DE NOVO ANTIBODY SEQUENCING BASED ON MASS SPECTROMETRY

Weiwei Peng（彭炜玮）

DE NOVO ANTIBODY SEQUENCING BASED ON MASS SPECTROMETRY
Weiwei Peng（彭炜玮） 2024

De Novo Antibody Sequencing Based On Mass Spectrometry
PhD Thesis, University Utrecht

Weiwei Peng（彭炜玮）, **2024, Utrecht, the Netherlands**

ISBN: 978-94-6483-778-0

De Novo Antibody Sequencing Based On Mass Spectrometry

Weiwei Peng (彭炜玮)

ISBN: 978-94-6483-778-0

Printed by: Ridderprint | www.ridderprint.nl

The research in this thesis was performed in the Biomolecular Mass Spectrometry & Proteomics Group, Utrecht University, Utrecht, The Netherland

The research was financially supported by the Dutch Research Council NWO Gravitation 2013 BOO, Institute for Chemical Immunology (ICI; 024.002.009).

De novo antibody sequencing based on mass spectrometry

de novo antilichaam sequentie bepaling met behulp van massaspectrometrie

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 26 februari 2024 des middags te 2.15 uur

door

Weiwei Peng

geboren op 12 juli 1992
te Hubei, China

Promotor:

Prof. dr. A.J.R. Heck

Copromotor:

Dr. J. Snijder

Beoordelingscommissie:

Prof. dr. C.R. Berkers
Dr. M.G.M. Huijbers
Prof. dr. F.J.M. van Kuppeveld
Prof. dr. R.E.M. Toes
Prof. dr. G. Vidarsson

Chapter 1

Introduction 6

Chapter 2

Mass spectrometry-based *de novo* sequencing of monoclonal antibodies using multiple proteases and a dual fragmentation scheme 47

Chapter 3

Template-based assembly of proteomic short reads for *de novo* antibody sequencing and repertoire profiling 75

Chapter 4

Reverse engineering the anti-MUC1 hybridoma antibody 139H2 by mass spectrometry-based *de novo* sequencing 101

Chapter 5

Structural basis for postfusion-specific binding to Respiratory Syncytial Virus F protein by the antigenic site-I antibody 131-2a. 137

Chapter 6

Direct mass spectrometry-based detection and antibody sequencing of Monoclonal Gammopathy of Undetermined Significance from patient serum – a case study. 157

Chapter 7

Summary, Outlook, samenvatting (Nederlands), 总结 （中文）, Zusammenfassung (Deutsch), Acknowledgements and Curriculum vitae 179

CHAPTER

1

Introduction

Antibodies are the main mediators of adaptive humoral immunity originating from B lymphocytes, elicited primarily by foreign antigens, for example, bacteria, viruses, parasites and other microorrganisms[1]. Relying on their sequence diversity, antibodies protect our body from pathogens via neutralization of infectivity, mediating phagocytosis, eliciting antibody-dependent cellular cytotoxicity (ADCC), and triggering complement-mediated lysis of pathogens and infected cells[2]. The exceptional specificity and efficiency of antibodies have positioned them at the forefront of drug discovery and therapeutics development. Since the approval of the first therapeutic monoclonal antibody by the FDA in 1986, there has been a resurgence of antibody-based therapies in the 21st century, with an annual average of nearly 10 newly approved antibodies by the FDA since 2014[3].

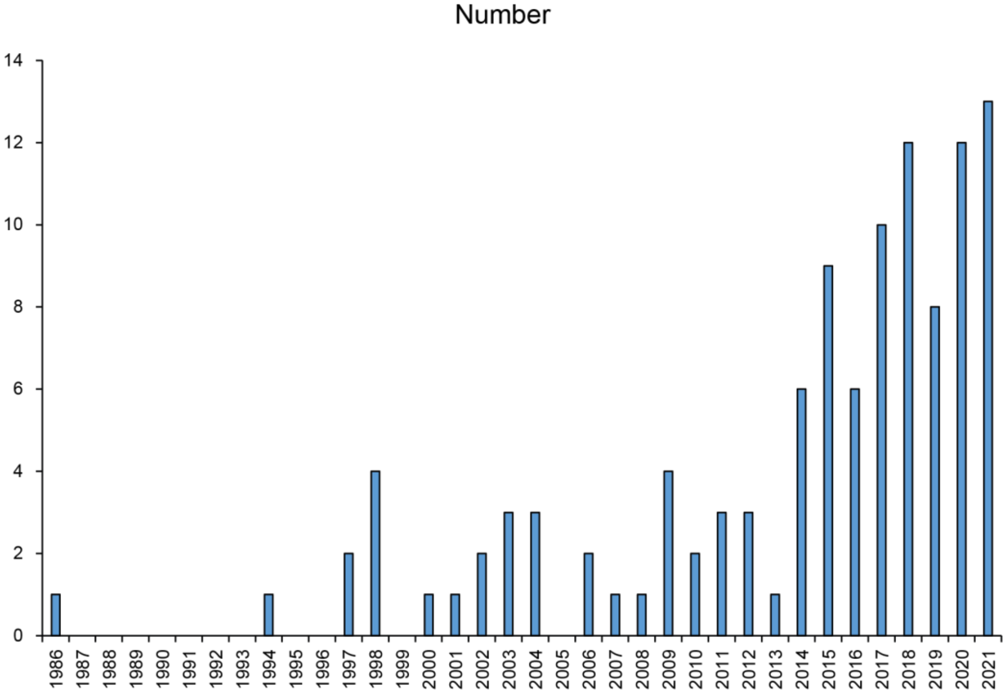


Figure 1: The number of FDA approved antibody therapeutics from 1986 to 2021[3]

There are five isotypes of antibodies in human: IgM, IgG, IgA, IgE and IgD. While all naïve, immature B cells, before stimulation by antigens, present IgM on their surface, the remaining subclasses are generated during B cell maturation through class switching[1,4–6]. Among these isotypes, IgG is the most abundant and versatile antibody in circulation. It plays a major role in adaptive immunity and is involved in long-term immune memory.

IgG is also the most used isotype in immunotherapies [7–10]. In the 1950s, Porter found that the intact IgG can be proteolytic cleaved into two distinct fragments: the fragment antigen binding (Fab) region and the Fragment crystallizable (Fc) region[11]. Decades later, crystallographic studies revealed the molecular structure of IgG. IgG consists of four polypeptide chains: two identical heavy chains and two identical light chains, yielding a molecular weight of around 150 kDa[12,13]. Both chains consist of multiple domains with the characteristic Immunoglobulin (Ig) fold, connected by flexible linkers. The heavy chain comprises four Ig domains, while the light chain comprises two Ig domains. The heavy chain and light chain are covalently linked via disulfide bonds. Two heavy chains form disulfide bonds within the flexible linker region connecting the 2nd and 3rd Ig domain. This region, known as hinge region, serves as a distinct boundary, delineating the Fab and Fc regions. Functionally, both heavy chain and light chain can be divided into a variable (V) region and a constant (C) region. The V-region consists of the first Ig-domain in both heavy and light chain, and is responsible for antigen binding, while the C region is formed by the remaining Ig-domains and mediates interactions with immune cells through a host of Fc receptors, is involved in oligomerization of higher-order Ig complexes, interacts with the complement system, and determines the antibody isotype. The Fab region is made up of the full light chain and the heavy chain above the hinge region, while the Fc region derives solely from the heavy chain below the hinge region. There is one conserved N-linked glycan located in the Fc region of each heavy chain, which modulates oligomerization and immune effector functions.[14] The Ig-domains of the Variable regions each expose three highly variable loops called complementarity-determining regions (CDRs), which play a key role in antigen recognition, binding and neutralization. The three CDR loops on the V-regions are situated within four relatively conserved sequences called the framework. Antibodies can be expressed as B cell antigen receptor (BCR) in a membrane bounded form, but can also be secreted as soluble protein. The main difference between the BCR and secreted immunoglobulin lies in the C-terminus: the BCR has a hydrophic transmembrane peptide while the secreted antibody has a hydrophilic peptide[15]. In contrast with the monomeric BCR, the secreted antibody can form higher order oligomers, from dimers to hexamers, which are mediated by direct interactions in the Fc region[16].

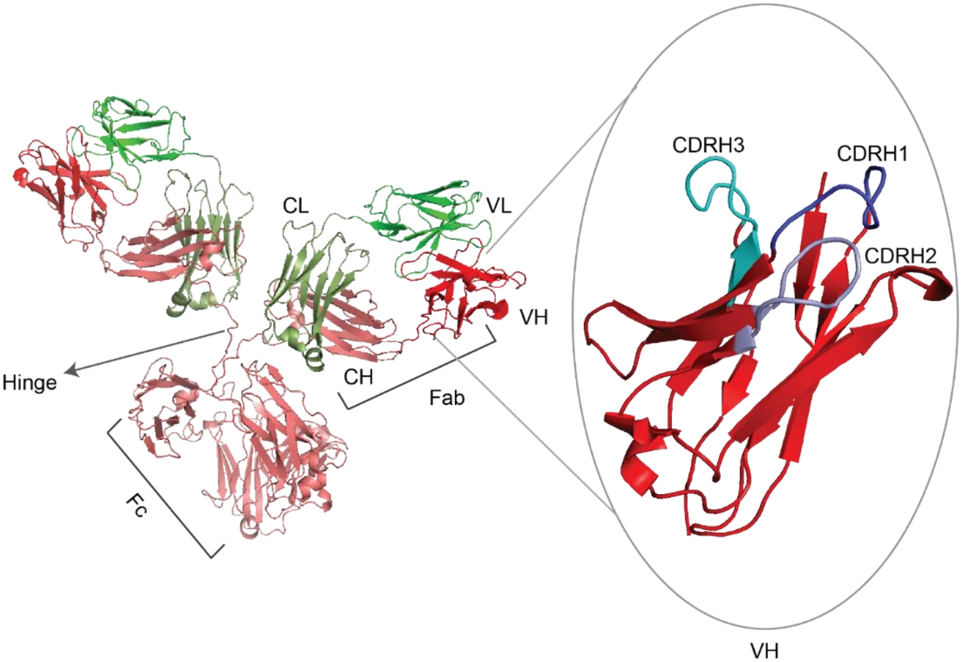


Figure 2: The crystal structure of a IgG1 molecule (PDB: 1IGY): A IgG molecule consists of two identical heavy chains and two identical light chains[13]. Three CDRs loops are presented on the surface of the V-regions of each heavy and light chain. The figure was made with PYMOL using IMGT numbering scheme[17].

Antibody sequence diversity stands as a hallmark of the adaptive immune system, enabling the recognition and response to an almost infinite array of antigens. This sequence diversity stems from somatic recombination and hypermutation of the coding gene segments (called V-, D- and J-segments), as well as class switching and heavy-light chain pairing. As the name suggests, antibody sequence diversity occurs primarily in the Variable region of the heavy and light chains. Around 2.7×10^6 different antibody molecules can be generated due to the gene rearrangement and heavy-light chain pairings alone (see table 1)[18]. During the rearrangement of the heavy chain, a D segment is firstly joined to a J segment, followed by rearrangement of a V segment to the combined DJ segment to complete the V region. The main antigen-binding loop CDR3 is situated right at the junction of the recombined gene segments, which consequently is the region of highest variability in typical antibodies. The highly variable D segment is exclusively present in the heavy chain, whereas in the light chain the V-segment is connected straight to the J-segment. Consequently, sequence diversity is typically higher in the heavy chain than the light chain. There are two main types of light chain, denoted as kappa (κ) and lambda (λ), which each draw from their own repertoire of coding V/J/C segments. After the gene of the heavy chain is rearranged, the rearrangement of a κ light chain will start. If this fails, the B cell will proceed to the rearrangement in the λ light chain. As depicted in figure 3, this somatic recombination of gene segments happens in the bone marrow, when the progenitor B cells develop to the immature B cells.

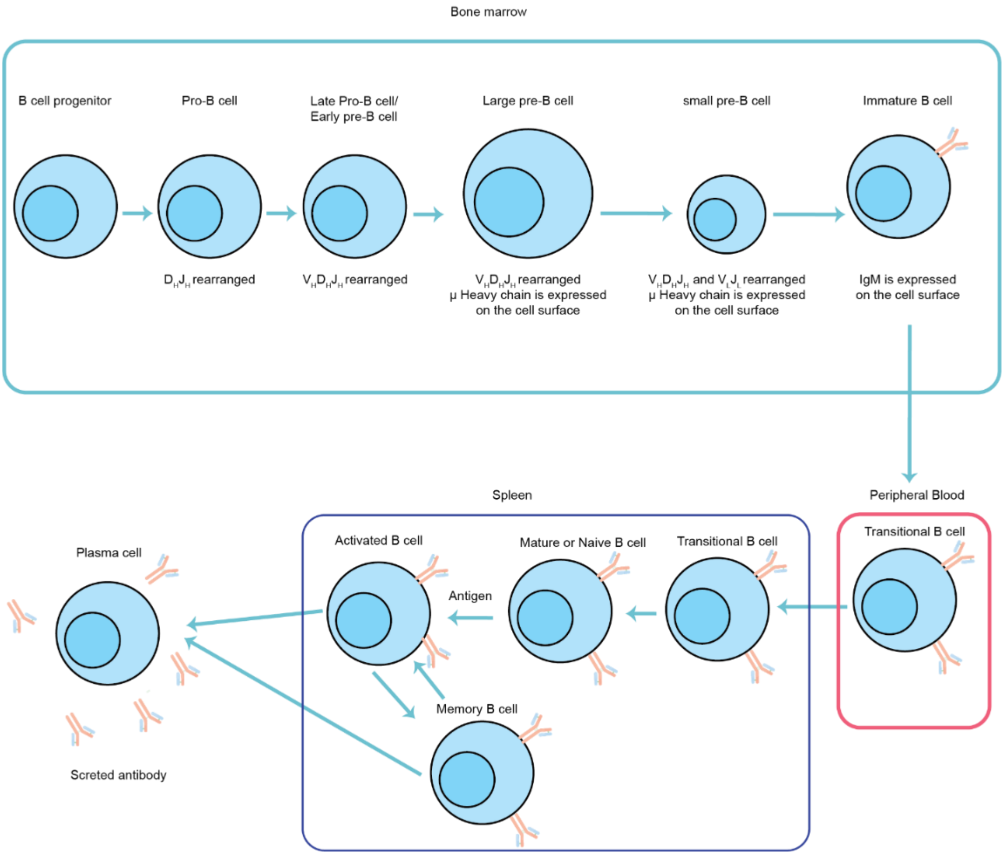


Figure 3: Schematic overview of B cell development: B cell progenitor process to immature B cells in the bone marrow, accompanied with V-(D)-J gene recombination; The immature B cells are activated by the antigen in spleen and developed to the plasma cells and secrete antibody.

Table 1: Number of human V-(D)-J gene segment of heavy chain and light chain, data collected from IMGT database[18]

Segment	Heavy chains	Light chains	
	H	κ	λ
V	51	40	30
D	27	0	0
J	6	5	4

The immature B cells are transported through the peripheral blood into lymphoid tissue, where they undergo maturation into activated B cells and memory B cells upon exposure to antigen. A few days after antigen exposure, small B cells clusters are formed in the so-called germinal center (GC) of lymphoid tissues. The GC can be divided into distinct zones: a proximal dark zone (DZ) and a distal light zone (LZ). T follicular helper cells (Tfh), follicular dendritic cells (FDCs), and the B cells that are close to exiting the cell cycle are presented in the LZ[19]. In the DZ, a heightened mutation rate of approximately 10-3 per base pair occurs at the coding regions for the heavy and light chains, which is thousand times higher than in normal cells[20–22]. This process, termed somatic hypermutation (SHM), is facilitated by activation-induced deaminase (AID), leading to single point mutations primarily in the V regions of antibodies[20,23]. Though during the SHM, low affinity antibodies can also be produced, the B cells that produce higher affinity antibodies have an advantage in competing for limiting growth resources. B cells are competing for binding to the FDCs that present antigen on their cell surface. Following binding, the antigen is taken into the B cell and processed through the immunoproteasome, then presented using the major histocompatibility complex class-II molecules (MHC-II)[24]. Consequently, the B cells that express BCRs with higher affinity are able to present more antigen-loaded MHC-II to have better access to the Tfh's, which helps them to continue participating in the cyclic reentry into the GC, while low affinity B cells face limited assistance and ultimately undergo apoptosis[25–27]. The proliferation rate of B cells in the GC is influenced by the amount of Tfh binding to the B cells as well, allowing B cell clones with higher affinity antibodies to expand to a greater degree and further diversify their antibody genes through prolonged SHM[28]. Memory B cells can persist after the antigen stimulation and rapidly expand during secondary responses and differentiate into antibody–secreting plasma cells. The majority of these long-lived plasma cells are hosted in the bone marrow, also in lymphoid organs and in non-lymphoid organs in disease. The high affinity antibodies, with half-lives of 40-80 days, are rapidly secreted by these plasma cells [29,30]. Somatic hypermutation tends to accumulate within the CDRs, rather than the framework region. From an evolutionary standpoint, mutations in the CDR are more likely to improve affinity for the antigen to enhance survival in the GC, while excessive SHM in the framework region has the potential to disrupt antibody structural integrity, potentially leading to non-functional BCRs and subsequent cellular apoptosis. Furthermore, mutable codons, such as serine AGY, are more frequently utilized in the CDR regions, while conserved codons, such as TCN, are favored within the framework region[33]; Additionally, the AID hotspots in the CDRs and the enhancer elements may also help to promote the SHM in the CDR[34–38]. Exceptions to this general pattern of SHM have been reported for HIV broadly neutralizing antibodies, which can accommodate as many as 50 mutations in the framework regions that also contribute to antigen binding and neutralizing activity[29,30].

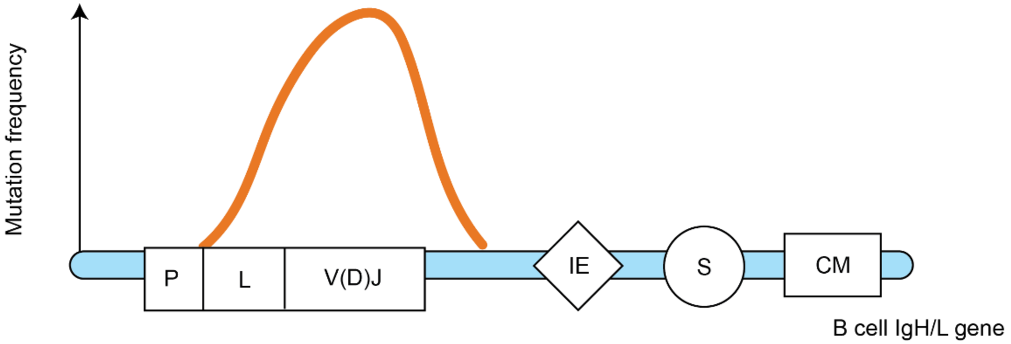


Figure 4: Somatic hypermutation in B cell Ig gene. Mutation frequency along the Ig genes (Leader (L), Promoter (P) , V(D)J, intronic enhancer (IE) , the switch region (S) and the C region genes (CM)) are illustrated: somatic hypermutation occurs in the V(D)J genes and its mutation rate is estimated to be 10-3 per base pair compared with 10-6 per base pair in normal[1,20].

Antibody isotypes and class switch recombination

There are nine antibody subclasses in human: IgM, IgD, IgG1, IgG2, IgG3, IgG4, IgA1, IgA2 and IgE. Before encountering antigens, the immature B cells only generate IgM on the cell surface as receptor. Once the B cells meet the antigen, B cells can switch the antibody subclasses under the influence of T cell and cytokine while retaining the same V region of the heavy chain, as well as the full light chain. Thus, the antigen specificity remains identical. Figure 5 illustrates how IgM is switched to IgG2 via gene recombination[1]. Normally the class switching happens in the mature B cells in the secondary lymphoid tissues. Like SHM, the class switching is also initiated by the AID[39]. As shown in table 2 below, different antibody subclasses have unique biological functions. Though sharing the same antigen-binding variable region, diverse antibody subclasses can also influence the neutralization breadth towards antigen such as SARS-COV-2 spike protein or HIV glycoprotein[40–43]. Despite the fact that most described anti-viral monoclonal antibodies are of the IgG1 subclass, several studies show that compared with IgG1, IgG3 and IgA may exhibits stronger neutralization breadth[41,43–47]. Therefore, the selection of the correct antibody subclasses is also a crucial consideration for antibody therapeutics development.

