).

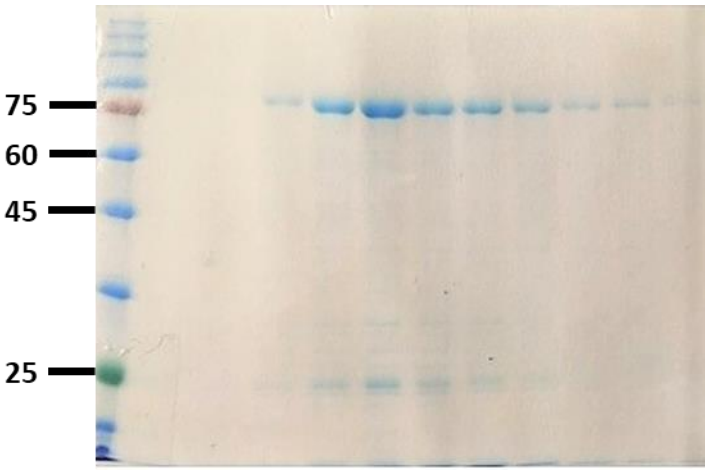


Figure 35. ***Human Transglutaminase 2.*** *Elution fractions, from the purification column of TG2 protein were collected and loaded in SDS-PAGE gel 12% under denaturing conditions. The expected Molecular Weight is 78KDa, the other bands are degradation forms.*

5 Conclusion

In 1985 Smith published an article entitled: ”Filamentous fusion phages; novel expression vectors that displayed cloned antigen on the virion surface”[132]. Three years later, the scFv antibody fragment was projected, and in 1890 the scFv was fused with the phage pIII gene in a phagemid vector. It has been demonstrated that the so-cloned antibody fragments in a phagemid vector, and that after transformation of the recombinant vector DNA into bacteria cells the scFvs displayed on phages are specific for the antigen, once tested in ELISA[135]. The phage display concept evolved since 1985, and phage display of antibody fragments, antibodies for therapy, diagnoses and research have been generated.

CD is an intestinal malabsorption characterized by intolerance to cereal proteins by immunological responses to gliadins of gluten and TG2. Analyses of the antibodies from CD patients' peripheral blood lymphocyte have been performed using phage antibody libraries. After selection cycles α-gliadin Abs have been identified, whereas no positive clones have been selected for TG2 binding[170]. The following year, libraries from both peripheral blood lymphocyte (PBL) and intestinal biopsy lymphocytes (IBL) have been generated[52]. Where Abs fragments specific for gliadin could be selected from all libraries, Abs fragments specific for TG2 have only been selected from IBL libraries, suggesting that anti-TG2 antibodies are primarily synthetized at the intestinal level. A further study highlights that untreated CD patients show a high abundance of TG2 specific antigen presenting cells in the small intestine, whereas subjects on a gluten free diet show fewer-TG2 specific antibody secreting cells[113].

Over the years, anti-TG2 antibodies have been studied. It was demonstrated that antibodies anti-TG2 preferentially use the *IGHV5* gene family[52,113,118,119], which is not the most used in a naïve repertoire[121,152]. Moreover, IGHV5 chains of anti-TG2 Abs from CD patients show low mutations levels, comparable to a naïve Abs library, whereas other diseases lead to generate Abs with a high number of somatic mutations[113]. High mutations levels have been observed even in autoimmune pathology. Indeed, Abs anti-rheumatoid factors in patients with rheumatoid arthritis, (rheumatoid arthritis, as CD is an autoimmune pathology) are characterized by mutations[171]. Naïve antibodies are polyreactive, and poly-reactive antibodies are often self-reactive[172]. It has been suggested that self/polyreactive IgA in the gut environment could move benefits[172]. Moreover, the frequencies of naïve B cells that can bind a given antigen is reasonably consistent[172]. However, naïve antibodies

show a low affinity, resulting undetectable in vitro[169]. Surprisingly, anti-TG2 antibodies have a high affinity, but polyreactivity for neither other transglutaminase family members[118] nor gliadin[113] has been detected. These characteristics are not common for all Abs produced by CD patients upon gluten ingestion. Indeed it has been demonstrated that anti-Gliadin Abs from CD patients with low affinity are polyreactive[170]. Notably some mutations that occur in *IGHV5* genes are recurrent in CD patients, and retro mutation in anti-TG2 from CD patients lead to a binding affinity reduction[119]. Anti-TG2 Abs from CD patients bearing the IGHV5 chain bind TG2 on the so-called Epitope 1. This interaction is well characterized, which has been possible thanks to conserved characteristics, whereas antibodies from other autoimmune diseases are not so well conserved. The epitope recognized by IGHV5 is a conformational epitope and we could highlight that as for other autoimmune diseases, like diabetes[173], the epitope is conformational[174]. Moreover, we took advantage of the crystallography structures of the interaction Ab-TG2 available on PDB[65].

These anti-TG2 antibodies are well characterized, but the fine interaction with the antigen needs further investigation. Moreover, some details make these Abs very peculiar and intriguing. We first investigated *in vitro* the fine structure of the VH of these antibodies, to try to draw out the reasons of the generation of these so peculiar and special antibodies. Furthermore, the recurrent characteristics make these antibodies perfect subjects for *in silico* approaches. *In silico* approaches are based on the analyses of anti-TG2 antibodies, which allows the generation the perfect sequence involved in the binding with the antigen.

We could exploit these characteristics to generate a defined phage display library of scFvs. Moreover, since 1% of the population is affected by CD [175], we used CD patients intestinal biopsy to isolate lymphocytes. Biopsy from 55 CD patients could be used to generate a scFv library, the high number of CD patients allowing for the generation of a library without bias due to a low sample. As the *IGHV5* gene family is the most represented among CD patients, only this gene family was used and assembled with VLs. Referred VL chains belong to anti-TG2 scFvs from a previous study [52]. 4 VL chains have been chosen, and they were from three VL gene families: *IGKV1-5*, *IGKV1-39* and *IGL2-14*. This choice was taken based on the study of the preferred pairing of VH and VL[113,119]. We decided to generate libraries with VL from anti-TG2 scFv because previous study indicate that IGHV5 chains require specific VL chains to bind to the antigen[150].

ScFv displayed on phage surface were selected for TG2 reactivity, and TG2 binding clones were deeply analyzed. 77% positive clones were characterized by the use of IGHV5 and 23% by that of

IGHV1 (the presence of this gene family is due to the degeneration and poly-specificity of the primers used to generate the library[149]). The TG2 positive clones characterized by IGHV1 have been considered for the comparison for specific characteristics of IGHV5. The analyses of VL revealed that using phage display technology, the preferential pairing of IGHV5-51 with IGKV1-5 is maintained. In fact, 79% of the scFv reactive with TG2 show this pairing, 18% show a pairing with IGKV1-39, whereas only 3% of IGHV5 which generate TG2 specific clones were pared with IGLV2-14. These data are in agreement with previous study in which anti-TG2 antibodies from CD patients, isolated from gut lesion plasma cells, had been analyzed [119]. Roy and colleagues identified IGHV5-51 as preferentially paired with IGKV1-5. Moreover, IGKV1-39 was also found, meaning that κ Ls were preferentially used, whereas IGLV2-12 was identified paired, although with lower frequency than other VLs. These data suggest that the usage of phage display technology preserves the preferred gene usage.

The analyses of VH5 chains from TG2 reactive clones reveal features characteristic of anti-TG2 from CD patients as well as a limited number of somatic hypermutations, comparable to a Naïve library[113,119]. These data confirm the goodness of our samples and technology. Here, we deeply analyzed VH. First, we underlined that the less mutated regions of VH5 are less mutated than VH1, and not only than non-anti-TG2 Abs. After we focused specifically on three hypervariable loops, representing the antibody recognition site. H1 and H2 were conserved, and no significative differences were identified, neither in the amino acid sequence nor in the three-dimensional structure. This is probably due to a structural role in the maintenance of the correct structure of the binding domain. These data reflect the canonical antibody structure [105,164–166]. Therefore, we analyzed the most variable region, i.e. the loop H3, that is the most variable region derived by the rearrangement of VDJ segments. Our TG2 positive clones have a fixed IGHV5 fragment, so we analyzed DJ pairing. The analyses suggested a preferred pairing never seen before [119], which could however depend on the fixed VL. Analyzing the position on the chromosome of the preferred D and J segments, we showed that this preferred pairing could not be due to the position. It could however be due to the sequences for rearrangements, but further studies will be done.

Our data are in agreement with previous ones which showed stronger bias in CDRH3 lengthening than naïve libraries, and 14 amino acids is the most identified CDRH3 length [119]. Hence, we focused on the sequence of this VH CDRH3 14 amino acids long. As is easily observable, there are 4 positions which are recurring in more than 44% of the CDRH3 sequences (Fig. 18), whereas, CDRH3 from naïve

sequences were characterized by a different profile (Fig. 18). Amazingly, one of the conserved amino acid of the CDRH3 is generated by the nucleotides of the junction between IGHV and IGHD. These fiundings pave the street to further experiments and analyses of a larger number of positive clones. The presence of recurrent amino acids of CDRH3 make them ideal to generate *in silico* a "*consensus*" sequence. An ideal TG2-positive H3 sequence for experimental testing was obtained as a plurality-rule consensus sequence based on the HMM of the set of CDRH3 from TG2 reactive sequences. Amazingly, the ideal "*consensus*" sequence was not present neither in the panel of anti-TG2 antibodies nor in TG2 positive clones identified by Roy and colleagues[119] or in antibody repertoires from non-CD donor[121]. This consensus CDRH3 was cloned in a well characterized antibody: 679-14-E06[113]. ScFv with consensus CDRH3 was assessed for TG2 reactivity. The consensus clone shows a specificity for Epitope 1 on TG2 (as the parental 679-14-E06), and this antibody traces the features of anti-TG2 antibodies from CD patients intestinal biopsy with the IGHV5 chain[113,118]. The phage display technology allows a rapid generation of scFv Ab formats. However, these formats do not reflect the physiological condition. Hence more physiological Ab formats were generated, and the reactivity was maintained even in more physiological formats, as IgG and IgA. IgA is the isotype mostly produced by CD patients. IgA with the consensus sequence confirms the specificity for TG2, whereas no polyreactive with TG3 (transglutaminase 3) was revealed (data not shown). These data indicate that is possible to generate and screen a phage display library, and, thanks to identified clones, generate *in silico* CDRH3 with specificity for the antigen, a sequence that is not present in physiological conditions. This approach could be used to identify a *consensus* CDRH3 and key amino acids even in antibodies specific for other antigens. The power of the phage display to be used to identify antibodies against toxic or self-antigens, combined with the *in silico* study, allows the identification and generation of antibodies against different targets. Moreover, after the identification of the scFv, VH and VL can be cloned in vectors for the expression of Igs.