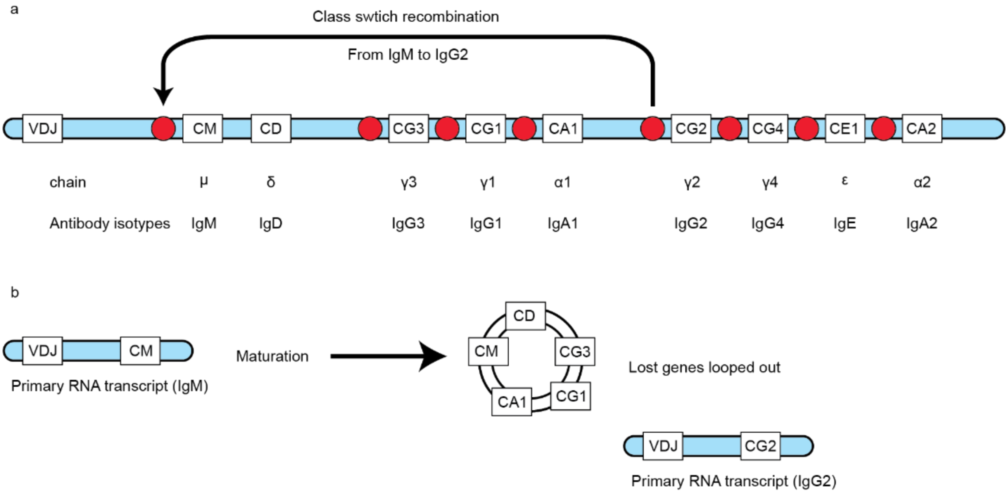


Figure 5: Class switch recombination in human[1,2,48] *Here illustrates an example how IgM is switched to IgG2. Seen in panel b left, the VDJ region undergoes initial transcription along with the M gene, producing the mRNA for IgM. During the maturation of B cells, the class switching can occur under the influence of T cells and cytokines. The intervening region, comprising the genes encoding IgM, IgD, IgG3, IgG1, and IgA1, is looped out and cleaved. Afterwards the two switch regions are joined together and leads to the transcription of the VDJ region with the IgG2 gene, ultimately producing IgG2. The red dots indicate the switching sequence.*

Current methods for antibody discovery and repertoire profiling

Given the crucial role of antibody sequence diversity in shaping adaptive immunity, methodology to select and sequence (antigen-specific) antibodies of interest has been key to the understanding of the immune repertoire and the discovery of monoclonal antibodies for therapeutic, diagnostic, and research applications. Over the years, an assortment of innovative antibody discovery methods has emerged, each offering unique advantages in terms of specificity, efficiency, and diversity.

Hybridoma technology is historically one of the first methods for monoclonal antibody development. It starts with injecting mice or other animals with the target antigen to initiate an immune response[49]. Antigen-specific B cells are then harvested from the immunized animal and subsequently fused with myeloma cells to immortalize the cell line and culture it at large scale for the production of monoclonal antibodies. This method has been utilized for decades and generated numerous antibodies that are not only therapeutics, but also research tools for ELISA, western blot, immunofluorescence microscopy, immunohistochemistry, and other research applications. This method is highly reproducible and scalable and if the hybridoma cell line can be maintained in culture, an unlimited production of mAb can be provided. However, because of drawbacks in terms of its long preparation time requirements, high risk of contamination, and limited applicability to primarily mouse antibodies, it has gradually been replaced by other technologies[50].

Phage, yeast, and mammalian display technology has emerged as a versatile and efficient alternative method for antibody selection. Phage display, for example, expresses a library of antibody sequences on the surface of a filamentous phage through fusion with the phage coat protein. The phage library is then selectively propagated by panning for phage particles with high affinity for the target antigen, followed by selection of individual clones to reconstruct the monoclonal antibodies.[51]. The small size and solubility of the phage particle (up to 10^{13} particles/ml) has allowed repertoire sizes of up to 10^{12} to be efficiently displayed and manipulated, which improves the likelihood of finding suitable antibodies[51,52]. However, it's noteworthy that display technology is still relying on the available antibody library.

With the advent of next-generation sequencing technology, single B-cell sequencing has emerged as the predominant method for the comprehensive investigation of B-cell repertoires[53,54]. Seen in Figure 6 panel B, the single B cell technology first isolates the B cells from immunized animals or plasma cells from human sera and sorts the cells on the multi-well plates to achieve one cell/well[55]. Then cDNA from the sorted single cell is amplified and barcoded by PCR, followed by high throughput sequencing. This method allows high-throughput B cell sequencing to become true: the paired heavy and light chain sequences of millions of B cells can be analyzed in one single experiment[56,57]. Thus, this method greatly facilitates the characterization of B-cell populations, analysis of individual clinical profiles, and supports extensive cohort studies[58–61]. Currently, this technology plays an increasingly vital role in the discovery of new antibody therapeutics, particularly fully human antibodies. For instance, Mab114, an antibody therapeutics for Ebola virus disease, was selected after isolation and screening of a panel of memory B cells from a 1995 Kikwit Ebola virus disease survivor[62].

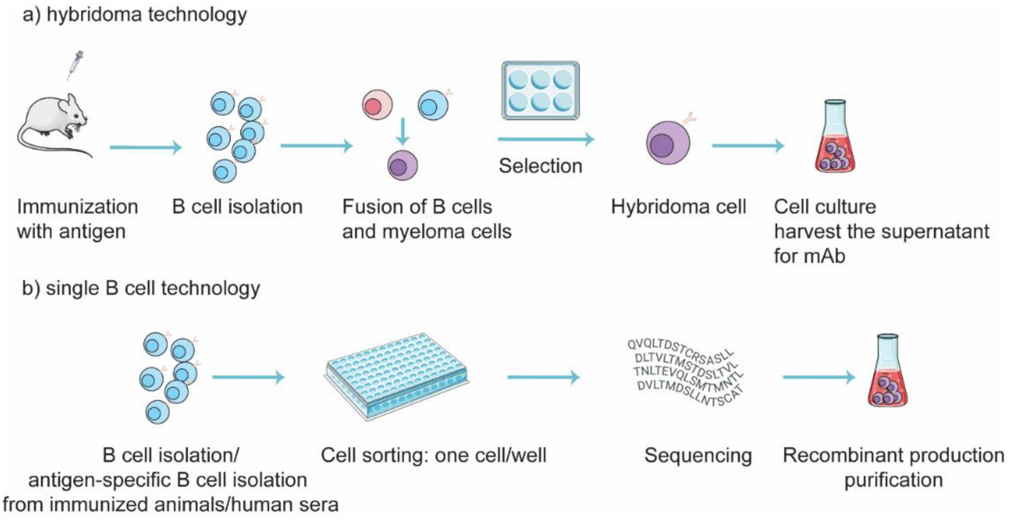

a) hybridoma technology
Selection
Immunization with antigen
B cell isolation
Fusion of B cells and myeloma cells
Hybridoma cell
Cell culture harvest the supernatant for mAb
b) single B cell technology
QVQLTDSTCRSASLL
DLTVLTMSTDSLTVL
TNLTEVQLSMTMNTL
DVLTMDSLLNTSCAT
B cell isolation/ antigen-specific B cell isolation from immunized animals/human sera
Cell sorting: one cell/well
Sequencing
Recombinant production purification

Figure 6: Comparison of Hybridoma technology and single B cell technology for antibody discovery: a illustrates the hybridoma technology that harvests the B cells from immunized animals and merges the B cells and myeloma cells. The hybridoma cells are selected for culture and produce mAbs. b illustrates the single B cell technology: B cells are isolated from immunized animals or human sera and sorted in a multi-well plate to achieve one cell/plate. The sequence cDNA of the cell is obtained for mAb production. The Figure was partly

generated using Servier Medical Art, provided by Servier, licensed under a Creative Commons Attribution 3.0 unported license

However, as previously noted, the long-lived plasma B cells that produce a wide array of high-affinity antibodies for years after infection are predominantly located within the bone marrow and spleen, the sampling of which is highly invasive so they are not normally accessible, especially in human subjects[60,63]. Therefore, single B-cell technology primarily focuses on the analysis of peripheral B cells and can't identify and determine the relative concentrations of the secreted antibodies that compose the serum/milk polyclonal pool elicited in response to infection or vaccination. Meanwhile, functional serological studies that evaluate serum antibody binding and neutralization titers all sample the bulk of secreted antibodies rather than the corresponding B- and plasma cell populations. It remains uncertain to what extent the peripheral B cells reflect the diversity of peripheral antibodies. Expansion of memory B-cells into the plasma cell populations, secretion levels of the antibody, and their lifetimes in circulation all provide multiple layers of biological regulation that may results in orders of magnitude difference between the secreted pool of antibodies and the observed frequency of clones in peripheral B-cell populations. Thus, a full understanding of the antibody repertoire requires analytical methods that detect and sequence immunoglobulins straight at the protein level. For this purpose, bottom-up proteomics emerges as the most potent tool for protein identification, sequencing, characterization and quantification.

The fundamentals of bottom-up proteomics

Proteomics is the study of proteins including detection, identification, quantification, post-translation modification, protein-protein interaction and so on. Proteomics employs two primary strategies: the top-down and bottom-up approaches, distinguished by their target analytes. In the top-down approach, intact proteins are the subject of study, whereas the bottom-up approach involves the enzymatic digestion of proteins, typically using trypsin, to generate peptides[64,65]. These peptides are subsequently separated via high-performance liquid chromatography (HPLC) and analyzed using mass spectrometry (MS), as depicted in Figure 7.The reversed-phase HPLC that uses a non-polar stationary phase and a polar mobile phase is the most commonly used LC for peptide separation, where peptides are separated based on their hydrophobicity[66]. This bottom-up approach takes advantage of the fact that peptides can be better separated on the HPLC and more easily ionized and fragmented to be identified by MS over intact protein[67]. Consequently, the bottom-up approach has gained widespread adoption in protein characterization, particularly for high-throughput analyses. Depending on the purpose of the study, various mass analyzers and peptide fragmentation techniques are available to facilitate the bottom-up approach.

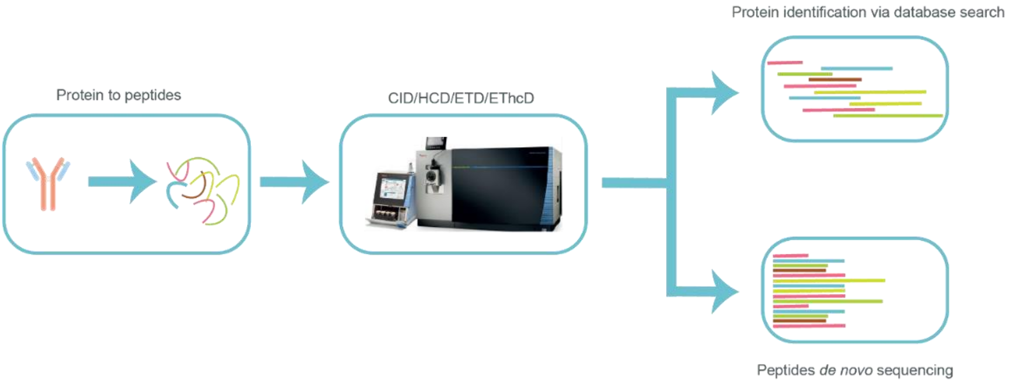

Protein to peptides
CID/HCD/ETD/EThcD
Protein identification via database search
Peptides *de novo* sequencing

Figure 7: Overall workflow bottom-up proteomics: proteins are digested into peptides, followed with separation and analysis by LC-MS/MS. Depending on purpose, the MS spectra can be converted to sequence either with database search or de novo sequencing.

In MS, peptides ions are detected in the mass analyzer based on their mass-charge-ratio (m/z). Since last century, multiple mass analyzer are available on the market, e.g. linear ion trap (LIT), time of flight (TOF), quadrupole, Orbitrap, and Fourier Transform Ion Cyclotron Resonance analyzers. Each analyzer applies a different strategy on isolation and measurement of peptide masses. A quadrupole mass analyzer, for instance, discriminates and filters ions of different m/z based on its trajectory in the electric field under certain direct current potential and radio frequency (RF)[68,69]. The ions that have unstable trajectory are filtered out. LIT has very similar mechanism like quadrupole, but consists of only two hyperbolic electrode plates instead of four hyperbolic rods[70]. Ions of different m/z are trapped by varying the RF potential. While both quadrupole and LIT are very popular mass analyzer for its relative low cost and high scan speed, they have a big disadvantage for the low resolution power[71]. TOF is made up of a flight tube and an acceleration grid[72,73]. Ions are accelerated from the ion source and fly in the tube, then they are distinguished based on their arrival times at the detector. TOF analyzer exhibits high ion transmission efficiency and achieves wide mass range[74]. Orbitrap, in contrast, is built of an inner spindle electrode and two outer hollow concave electrodes facing each other[75,76]. Ions are introduced into the orbitrap when a voltage potential is applied between inner and outer electrodes. The ions will have a harmonic axial oscillation and the motion of the ions are detected. The signal of the time domain is converted to a frequency domain by Fourier transform, thus the m/z of ions are determined. Orbitrap gained significant attention since its integration into commercial instruments (LTQ Orbitrap, Thermo Fisher Scientific) in 2005, owing to its high resolving power. Modern tandem MS utilizes various mass analyzers, enabling analyses of a broader mass range and enhanced accuracy.

In tandem MS, the m/z of peptide ions are analyzed by the first mass analyzer and the selected ions are further fragmented into product ions. From the mass analysis of the resulting peptide fragments, the sequence and post translation modifications of the peptides can be characterized. Nowadays, there are three main fragmentation strategy: collision-based fragmentation (collision induced/activated dissociation (CID/CAD)); electron-based fragmentation (electron transfer /capture dissociation (ETD/ECD)); and photon-based fragmentation (ultraviolet or infrared photodissociation (UVPD/IRPD)). The collision-based dissociation strategy is the most commonly utilized fragmentation scheme in bottom-up proteomics for its easy implementation in MS and high fragmentation efficiency on the tryptic digested peptides. In CID, peptide ions collide with neutral gas atoms and induces the C-N (peptide) bonds cleavage, generating what are called b/y ions as fragments (seen panel c Figure 8). Higher-energy collisional dissociation (HCD) is a beam-type CID technology associated with orbitrap instrument. Currently the stepped HCD has been applied, which fragments the precursor ions with low, medium and high energy[77]. Compared with using a single collisional energy, stepped HCD shows superior protein identification and enhanced peptide sequence coverage.

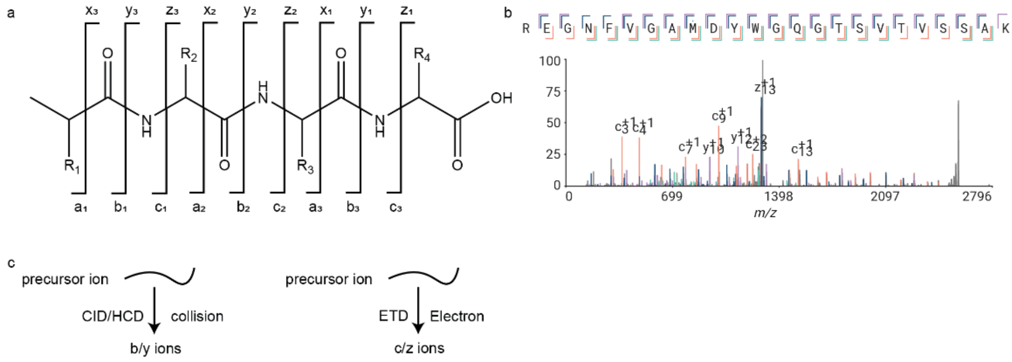


Figure 8: Peptide fragmentation by bottom-up approach. a) The Roepstorff-Fohlman-Johnson nomenclature for peptide fragment ions: the charge of a,b,c ions are retained on the N terminal and x,y,z ions are retained on the C terminal[78,79]*. b) an EThcD fragmentation spectrum, b/y ions are indicated in green and purple, while c/z ions are indicated in orange and dark blue. c) two commonly used fragmentation techniques*

In contrast with collision-based methods, ETD and ECD employ a distinct mechanism to fragment precursor ions, namely electron-mediated dissociation. In ECD, multiple positive charged ions interact with free low-energy electrons, while ETD utilizes gas-phase reagent consisting of radical anions. The reaction leads to the N-Cα bond cleavage in the peptide backbone and generates c/z ions, whilst preserving labile post-translation modification, e.g. glycosylation and phosphorylation[80]. Though ETD was introduced several years later than ECD, it has been more wildly applied mainly due to its implementation on more affordable instruments. Compared with collision based method. ETD/ECD has a much lower efficacy and requires longer activation time, as well as more precursor ions as input. Especially with the lower charge peptide ions, the non-dissociative electron transfer dissociation (ETnoD) often happens: the cleavage of the N-Cα bonds occurs but the c- and z•-type product ions are held together by noncovalent interactions present in the more compact structures, thus no sequence information is provided[81]. To tackle this challenge, ways to increase the peptides charge state have been studied: chemical modification on the peptides or using proteases to generate long peptides.

UVPD relies on lasers to increase the internal energy of selected precursor ions, leading to fragmentation and the production of various ion types (a/b/c/x/y/z), particularly advantageous for intact protein analysis[82]. Unlike electron-based dissociation, UVPD does not require high charge states but presents challenges in precise ions assignment.

In conclusion, CID and HCD are still the most popular technique in bottom-up proteomics workflows for their high efficiency and affordability, particularly for high-throughput screening. Electron-mediated dissociation techniques are employed more for protein post-translation modification and long peptides or intact protein study for its requirement on multiply charged ions.

Following fragmentation, MS spectra are acquired and analyzed by two main strategies: database search and *de novo* sequencing. The database search method is the most widespread: the protein sequences from a given database are digested in silico into peptides and the theoretical spectra are generated for comparison with the experimental spectra. In contrast, *de novo* sequencing tools directly convert MS spectra into sequence information without relying on a predefined database. While accurate *de novo* sequencing requires rich fragmentation spectra with full coverage of b/y or c/z ions at every possible position along the peptide backbone, database searching is more tolerant to missing peaks. On the other hand, *de novo* sequencing can be performed without any prior knowledge on the underlying protein sequences, while database searching is limited to the previously determined protein sequences. As a result of this inherent tradeoff, several bottom-up proteomics strategies have been employed, combining both database searching and *de novo* sequencing to characterize antibodies directly at the protein level.

Proteogenomics approach

Depending on sample material and data analysis, antibody discovery methods based on bottom up approach can be classified into two strategies: Proteogenomics approach and direct antibody sequencing. Seen in figure 8 panel a, the proteogenomics approach, sometimes refered to as “Ig-seq”, combines both single B cell sequencing technology and bottom-up proteomics, while the direct antibody *de novo* sequencing method studies the antibody without any cDNA-based database.

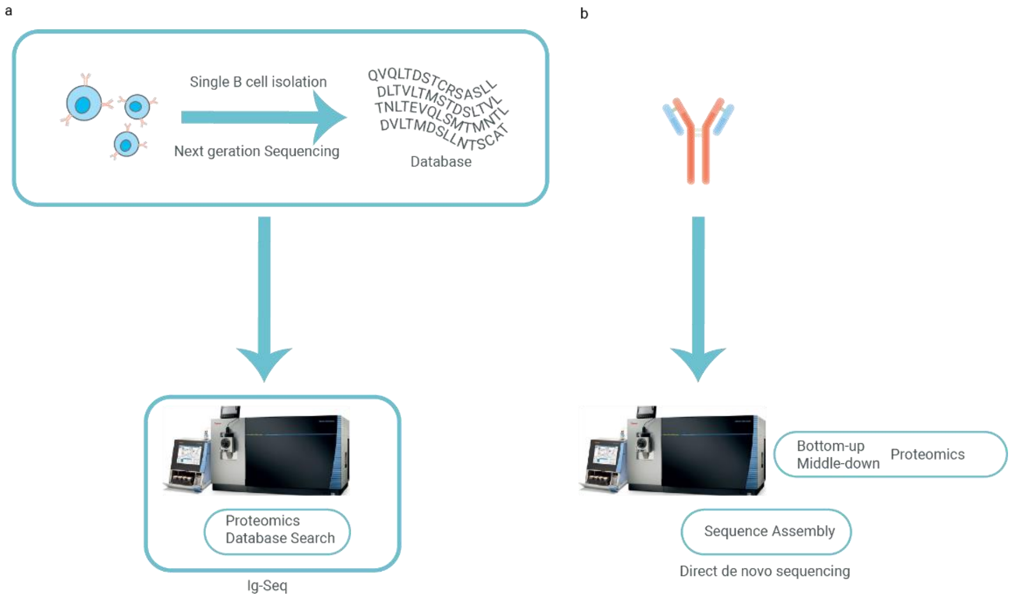


Figure 9: Two methods in antibody de novo sequencing. a) uses a customized database generated by single B cell sequencing; b) Study the antibody sequence directly on the protein level without any knowledge of DNA sequence

Lavinder et al have applied the “Ig-seq” method to study the human serum antibody repertoire after vaccination[83]. In the Ig-seg method, a paired VH-VL gene database is collected by single B cell sequencing and the antigen-specific antibody is purified from sera. The antigen-specific antibodies are enzymatically digested by trypsin to generate peptides for bottom-up proteomics analysis. Owing to the high frequency of R/K in the beginning of the CDR3 regions, trypsin often generates the peptides that cover the whole CDR3 regions. CDR3 is the most heterogeneous region in the antibody sequence: a unique CDR3 often represents a unique clone[84]. Through the Ig-seg method, the gap between the B cell repertoire and serological antibody repertoire is bridged. Nevertheless, the incompleteness of the B cell sequence database remains a significant challenge, necessitating the development of novel analytical methods that study the antibody repertoire directly at the protein-level, independently from databases constructed with cDNA sequences.