Furthermore, to deeply investigate the role of amino acids in CDRH3, random CDRH3 sequences were generated by phage display in the same scaffold of 679-14-E06 sequences upon fixing the above positivity conditions and tested for positivity to the TG2 binding. Reactivity assays demonstrated the ability of the above criteria to produce TG2-directed Abs (Fig. 30).

As an experimental 3D structure is available for a celiac antibody matching the conditions of our positive consensus sequence[65], we could model with high reliability the newly obtained Abs experimentally shown to be positive or negative to the TG2 recognition. Based on the analysis of

these 3D models we could identify a fifth positivity criterion. The combined computational and experimental approach we adopted here led to an unprecedented successful design of TG2-directed Abs, based on a detailed insight of the structural determinants required for Ab binding to TG2 Epitope 1.

Finally, binding affinity measurements indicate that our strategy for the design of an ideal positive sequence produced a sequence with enhanced affinity for TG2 binding. Furthermore, anti-TG2 antibodies from CD patients shown a higher affinity than other autoantibodies [176]. These data suggest that the combination of *in vitro* and *in silico* technologies is a good instrument to generate specific and high affinity antibodies.

The contribution of each H3 position to the Abs affinity for TG2 Epitope 1 was investigated by site-directed mutagenesis. 8 mutants have been generated, four with high affinity and 4 with low affinity, and we could identify amino acids in defined positions that admit the binding. The data of scFv displayed on phage surface was in support of our hypotheses (Fig. 32). However, some mutants displayed on phage surface could be less expressed than others and phage ELISA is less sensitive than ELISA performed with scFv-Fc or Igs. In fact, after producing the more stable and quantified scFv-Fc format, the data revealed no significant reduction in binding affinity (Fig. 35), whereas MutPP shows a significant reduction, supporting in a new and different way the importance of the CDRH3 in the specific binding between these antibodies and Epitope 1 on TG2.

The combination of *in vitro* and *in silico* analyses generate a powerful tool to identify high affinity antibodies, allowing the generation of antibodies that could be used in a spread fields in antibody research.

Future studies could be more focused on the *is silico* analyses, thanks to technology of next-generation sequencing, that allow to sequence millions of clones. After a selection cycle, all the clones could be sequenced, and a list of hypothetic *consensus* sequence identified. This could be a wonderful technology that could allow the generation of high specificity antibodies in a very fast way.

6 Acknowledgement

I would like to thank *IRCCS materno infantile Burlo Garofolo, Trieste,* to provide the CD patients' intestinal biopsy, and all CD patients that allowed this study.

I would like to thank Fabiana Ziberna from *IRCCS materno infantile Burlo Garofolo, Trieste,* for the help in the generation of the phage display libraries and the sequencing.

I would like to thank Maria Felicia Soluri from *Department of health Sciences university of Eastern Piedmont Novara,* for the help in the generation of the phage display libraries.

I would like to thank Anna Vangone and Romina Oliva, respectively from *KAUST Catalysis center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia and Department of Sciences and Technologies, University "Parthenope" of Naples,* for bioinformatic analyses.

I would like to thank my PhD colleague Francesca Marano from *Department of Life Sciences, University of Trieste,* for the construction of the Naïve library and the vectors to generate Full length Igs.

I would like to thank Deborah Ferrante from *DiaSorin, Gerenzano*, to help me to validate affinity of the antibodies using a different technology (data not shown).

Finally, I would like to thank Daniele Mistretta, from *Department of Life Sciences, University of Trieste,* for the help in the analyses of the mutant clones.

7 Bibliography

1. Losowsky, M. S. and Gee, S. A History of Coeliac Disease. 112–120 (2008).

2. Dicke, W. K. *et al.* Coeliac Disease 11. The Presence in Wheat of a Factor Having a Deleterious Effect in Cases of Coeliac Disease. *Acta Pediatr.* **42**, 34–42 (1953).

3. Goggins, M. and Kelleher, D. Celiac Disease and Other Nutrient Related Injuries to the Gastrointestinal Tract. *The Amer. J. of Gastroenterol.* **89**, s2–s17 (1994).

4. Karell, K. *et al.* HLA Types in Celiac Disease Patients not Carrying the DQA1 * 05-DQB1 * 02 (DQ2) Heterodimer : Results From the European Genetics Cluster on Celiac Disease. *Hum. Immunol.* **02**, (2003).

5. Dieterich, W. *et al.* Identification of tissue transglutaminase as the autoantigen of celiac disease. *Nat. Med.* **3**, 797–801 (1997).

6. Schuppan, D. *et al.* Celiac Disease: From Pathogenesis to Novel Therapies. *YGAST* **137**, 1912–1933 (2009).

7. Laura, K. *et al.* Presentation of Celiac Disease in Finnish Children Is No Longer Changing: A 50-Year Perspective. *J. Pediatr.* 1109–1116 (2015).

8. Volta, U. *et al*. The changing clinical profile of celiac disease : a 15-year experience (1998-2012) in an Italian referral center. *BMC Gastroenterol*. 1–8 (2014).

9. Annette, K. T. *et al.* Celiac Disease. *Gene Rev*.1–20 (2019).

10. Husby, S. and Bai, J. C. Follow-up of Celiac Disease. *Gastroenterol. Clin. NA* (2018).

11. Lindfors, K. *et al.* Coeliac disease. *Nat. Rev. Dis. Prim.* doi:10.1038/s41572-018-0054-z

12. Ciclitira, P. J. AGA technical review on celiac sprue. *Gastroenterol.* (2001).

13. Fasano, A. and Catassi, C. Current approaches to diagnosis and treatment of celiac disease: An evolving spectrum. *Gastroenterol.* **120**, 636–651 (2001).

14. Green, P. H. R. *et al.* Characteristics of adult celiac disease in the USA: Results of a national survey. *Am. J. Gastroenterol.* (2001).

15. Singh, P. *et al.* Global Prevalence of Celiac Disease : Systematic Review. *Clin. Gastroenterol. Hepatol.* **16**, 823–836.e2. (2018).

16. Mustalahti, K. *et al.* The prevalence of celiac disease in Europe: Results of a centralized, international mass screening project. *Ann. Med.* (2010).

17. Ramakrishna, B. S. *et al.* Prevalence of Adult Celiac Disease in India: Regional Variations and Associations. *Am. J. Gastroenterol.* **Jan**, 115–23 (2016).

18. Shamir, R. *et al.* The use of a single serological marker underestimates the prevalence of celiac disease in Israel: A study of blood donors. *Am. J. Gastroenterol.* (2002).

19. Falchuk, Z. M. *et al.* Predominance of histocompatibility antigen HL-A8 in patients with gluten-sensitive enteropathy. *J. Clin. Invest.* (1972).

20. Stokes, P. L. *et al.* Histocompatibility antigens associated with adult cœliac disease. *Lancet* (1972).

21. Singh, P. *et al*. Risk of celiac disease in the first- and second-degree relatives of patients with celiac disease: A systematic review and meta-analysis. *Amer. J. of Gastroenterol.* (2015).

22. Van Heel, D. A. *et al.* A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat. Genet.* (2007).

23. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* (2011).

24. Gutierrez-Achury, J. *et al.* Fine mapping in the MHC region accounts for 18% additional genetic risk for celiac disease. *Nat. Genet.* (2015).

25. Wolters, V. M. and Wijmenga, C. Genetic background of celiac disease and its clinical implications. *Amer. J. of Gastroenterol.* (2008).

26. MacDonald, T. T. and Monteleone, G. Immunity, inflammation, and allergy in the gut. *Science* (2005).

27. Kondrashova, A. *et al.* Lower economic status and inferior hygienic environment may protect against celiac disease. *Ann. Med.* (2008).

28. Kemppainen.M., K. *et al.* Factors That Increase Risk of Celiac Disease Autoimmunity After a Gastrointestinal Infection in Early Life. *Clinical Gastroenterology and Hepatology* (2017).

29. Andrén Aronsson, C. *et al.* Effects of Gluten Intake on Risk of Celiac Disease: A Case-Control Study on a Swedish Birth Cohort. *Clin. Gastroenterol. Hepatol.* (2016).

30. Sollid, L. M. and Jabri, B. Triggers and drivers of autoimmunity: Lessons from coeliac disease. *Nat. Rev. Immunol.* **13**, 294–302 (2013).

31. Caio, G. *et al.* Celiac disease : a comprehensive current review. 1–20 (2019).

32. Shan, L. *et al.* Structural basis for gluten intolerance in Celiac Sprue. *Science (80-.).* (2002).

33. Stamnaes, J. *et al.* Enhanced B-cell receptor recognition of the autoantigen transglutaminase 2 by efficient catalytic self-multimerization. *PLoS One* **10**, 1–19 (2015).

34. Fasano, A. *et al.* Prevalence of Celiac disease in at-risk and not-at-risk groups in the United States: A large multicenter study. *Arch. Intern. Med.* (2003).

35. Sollid, L. *et al.* Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer. *J. Exp. Med.* **169**, 345–350 (1989).

36. Dieli-Crimi, R. *et al.* The genetics of celiac disease: A comprehensive review of clinical implications. *J. of Autoimmun.*(2015).

37. Pisapia, L. *et al.* HLA-DQ2.5 genes associated with celiac disease risk are preferentially expressed with respect to non-predisposing HLA genes: Implication for anti-gluten T cell response. *J. Autoimmun.* (2016).

38. Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* (2010).

39. Serena, G. *et al*. Genetic and Environmental Contributors for Celiac Disease. (2019).

40. Hunt, K. A. and Van Heel, D. A. Recent advances in coeliac disease genetics. *Gut* (2009).

41. Schirru, E. *et al.* High frequency of low-risk human leukocyte antigen class II genotypes in latent celiac disease. *Hum. Immunol.* (2011).

42. Molberg, Ø. *et al.* Tissue transglutaminase selctively modifies gliadin peptides that are recognized by gut-derived T cells in celiac disease. *Nat. Med.* **4**, 713–717 (1998).

43. Tollefsen, S. *et al.* HLA-DQ2 and -DQ8 signatures of gluten T cell epitopes in celiac disease. *J. Clin. Invest.* (2006).

44. Parra-Medina, R. *et al.* Prevalence of celiac disease in Latin America: A systematic review and meta-regression. *PLoS ONE* (2015).

45. van de Wal, Y. *et al.* Small intestinal T cells of celiac disease patients recognize a natural pepsin fragment of gliadin. *Proc. Natl. Acad. Sci.* (2002).