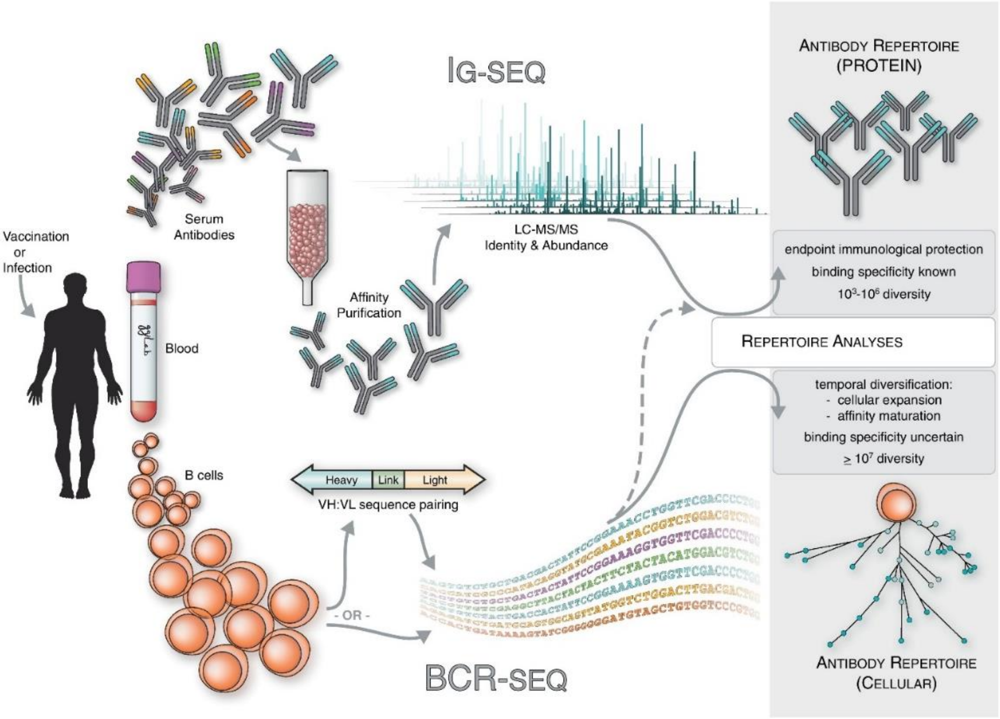


Figure 10: Workflow of Ig-seq method: Generation of customized database by single B cell sequencing, followed with tryptic digestion of serum IgG and quantitatively search for CDR3 region peptides. Figure adapted from Lavinder et al[83]. Copyright © 2014 Elsevier Ltd.

A direct *de novo* antibody sequencing approach

In contrast to the previously discussed methods, the direct *de novo* antibody sequencing method based on MS stands out for not relying on database search, but rather on estimating the peptide sequence from every individual fragmentation spectrum and assembling the full heavy and light chain sequences from these shorter fragments. Like the conventional bottom-up proteomics experiment, it starts with the proteolytic digestion of antibody to peptides, but the key challenge is to obtain full coverage, as missing regions can no longer be inferred from matched database entries. In general, trypsin is the gold standard in the proteomics work due to its extraordinary performance in generating ideal $2^+/3^+$ peptides with suitable length (10-15 amino acids) for CID or HCD fragmentation[66,85]. Trypsin cleaves the protein at the C- terminus of the basic amino acids arginine and lysine. The C-terminus basic residue are ideal for positive ionization[86].

Moreover, the CID/HCD fragmentation of tryptic peptides benefits from the balance of the N-terminus amine and the C-terminus basic residues[87]. Most peptide identification algorithms are developed based on tryptic digestion and the nature of the tryptic peptides are extensively studied, thus the identification of tryptic peptides notably superior to those from other proteases.[88–90]. However, even trypsin alone still cannot provide a full protein sequence coverage: The uneven distribution of the target sites (R and K) leads to yield either too long or too short peptides that cannot be identified or assembled into the full protein sequence. Using multiple proteases to yield overlapping peptides can significantly increase the antibody sequence coverage[91–94]. Several alternative proteases have been studied in the past few years to optimize the sample preparation for proteomics work, including chymotrypsin, thermolysin, alpha-lytic protease and so on. Each of these proteases have their unique digestion sites and exhibit different benefits.

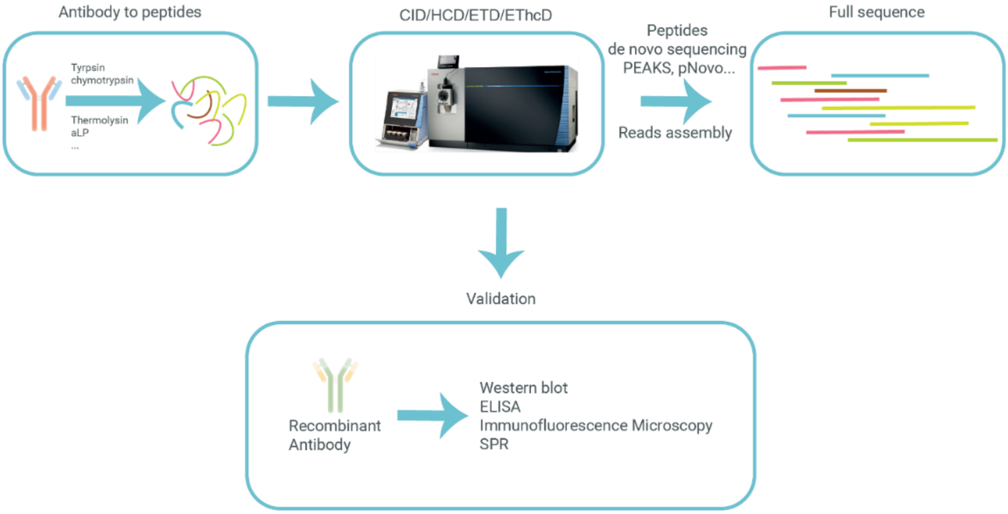


Figure 11: Strategy for the direct antibody de novo sequencing based on bottom-up proteomics: Antibody is digested by multiple proteases, the digested peptides are separated and analyzed by LC-MS/MS. The peptides can be de novo sequenced and assembled into full sequence. The sequence accuracy can be validated by production of the recombinant antibody and characterization its binding with antigen via western blot, ELISA, immunofluorescence microscopy.

For example, chymotrypsin is one protease that is often used beyond trypsin, which shows preference for cleaving at hydrophobic sites (leucine and methionine) as well as aromatic sites (tryptophan, tyrosine and phenylalanine), which fills the sequence gap that tryptic peptides leave. As an added benefit, the specificity of chymotrypsin to cleave at leucine, but not isoleucine, provides a great opportunity to distinguish between these isobaric residues, offering a partial solution to this classical problem in mass spectrometry based peptide sequencing[95,96].

Some unspecific proteases like elastase, thermolysin, and pepsin also show advantages by generating numerous overlapping peptides[97]. The non-trypsin proteases often generate peptides with higher charge state because of the internal basic amino acids, which are not optimal for HCD/CID but better for ETD fragmentation[91,98]. For example, application of ETD fragmentation on the LysN digested peptides produced spectra dominated by an extensive series of c-ions, which showed great potential for *de novo* sequencing[99]. Though ETD is extremely inefficient for the fragmentation of doubly charged peptides[81,98,100], the problem can be solved by application of data-dependent decision-tree-based fragmentation schemes, or supplemental activation of the ions with collision-based methods, commonly referred to as activated ion electron transfer dissociation (AI-ETD), collision activated ETD (ETcaD and EThcD)[92,101–103]. By generation of both b/y and c/z series of ions, EThcD achieves very extensive peptide sequence coverage across a wide range of peptide charge states generated by a wide range of proteases[104]. Several studies demonstrate that EThcD also shows great potential in distinguish leucine and isoleucine by the production of w ions, which are secondary fragments with signature losses that differ between the two isobaric residues[105,106].

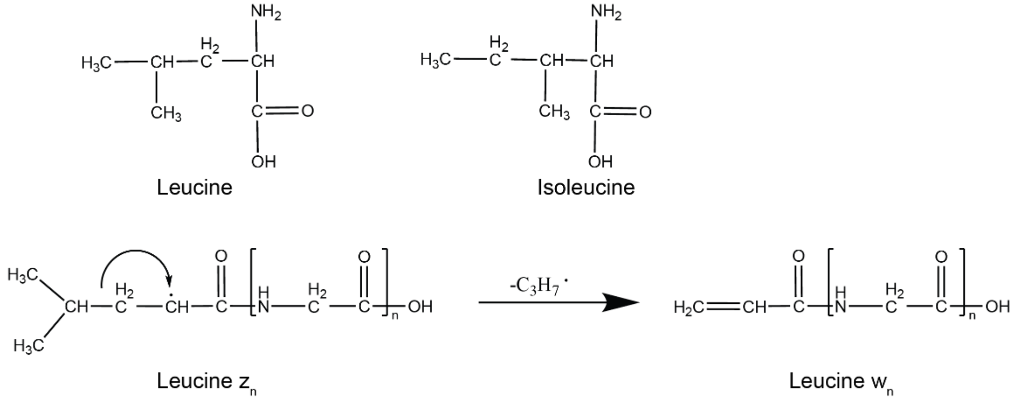


Figure 12: Mechanism of production of w ions from Leucine residue, which demonstrate the chance of distinguish leucine and isoleucine with MS[107]

In classical bottom-up proteomics work, MS spectra are conventionally matched against a specific protein database containing relevant proteins. Various search engines, such as Maxquant, utilize artificially generated spectra for computational comparison with experimental spectra, yielding a score for the identified peptide. The drawback of the database search is it is relying on the database availability. Human antibodies, for example, are highly variable and a comprehensive database is not only absent, but also wildly impractical given the enormous sequence space that antibodies can sample from.

Table 2: List of some published de novo sequencing algorithms

Software	*License*	*Support MS Fragmentation Method*	*Speed*	*Open-source*	*Customize Modification /Digestion*
Supernovo[108]	Commercial	HCD/ETD/EThcD	Days	No	Yes
PEAKS[109]	Commercial	HCD/ETD/EThcD	Hours	No	Yes
pNovo[110–112]	free	HCD/ETD	Minutes	No	Yes
MaxNovo[113]	free	HCD	Hours	No	Yes
Novor[114]	free	HCD/ETD/EThcD	Minutes	No	Yes
Casanovo[115,116]	free		Minutes with GPU Weeks without GPU	Yes	Needs to retrain the model

De novo sequencing, in contrast, provides the sequence information only from the MS spectra, is a complementary approach to overcome this limitation, and especially suitable for this direct antibody *de novo* sequencing. Successful implementation hinges on both the high-resolution mass spectrometer providing quality spectra and software capable of directly translating these spectra into sequences. Table 2 outlines the main concerns and capabilities of currently available and commonly used *de novo* sequencing tools. Distinguishing itself from other algorithms listed in the table, Supernovo is an automated tool tailored for monoclonal antibody *de novo* sequencing. Supernovo leverages the antibody germline sequences from the IMGT database and utilizes the Byonic search engine for V- and C-region candidate identification separately, based on a database matching step[18,117]. After the most probable J-region is found by wildcard search, Supernovo assemble the V-,J- and C- region into the full template for the further refinement via both *de novo* sequencing and repeated wildcard search[118,119].

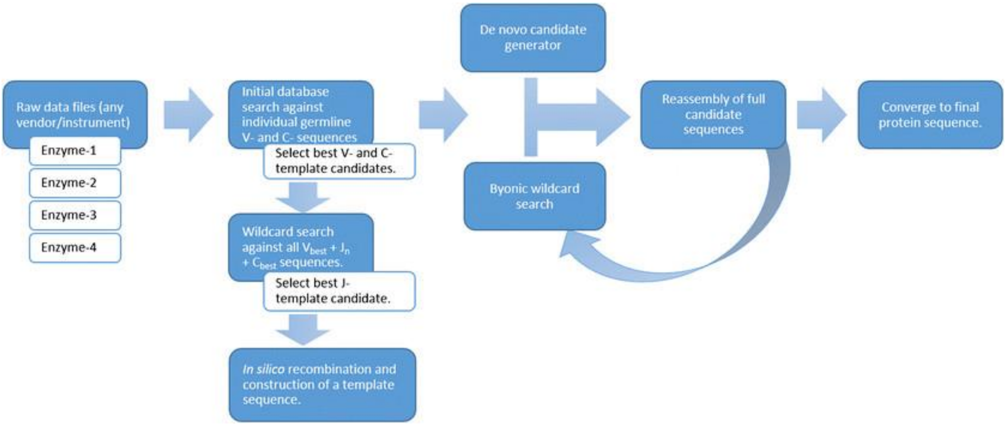

Raw data files (any vendor/instrument)
Enzyme-1
Enzyme-2
Enzyme-3
Enzyme-4
Initial database search against individual germline V- and C- sequences
Select best V- and C- template candidates.
Wildcard search against all V_{best} + J_n + C_{best} sequences.
Select best J- template candidate.
In silico recombination and construction of a template sequence.
De novo candidate generator
Byonic wildcard search
Reassembly of full candidate sequences
Converge to final protein sequence.

Figure 13: The monoclonal antibody de novo sequencing workflow of Supernovo[108]*. Copyright © 2017 © American Society for Mass Spectrometry*

Besides Supernovo, all other peptide sequencing software listed in table 2 focus on *de novo* peptide sequencing, which interpret the MS spectra to peptide sequence. Each tool possesses unique strengths and weaknesses. As discussed in the sample preparation for *de novo* sequencing, atypical chemical reagents and multiple proteases are often employed to enhance sequencing accuracy. Thus, it is crucial for the search engines to accommodate customized modifications and digestion parameters. While most tools offer this flexibility, it's noteworthy that Casanovo and all other deep-learning based tools, lean heavily on training the model with available datasets of peptide identifications. As a consequence, target peptides that are generated by uncommon proteases, carry uncommon modifications, or are fragmented with uncommon methods, all require costly retraining of the model with appropriate data. Support for uncommon fragmentation techniques like EThcD, is currently limited in most types of *de novo* sequencing algorithms, while it stands out in its performance for peptide fragmentation and potential for *de novo* peptide sequencing. Among the listed tools, PEAKS and Novor stand out as the two algorithms capable of handling EThcD spectra, contributing to their enhanced peptide sequence coverage[120,121]. Although pNovo3 supports both HCD and ETD fragmentation schemes, it lacks support for EThcD. Notably, PEAKS is the sole commercial tool among these *de novo* peptide sequencing tools, marked by its consistent updates and robust maintenance.

Both MaxNovo and pNovo adopt the strategy of converting MS/MS spectra into graphs prior to computation [122]. In this method, one peak represents one pair of vertices: b/y ions in HCD spectrum and c/z ions in ETD spectrum. When the mass difference between two vertices generated from two peaks matches the mass of the amino acid, they are connected into an edge. Thus, the path connecting the edges stands for the peptide sequence. However, the missing ions between the path from N to C terminus break up the connection and leads to the mistake in the sequences. Moreover, situations where only b or y ions are generated in HCD/CID spectra further compound the limitations of this graph-based strategy[123].

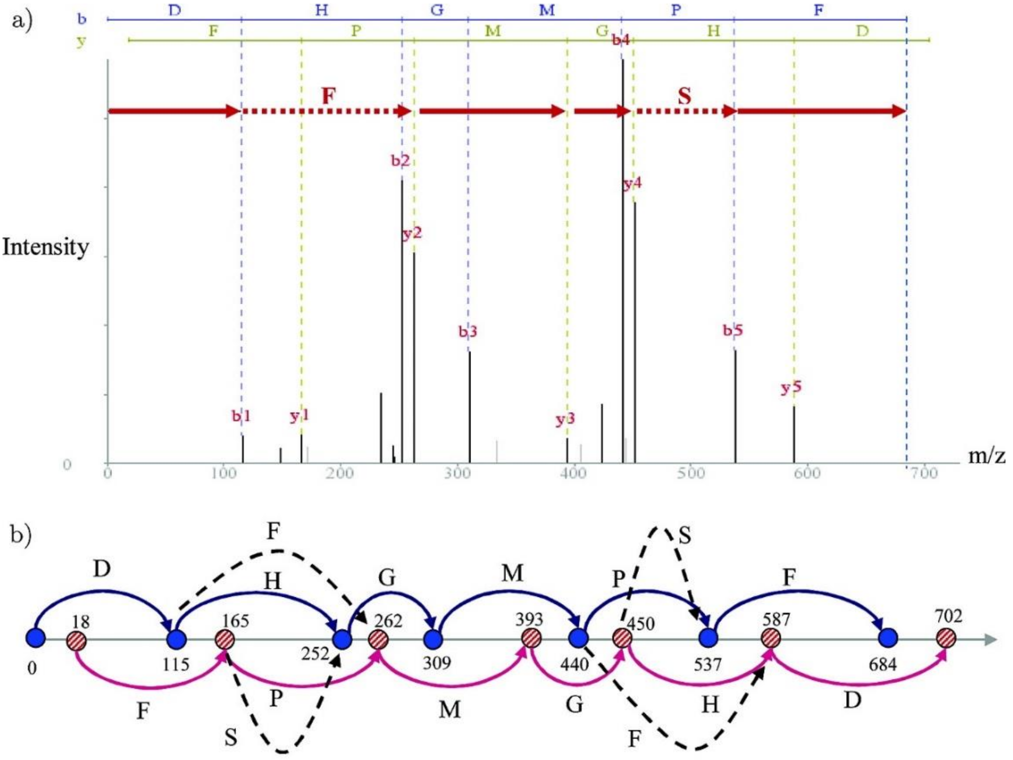


Figure 14: An example exhibits how spectrum-graph strategy works. a) illustrates the MS spectrum of peptide DHGMPF. b) shows the graph converted from the MS spectrum of peptide DHGMPF. Figure adopted from Frank et al[124]. Copyright © 2007 American Chemical Society

In contrast, PEAKS and Novor adopt alternative strategies that bypass spectrum-to-graph conversion. PEAKS, for instance, preprocess the raw file data by noise filtering, peak centering and deconvolution higher charged PEAKS into singly charged ions.

Subsequently, it employs a dynamic programming algorithm to predict up to 10,000 potential peptide sequences matching a precursor ion mass. These sequences are evaluated based on proximity of predicted amino acid masses to observed peaks in the spectrum, thereby assigning scores to both sequences and individual amino acids. In the last few years, notably, PEAKS has integrated the deepNovo algorithm , enhancing analysis accuracy and speed[125]. It has also extended its capabilities to accommodate data-independent acquisition mode[126]. Notably, Novor has made the real-time *de novo* sequencing possible: It reaches a speed of over 300 spectra per second on a laptop computer[114,127,128].

De novo antibody sequencing based on middle-down approach

While traditional bottom-up strategies have yielded valuable information, they face challenges in handling complex polyclonal samples. The polyclonal antibodies derived from human sera contain more than thousands of clones, even after antigen specific enrichment. Via bottom-up approach, antibodies are digested into peptides. However, the great heterogeneity in the CDR regions and high level of conservation in the framework poses a formidable barrier to confidently reconstruct complete antibody sequences. Two identical heavy chains and two identical light chains are attached to each other via disulfide bonds to form an intact antibody. But they are fallen apart for the digestion, so that heavy chain and light chain pairing cannot be obtained in the bottom up approach.

To tackle this challenge, the middle-down approach has emerged as a promising complementary method to study the antibody in a protein-centric manner. Compared with the top-down approach that analyzes the intact antibody, the middle-down approach first uses a protease to remove the Fc region and generate either the Fab (50 kDa) or Fab2 (100 kDa) for reducing the complexity and molecular weight of the intact antibodies[129-132]. Recently, Bondt et al published a workflow using the protease IdgE to generate Fab from the whole serum IgG1 repertoire and quantitatively profiled and compared the IgG1 Fab repertoire of both sepsis patients and healthy donors[133]. Mass spectrometry provides the accurate mass of each clone, thereby facilitating more accurate antibody sequence assembly. Furthermore, the middle-down approach provides insights into the pairing of heavy and light chains.