46. Rossjohn, J. and Koning, F. A biased view toward celiac disease. *Mucosal Immunol.* (2016).

47. Bodd, M. *et al.* HLA-DQ2-restricted gluten-reactive T cells produce IL-21 but not IL-17 or IL-22. *Mucosal Immunol.* (2010).

48. Stamnaes, J. and Sollid, L. M. Celiac disease: Autoimmunity in response to food antigen. *Seminars in Immunol.* (2015).

49. Sollid, L. M. *et al*. Autoantibodies in coeliac disease: tissue transglutaminase - guilt by association? *Gut* **41**, 851–852 (1997).

50. Martucciello, S. *et al*. Anti-type 2 transglutaminase antibodies as modulators of type 2 transglutaminase functions: a possible pathological role in celiac disease. *Cell. and Molec. Life Scien.* (2018).

51. Korponay-Szabó, I. R. *et al.* In vivo targeting of intestinal and extraintestinal transglutaminase 2 by coeliac autoantibodies. *Gut* (2004).

52. Marzari, R. *et al.* Molecular dissection of the tissue transglutaminase autoantibody response in celiac disease. *J. Immunol.* **166**, 4170–4176 (2001).

53. Di Niro, R. *et al.* Responsive population dynamics and wide seeding into the duodenal lamina propria of transglutaminase-2-specific plasma cells in celiac disease. *Mucosal Immunol.* (2016).

54. Iversen, R. *et al.* Strong Clonal Relatedness between Serum and Gut IgA despite Different Plasma Cell Origins. *Cell Rep.* **20**, 2357–2367 (2017).

55. Rauhavirta, T. *et al.* Transglutaminase 2 and Transglutaminase 2 Autoantibodies in Celiac Disease: a Review. *Clinical Reviews in Allergy and Immunology* (2019).

56. Korponay-Szabo´, I. R. *et al*. Elevation of IgG antibodies against tissue transglutaminase as a diagnostic tool for coeliac disease in selective IgA deficiency. *Gut* **52**, 1567–1571 (2003).

57. Husby, S. *et al.* European society for pediatric gastroenterology, hepatology, and nutrition guidelines for the diagnosis of coeliac disease. *J. of Pediat. Gastroenterol. and Nutri.* (2012).

58. Collin, P. *et al*. The safe threshold for gluten contamination in gluten-free products. Can trace amounts be accepted in the treatment of coeliac disease? *Aliment. Pharmacol. Ther.* (2004).

59. Griffin, M. *et al.* Transglutaminases: Nature's biological glues. *Biochemical J.* (2002).

60. Lee, C. S. and Park, H. H. Structural aspects of transglutaminase 2: functional, structural, and regulatory diversity. *Apoptosis* (2017).

61. Gentile, V. *et al.* The human tissue transglutaminase gene maps on chromosome 20q12 by in situ fluorescence hybridization. *Genom.* (1994).

62. Iversen, R. *et al*. Activity-regulating structural changes and autoantibody epitopes in transglutaminase

2 assessed by hydrogen/deuterium exchange. *Proc. Natl. Acad. Sci.* **111**, 17146–17151 (2014).

63. Liu, S. *et al.* Structural basis for the guanine nucleotide-binding activity of tissue transglutaminase and its regulation of transamidation activity. *Proc. Natl. Acad. Sci. U. S. A.* (2002).

64. Stamnaes, J. *et al*. Redox regulation of transglutaminase 2 activity. *J. Biol. Chem.* **285**, 25402–25409 (2010).

65. Chen, X. *et al.* Structural basis for antigen recognition by transglutaminase 2-specific autoantibodies in celiac disease. *J. Biol. Chem.* **290**, 21365–21375 (2015).

66. Siegel, M. *et al.* Extracellular transglutaminase 2 is catalytically inactive, but is transiently activated upon tissue injury. *PLoS One* (2008).

67. Folk, J. E. and Finlayson, J. S. The ε-(γ-Glutamyl)Lysine Crosslink and the Catalytic Role of Transglutaminases. *Adv. Protein Chem.* (1977).

68. Lorand, L. and Graham, R. M. Transglutaminases: Crosslinking enzymes with pleiotropic functions. *Nature Reviews Molecular Cell Biology* (2003).

69. Nakaoka, H. *et al.* Gh: A GTP-Binding protein with transglutaminase activity and receptor signaling function. *Science (80-.).* (1994).

70. Iismaa, S. E. *et al.* The core domain of the tissue transglutaminase G(h) hydrolyzes GTP and ATP. *Biochemistry* (1997).

71. Mishra, S. and Murphy, L. J. Tissue transglutaminase has intrinsic kinase activity. Identification of transglutaminase 2 as an insulin-like growth factor-binding protein-3 kinase. *J. Biol. Chem.* (2004).

72. Akimov, S. S. *et al.* Tissue transglutaminase is an integrin-binding adhesion coreceptor for fibronectin. *J. Cell Biol.* (2000).

73. Akimov, S. S. and Belkin, A. M. Cell surface tissue transglutaminase is involved in adhesion and migration of monocytic cells on fibronectin. *Blood* (2001).

74. Wang, Z. *et al.* RGD-independent cell adhesion via a tissue transglutaminase-fibronectin matrix promotes fibronectin fibril deposition and requires syndecan-4/2 and α5β1 integrin co-signaling. *J. Biol. Chem.* (2010).

75. Villanacci, V. *et al.* Mucosal tissue transglutaminase expression in celiac disease. *J. Cell. Mol. Med.* (2009).

76. Fleckenstein, B. *et al.* Gliadin T cell epitope selection by tissue transglutaminase in celiac disease. Role

of enzyme specificity and pH influence on the transamidation versus deamidation reactions. *J. Biol. Chem.* (2002).

77. Rauhavirta, T. *et al.* Epithelial transport and deamidation of gliadin peptides: A role for coeliac disease patient immunoglobulin A. *Clin. Exp. Immunol.* (2011).

78. Du Pré, M. F. and Sollid, L. M. T-cell and B-cell immunity in celiac disease. *Best Practice and Research: Clinical Gastroenterology* (2015).

79. Murphy, K. *Immunobiology, 8th Edition*. *Garland Science* (2012).

80. Schroeder, H. W. J. and Cavacini, L. Structure and Function of Immunoglobulins (author manuscript). *J. Allergy Clin. Immunol.* (2010).

81. Ribatti, D. Edelman's view on the discovery of antibodies. *Immunology Letters* (2015).

82. Pauling, L. A Theory of the Structure and Process of Formation of Antibodies. *J. Am. Chem. Soc.* (1940).

83. Jerne, N. K. The natural-selection theory of antibody formation. *Proc. Natl. Acad. Sci.* (1955).

84. Porter, R. R. The hydrolysis of rabbit y-globulin and antibodies with crystalline papain. *Biochem. J.* (1959).

85. Edelman, G. M. Dissociation of Γ-globulin. *J. of the American Chemical Society* (1959).

86. Edelman, G. M. Biochemistry and the Sciences of Recognition. *J. of Biological Chemistry* (2004).

87. Edelman, G. M. *et al.* The covalent structure of an entire gammaG immunoglobulin molecule. *Proc. Natl. Acad. Sci. U. S. A.* (1969).

88. Press, E. M. and Hogg, N. M. Comparative study of two immunoglobulin G Fd-fragments. *Nature* (1969).

89. Regueiro Gonzalez, J. R.*et al*. Inmunología: Biología y patología del sistema inmunitario. *Cirugía Española* (2009).

90. Korngold, L. and Lipari, R. Multiple-myeloma proteins. III. The antigenic relationship of Bence Jones proteins to normal gamma-globulin and multiple-myeloma serum proteins. *Cancer* (1956).

91. Maverakis, E. *et al.* Glycans in the immune system and The Altered Glycan Theory of Autoimmunity: A critical review. *J. of Autoimmunity* (2015).

92. Leder, P. The genetics of antibody diversity. *Sci. Am.* (1982).

93. Tonegawa, S. Somatic generation of antibody diversity. *Nature* (1983).

94. Lefranc, M. P. Nomenclature of the human immunoglobulin kappa (IGK) genes. *Exp. Clin. Immunogenet.* (2001).

95. Lefranc, M. P. Nomenclature of the Human Immunoglobulin Heavy (IGH) Genes. *Exp. Clin. Immunogenet.* (2001).

96. LeFranc, M. P. Nomenclature of the Human Immunoglobulin Lambda (IGL) genes. *Exp. Clin. Immunogenet.* (2001).

97. Lefranc, M.-P. IMGT Locus on Focus. *Exp. Clin. Immunogenet.* (1998).

98. Matsuda, F. *et al.* The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *J. Exp. Med.* (1998).

99. Corbett, S. J. *et al.* Sequence of the human immunoglobulin diversity (D) segment locus: A systematic analysis provides no evidence for the use of DIR segments, inverted D segments, 'Minor' D segments or D-D recombination. *J. Mol. Biol.* (1997).

100. Zachau, H. G. The immunoglobulin κ gene families of human and mouse: A cottage industry approach. *Biol. Chem.* (2000).

101. Lefranc, M. P. IMGT® and 30 Years of Immunoinformatics Insight in Antibody V and C Domain Structure and Function. *Antibodies* (2019).

102. Lefranc, M. P. Nomenclature of the Human Immunoglobulin Genes. *Immunology* 1–37 (2000).

103. Te Wu, T. and Kabat, E. A. An analysis of the sequences of the variable regions of bence jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Exp. Med.* (1970).

104. Al-Lazikani, B. et al. Standard conformations for the canonical structures of immunoglobulins. *J. Mol. Biol.* **273**, 927–948 (1997).

105. Chothia, C. and Lesk, A. M. Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196**, 901–917 (1987).

106. Xu, J. L. and Davis, M. M. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* (2000).

107. Hoogenboom, H. R. and Winter, G. By-passing immunisation. Human antibodies from synthetic repertoires of germline VH gene segments rearranged in vitro. *J. Mol. Biol.* **227**, 381–388 (1992).

108. Barbas, C. F.et al. Semisynthetic combinatorial antibody libraries: A chemical solution to the diversity problem. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 4457–4461 (1992).

109. D'Angelo, S. *et al.* Many routes to an antibody heavy-chain CDR3: Necessary, yet insufficient, for specific binding. *Front. Immunol.* (2018).

110. Black, C. A. A brief history of the discovery of the immunoglobulins and the origin of the modern immunoglobulin nomenclature. *Immunology and Cell Biology* (1997).