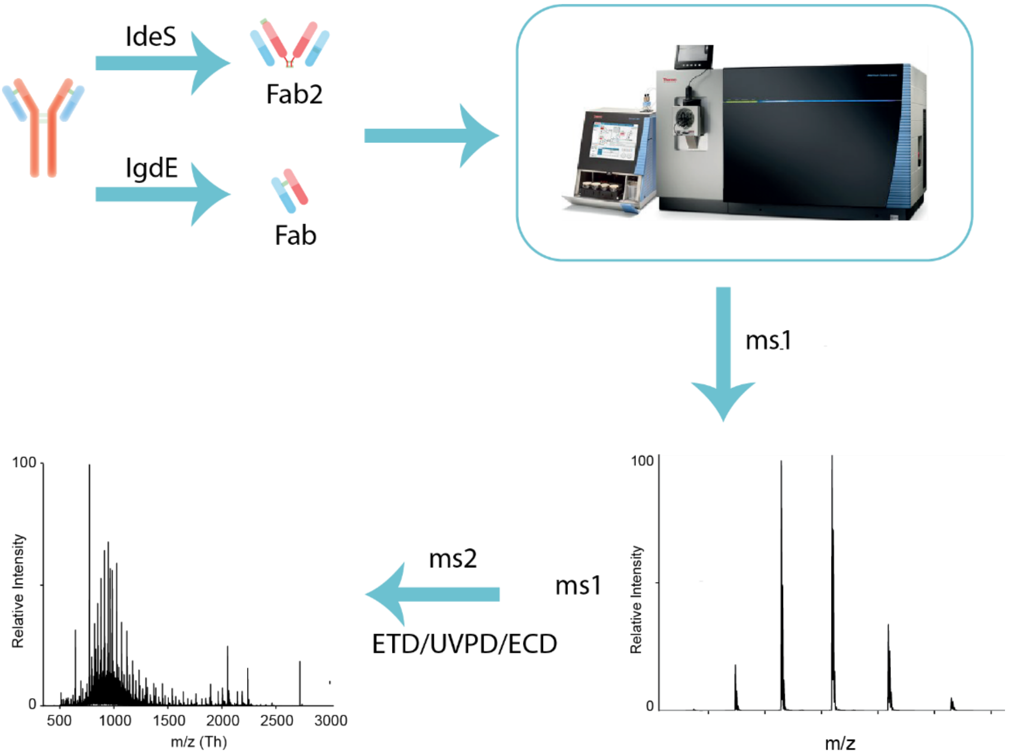


Figure 15: the scheme of middle-down approach of antibody de novo sequencing: Generation of Fab or Fab2 by IgdE or IdeS enzyme, the exact molecular weight of Fab/Fab2 can be analyzed by LC/MS and it can be further fragmented by ETD/UVPD/ECD to obtain the sequence information.

Remarkable progress has been made in mass analyzers and ion activation/dissociation techniques, such as ETD, ECD and UVPD. These advancements have significantly enhanced the sequence coverage achievable for Fab and Fab2 fragments, or at least the CDR3 regions[134]. *Den Boer et al* demonstrated that the application of the ECD on the whole Fab2 revealing CDR3 and FR4 coverage of both heavy chain and light chain while preserving disulfide bonds (Figure 16)[135]. Some attempts have also been made on more complicated samples: Mathieu el au applied ETD/HCD/EThcD/UVPD on the light chain from patients with multiple myeloma and achieved nearly full sequence coverage [136].

Despite the promising strides in middle-down approaches, several challenges persist, e.g. the separation based on LC of protein is much less established than peptides. Additionally, the ionization of Fab/Fab2 tends to be less efficient than peptides, posing difficulties in achieving high sequence coverage. The complexity of the resulting spectra further complicates data analysis. The Kelleher group and Gibbons group have developed some software, e.g. Prosight and LC-MS Spectator for processing the complicated data[90,137,138]. However, all the current algorithms require database search and cannot yet analyze the middle-down data in a *de novo* manner.

To date, the predominant approach for novel antibody therapeutics discovery is still relying on isolation of the peripheral B cells from human survivors or seropositive individual that has immunoresponse to certain antigen and obtain the sequence via the single B cell sequencing technique. However, a considerable reservoir of potential antibodies remains undiscovered owing to the incomplete B cell databases. Nonetheless, the integration of bottom-up and middle-down approaches holds the potential to bridge this gap. The bottom-up strategy provides an initial glimpse of the full-length sequence, which can then be validated and refined using the middle-down approach. This hybrid methodology provides a more comprehensive and accurate representation of antibody repertoires for therapeutic exploration. The remaining chapters in this thesis will showcase the current state-of-the-art in antibody sequencing by bottom-up proteomics technology, illustrating on several accounts that monoclonal antibodies, produced recombinantly, from hybridoma cell lines, or straight from human serum, can be sequenced with sufficient accuracy to reverse engineer functional antibody products.

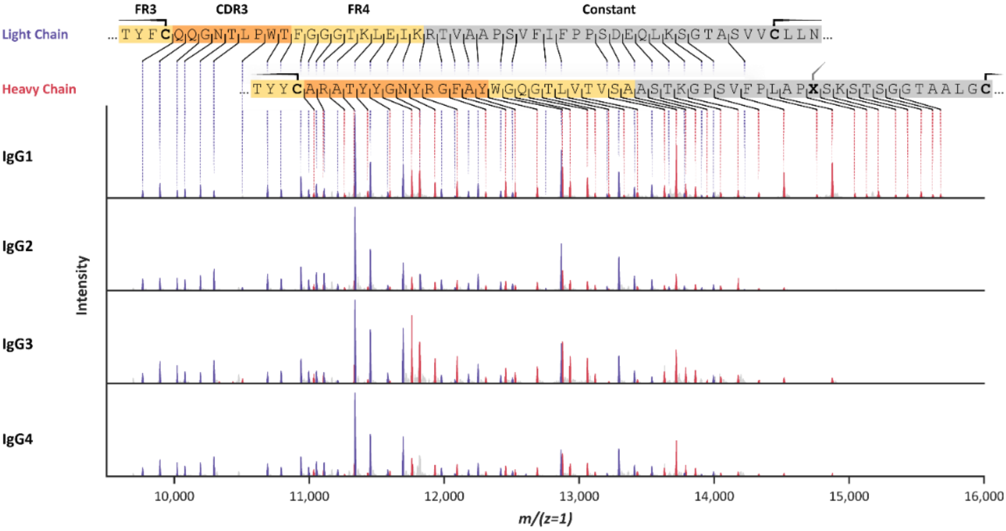


Figure 16: ECD spectra of 4 Fab2: illustration of the coverage of CDR3 regions of both HC and LC. Figure adopted from Den Boer et al[135]*. Copyright 2021 Royal Society of Chemistry.*

Reference

(1) *B-Cell Development and the Antibody Response - ClinicalKey*. https://www.clinicalkey.com/#!/content/book/3-s2.0-B978070207844600009X (accessed 2023-02-21).

(2) Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. B Cells and Antibodies. In *Molecular Biology of the Cell. 4th edition*; Garland Science, 2002.

(3) Mullard, A. FDA Approves 100th Monoclonal Antibody Product. *Nature Reviews Drug Discovery* **2021**, *20* (7), 491–495. https://doi.org/10.1038/d41573-021-00079-7.

(4) Sathe, A.; Cusick, J. K. Biochemistry, Immunoglobulin M. In *StatPearls*; StatPearls Publishing: Treasure Island (FL), 2023.

(5) Stavnezer, J. Immunoglobulin Class Switching. *Current Opinion in Immunology* **1996**, *8* (2), 199–205. https://doi.org/10.1016/S0952-7915(96)80058-6.

(6) Li, Z.; Woo, C. J.; Iglesias-Ussel, M. D.; Ronai, D.; Scharff, M. D. The Generation of Antibody Diversity through Somatic Hypermutation and Class Switch Recombination. *Genes Dev.* **2004**, *18* (1), 1–11. https://doi.org/10.1101/gad.1161904.

(7) Buss, N. A.; Henderson, S. J.; McFarlane, M.; Shenton, J. M.; de Haan, L. Monoclonal Antibody Therapeutics: History and Future. *Current Opinion in Pharmacology* **2012**, *12* (5), 615–622. https://doi.org/10.1016/j.coph.2012.08.001.

(8) Muhammed, Y. The Best IgG Subclass for the Development of Therapeutic Monoclonal Antibody Drugs and Their Commercial Production: A Review. *Immunome Research* **2020**, *16* (1), 1–12. https://doi.org/10.35248/1745-7580.20.16.173.

(9) Yu, J.; Song, Y.; Tian, W. How to Select IgG Subclasses in Developing Anti-Tumor Therapeutic Antibodies. *J Hematol Oncol* **2020**, *13* (1), 45. https://doi.org/10.1186/s13045-020-00876-4.

(10) Nissim, A.; Chernajovsky, Y. Historical Development of Monoclonal Antibody Therapeutics. In *Therapeutic Antibodies*; Chernajovsky, Y., Nissim, A., Eds.; Handbook of Experimental Pharmacology; Springer: Berlin, Heidelberg, 2008; pp 3–18. https://doi.org/10.1007/978-3-540-73259-4_1.

(11) Porter, R. R. The Hydrolysis of Rabbit γ-Globulin and Antibodies with Crystalline Papain. *Biochem J* **1959**, *73* (1), 119–127.

(12) Davies, D. R.; Padlan, E. A.; Segal, D. M. Three-Dimensional Structure of Immunoglobulins. *Annual Review of Biochemistry* **1975**, *44* (1), 639–667. https://doi.org/10.1146/annurev.bi.44.070175.003231.

(13) Harris, L. J.; Skaletsky, E.; McPherson, A. Crystallographic Structure of an Intact IgG1 Monoclonal antibody11Edited by I. A. Wilson. *Journal of Molecular Biology* **1998**, *275* (5), 861–872. https://doi.org/10.1006/jmbi.1997.1508.

(14) Cobb, B. A. The History of IgG Glycosylation and Where We Are Now. *Glycobiology* **2020**, *30* (4), 202–213. https://doi.org/10.1093/glycob/cwz065.

(15) Kehry, M.; Ewald, S.; Douglas, R.; Sibley, C.; Raschke, W.; Fambrough, D.; Hood, L. The Immunoglobulin μ Chains of Membrane-Bound and Secreted IgM Molecules Differ in Their C-Terminal Segments. *Cell* **1980**, *21* (2), 393–406. https://doi.org/10.1016/0092-8674(80)90476-6.

(16) Luo, Y.; Lu, Z.; Raso, S. W.; Entrican, C.; Tangarone, B. Dimers and Multimers of Monoclonal IgG1 Exhibit Higher in Vitro Binding Affinities to Fcγ Receptors. *MAbs* **2009**, *1* (5), 491–504.

(17) Ehrenmann, F.; Kaas, Q.; Lefranc, M.-P. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: A Database and a Tool for Immunoglobulins or Antibodies, T Cell Receptors, MHC, IgSF and MhcSF. *Nucleic Acids Research* **2010**, *38* (suppl_1), D301–D307. https://doi.org/10.1093/nar/gkp946.

(18) Lefranc, M.-P. IMGT® Databases, Web Resources and Tools for Immunoglobulin and T Cell Receptor Sequence Analysis, Http://Imgt.Cines.Fr. *Leukemia* **2003**, *17* (1), 260–266. https://doi.org/10.1038/sj.leu.2402637.

(19) Bannard, O.; Cyster, J. G. Germinal Centers: Programmed for Affinity Maturation and Antibody Diversification. *Current Opinion in Immunology* **2017**, *45*, 21–30. https://doi.org/10.1016/j.coi.2016.12.004.

(20) Peled, J. U.; Kuang, F. L.; Iglesias-Ussel, M. D.; Roa, S.; Kalis, S. L.; Goodman, M. F.; Scharff, M. D. The Biochemistry of Somatic Hypermutation. *Annu. Rev. Immunol.* **2008**, *26* (1), 481–511. https://doi.org/10.1146/annurev.immunol.26.021607.090236.

(21) Methot, S. P.; Di Noia, J. M. Chapter Two - Molecular Mechanisms of Somatic Hypermutation and Class Switch Recombination. In *Advances in Immunology*; Alt, F. W., Ed.; Academic Press, 2017; Vol. 133, pp 37–87. https://doi.org/10.1016/bs.ai.2016.11.002.

(22) *Evolutionary and Somatic Selection of the Antibody Repertoire in the Mouse*. https://doi.org/10.1126/science.3317826.

(23) Cattoretti, G.; Büttner, M.; Shaknovich, R.; Kremmer, E.; Alobeid, B.; Niedobitek, G. Nuclear and Cytoplasmic AID in Extrafollicular and Germinal Center B Cells. *Blood* **2006**, *107* (10), 3967–3975. https://doi.org/10.1182/blood-2005-10-4170.

(24) Barroso, M.; Tucker, H.; Drake, L.; Nichol, K.; Drake, J. R. Antigen-B Cell Receptor Complexes Associate with Intracellular Major Histocompatibility Complex (MHC) Class II Molecules*. *Journal of Biological Chemistry* **2015**, *290* (45), 27101–27112. https://doi.org/10.1074/jbc.M115.649582.

(25) Allen, C. D. C.; Okada, T.; Cyster, J. G. Germinal-Center Organization and Cellular Dynamics. *Immunity* **2007**, *27* (2), 190–202. https://doi.org/10.1016/j.immuni.2007.07.009.

(26) Batista, F. D.; Neuberger, M. S. B Cells Extract and Present Immobilized Antigen: Implications for Affinity Discrimination. *EMBO Journal* **2000**, *19* (4), 513–520. https://doi.org/10.1093/emboj/19.4.513.

(27) Victora, G. D.; Schwickert, T. A.; Fooksman, D. R.; Kamphorst, A. O.; Meyer-Hermann, M.; Dustin, M. L.; Nussenzweig, M. C. Germinal Center Dynamics Revealed by Multiphoton Microscopy with a Photoactivatable Fluorescent Reporter. *Cell* **2010**, *143* (4), 592–605. https://doi.org/10.1016/j.cell.2010.10.032.

(28) Tas, J. M. J.; Mesin, L.; Pasqual, G.; Targ, S.; Jacobsen, J. T.; Mano, Y. M.; Chen, C. S.; Weill, J.-C.; Reynaud, C.-A.; Browne, E. P.; Meyer-Hermann, M.; Victora, G. D. Visualizing Antibody Affinity Maturation in Germinal Centers. *Science* **2016**, *351* (6277), 1048–1054. https://doi.org/10.1126/science.aad3439.

(29) Klein, F.; Diskin, R.; Scheid, J. F.; Gaebler, C.; Mouquet, H.; Georgiev, I. S.; Pancera, M.; Zhou, T.; Incesu, R.-B.; Fu, B. Z.; Gnanapragasam, P. N. P.; Oliveira, T. Y.; Seaman, M. S.; Kwong, P. D.; Bjorkman, P. J.; Nussenzweig, M. C. Somatic Mutations of the Immunoglobulin Framework Are Generally Required for Broad and Potent HIV-1 Neutralization. *Cell* **2013**, *153* (1), 126–138. https://doi.org/10.1016/j.cell.2013.03.018.

(30) Kepler, T. B.; Liao, H.-X.; Alam, S. M.; Bhaskarabhatla, R.; Zhang, R.; Yandava, C.; Stewart, S.; Anasti, K.; Kelsoe, G.; Parks, R.; Lloyd, K. E.; Stolarchuk, C.; Pritchett, J.; Solomon, E.; Friberg, E.; Morris, L.; Karim, S. S. A.; Cohen, M. S.; Walter, E.; Moody, M. A.;

Wu, X.; Altae-Tran, H. R.; Georgiev, I. S.; Kwong, P. D.; Boyd, S. D.; Fire, A. Z.; Mascola, J. R.; Haynes, B. F. Immunoglobulin Gene Insertions and Deletions in the Affinity Maturation of HIV-1 Broadly Reactive Neutralizing Antibodies. *Cell Host & Microbe* **2014**, *16* (3), 304–313. https://doi.org/10.1016/j.chom.2014.08.006.

(31) Zhou, J. O.; Zaidi, H. A.; Ton, T.; Fera, D. The Effects of Framework Mutations at the Variable Domain Interface on Antibody Affinity Maturation in an HIV-1 Broadly Neutralizing Antibody Lineage. *Frontiers in Immunology* **2020**, *11*.

(32) Davenport, T. M.; Gorman, J.; Joyce, M. G.; Zhou, T.; Soto, C.; Guttman, M.; Moquin, S.; Yang, Y.; Zhang, B.; Doria-Rose, N. A.; Hu, S.-L.; Mascola, J. R.; Kwong, P. D.; Lee, K. K. Somatic Hypermutation-Induced Changes in the Structure and Dynamics of HIV-1 Broadly Neutralizing Antibodies. *Structure* **2016**, *24* (8), 1346–1357. https://doi.org/10.1016/j.str.2016.06.012.

(33) Wagner, S. D.; Milstein, C.; Neuberger, M. S. Codon Bias Targets Mutation. *Nature* **1995**, *376* (6543), 732–732. https://doi.org/10.1038/376732a0.

(34) Wei, L.; Chahwan, R.; Wang, S.; Wang, X.; Pham, P. T.; Goodman, M. F.; Bergman, A.; Scharff, M. D.; MacCarthy, T. Overlapping Hotspots in CDRs Are Critical Sites for V Region Diversification. *Proceedings of the National Academy of Sciences* **2015**, *112* (7), E728–E737. https://doi.org/10.1073/pnas.1500788112.

(35) Yeap, L.-S.; Meng, F.-L. Chapter Three - Cis- and Trans-Factors Affecting AID Targeting and Mutagenic Outcomes in Antibody Diversification. In *Advances in Immunology*; Alt, F., Ed.; Academic Press, 2019; Vol. 141, pp 51–103. https://doi.org/10.1016/bs.ai.2019.01.002.

(36) Dinesh, R. K.; Barnhill, B.; Ilanges, A.; Wu, L.; Michelson, D. A.; Senigl, F.; Alinikula, J.; Shabanowitz, J.; Hunt, D. F.; Schatz, D. G. Transcription Factor Binding at Ig Enhancers Is Linked to Somatic Hypermutation Targeting. *European Journal of Immunology* **2020**, *50* (3), 380–395. https://doi.org/10.1002/eji.201948357.

(37) Perlot, T.; Alt, F. W.; Bassing, C. H.; Suh, H.; Pinaud, E. Elucidation of IgH Intronic Enhancer Functions via Germ-Line Deletion. *Proceedings of the National Academy of Sciences* **2005**, *102* (40), 14362–14367. https://doi.org/10.1073/pnas.0507090102.

(38) Inlay, M. A.; Gao, H. H.; Odegard, V. H.; Lin, T.; Schatz, D. G.; Xu, Y. Roles of the Ig κ Light Chain Intronic and 3′ Enhancers in Igk Somatic Hypermutation1. *The Journal of Immunology* **2006**, *177* (2), 1146–1151. https://doi.org/10.4049/jimmunol.177.2.1146.

(39) Stavnezer, J.; Guikema, J. E. J.; Schrader, C. E. Mechanism and Regulation of Class Switch Recombination. *Annu Rev Immunol* **2008**, *26*, 261–292. https://doi.org/10.1146/annurev.immunol.26.021607.090248.

(40) Onodera, T.; Kita, S.; Adachi, Y.; Moriyama, S.; Sato, A.; Nomura, T.; Sakakibara, S.; Inoue, T.; Tadokoro, T.; Anraku, Y.; Yumoto, K.; Tian, C.; Fukuhara, H.; Sasaki, M.; Orba, Y.; Shiwa, N.; Iwata, N.; Nagata, N.; Suzuki, T.; Sasaki, J.; Sekizuka, T.; Tonouchi, K.; Sun, L.;

Fukushi, S.; Satofuka, H.; Kazuki, Y.; Oshimura, M.; Kurosaki, T.; Kuroda, M.; Matsuura, Y.; Suzuki, T.; Sawa, H.; Hashiguchi, T.; Maenaka, K.; Takahashi, Y. A SARS-CoV-2 Antibody Broadly Neutralizes SARS-Related Coronaviruses and Variants by Coordinated Recognition of a Virus-Vulnerable Site. *Immunity* **2021**, *54* (10), 2385-2398.e10. https://doi.org/10.1016/j.immuni.2021.08.025.

(41) Jia, M.; Liberatore, R. A.; Guo, Y.; Chan, K.-W.; Pan, R.; Lu, H.; Waltari, E.; Mittler, E.; Chandran, K.; Finzi, A.; Kaufmann, D. E.; Seaman, M. S.; Ho, D. D.; Shapiro, L.; Sheng, Z.; Kong, X.-P.; Bieniasz, P. D.; Wu, X. VSV-Displayed HIV-1 Envelope Identifies Broadly Neutralizing Antibodies Class-Switched to IgG and IgA. *Cell Host & Microbe* **2020**, *27* (6), 963-975.e5. https://doi.org/10.1016/j.chom.2020.03.024.

(42) Tudor, D.; Yu, H.; Maupetit, J.; Drillet, A.-S.; Bouceba, T.; Schwartz-Cornil, I.; Lopalco, L.; Tuffery, P.; Bomsel, M. Isotype Modulates Epitope Specificity, Affinity, and Antiviral Activities of Anti–HIV-1 Human Broadly Neutralizing 2F5 Antibody. *Proceedings of the National Academy of Sciences* **2012**, *109* (31), 12680–12685. https://doi.org/10.1073/pnas.1200024109.

(43) Richardson, S. I.; Ayres, F.; Manamela, N. P.; Oosthuysen, B.; Makhado, Z.; Lambson, B. E.; Morris, L.; Moore, P. L. HIV Broadly Neutralizing Antibodies Expressed as IgG3 Preserve Neutralization Potency and Show Improved Fc Effector Function. *Frontiers in Immunology* **2021**, *12*.

(44) Richardson, S. I.; Lambson, B. E.; Crowley, A. R.; Bashirova, A.; Scheepers, C.; Garrett, N.; Karim, S. A.; Mkhize, N. N.; Carrington, M.; Ackerman, M. E.; Moore, P. L.; Morris, L. IgG3 Enhances Neutralization Potency and Fc Effector Function of an HIV V2-Specific Broadly Neutralizing Antibody. *PLOS Pathogens* **2019**, *15* (12), e1008064. https://doi.org/10.1371/journal.ppat.1008064.