111. Geisberger, R. *et al.* The riddle of the dual expression of IgM and IgD. *Immunology* (2006).

112. Elli, L. *et al.* Transglutaminases in inflammation and fibrosis of the gastrointestinal tract and the liver. *Dig. Liver Dis.* **41**, 541–550 (2009).

113. Di Niro, R. *et al.* High abundance of plasma cells secreting transglutaminase 2–specific IgA autoantibodies with limited somatic hypermutation in celiac disease intestinal lesions. *Nat. Med.* **18**, 441–445 (2012).

114. Chan, O. and Shlomchik, M. J. A new role for B cells in systemic autoimmunity: B cells promote spontaneous T cell activation in MRL-lpr/lpr mice. *J. Immunol.* (1998).

115. Høydahl, L. S. *et al.* Plasma Cells Are the Most Abundant Gluten Peptide MHC-expressing Cells in Inflamed Intestinal Tissues From Patients With Celiac Disease. *Gastroenterology* (2019).

116. Vader, L. W. *et al.* Specificity of tissue transglutaminase explains cereal toxicity in celiac disease. *J. Exp. Med.* **195**, 643–9 (2002).

117. Iversen, R. *et al.* Efficient T cell–B cell collaboration guides autoantibody epitope bias and onset of celiac disease. *Proc. Natl. Acad. Sci.* **116**, 15134–15139 (2019).

118. Iversen, R. *et al.* Transglutaminase 2-specific autoantibodies in celiac disease target clustered, N-terminal epitopes not displayed on the surface of cells. *J. Immunol.* **190**, 5981–91 (2013).

119. Roy, B. *et al.* High-Throughput Single-Cell Analysis of B Cell Receptor Usage among Autoantigen-Specific Plasma Cells in Celiac Disease. *J. Immunol.* **199**, 782–791 (2017).

120. DeWitt, W. S. *et al.* A public database of memory and naive B-cell receptor sequences. *PLoS One* **11**, 1–18 (2016).

121. DeKosky, B. J. *et al.* In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat. Med.* **21**, 86–91 (2015).

122. Steinsbø, O. *et al.* Restricted VH/VL usage and limited mutations in gluten-specific IgA of coeliac disease lesion plasma cells. *Nat. Commun.* **5**, 4041 (2014).

123. Simon-Vecsei, Z. *et al.* A single conformational transglutaminase 2 epitope contributed by three

domains is critical for celiac antibody binding and effects. *Proc. Natl. Acad. Sci.* **109**, 431–436 (2012).

124. Hnida, K. *et al.* Epitope-dependent functional effects of celiac disease autoantibodies on transglutaminase 2. *J. Biol. Chem.* **291**, 25542–25552 (2016).

125. Stamnaes, J. *et al.* Transglutaminase 2 strongly binds to an extracellular matrix component other than fibronectin via its second C-terminal beta-barrel domain. *FEBS J.* **283**, 3994–4010 (2016).

126. Jang, T. H. *et al.* Crystal structure of transglutaminase 2 with GTP complex and amino acid sequence evidence of evolution of GTP binding site. *PLoS One* (2014).

127. Pinkas, D. M. *et al*. Transglutaminase 2 undergoes a large conformational change upon activation. *PLoS Biol.* (2007).

128. Geyer, C. R.*et al.* Recombinant antibodies and in vitro selection technologies. *Methods in Molecular Biology* (2012).

129. Köhler, G. and Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* (1975).

130. Winter, G. Harnessing Evolution to Make Medicines (Nobel Lecture). *Angew. Chemie Int. Ed.* **58**, 14438–14445 (2019).

131. Winter, G. and Milstein, C. Man-made antibodies. *Nature* (1991).

132. Smith, G. P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228**, 1315–1317 (1985).

133. Orlandi, R. *et al*. Cloning immunoglobulin variable domains for expression by the polymerase chain reaction. *Proc. Natl. Acad. Sci. U. S. A.* (1989).

134. Huse, W. D. *et al.* Generation of a large combinatorial library of the immunoglobulin repertoire in phage lambda. *Science (80-.).* (1989).

135. McCafferty, J. *et al*. Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* **348**, 552–554 (1990).

136. Smith, G. P. Phage Display: Simple Evolution in a Petri Dish (Nobel Lecture). *Angew. Chemie Int. Ed.* **58**, 14428–14437 (2019).

137. Ledsgaard, L. *et al.* Basics of antibody phage display technology. *Toxins (Basel).* **10**, (2018).

138. Azzazy, H. M. E. and Highsmith, W. E. Phage display technology: Clinical applications and recent innovations. *Clin. Biochem.* **35**, 425–445 (2002).

139. Lim, C. C. *et al*. Cognizance of molecular methods for the generation of mutagenic phage display antibody libraries for affnity maturation. *International J. of Molecular Sciences* **20**, (2019).

140. Bird, R. E. *et al.* Single-chain antigen-binding proteins. *Science (80-.).* (1988).

141. Di Niro, R. *et al.* Construction of miniantibodies for the in vivo study of human autoimmune diseases in animal models. *BMC Biotechnol.* **7**, 46 (2007).

142. Dal Ferro, M. *et al.* Chapter 15. *Hum. Monoclon. antibodies* **1904**, 319–338 (2019).

143. Kumar, R. *et al*. Phage display antibody libraries: A robust approach for generation of recombinant human monoclonal antibodies. *Int. J. Biol. Macromol.* **135**, 907–918 (2019).