(45) Cavacini, L. A.; Emes, C. L.; Power, J.; Desharnais, F. D.; Duval, M.; Montefiori, D.; Posner, M. R. Influence of Heavy Chain Constant Regions on Antigen Binding and HIV-1 Neutralization by a Human Monoclonal Antibody. *The Journal of Immunology* **1995**, *155* (7), 3638–3644. https://doi.org/10.4049/jimmunol.155.7.3638.

(46) *Epitope convergence of broadly HIV-1 neutralizing IgA and IgG antibody lineages in a viremic controller | Journal of Experimental Medicine | Rockefeller University Press*. https://rupress.org/jem/article/219/3/e20212045/213042/Epitope-convergence-of-broadly-HIV-1-neutralizing (accessed 2023-08-05).

(47) Scheepers, C.; Richardson, S. I.; Moyo-Gwete, T.; Moore, P. L. Antibody Class-Switching as a Strategy to Improve HIV-1 Neutralization. *Trends in Molecular Medicine* **2022**, *28* (11), 979–988. https://doi.org/10.1016/j.molmed.2022.08.010.

(48) Hardy, R. R.; Hayakawa, K. B Cell Development Pathways. *Annu. Rev. Immunol.* **2001**, *19* (1), 595–621. https://doi.org/10.1146/annurev.immunol.19.1.595.

(49) Köhler, G.; Milstein, C. Continuous Cultures of Fused Cells Secreting Antibody of Predefined Specificity. *Nature* **1975**, *256* (5517), 495–497. https://doi.org/10.1038/256495a0.

(50) Mitra, S.; Tomar, P. C. Hybridoma Technology; Advancements, Clinical Significance, and Future Aspects. *Journal of Genetic Engineering and Biotechnology* **2021**, *19* (1), 159. https://doi.org/10.1186/s43141-021-00264-6.

(51) Ledsgaard, L.; Kilstrup, M.; Karatt-Vellatt, A.; McCafferty, J.; Laustsen, A. H. Basics of Antibody Phage Display Technology. *Toxins* **2018**, *10* (6), 236. https://doi.org/10.3390/toxins10060236.

(52) Hoover, J. M.; Prinslow, E. G.; Teigler, J. E.; Truppo, M. D.; La Porte, S. L. Chapter 23 - Therapeutic Antibody Discovery. In *Remington (Twenty-third Edition)*; Adejare, A., Ed.; Academic Press, 2021; pp 417–436. https://doi.org/10.1016/B978-0-12-820007-0.00023-4.

(53) Shendure, J.; Ji, H. Next-Generation DNA Sequencing. *Nat Biotechnol* **2008**, *26* (10), 1135–1145. https://doi.org/10.1038/nbt1486.

(54) Metzker, M. L. Sequencing Technologies — the next Generation. *Nat Rev Genet* **2010**, *11* (1), 31–46. https://doi.org/10.1038/nrg2626.

(55) Compañón, I.; Guerreiro, A.; Mangini, V.; Castro-López, J.; Escudero-Casao, M.; Avenoza, A.; Busto, J. H.; Castillón, S.; Jiménez-Barbero, J.; Asensio, J. L.; Jiménez-Osés, G.; Boutureira, O.; Peregrina, J. M.; Hurtado-Guerrero, R.; Fiammengo, R.; Bernardes, G. J. L.; Corzana, F. Structure-Based Design of Potent Tumor-Associated Antigens: Modulation of Peptide Presentation by Single-Atom O/S or O/Se Substitutions at the Glycosidic Linkage. *J. Am. Chem. Soc.* **2019**, *141* (9), 4063–4072. https://doi.org/10.1021/jacs.8b13503.

(56) DeKosky, B. J.; Ippolito, G. C.; Deschner, R. P.; Lavinder, J. J.; Wine, Y.; Rawlings, B. M.; Varadarajan, N.; Giesecke, C.; Dörner, T.; Andrews, S. F.; Wilson, P. C.; Hunicke-Smith, S. P.; Willson, C. G.; Ellington, A. D.; Georgiou, G. High-Throughput Sequencing of the Paired Human Immunoglobulin Heavy and Light Chain Repertoire. *Nat Biotechnol* **2013**, *31* (2), 166–169. https://doi.org/10.1038/nbt.2492.

(57) Tan, Y.-C.; Blum, L. K.; Kongpachith, S.; Ju, C.-H.; Cai, X.; Lindstrom, T. M.; Sokolove, J.; Robinson, W. H. High-Throughput Sequencing of Natively Paired Antibody Chains Provides Evidence for Original Antigenic Sin Shaping the Antibody Response to Influenza Vaccination. *Clinical Immunology* **2014**, *151* (1), 55–65. https://doi.org/10.1016/j.clim.2013.12.008.

(58) Hou, X.-L.; Wang, L.; Ding, Y.-L.; Xie, Q.; Diao, H.-Y. Current Status and Recent Advances of next Generation Sequencing Techniques in Immunological Repertoire. *Genes Immun* **2016**, *17* (3), 153–164. https://doi.org/10.1038/gene.2016.9.

(59) Boyd, S. D.; Crowe, J. E. Deep Sequencing and Human Antibody Repertoire Analysis. *Current Opinion in Immunology* **2016**, *40*, 103–109. https://doi.org/10.1016/j.coi.2016.03.008.

(60) Georgiou, G.; Ippolito, G. C.; Beausang, J.; Busse, C. E.; Wardemann, H.; Quake, S. R. The Promise and Challenge of High-Throughput Sequencing of the Antibody Repertoire. *Nat Biotechnol* **2014**, *32* (2), 158–168. https://doi.org/10.1038/nbt.2782.

(61) Kashima, Y.; Sakamoto, Y.; Kaneko, K.; Seki, M.; Suzuki, Y.; Suzuki, A. Single-Cell Sequencing Techniques from Individual to Multiomics Analyses. *Exp Mol Med* **2020**, *52* (9), 1419–1427. https://doi.org/10.1038/s12276-020-00499-2.

(62) Misasi, J.; Gilman, M. S. A.; Kanekiyo, M.; Gui, M.; Cagigi, A.; Mulangu, S.; Corti, D.; Ledgerwood, J. E.; Lanzavecchia, A.; Cunningham, J.; Muyembe-Tamfun, J. J.; Baxa, U.; Graham, B. S.; Xiang, Y.; Sullivan, N. J.; McLellan, J. S. Structural and Molecular Basis for Ebola Virus Neutralization by Protective Human Antibodies. *Science* **2016**, *351* (6279), 1343–1346. https://doi.org/10.1126/science.aad6117.

(63) Hibi, T.; Dosch, H.-M. Limiting Dilution Analysis of the B Cell Compartment in Human Bone Marrow. *European Journal of Immunology* **1986**, *16* (2), 139–145. https://doi.org/10.1002/eji.1830160206.

(64) Catherman, A. D.; Skinner, O. S.; Kelleher, N. L. Top Down Proteomics: Facts and Perspectives. *Biochem Biophys Res Commun* **2014**, *445* (4), 683–693. https://doi.org/10.1016/j.bbrc.2014.02.041.

(65) Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M.-C.; Yates, J. R. I. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chem. Rev.* **2013**, *113* (4), 2343–2394. https://doi.org/10.1021/cr3003533.

(66) Kolsrud, H.; Malerod, H.; Ray, S.; Reubsaet, L.; Lundanes, E.; Greibrokk, T. A Critical Review of Trypsin Digestion for LC-MS Based Proteomics. In *Integrative Proteomics*; Leung, H.-C., Ed.; InTech, 2012. https://doi.org/10.5772/29326.

(67) M. Miller, R.; M. Smith, L. Overview and Considerations in Bottom-up Proteomics. *Analyst* **2023**, *148* (3), 475–486. https://doi.org/10.1039/D2AN01246D.

(68) Miller, P. E.; Denton, M. B. The Quadrupole Mass Filter: Basic Operating Concepts. *J. Chem. Educ.* **1986**, *63* (7), 617. https://doi.org/10.1021/ed063p617.

(69) El-Aneed, A.; Cohen, A.; Banoub, J. Mass Spectrometry, Review of the Basics: Electrospray, MALDI, and Commonly Used Mass Analyzers. *Applied Spectroscopy Reviews* **2009**, *44* (3), 210–230. https://doi.org/10.1080/05704920902717872.

(70) Douglas, D. J.; Frank, A. J.; Mao, D. Linear Ion Traps in Mass Spectrometry. *Mass Spectrometry Reviews* **2005**, *24* (1), 1–29. https://doi.org/10.1002/mas.20004.

(71) Haag, A. M. Mass Analyzers and Mass Spectrometers. In *Modern Proteomics – Sample Preparation, Analysis and Practical Applications*; Mirzaei, H., Carrasco, M., Eds.;

Advances in Experimental Medicine and Biology; Springer International Publishing: Cham, 2016; pp 157–169. https://doi.org/10.1007/978-3-319-41448-5_7.

(72) Boesl, U. Time-of-Flight Mass Spectrometry: Introduction to the Basics. *Mass Spectrometry Reviews* **2017**, *36* (1), 86–109. https://doi.org/10.1002/mas.21520.

(73) Mamyrin, B. A. Time-of-Flight Mass Spectrometry (Concepts, Achievements, and Prospects). *International Journal of Mass Spectrometry* **2001**, *206* (3), 251–266. https://doi.org/10.1016/S1387-3806(00)00392-4.

(74) Hillenkamp, F.; Karas, M. [12] Mass Spectrometry of Peptides and Proteins by Matrix-Assisted Ultraviolet Laser Desorption/Ionization. In *Methods in Enzymology*; Mass Spectrometry; Academic Press, 1990; Vol. 193, pp 280–295. https://doi.org/10.1016/0076-6879(90)93420-P.

(75) Hu, Q.; Noll, R. J.; Li, H.; Makarov, A.; Hardman, M.; Graham Cooks, R. The Orbitrap: A New Mass Spectrometer. *Journal of Mass Spectrometry* **2005**, *40* (4), 430–443. https://doi.org/10.1002/jms.856.

(76) Zubarev, R. A.; Makarov, A. Orbitrap Mass Spectrometry. *Anal. Chem.* **2013**, *85* (11), 5288–5296. https://doi.org/10.1021/ac4001223.

(77) Diedrich, J. K.; Pinto, A. F. M.; Yates, J. R. I. Energy Dependence of HCD on Peptide Fragmentation: Stepped Collisional Energy Finds the Sweet Spot. *J. Am. Soc. Mass Spectrom.* **2013**, *24* (11), 1690–1699. https://doi.org/10.1007/s13361-013-0709-7.

(78) Roepstorff, P.; Fohlman, J. Letter to the Editors. *Biomedical Mass Spectrometry* **1984**, *11* (11), 601–601. https://doi.org/10.1002/bms.1200111109.

(79) Johnson, R. S.; Martin, S. A.; Biemann, Klaus.; Stults, J. T.; Watson, J. Throck. Novel Fragmentation Process of Peptides by Collision-Induced Decomposition in a Tandem Mass Spectrometer: Differentiation of Leucine and Isoleucine. *Anal. Chem.* **1987**, *59* (21), 2621–2625. https://doi.org/10.1021/ac00148a019.

(80) Coon, J. J. Collisions or Electrons? Protein Sequence Analysis in the 21st Century. *Anal. Chem.* **2009**, *81* (9), 3208–3215. https://doi.org/10.1021/ac802330b.

(81) Pitteri, S. J.; Chrisman, P. A.; McLuckey, S. A. Electron-Transfer Ion/Ion Reactions of Doubly Protonated Peptides: Effect of Elevated Bath Gas Temperature. *Anal. Chem.* **2005**, *77* (17), 5662–5669. https://doi.org/10.1021/ac050666h.

(82) Brodbelt, J. S.; Morrison, L. J.; Santos, I. Ultraviolet Photodissociation Mass Spectrometry for Analysis of Biological Molecules. *Chem Rev* **2020**, *120* (7), 3328–3380. https://doi.org/10.1021/acs.chemrev.9b00440.

(83) Lavinder, J. J.; Horton, A. P.; Georgiou, G.; Ippolito, G. C. Next-Generation Sequencing and Protein Mass Spectrometry for the Comprehensive Analysis of Human Cellular and Serum Antibody Repertoires. *Current Opinion in Chemical Biology* **2015**, *24*, 112–120. https://doi.org/10.1016/j.cbpa.2014.11.007.

(84) Lavinder, J. J.; Wine, Y.; Giesecke, C.; Ippolito, G. C.; Horton, A. P.; Lungu, O. I.; Hoi, K. H.; DeKosky, B. J.; Murrin, E. M.; Wirth, M. M.; Ellington, A. D.; Dörner, T.; Marcotte, E. M.; Boutz, D. R.; Georgiou, G. Identification and Characterization of the Constituent Human Serum Antibodies Elicited by Vaccination. *Proceedings of the National Academy of Sciences* **2014**, *111* (6), 2259–2264. https://doi.org/10.1073/pnas.1317793111.

(85) Vandermarliere, E.; Mueller, M.; Martens, L. Getting Intimate with Trypsin, the Leading Protease in Proteomics. *Mass Spectrometry Reviews* **2013**, *32* (6), 453–465. https://doi.org/10.1002/mas.21376.

(86) Dongré, A. R.; Jones, J. L.; Somogyi, Á.; Wysocki, V. H. Influence of Peptide Composition, Gas-Phase Basicity, and Chemical Modification on Fragmentation Efficiency: Evidence for the Mobile Proton Model. *J. Am. Chem. Soc.* **1996**, *118* (35), 8365–8374. https://doi.org/10.1021/ja9542193.

(87) Tabb, D. L.; Huang, Y.; Wysocki, V. H.; Yates, J. R. Influence of Basic Residue Content on Fragment Ion Peak Intensities in Low-Energy Collision-Induced Dissociation Spectra of Peptides. *Anal. Chem.* **2004**, *76* (5), 1243–1248. https://doi.org/10.1021/ac0351163.

(88) Michalski, A.; Neuhauser, N.; Cox, J.; Mann, M. A Systematic Investigation into the Nature of Tryptic HCD Spectra. *J. Proteome Res.* **2012**, *11* (11), 5479–5491. https://doi.org/10.1021/pr3007045.

(89) Sun, W.; Wu, S.; Wang, X.; Zheng, D.; Gao, Y. A Systematical Analysis of Tryptic Peptide Identification with Reverse Phase Liquid Chromatography and Electrospray Ion Trap Mass Spectrometry. *Genomics Proteomics Bioinformatics* **2004**, *2* (3), 174–183. https://doi.org/10.1016/S1672-0229(04)02023-6.

(90) Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A.; Reimer, U.; Ehrlich, H.-C.; Aiche, S.; Kuster, B.; Wilhelm, M. Prosit: Proteome-Wide Prediction of Peptide Tandem Mass Spectra by Deep Learning. *Nat Methods* **2019**, *16* (6), 509–518. https://doi.org/10.1038/s41592-019-0426-7.

(91) Tsiatsiani, L.; Heck, A. J. R. Proteomics beyond Trypsin. *The FEBS Journal* **2015**, *282* (14), 2612–2626. https://doi.org/10.1111/febs.13287.

(92) Meyer, J. G.; Kim, S.; Maltby, D. A.; Ghassemian, M.; Bandeira, N.; Komives, E. A. Expanding Proteome Coverage with Orthogonal-Specificity α-Lytic Proteases. *Mol Cell Proteomics* **2014**, *13* (3), 823–835. https://doi.org/10.1074/mcp.M113.034710.

(93) Dau, T.; Bartolomucci, G.; Rappsilber, J. Proteomics Using Protease Alternatives to Trypsin Benefits from Sequential Digestion with Trypsin. *Anal Chem* **2020**, *92* (14), 9523–9527. https://doi.org/10.1021/acs.analchem.0c00478.

(94) Giansanti, P.; Tsiatsiani, L.; Low, T. Y.; Heck, A. J. R. Six Alternative Proteases for Mass Spectrometry–Based Proteomics beyond Trypsin. *Nat Protoc* **2016**, *11* (5), 993–1006. https://doi.org/10.1038/nprot.2016.057.

(95) Blow, D. M. 6 The Structure of Chymotrypsin. In *The Enzymes*; Boyer, P. D., Ed.; Hydrolysis: Peptide Bonds; Academic Press, 1971; Vol. 3, pp 185–212. https://doi.org/10.1016/S1874-6047(08)60397-2.

(96) Poston, C. N.; Higgs, R. E.; You, J.; Gelfanova, V.; Hale, J. E.; Knierman, M. D.; Siegel, R.; Gutierrez, J. A. A Quantitative Tool to Distinguish Isobaric Leucine and Isoleucine Residues for Mass Spectrometry-Based *De novo* Monoclonal Antibody Sequencing. *J. Am. Soc. Mass Spectrom.* **2014**, *25* (7), 1228–1236. https://doi.org/10.1007/s13361-014-0892-1.

(97) Meyer, B.; Papasotiriou, D. G.; Karas, M. 100% Protein Sequence Coverage: A Modern Form of Surrealism in Proteomics. *Amino Acids* **2011**, *41* (2), 291–310. https://doi.org/10.1007/s00726-010-0680-6.

(98) Good, D. M.; Wirtala, M.; McAlister, G. C.; Coon, J. J. Performance Characteristics of Electron Transfer Dissociation Mass Spectrometry*. *Molecular & Cellular Proteomics* **2007**, *6* (11), 1942–1951. https://doi.org/10.1074/mcp.M700073-MCP200.

(99) Taouatas, N.; Drugan, M. M.; Heck, A. J. R.; Mohammed, S. Straightforward Ladder Sequencing of Peptides Using a Lys-N Metalloendopeptidase. *Nat Methods* **2008**, *5* (5), 405–407. https://doi.org/10.1038/nmeth.1204.

(100) Pitteri, S. J.; Chrisman, P. A.; Hogan, J. M.; McLuckey, S. A. Electron Transfer Ion/Ion Reactions in a Three-Dimensional Quadrupole Ion Trap: Reactions of Doubly and Triply Protonated Peptides with SO2·−. *Anal Chem* **2005**, *77* (6), 1831–1839. https://doi.org/10.1021/ac0483872.

(101) Swaney, D. L.; McAlister, G. C.; Wirtala, M.; Schwartz, J. C.; Syka, J. E. P.; Coon, J. J. Supplemental Activation Method for High-Efficiency Electron-Transfer Dissociation of Doubly Protonated Peptide Precursors. *Anal. Chem.* **2007**, *79* (2), 477–485. https://doi.org/10.1021/ac061457f.

(102) Mihalca, R.; van der Burgt, Y. E. M.; McDonnell, L. A.; Duursma, M.; Cerjak, I.; Heck, A. J. R.; Heeren, R. M. A. Combined Infrared Multiphoton Dissociation and Electron-Capture Dissociation Using Co-Linear and Overlapping Beams in Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Rapid Communications in Mass Spectrometry* **2006**, *20* (12), 1838–1844. https://doi.org/10.1002/rcm.2520.

(103) Tsybin, Y. O.; Witt, M.; Baykut, G.; Kjeldsen, F.; Håkansson, P. Combined Infrared Multiphoton Dissociation and Electron Capture Dissociation with a Hollow Electron Beam in Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Rapid Communications in Mass Spectrometry* **2003**, *17* (15), 1759–1768. https://doi.org/10.1002/rcm.1118.

(104) Frese, C. K.; Altelaar, A. F. M.; van den Toorn, H.; Nolting, D.; Griep-Raming, J.; Heck, A. J. R.; Mohammed, S. Toward Full Peptide Sequence Coverage by Dual Fragmentation Combining Electron-Transfer and Higher-Energy Collision Dissociation Tandem Mass Spectrometry. *Anal. Chem.* **2012**, *84* (22), 9668–9673. https://doi.org/10.1021/ac3025366.

(105) Lebedev, A. T.; Damoc, E.; Makarov, A. A.; Samgina, T. Yu. Discrimination of Leucine and Isoleucine in Peptides Sequencing with Orbitrap Fusion Mass Spectrometer. *Anal. Chem.* **2014**, *86* (14), 7017–7022. https://doi.org/10.1021/ac501200h.

(106) Samgina, T. Y.; Kovalev, S. V.; Tolpina, M. D.; Trebse, P.; Torkar, G.; Lebedev, A. T. EThcD Discrimination of Isomeric Leucine/Isoleucine Residues in Sequencing of the Intact Skin Frog Peptides with Intramolecular Disulfide Bond. *J. Am. Soc. Mass Spectrom.* **2018**, *29* (5), 842–852. https://doi.org/10.1007/s13361-017-1857-y.

(107) Fung, Y. M. E.; Chan, T.-W. D. Experimental and Theoretical Investigations of the Loss of Amino Acid Side Chains in Electron Capture Dissociation of Model Peptides. *Journal of the American Society for Mass Spectrometry* **2005**, *16* (9), 1523–1535. https://doi.org/10.1016/j.jasms.2005.05.001.

(108) Sen, K. I.; Tang, W. H.; Nayak, S.; Kil, Y. J.; Bern, M.; Ozoglu, B.; Ueberheide, B.; Davis, D.; Becker, C. Automated Antibody *De novo* Sequencing and Its Utility in Biopharmaceutical Discovery. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (5), 803–810. https://doi.org/10.1007/s13361-016-1580-0.