144. Eddy, S. R. What is a hidden Markov model? *Nat. Biotechnol.* **22**, 1315–1316 (2004).

145. Weitzner, B. D. *et al.* Modeling and docking antibody structures with Rosetta HHS Public Access Author manuscript. *Nat Protoc* **12**, 401–416 (2017).

146. Entzminger, K. C. *et al.* De novo design of antibody complementarity determining regions binding a FLAG tetra-peptide. *Sci. Rep.* (2017).

147. Chothia, C. and Lesk, A. M. Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196**, 901–917 (1987).

148. Kuroda, D. *et al*. Computer-aided antibody design. *Protein Eng. Des. Sel.* **25**, 507–521 (2012).

149. Sblattero, D. and Bradbury, A. A definitive set of oligonucleotide primers for amplifying human V regions. *Immunotechnology* **3**, 271–278 (1998).

150. Sblattero, D. *et al.* One-step cloning of anti tissue transglutaminase scFv from subjects with celiac disease. *J Autoimmun* **22**, 65–72 (2004).

151. Krebber, A. *et al.* Reliable cloning of functional antibody variable domains from hybridomas and spleen cell repertoires employing a reengineered phage display system. *J. Immunol. Methods* **201**, 35–55 (1997).

152. Sblattero, D. and Bradbury, A. Exploiting recombination in single bacteria to make large phage antibody libraries. *Nat Biotechnol* **18**, 75-80. (2000).

153. Di Niro, R. *et al.* Anti-idiotypic response in mice expressing human autoantibodies. *Mol. Immunol.* **45**, 1782–1791 (2008).

154. Fang, J. *et al.* Stable antibody expression at therapeutic levels using the 2A peptide. *Nat. Biotechnol.* **23**, 584–590 (2005).

155. Keele, B. F. *et al.* Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc. Natl. Acad. Sci. U. S. A.* (2008).

156. Crooks, G.*et al*. NCBI GenBank FTP Site\nWebLogo: a sequence logo generator. *Genome Res* **14**, 1188–1190 (2004).

157. Sblattero, D. and Bradbury, A. Exploiting recombination in single bacteria to make large phage antibody libraries. *Nat. Biotechnol.* **18**, 75–80 (2000).

158. Dörner, T. and Lipsky, P. E. Molecular basis of immunoglobulin variable region gene usage in systemic autoimmunity. *Clin. Exp. Med.* **4**, 159–169 (2005).

159. Krebber, A. *et al.* Reliable cloning of functional antibody variable domains from hybridomas and spleen cell repertoires employing a reengineered phage display system. *J Immunol Methods* **201**, 35–55 (1997).

160. F oreman, A. L. et al. B cells in autoimmune diseases: insights from analyses of immunoglobulin variable (Ig V) gene usage. *Autoimmun. Rev*. **86**, 573–579 (2007).

161. Di Niro, R. *et al.* Rapid Generation of Rotavirus-Specific Human Monoclonal Antibodies from Small-Intestinal Mucosa. *J. Immunol.* **185**, 5377–5383 (2010).

162. Wrammert, J. *et al.* Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature* (2008).

163. Scheid, J. F. *et al.* Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals. *Nature* (2009).

164. Chothia, C. *et al.* Conformations of immunoglobulin hypervariable regions. *Nature* **342**, 877–883 (1989).

165. Chothia, C. Structural repertoire of the human VH segments. *J. Mol. Biol.* **227**, 799–817 (1992).

166. North, B., Lehmann, A. and Jr, R. L. D. NIH Public Access. **406**, 228–256 (2012).

167. Sblattero, D. *et al.* One-step cloning of anti tissue transglutaminase scFv from subjects with celiac disease. *J. Autoimmun.* **22**, 65–72 (2004).

168. Fang, J. *et al.* Stable antibody expression at therapeutic levels using the 2A peptide. *Nat Biotechnol* **23**, 584–590 (2005).

169. Di Niro, R. *et al.* SalmOnella Infection Drives Promiscuous B Cell Activation Followed By Extrafollicular Affinity Maturation. *Immunity* (2015). doi:10.1016/j.immuni.2015.06.013

170. Sblattero, D. *et al.* Analyzing the peripheral blood antibody repertoire of a celiac disease patient using phage antibody libraries. *Hum. Antibodies* **9**, 199–205 (2000).

171. Van Esch, W. J. E. *et al.* Human IgG Fc-binding phage antibodies constructed from synovial fluid CD38+ B cells of patients with rheumatoid arthritis show the imprints of an antigen-dependent process of somatic hypermutation and clonal selection. *Clin. Exp. Immunol.* **131**, 364–376 (2003).

172. Zuo, T. *et al*. Affinity war: forging immunoglobulin repertoires. *Current Opinion in Immunology* (2019).

173. Myers, M. A. *et al.* Conformational Epitopes on the Diabetes Autoantigen GAD65 Identified by Peptide Phage Display and Molecular Modeling. *J. Immunol.* (2000).

174. Sblattero, D. *et al.* The analysis of the fine specificity of celiac disease antibodies using tissue transglutaminase fragments. *Eur. J. Biochem.* **269**, 5175–5181 (2002).

175. Green, P. H. R. The many faces of celiac disease: Clinical presentation of celiac disease in the adult population. *Gastroenterology* **128**, (2005).

176. Szarka, E. *et al.* Affinity purification and comparative biosensor analysis of citrulline-peptide-specific antibodies in rheumatoid arthritis. *Int. J. Mol. Sci.* **19**, (2018).