(109) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. PEAKS: Powerful Software for Peptide *de novo* Sequencing by Tandem Mass Spectrometry. *Rapid Communications in Mass Spectrometry* **2003**, *17* (20), 2337–2342. https://doi.org/10.1002/rcm.1196.

(110) Chi, H.; Sun, R.-X.; Yang, B.; Song, C.-Q.; Wang, L.-H.; Liu, C.; Fu, Y.; Yuan, Z.-F.; Wang, H.-P.; He, S.-M.; Dong, M.-Q. pNovo: *De novo* Peptide Sequencing and Identification Using HCD Spectra. *J. Proteome Res.* **2010**, *9* (5), 2713–2724. https://doi.org/10.1021/pr100182k.

(111) Chi, H.; Chen, H.; He, K.; Wu, L.; Yang, B.; Sun, R.-X.; Liu, J.; Zeng, W.-F.; Song, C.-Q.; He, S.-M.; Dong, M.-Q. pNovo+: *De novo* Peptide Sequencing Using Complementary HCD and ETD Tandem Mass Spectra. *J. Proteome Res.* **2013**, *12* (2), 615–625. https://doi.org/10.1021/pr3006843.

(112) Yang, H.; Chi, H.; Zeng, W.-F.; Zhou, W.-J.; He, S.-M. pNovo 3: Precise *de novo* Peptide Sequencing Using a Learning-to-Rank Framework. *Bioinformatics* **2019**, *35* (14), i183–i190. https://doi.org/10.1093/bioinformatics/btz366.

(113) Gutenbrunner, P.; Kyriakidou, P.; Welker, F.; Cox, J. Spectrum Graph-Based de-Novo Sequencing Algorithm MaxNovo Achieves High Peptide Identification Rates in Collisional Dissociation MS/MS Spectra. bioRxiv September 6, 2021, p 2021.09.04.458985. https://doi.org/10.1101/2021.09.04.458985.

(114) Ma, B. Novor: Real-Time Peptide *de novo* Sequencing Software. *J. Am. Soc. Mass Spectrom.* **2015**, *26* (11), 1885–1894. https://doi.org/10.1007/s13361-015-1204-0.

(115) Yilmaz, M.; Fondrie, W.; Bittremieux, W.; Oh, S.; Noble, W. S. *De novo* Mass Spectrometry Peptide Sequencing with a Transformer Model. In *Proceedings of the 39th International Conference on Machine Learning*; PMLR, 2022; pp 25514–25522.

(116) Yilmaz, M.; Fondrie, W. E.; Bittremieux, W.; Nelson, R.; Ananth, V.; Oh, S.; Noble, W. S. *Sequence-to-Sequence Translation from Mass Spectra to Peptides with a Transformer Model*; preprint; Bioinformatics, 2023. https://doi.org/10.1101/2023.01.03.522621.

(117) Bern, M.; Kil, Y. J.; Becker, C. Byonic: Advanced Peptide and Protein Identification Software. *Current Protocols in Bioinformatics* **2012**, *40* (1), 13.20.1-13.20.14. https://doi.org/10.1002/0471250953.bi1320s40.

(118) Bhatia, S.; Kil, Y. J.; Ueberheide, B.; Chait, B. T.; Tayo, L.; Cruz, L.; Lu, B.; Yates, J. R. I.; Bern, M. Constrained *De novo* Sequencing of Conotoxins. *J. Proteome Res.* **2012**, *11* (8), 4191–4200. https://doi.org/10.1021/pr300312h.

(119) Bern, M.; Cai, Y.; Goldberg, D. Lookup Peaks: A Hybrid of *de novo* Sequencing and Database Search for Protein Identification by Tandem Mass Spectrometry. *Anal. Chem.* **2007**, *79* (4), 1393–1400. https://doi.org/10.1021/ac0617013.

(120) *de novo Peptide Sequencing | LC-MS/MS Software*. Bioinformatics Solutions Inc. https://www.bioinfor.com/de-novo-sequencing/ (accessed 2023-08-16).

(121) *novor.cloud*. https://app.novor.cloud/request (accessed 2023-08-16).

(122) Bartels, C. Fast Algorithm for Peptide Sequencing by Mass Spectroscopy. *Biomedical & Environmental Mass Spectrometry* **1990**, *19* (6), 363–368. https://doi.org/10.1002/bms.1200190607.

(123) Hughes, C.; Ma, B.; Lajoie, G. A. *De novo* Sequencing Methods in Proteomics. In *Proteome Bioinformatics*; Hubbard, S. J., Jones, A. R., Eds.; Methods in Molecular Biology™; Humana Press: Totowa, NJ, 2010; pp 105–121. https://doi.org/10.1007/978-1-60761-444-9_8.

(124) Frank, A. M.; Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A.; Pevzner, P. A. *De novo* Peptide Sequencing and Identification with Precision Mass Spectrometry. *J. Proteome Res.* **2007**, *6* (1), 114–123. https://doi.org/10.1021/pr060271u.

(125) Tran, N. H.; Zhang, X.; Xin, L.; Shan, B.; Li, M. *De novo* Peptide Sequencing by Deep Learning. *Proceedings of the National Academy of Sciences* **2017**, *114* (31), 8247–8252. https://doi.org/10.1073/pnas.1705691114.

(126) Tran, N. H.; Qiao, R.; Xin, L.; Chen, X.; Liu, C.; Zhang, X.; Shan, B.; Ghodsi, A.; Li, M. Deep Learning Enables *de novo* Peptide Sequencing from Data-Independent-Acquisition Mass Spectrometry. *Nat Methods* **2019**, *16* (1), 63–66. https://doi.org/10.1038/s41592-018-0260-3.

(127) Mitchell, T. M. *Machine Learning*; McGraw-Hill series in computer science; McGraw-Hill: New York, 1997.

(128) Bailey, D. J.; Rose, C. M.; McAlister, G. C.; Brumbaugh, J.; Yu, P.; Wenger, C. D.; Westphall, M. S.; Thomson, J. A.; Coon, J. J. Instant Spectral Assignment for Advanced Decision Tree-Driven Mass Spectrometry. *Proc Natl Acad Sci U S A* **2012**, *109* (22), 8411–8416. https://doi.org/10.1073/pnas.1205292109.

(129) Spoerry, C.; Hessle, P.; Lewis, M. J.; Paton, L.; Woof, J. M.; Pawel-Rammingen, U. von. Novel IgG-Degrading Enzymes of the IgdE Protease Family Link Substrate Specificity to Host Tropism of Streptococcus Species. *PLOS ONE* **2016**, *11* (10), e0164809. https://doi.org/10.1371/journal.pone.0164809.

(130) Rosenstein, S.; Vaisman-Mentesh, A.; Levy, L.; Kigel, A.; Dror, Y.; Wine, Y. Production of F(Ab′)2 from Monoclonal and Polyclonal Antibodies. *Current Protocols in Molecular Biology* **2020**, *131* (1), e119. https://doi.org/10.1002/cpmb.119.

(131) Johansson, B. P.; Shannon, O.; Björck, L. IdeS: A Bacterial Proteolytic Enzyme with Therapeutic Potential. *PLoS One* **2008**, *3* (2), e1692. https://doi.org/10.1371/journal.pone.0001692.

(132) Deveuve, Q.; Lajoie, L.; Barrault, B.; Thibault, G. The Proteolytic Cleavage of Therapeutic Monoclonal Antibody Hinge Region: More Than a Matter of Subclass. *Frontiers in Immunology* **2020**, *11*.

(133) Bondt, A.; Hoek, M.; Tamara, S.; de Graaf, B.; Peng, W.; Schulte, D.; van Rijswijck, D. M. H.; den Boer, M. A.; Greisch, J.-F.; Varkila, M. R. J.; Snijder, J.; Cremer, O. L.; Bonten, M. J. M.; Heck, A. J. R. Human Plasma IgG1 Repertoires Are Simple, Unique, and Dynamic. *Cell Systems* **2021**, *12* (12), 1131-1143.e5. https://doi.org/10.1016/j.cels.2021.08.008.

(134) Shaw, J. B.; Liu, W.; Vasil′ev, Y. V.; Bracken, C. C.; Malhan, N.; Guthals, A.; Beckman, J. S.; Voinov, V. G. Direct Determination of Antibody Chain Pairing by Top-down and Middle-down Mass Spectrometry Using Electron Capture Dissociation and Ultraviolet Photodissociation. *Anal. Chem.* **2020**, *92* (1), 766–773. https://doi.org/10.1021/acs.analchem.9b03129.

(135) Den Boer, M. A.; Greisch, J.-F.; Tamara, S.; Bondt, A.; Heck, A. J. R. Selectivity over Coverage in *de novo* Sequencing of IgGs. *Chem. Sci.* **2020**, *11* (43), 11886–11896. https://doi.org/10.1039/D0SC03438J.

(136) Dupré, M.; Duchateau, M.; Sternke-Hoffmann, R.; Boquoi, A.; Malosse, C.; Fenk, R.; Haas, R.; Buell, A. K.; Rey, M.; Chamot-Rooke, J. *De novo* Sequencing of Antibody Light Chain Proteoforms from Patients with Multiple Myeloma. *Anal. Chem.* **2021**, *93* (30), 10627–10634. https://doi.org/10.1021/acs.analchem.1c01955.

(137) Fellers, R. T.; Greer, J. B.; Early, B. P.; Yu, X.; LeDuc, R. D.; Kelleher, N. L.; Thomas, P. M. ProSight Lite: Graphical Software to Analyze Top-down Mass Spectrometry Data. *PROTEOMICS* **2015**, *15* (7), 1235–1238. https://doi.org/10.1002/pmic.201400313.

(138) Park, J.; Piehowski, P. D.; Wilkins, C.; Zhou, M.; Mendoza, J.; Fujimoto, G. M.; Gibbons, B. C.; Shaw, J. B.; Shen, Y.; Shukla, A. K.; Moore, R. J.; Liu, T.; Petyuk, V. A.; Tolic, N.; Pasa-Tolic, L.; Smith, R. D.; Payne, S. H.; Kim, S. Informed-Proteomics: Open Source Software Package for Top-down Proteomics. *Nat Methods* **2017**, *14* (9), 909–914. https://doi.org/10.1038/nmeth.4388.

Chapter 2

Mass spectrometry-based *de novo* sequencing of monoclonal antibodies using multiple proteases and a dual fragmentation scheme

Authors:

Weiwei Peng[1#], Matti F. Pronker[1#], Joost Snijder[1*]

[#]equal contribution

*corresponding author: j.snijder@uu.nl

Affiliation:

[1] Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute of Pharmaceutical Sciences, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

Published as:

https://doi.org/10.1021/acs.jproteome.1c00169

Abstract:

Antibody sequence information is crucial to understanding the structural basis for antigen binding and enables the use of antibodies as therapeutics and research tools. Here we demonstrate a method for direct *de novo* sequencing of monoclonal IgG from the purified antibody products. The method uses a panel of multiple complementary proteases to generate suitable peptides for *de novo* sequencing by LC-MS/MS in a bottom-up fashion. Furthermore, we apply a dual fragmentation scheme, using both stepped high-energy collision dissociation (stepped HCD) and electron transfer high-energy collision dissociation (EThcD) on all peptide precursors. The method achieves full sequence coverage of the monoclonal antibody Herceptin, with an accuracy of 99% in the variable regions. We applied the method to sequence the widely used anti-FLAG™-M2 mouse monoclonal antibody, which we successfully validated by remodeling a high-resolution crystal structure of the Fab and demonstrating binding to a FLAG™-tagged target protein in Western blot analysis. The method thus offers robust and reliable sequences of monoclonal antibodies.

Keywords:

mass spectrometry, antibody, *de novo* sequencing, EThcD, stepped HCD, Herceptin, FLAG tag, anti-FLAG-M2.

Introduction

Antibodies can bind a great molecular diversity of antigens, owing to the high degree of sequence diversity that is available through somatic recombination, hypermutation, and heavy-light chain pairings (1, 2). Sequence information on antibodies therefore is crucial to understanding the structural basis of antigen binding, how somatic hypermutation governs affinity maturation, and an overall understanding of the adaptive immune response in health and disease, by mapping out the antibody repertoire. Moreover, antibodies have become invaluable research tools in the life sciences and ever more widely developed as therapeutic agents (3, 4). In this context, sequence information is crucial for the use, production and validation of these important research tools and biopharmaceutical agents (5, 6).

Antibody sequences are typically obtained through cloning and sequencing of the coding mRNAs of the paired heavy and light chains (7-9). The sequencing workflows thereby rely on isolation of the antibody-producing cells from peripheral blood monocytes, or spleen and bone marrow tissues. These antibody-producing cells are not always readily available however, and cloning/sequencing of the paired heavy and light chains is a non-trivial task with a limited success rate (7-9). Moreover, antibodies are secreted in bodily fluids and mucus. Antibodies are thereby in large part functionally disconnected from their producing B-cell, which raises questions on how the secreted antibody pool relates quantitatively to the underlying B-cell population and whether there are potential sampling biases in current antibody sequencing strategies.

Direct mass spectrometry (MS)-based sequencing of the secreted antibody products is a useful complementary tool that can address some of the challenges faced by conventional sequencing strategies relying on cloning/sequencing of the coding mRNAs (10-17). MS-based methods do not rely on the availability of the antibody-producing cells, but rather target the polypeptide products directly, offering the prospect of a next generation of serology, in which secreted antibody sequences might be obtained from any bodily fluid. Whereas MS-based *de novo* sequencing still has a long way to go towards this goal, owing to limitations in sample requirements, sequencing accuracy, read length and sequence assembly, MS has been successfully used to profile the antibody repertoire and obtain (partial) antibody sequences beyond those available from conventional sequencing strategies based on cloning/sequencing of the coding mRNAs (10-17).

Most MS-based strategies for antibody sequencing rely on a proteomics-type bottom-up LC-MS/MS workflow, in which the antibody product is digested into smaller peptides for MS analysis (14, 18-23). Available germline antibody sequences are then often used either as a template to guide assembly of *de novo* peptide reads (such as in PEAKS Ab) (23), or used as a starting point to iteratively identify somatic mutations to arrive at the mature antibody sequence (such as in Supernovo) (21). To maximize sequence coverage and aid read assembly, these MS-based workflows typically use a combination of complementary proteases and unspecific digestion to generate overlapping peptides. The most straightforward application of these MS-based sequencing workflows is the successful sequencing of monoclonal antibodies from (lost) hybridoma cell lines, but it also forms the basis of more advanced and challenging applications to characterize polyclonal antibody mixtures and profile the full antibody repertoire from serum.

Here we describe an efficient protocol for MS-based sequencing of monoclonal antibodies. The protocol requires approximately 200 picomol of the antibody product and sample preparation can be completed within one working day. We selected a panel of 9 proteases with complementary specificities, which are active in the same buffer conditions for parallel digestion of the antibodies. We developed a dual fragmentation strategy for MS/MS analysis of the resulting peptides to yield rich sequence information from the fragmentation spectra of the peptides. The protocol yields full and deep sequence coverage of the variable domains of both heavy and light chains as demonstrated on the monoclonal antibody Herceptin. As a test case, we used our protocol to sequence the widely used anti-FLAG™-M2 mouse monoclonal antibody, for which no sequence was publicly available despite its described use in 5000+ peer-reviewed publications (24, 25). The protocol achieved full sequence coverage of the variable domains of both heavy and light chains, including all complementarity determining regions (CDRs). The obtained sequence was successfully validated by remodeling the published crystal structure of the anti-FLAG™-M2 Fab and demonstrating binding of the synthetic recombinant antibody following the experimental sequence to a FLAG™-tagged protein in Western blot analysis. The protocol developed here thus offers robust and reliable sequencing of monoclonal antibodies with prospective applications for sequencing secreted antibodies from bodily fluids.

Method

Sample preparation

Anti-FLAG™ M2 antibody was purchased from Sigma (catalogue number F1804). Herceptin was provided by Roche (Penzberg, Germany). 27 µg of each sample was denatured in 2% sodium deoxycholate (SDC), 200 mM Tris-HCl, 10 mM tris(2-carboxyethyl)phosphine (TCEP), pH 8.0 at 95°C for 10 min, followed with 30 min incubation at 37°C for reduction. Sample was then alkylated by adding iodoacetic acid to a final concentration of 40 mM and incubated in the dark at room temperature for 45 min. 3 µg Sample was then digested by one of the following proteases: trypsin, chymotrypsin, lysN, lysC, gluC, aspN, aLP, thermolysin and elastase in a 1:50 ratio (w:w) in a total volume of 100 uL of 50 mM ammonium bicarbonate at 37°C for 4 h. After digestion, SDC was removed by adding 2 uL formic acid (FA) and centrifugation at 14000 g for 20 min. Following centrifugation, the supernatant containing the peptides was collected for desalting on a 30 µm Oasis HLB 96-well plate (Waters). The Oasis HLB sorbent was activated with 100% acetonitrile and subsequently equilibrated with 10% formic acid in water. Next, peptides were bound to the sorbent, washed twice with 10% formic acid in water and eluted with 100 µL of 50% acetonitrile/5% formic acid in water (v/v). The eluted peptides were vacuum-dried and reconstituted in 100 µL 2% FA.

Mass Spectrometry

The digested peptides (single injection of 0.2 ug) were separated by online reversed phase chromatography on an Agilent 1290 UHPLC (column packed with Poroshell 120 EC C18; dimensions 50 cm x 75 µm, 2.7 µm, Agilent Technologies) coupled to a Thermo Scientific Orbitrap Fusion mass spectrometer. Samples were eluted over a 90 min gradient from 0% to 35% acetonitrile at a flow rate of 0.3 µL/min. Peptides were analyzed with a resolution setting of 60000 in MS1. MS1 scans were obtained with standard AGC target, maximum injection time of 50 ms, and scan range 350-2000. The precursors were selected with a 3 m/z window and fragmented by stepped HCD as well as EThcD. The stepped HCD fragmentation included steps of 25%, 35% and 50% NCE. EThcD fragmentation was performed with calibrated charge-dependent ETD parameters and 27% NCE supplemental activation. For both fragmentation types, ms2 scans were acquired at 30000 resolution, 4e5 AGC target, 250 ms maximum injection time, scan range 120-3500.

MS Data Analysis

Automated *de novo* sequencing was performed with Supernovo (version 3.10, Protein Metrics Inc.). Supernovo identifies closely matching antibody germline sequences from an initial database search, followed by iterative substitutions to the recombined V-J-C sequences of heavy and light chain by wildcard searches on the MS/MS spectra to converge to the final output sequence (21). Custom parameters were used as follows: non-specific digestion; precursor and product mass tolerance was set to 12 ppm and 0.02 Da respectively; carboxymethylation (+58.005479) on cysteine was set as fixed modification; oxidation on methionine and tryptophan was set as variable common 1 modification; carboxymethylation on the N-terminus, pyroglutamic acid conversion of glutamine and glutamic acid on the N-terminus, deamidation on asparagine/glutamine were set as variable rare 1 modifications. Peptides were filtered for score >=500 for the final evaluation of spectrum quality and (depth of) coverage. The 'depth of coverage' was defined as the number of unique peptides with score >=500 that cover the position. Supernovo generates peptide groups for redundant MS/MS spectra, including also when stepped HCD and EThcD fragmentation on the same precursor both generate good peptide-spectrum matches. In these cases only the best-matched spectrum is counted as representative for that group. This criterium was used in counting the number of peptide reads reported in Table S1. Germline sequences and CDR boundaries were inferred using IMGT/DomainGapAlign (26, 27).

Revision of the anti-FLAG™-M2 Fab crystal structure model

As a starting point for model building, the reflection file and coordinates of the published anti-FLAG™-M2 Fab crystal structure were used (PDB ID: 2G60) (28). Care was taken to use the original R_{free} labels of the deposited reflection file for refinement, so as not to introduce extra model bias. Differential residues between this structure and our mass spectrometry-derived anti-FLAG™ sequence were manually mutated and fitted in the density using Coot (29). Many spurious water molecules that caused severe steric clashes in the original model were also manually removed in Coot. Densities for two sulfate and one chloride ion were identified and built into the model. The original crystallization solution contained 0.1 M ammonium sulfate. Iterative cycles of model geometry optimization in real space in Coot and reciprocal space refinement by Phenix were used to generate the final model, which was validated with Molprobity (30, 31).

Cloning and expression of synthetic recombinant anti-FLAG™-M2

To recombinantly express full-length anti-FLAG™-M2, the proteomic sequences of both the light and heavy chains were reverse-translated and codon optimized for expression in human cells using the Integrated DNA Technologies (IDT) web tool (*http://www.idtdna.com/CodonOpt*) (32). For the linker and Fc region of the heavy chain, the standard mouse Ig gamma-1 (IGHG1) amino acid sequence (Uniprot P01868.1) was used. An N-terminal secretion signal peptide derived from human IgG light chain (MEAPAQLLFLLLLWLPDTTG) was added to the N-termini of both heavy and light chains. BamHI and NotI restriction sites were added to the 5' and 3' ends of the coding regions, respectively. Only for the light chain, a double stop codon was introduced at the 3' site before the NotI restriction site. The coding regions were subcloned using BamHI and NotI restriction-ligation into a pRK5 expression vector with a C-terminal octahistidine tag between the NotI site and a double stop codon 3' of the insert, so that only the heavy chain has a C-terminal AAAHHHHHHHH sequence for Nickel-affinity purification (the triple alanine resulting from the NotI site). The L51I correction in the heavy chain was introduced later (after observing it in the crystal structure) by IVA cloning (33). Expression plasmids for the heavy and light chain were mixed in a 1:1 (w/w) ratio for transient transfection in HEK293 cells with polyethylenimine, following standard procedures. Medium was collected 6 days after transfection and cells were spun down by 10 minutes of centrifugation at 1000 *g*. Antibody was directly purified from the supernatant using Ni-sepharose excel resin (Cytiva Lifes Sciences), washing with 500 mM NaCl, 2 mM $CaCl_2$, 15 mM imidazole, 20 mM HEPES pH 7.8 and eluting with 500 mM NaCl, 2 mM $CaCl_2$, 200 mM imidazole, 20 mM HEPES pH 7.8.

Western blot validation of anti-FLAG™-M2 binding

To test binding of our recombinant anti-FLAG™-M2 to the FLAG™-tag epitope, compared to the commercially available anti-FLAG™-M2 (Sigma), we used both antibodies to probe Western blots of a FLAG™-tagged protein in parallel. Purified Rabies virus glycoprotein ectodomain (SAD B19 strain, UNIPROT residues 20-450) with or without a C-terminal FLAG™-tag followed by a foldon trimerization domain and an octahistidine tag was heated to 95 °C in XT sample buffer (Biorad) for 5 minutes. Samples were run twice on a Criterion XT 4-12% polyacrylamide gel (Biorad) in MES XT buffer (Biorad) before Western blot transfer to a nitrocellulose membrane in tris-glycine buffer (Biorad) with 20% methanol. The membrane was blocked with 5% (w/v) dry non-fat milk in phosphate-buffered saline (PBS) overnight at 4 °C. The membrane was cut in two (one half for the commercial and one half for the recombinant anti-FLAG™-M2) and each half was probed with either commercial (Sigma) or recombinant anti-FLAG™-M2 at 1 μg/mL in PBS for 45 minutes. After washing three times with PBST (PBS with 0.1% v/v Tween20), polyclonal goat anti-mouse fused to horseradish peroxidase (HRP) was used to detect binding of anti-FLAG™-M2 to the FLAG™-tagged protein for both membranes. The membranes were washed three more times with PBST before applying enhanced chemiluminescence (ECL; Pierce) reagent to image the blots in parallel.

Results

We used an in-solution digestion protocol, with sodium-deoxycholate as the denaturing agent, to generate peptides from the antibodies for LC-MS/MS analysis. Following heat denaturation and disulfide bond reduction, we used iodoacetic acid as the alkylating agent to cap free cysteines. Note that conventional alkylating agents like iodo-/chloroacetamide generate +57 Da mass differences on cysteines and primary amines, which may lead to spurious assignments as glycine residues in *de novo* sequencing. The +58 Da mass differences generated by alkylation with iodoacetic acid circumvents this potential pitfall.

We chose a panel of 9 proteases with activity at pH 7.5-8.5, so that the denatured, reduced and alkylated antibodies could be easily split for parallel digestion under the same buffer conditions. These proteases (with indicated cleavage specificities) included: trypsin (C-terminal of R/K), chymotrypsin (C-terminal of F/Y/W/M/L), α-lytic protease (C-terminal of T/A/S/V), elastase (unspecific), thermolysin (unspecific), lysN (N-terminal of K), lysC (C-terminal of K), aspN (N-terminal of D/E), and gluC (C-terminal of D/E). Correct placement or assembly of peptide reads is a common challenge in *de novo* sequencing, which can be facilitated by sufficient overlap between the peptide reads. This favors the occurrence of missed cleavages and longer reads, so we opted to perform a brief 4-hour digestion. Following digestion, SDC is removed by precipitation and the peptide supernatant is desalted, ready for LC-MS/MS analysis. The resulting raw data was used for automated *de novo* sequencing with the Supernovo software package.

As peptide fragmentation is dependent on many factors like length, charge state, composition and sequence (34), we needed a versatile fragmentation strategy to accommodate the diversity of antibody-derived peptides generated by the 9 proteases. We opted for a dual fragmentation scheme that applies both stepped high-energy collision dissociation (stepped HCD) and electron transfer high-energy collision dissociation (EThcD) on all peptide precursors (35-37). The stepped HCD fragmentation includes three collision energies to cover multiple dissociation regimes and the EThcD fragmentation works especially well for higher charge states, also adding complementary c/z ions for maximum sequence coverage.

We used the monoclonal antibody Herceptin (also known as Trastuzumab) as a benchmark to test our protocol (38, 39). From the total dataset of 9 proteases, we collected 4408 peptide reads (defined as peptides with score >=500, see methods for details), 2866 of which with superior stepped HCD fragmentation (compared to EThcD) and conversely 1722 peptide reads with superior EThcD fragmentation (see Table S1). Sequence coverage was 100% in both heavy and light chains across the variable and constant domains (see Figures S1 and S2). The median depth of coverage was 148 overall and slightly higher in the light chain (see Table S1 and Figure S1-2). The median depth of coverage in the CDRs of both chains ranged from 42 to 210.

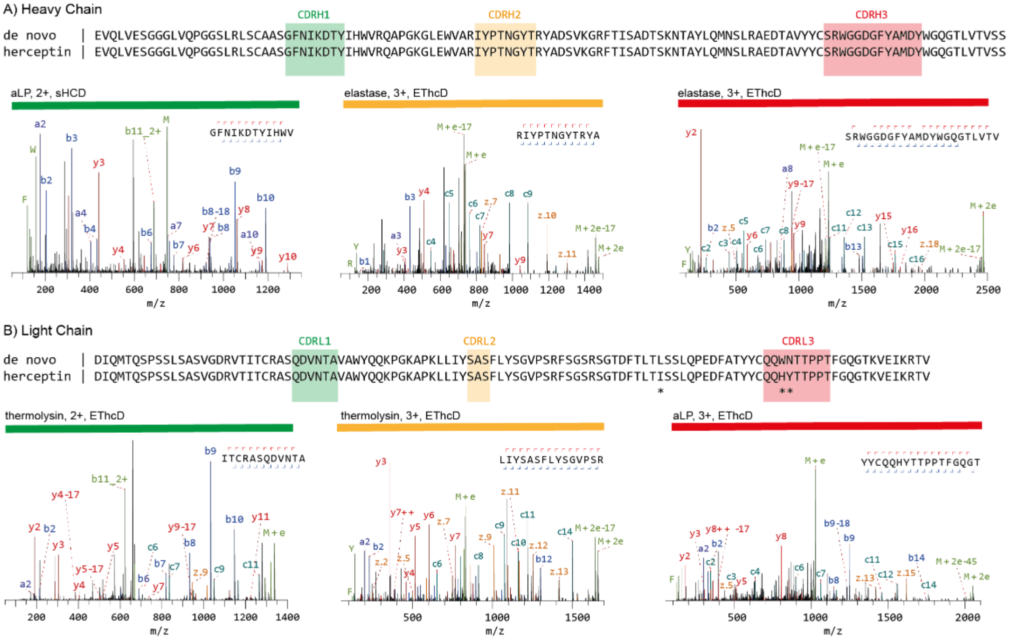


Figure 1. *Mass spectrometry-based de novo sequencing of the monoclonal antibody Herceptin. The variable regions of the Heavy (A) and Light Chains (B) are shown. The MS-based sequence is shown alongside the known Herceptin sequence, with differences highlighted by asterisks (*). Exemplary MS/MS spectra supporting the assigned sequences of the Heavy and Light Chain CDRs are shown below the alignments with protease, precursor charge state and fragmentation type indicated. Peptide sequence and fragment coverage are indicated on top of the spectra, with b/c ions indicated in blue and y/z ions in red. The same coloring is used to annotate peaks in the spectra, with additional peaks such as intact/charge reduced precursors, neutral losses and immonium ions indicated in green. Note that to prevent overlapping peak labels, only a subset of successfully matched peaks is annotated.*

The experimentally determined *de novo* sequence is shown alongside the known Herceptin sequence for the variable domains of both chains in Figure 1, with exemplary MS/MS spectra for the CDRs. We achieved an overall sequence accuracy of 99% with the automated sequencing procedure of Supernovo, with 3 incorrect assignments in the light chain. In framework 3 of the light chain, I75 was incorrectly assigned as the isomer Leucine (L), a common MS-based sequencing error. In CDRL3 of the light chain, an additional misassignment was made for the dipeptide H91/Y92, which was incorrectly assigned as W91/N92. The dipeptides HY and WN have identical masses, and the misassignment of W91/N92 (especially W91) was poorly supported by the fragmentation spectra, in contrast to the correct H91/Y92 assignment (see c6/c7 in fragmentation spectra, Figure 1). Overall, the protocol yielded highly accurate sequences at a combined 230/233 positions of the variable domains in Herceptin.

The combined use of both stepped HCD and EThcD fragmentation resulted in superior accuracy compared to the separate fragmentation techniques (see Figure S3). Likewise, the use of all nine proteases resulted in superior accuracy compared to a smaller subset of trypsin, chymotrypsin, thermolysin and elastase or single protease datasets (see Figure S3). Finally, compared to overnight digestion, the shorter 4-hour digestion of our protocol resulted in peptides of similar length (see Figure S4). However, specific proteases showed different effects of digestion time; overnight digestion gave a higher number of peptides for trypsin and chymotrypsin, but fewer for elastase and thermolysin digestion. From the subset of 4 proteases (trypsin, chymotrypsin, elastase and thermolysin) used for this comparison, the overnight digestion resulted in fewer errors compared to the 4 hour digestion overall, to an equivalent number as observed in the full dataset with 9 proteases. The main benefit of the shortened digestion is therefore that the sample preparation can be completed within one working day. These comparisons highlight that the key to accurate sequencing with our protocol is the dual fragmentation scheme in combination with the multitude of proteases for antibody digestion, rather than digestion time, and that the protocol could be further optimized by adapting digestion time for specific proteases individually.

We next applied our sequencing protocol to the mouse monoclonal anti-FLAG™-M2 antibody as a test case (24). Despite the widespread use of anti-FLAG™-M2 to detect and purify FLAG™-tagged proteins (40), the only publicly available sequences can be found in the crystal structure of the Fab (28). The modelled sequence of the original crystal structure had to be inferred from germline sequences that could match the experimental electron density and also includes many placeholder Alanines at positions that could not be straightforwardly interpreted. The full anti-FLAG™-M2 dataset from the 9 proteases included 3371 peptide reads (with scores >= 500); 1983 with superior stepped HCD fragmentation spectra (compared to EThcD) and conversely 1388 with superior EThcD spectra. We achieved full sequence coverage of the variable regions of both heavy and light chains, with a median depth of coverage in the CDRs ranging from 32 to 192 (see Table

S1). As for Herceptin, the depth of coverage was better in the light chain compared to the heavy chain (see Figure S1-S2). The full MS-based anti-FLAG™-M2 sequences can be found in FASTA format in the supplementary information.

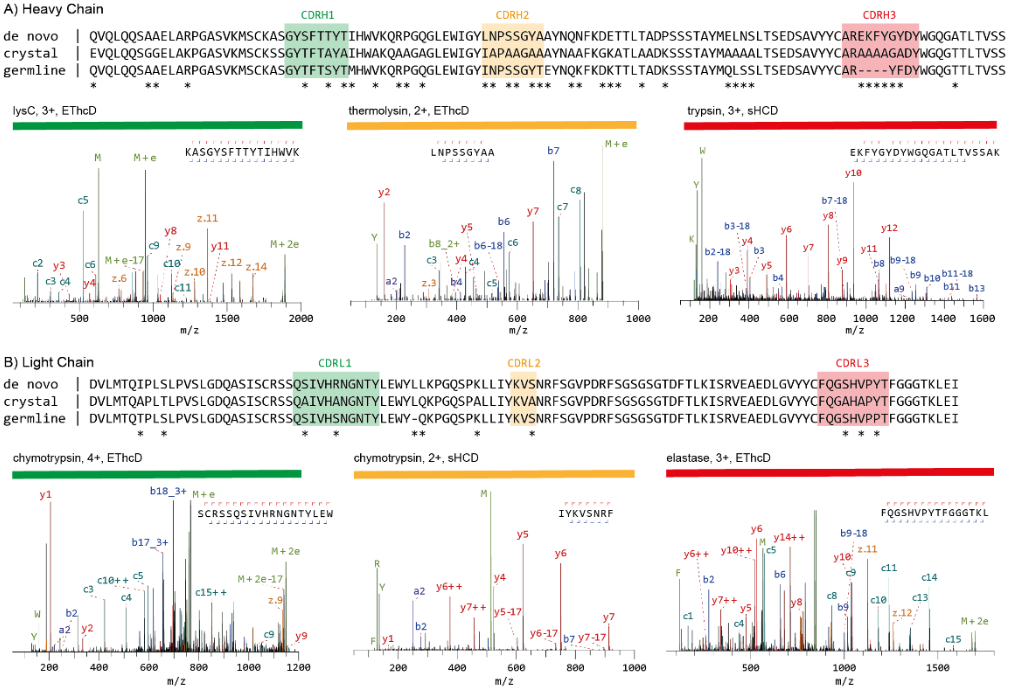


Figure 2. *Mass spectrometry based de novo sequence of the mouse monoclonal anti-FLAG™-M2 antibody. The variable regions of the Heavy (A) and Light Chains (B) are shown. The MS-based sequence is shown alongside the previously published sequenced in the crystal structure of the Fab (PDB ID: 2G60), and germline sequence (IMGT-DomainGapAlign; IGHV1-04/IGHJ2; IGKV1-117/IGKJ1). Differential residues are highlighted by asterisks (*). Exemplary MS/MS spectra in support of the assigned sequences are shown below the alignments with protease, precursor charge state and fragmentation type indicated. Peptide sequence and fragment coverage are indicated on top of the spectra, with b/c ions indicated in blue, y/z ions in red. The same coloring is used to annotate peaks in the spectra, with additional peaks such as intact/charge reduced precursors, neutral losses and immonium ions indicated in green. Note that to prevent overlapping peak labels, only a subset of successfully matched peaks is annotated.*

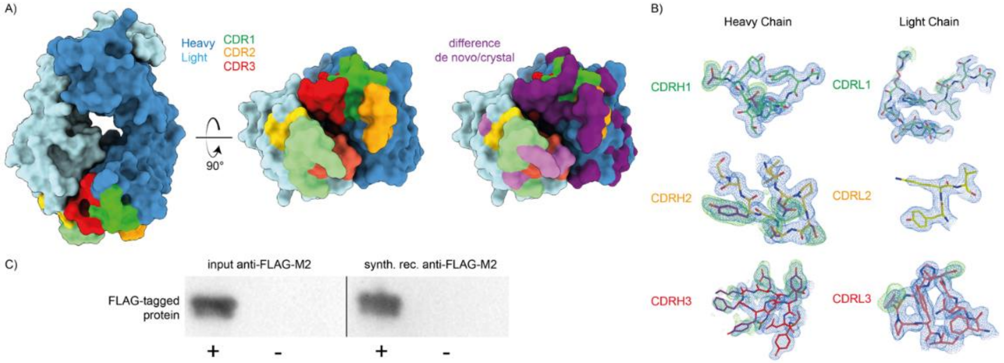


Figure 3. *Validation of the MS-based anti-FLAG™-M2 sequence. A) the previously published crystal structure of the anti-FLAG™-M2 FAb was remodeled with the experimentally determined sequence, shown in surface rendering with CDRs and differential residues highlighted in colors. B)* $2F_o$*-*F_c *electron density of the new refined map contoured at 1 RMSD is shown in blue and* F_o*-*F_c *positive difference density of the original deposited map contoured at 1.7 RMSD in green around the CDR loops of the heavy and light chains. Differential residues between the published crystal structure and the model based on our antibody sequencing are indicated in purple. C) Western blot validation of the synthetic recombinant anti-FLAG™-M2 antibody produced with the experimentally determined sequence demonstrate equivalent FLAG™-tag binding compared to commercial anti-FLAG™-M2 (see also Figure S3).*

The MS-based sequences of anti-FLAG™-M2 are shown alongside the crystal structure sequences and the inferred germline precursors with exemplary MS/MS spectra for the CDRs in Figure 2. The experimentally determined sequence reveals that anti-FLAG™-M2 is a mouse IgG1, with an IGHV1-04/IGHJ2 heavy chain and IGKV1-117/IGKJ1 kappa light chain. The experimentally determined sequence differs at 34 and 9 positions in the heavy and light chain of the Fab crystal structure, respectively. To validate the experimentally determined sequences, we remodeled the crystal structure using the MS-based heavy and light chains, resulting in much improved model statistics (see Figure 3 and Table S2). The experimental electron densities show excellent support of the MS-based sequence (as shown for the CDRs in Figure 3B). A notable exception is L51 in CDRH2 of the heavy chain. The MS-based sequence was assigned as Leucine, but the experimental electron density supports assignment of the isomer Isoleucine instead (see Figure S5). In contrast to the original model our new MS-based model reveals a predominantly positively charged paratope (see Figure S6), which potentially complements the -3 net charge of the FLAG™

tag epitope (DYKDDDDK) to mediate binding. The experimentally determined anti-FLAG™-M2 sequence, with the L51I correction, was further validated by testing binding of the synthetic recombinant antibody to a purified FLAG™-tagged protein in Western blot analysis (see Figure 3C and S7). The synthetic recombinant antibody showed equivalent binding compared to the original antibody sample used for sequencing, confirming that the experimentally determined sequence is reliable to obtain the recombinant antibody product with the desired functional profile.

Discussion

There are four other monoclonal antibody sequences against the FLAG™ tag publicly available through the ABCD (AntiBodies Chemically Defined) database (41-43). Comparison of the CDRs of anti-FLAG™-M2 with these additional four monoclonal antibodies reveals a few common motifs that may determine FLAG™-tag binding specificity (see Table S3). In the heavy chain, the only common motif between all five monoclonals is that the first three residues of CDRH1 follow a GXS sequence. In addition, the last three residues of CDRH3 of anti-FLAG™-M2 are YDY, similar to MDY in 2H8, and YDF in EEh13.6 (and EEh14.3 also ends CDRH3 with an aromatic F residue). In contrast to the heavy chain, the CDRs of the light chain are almost completely conserved in 4/5 monoclonals with only minimal differences compared to germline. The anti-FLAG™-M2 and 2H8 monoclonals were specifically raised in mice against the FLAG™-tag epitope (24, 42), whereas the computationally designed EEh13.6 and EEh14.3 monoclonals contain the same light chain from an EE-dipeptide tag directed antibody (41). This suggests that the IGKV1-117/IGKJ1 light chain may be a common determinant of binding to a small negatively charged peptide epitope like the FLAG™-tag and is readily available as a hardcoded germline sequence in the mouse antibody repertoire.

The availability of the anti-FLAG™-M2 sequences may contribute to the wider use of this important research tool, as well as the development and engineering of better FLAG™-tag directed antibodies. This example illustrates that our MS-based sequencing protocol yields robust and reliable monoclonal antibody sequences. The protocol described here also formed the basis of a recent application where we sequenced an antibody directly from patient-derived serum, using a combination with top-down fragmentation of the isolated Fab fragment (44). The dual fragmentation strategy yields high-quality spectra suitable for *de novo* sequencing and may further contribute to the exciting prospect of a new advanced serology in which antibody sequences can be directly obtained from bodily fluids.

Data Availability

The raw LC-MS/MS data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD023419. The coordinates and reflection file with phases for the remodeled crystal structure of the anti-FLAG™-M2 Fab have been deposited in the Protein Data Bank under accession code 7BG1.

Acknowledgements

Herceptin was a kind gift from Roche (Penzberg, Germany). We would like to acknowledge support by Protein Metrics Inc. through access to Supernovo software and helpful discussion on *de novo* antibody sequencing. We would like to thank everyone in the Biomolecular Mass Spectrometry and Proteomics group at Utrecht University for support and helpful discussions. This research was funded by the Dutch Research Council NWO Gravitation 2013 BOO, Institute for Chemical Immunology (ICI; 024.002.009).

Author Contributions

WP and JS conceived of the project. WP carried out the MS experiments. WP and JS analyzed the MS data. MFP remodeled the crystal structure. MFP cloned and produced the synthetic recombinant antibody and carried out Western blotting. JS supervised the project. JS wrote the first draft and all authors contributed to preparing the final version of the manuscript.

Competing Interests

The authors declare no competing interests

Supplementary Information

Contents:

anti-FLAG-M2 MS-based sequences in FASTA format

Table S1. Coverage statistics for the Herceptin benchmark and anti-FLAG™-M2 MAb sequences.

Table S2. Model statistics for Fab crystal structure.

Table S3. Comparison of CDR sequences of FLAG™-tag binding MAbs.

Figure S1. Coverage maps of and anti-FLAG™-M2 MAb sequences.

Figure S2. Depth of coverage profiles for Herceptin and anti-FLAG™-M2 sequences

Figure S3. Sequence accuracy of Herceptin by fragmentation type and protease

Figure S4. Herceptin peptide length distribution by digestion time

Figure S5. Isoleucine/Leucine assignment at Heavy Chain position 51 of anti-FLAG™-M2.

Figure S6. Electrostatic surface potential of the anti-FLAG™-M2 paratope.

Figure S7. Western blot validation of synthetic recombinant anti-FLAG™-M2

anti-FLAG-M2 MS-based sequence (with L51I correction)

>anti-FLAG-M2_MS_HeavyChain
QVQLQQSAAELARPGASVKMSCKASGYSFTTYTIHWVKQRPGQGLEWIGYINPSSGYAAYNQN
FKDETTLTADPSSSTAYMELNSLTSEDSAVYYCAREKFYGYDYWGQGATLTVSSAKTTPPSVYP
LAPGSAAQTNSMVTLGCLVKGYFPEPVTVTWNSGSLSSGVHTFPAVLQSDLYTLSSSVTVPSSP
RPSETVTCNVAHPASSTKVDKKIVPRDCGCKPCICTVPEVSSVFIFPPKPKDVLTITLTPKVTCVV
VDISKDDPEVQFSWFVDDVEVHTAQTQPREEQFNSTFRSVSELPIMHQDWLNGKEFKCRVNS
AAFPAPIEKTISKTKGRPKAPQVYTIPPPKEQMAKDKVSLTCMITDFFPEDITVEWQWNGQPA
ENYKNTQPIMNTNGSYFVYSKLNVQKSNWEAGNTFTCSVLHEGLHNHHTEKSLSHSPGK
>anti-FLAG-M2_MS_LightChain
DVLMTQIPLSLPVSLGDQASISCRSSQSIVHRNGNTYLEWYLLKPGQSPKLLIYKVSNRFSGVPD
RFSGSGSGTDFTLKISRVEAEDLGVYYCFQGSHVPYTFGGGTKLEIRRADAAPTVSIFPPSSEQLT
SGGASVVCFLNNFYPKDINVKWKIDGSERQNGVLNSWTDQDSKDSTYSMSSTLTLTKDEYERH
NSYTCEATHKTSTSPIVKSFNRNEC

Table S1. Coverage statistics for the Herceptin benchmark and anti-FLAG™-M2 MAb sequences.

		Herceptin	anti-FLAG-M2
	total	4408	3371

Refinement statistics			
Resolution (Å)		42.52-1.86	
No. of reflections		39988	
PDB		**2G60 (old)**	**7BG1 (new)**
Total number of atoms		3518	3497
Average atomic displacement parameter (Å^2)		45.0	52.0
R_{work}/R_{free}		0.235/0.278	0.217/0.255
Bond length RMSZ		0.93	0.28
Bond angle RMSZ		0.96	0.51
Ramachandran favored/outliers (%)		93.0/1.0	97.57/0.24
Molprobity score		3.37	1.60
Clashscore		56	3.61
# peptide reads (Byonic score >=500)	stepped HCD	2686	1983
	EThcD	1722	1388
depth-of-coverage (median [range])	total	148 [8-394]	84 [0-382]
	CDRH1	163 [158-176]	32 [22-47]
	CDRH2	94 [88-103]	39 [36-43]
	CDRH3	42 [18-67]	66 [50-75]
	CDRL1	210 [208-218]	192 [144-207]
	CDRL2	74 [71-84]	46 [40-60]
	CDRL3	140 [130-143]	127 [109-131]

Table S2. Model statistics for Fab crystal structure.

Table S3. Comparison of CDR sequences from anti-FLAG™-M2 to other known FLAG™-tag binding MAbs (see refs 41-42).

MAb	Heavy Chain CDRH1	 CDRH2	 CDRH3
anti-FLAG-M2	GYSFTTYT----	LNPSSGYA	AREKFYGYDY
2H8	GFSLNTSGRS--	IYWDDDK	ARRMDY
EEh13.6	GDSLSSFNAGVN	HGAVM-STR	AKSTGRYDF
EEh14.3	GDSLSSYNAGVN	HMAGV-STR	VRNEWSGAF
EEf15.4	GFSIK--GANVN	HVRGDASTR	ADRKMYSFYSGGEA

MAb	Light Chain CDRL1	 CDRL2	 CDRL3
anti-FLAG-M2	QSIVHRNGNTY	KVS	FQGSHVPYT
2H8	QSLVHSNGNTY	KVS	SQSTHVPYT
EEh13.6	QSIVHSNGNTY	KVS	FQGSLVPPT
EEh14.3	QSIVHSNGNTY	KVS	FQGSLVPPT
EEf15.4	NARSGS	DGN	SAFDQTNKYVG

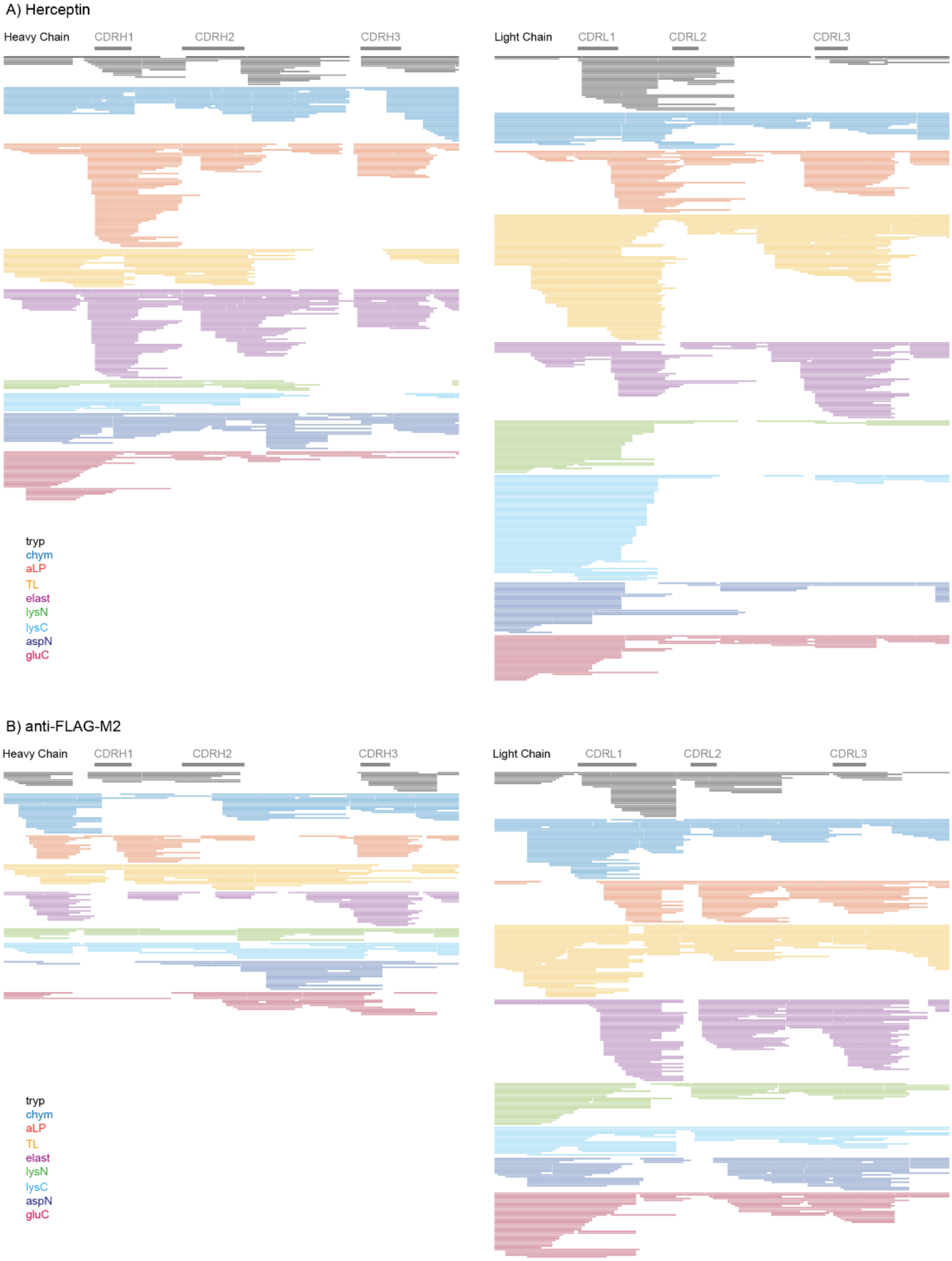


Figure S1. Coverage maps of Herceptin benchmark (A) and anti-FLAG™-M2 MAb (B) sequences. Peptides with Byonic scores of >=500 are shown.

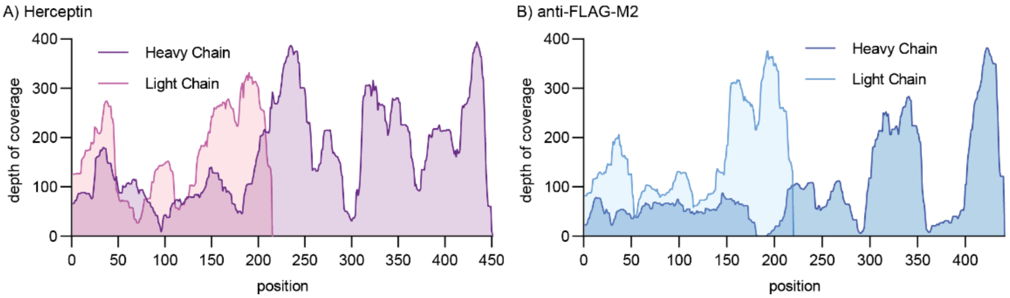
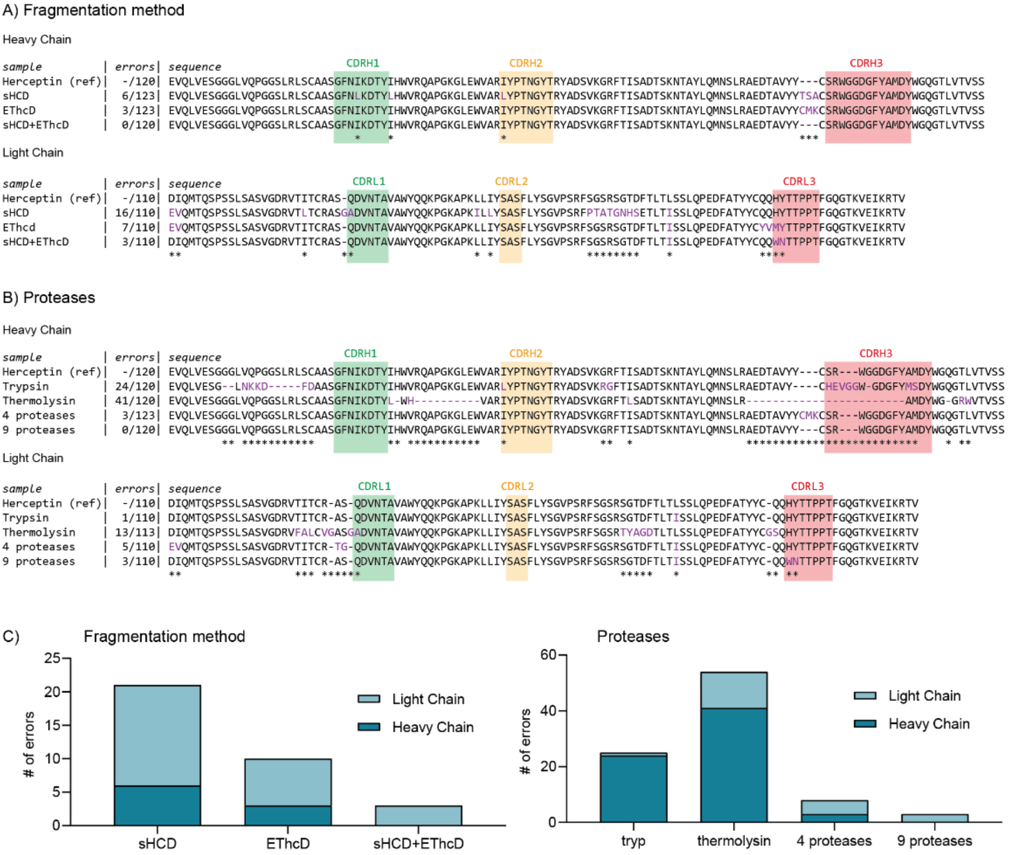


Figure S2. Depth of coverage profiles for Herceptin (A) and anti-FLAG™-M2 (B) sequences, based on peptides with Byonic score >=500, as in Figure S1.



A) Fragmentation method

Heavy Chain

```
                                                   CDRH1                    CDRH2                                             CDRH3
sample         | errors| sequence
Herceptin (ref)|  -/120| EVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYY---CSRWGGDGFYAMDYWGQGTLVTVSS
sHCD           |  6/123| EVQLVESGGGLVQPGGSLRLSCAASGFNLKDTYLHWVRQAPGKGLEWVARLYPTNGYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYYTSACSRWGGDGFYAMDYWGQGTLVTVSS
EThcD          |  3/123| EVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYYCMKCSRWGGDGFYAMDYWGQGTLVTVSS
sHCD+EThcD     |  0/120| EVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYY---CSRWGGDGFYAMDYWGQGTLVTVSS
                                                     *    *                *                                            ***
```

Light Chain

```
                                                     CDRL1                CDRL2                                      CDRL3
sample         | errors| sequence
Herceptin (ref)|  -/110| DIQMTQSPSSLSASVGDRVTITCRAS-QDVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPSRFSGSRSGTDFTLTLSSLQPEDFATYYCQQHYTTPPTFGQGTKVEIKRTV
sHCD           | 16/110| EVQMTQSPSSLSASVGDRVTLTCRASGADVNTAVAWYQQKPGKAPKILLYSASFLYSGVPSRFPTATGNHSETLTISSLQPEDFATYYCQQHYTTPPTFGQGTKVEIKRTV
EThcd          |  7/110| EVQMTQSPSSLSASVGDRVTITCRAS-QDVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPSRFSGSRSGTDFTLTISSLQPEDFATYYCYVMYTTPPTFGQGTKVEIKRTV
sHCD+EThcD     |  3/110| DIQMTQSPSSLSASVGDRVTITCRAS-QDVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPSRFSGSRSGTDFTLTISSLQPEDFATYYCQQWNTTPPTFGQGTKVEIKRTV
                         **                  *     **                  * *              ********    *             ****
```

B) Proteases

Heavy Chain

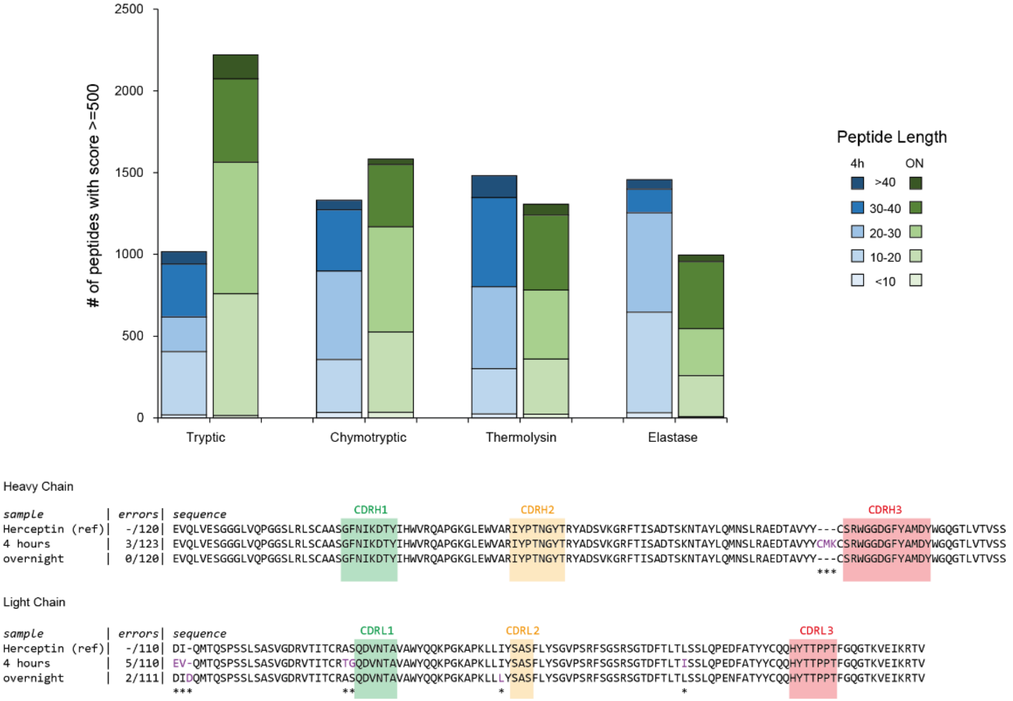
```
                                                   CDRH1                    CDRH2                                                CDRH3
sample         | errors| sequence
Herceptin (ref)|  -/120| EVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYY---CSR---WGGDGFYAMDYWGQGTLVTVSS
Trypsin        | 24/120| EVQLVESG--LNKKD-----FDAASGFNIKDTYIHWVRQAPGKGLEWVARLYPTNGYTRYADSVKRGFTISADTSKNTAYLQMNSLRAEDTAVY----CHEVGGW-GDGFYMSDYWGQGTLVTVSS
Thermolysin    | 41/120| EVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYL-WH----------VARIYPTNGYTRYADSVKGRFTLSADTSKNTAYLQMNSLR------------------------AMDYWG-GRWVTVSS
4 proteases    |  3/123| EVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYYCMKCSR---WGGDGFYAMDYWGQGTLVTVSS
9 proteases    |  0/120| EVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYY---CSR---WGGDGFYAMDYWGQGTLVTVSS
                                 ** ***********           ** ***********   *              **  *                 **************************    * **
```

Light Chain

```
                                                      CDRL1                CDRL2                                       CDRL3
sample         | errors| sequence
Herceptin (ref)|  -/110| DIQMTQSPSSLSASVGDRVTITCR-AS-QDVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPSRFSGSRSGTDFTLTLSSLQPEDFATYYC-QQHYTTPPTFGQGTKVEIKRTV
Trypsin        |  1/110| DIQMTQSPSSLSASVGDRVTITCR-AS-QDVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPSRFSGSRSGTDFTLTISSLQPEDFATYYC-QQHYTTPPTFGQGTKVEIKRTV
Thermolysin    | 13/113| DIQMTQSPSSLSASVGDRVFALCVGASGADVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPSRFSGSRTYAGDTLTLSSLQPEDFATYYCGSQHYTTPPTFGQGTKVEIKRTV
4 proteases    |  5/110| EVQMTQSPSSLSASVGDRVTITCR-TG-QDVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPSRFSGSRSGTDFTLTISSLQPEDFATYYC-QQHYTTPPTFGQGTKVEIKRTV
9 proteases    |  3/110| DIQMTQSPSSLSASVGDRVTITCR-AS-QDVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPSRFSGSRSGTDFTLTISSLQPEDFATYYC-QQWNTTPPTFGQGTKVEIKRTV
                         **                 *** ******                                       *****   *             ** **
```

C)

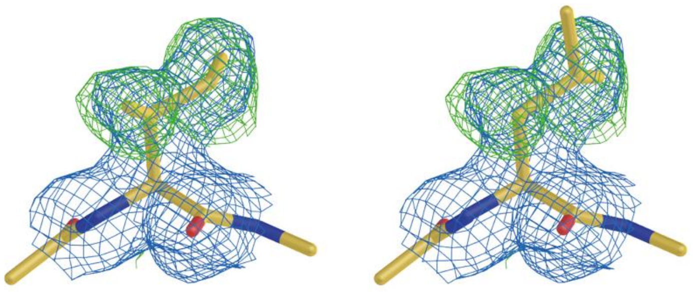
Figure S3. Sequence accuracy of Herceptin by fragmentation type (A) and use of proteases (B). Supernovo analysis was performed using only the specified fragmentation type or proteases as input data. Resulting sequences were aligned to the Herceptin reference sequences to count the number of errors. Every substitution, insertion or deletion was counted as an error as listed before the output sequence; *i.e.* all positions labeled in purple and marked with an asterisks are counted. The '4 proteases' dataset consists of trypsin, chymotrypsin, thermolysin and elastase. The total number of errors is shown for fragmentation strategy and protease datasets in panel C.



```
sample          | errors| sequence
Herceptin (ref)|  -/120| EVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYY---CSRWGGDGFYAMDYWGQGTLVTVSS
4 hours        |  3/123| EVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYYCMKCSRWGGDGFYAMDYWGQGTLVTVSS
overnight      |  0/120| EVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYY---CSRWGGDGFYAMDYWGQGTLVTVSS
                                                                                                                        ***
```

```
sample          | errors| sequence
Herceptin (ref)|  -/110| DI-QMTQSPSSLSASVGDRVTITCRASQDVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPSRFSGSRSGTDFTLTLSSLQPEDFATYYCQQHYTTPPTFGQGTKVEIKRTV
4 hours        |  5/110| EV-QMTQSPSSLSASVGDRVTITCRTGQDVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPSRFSGSRSGTDFTLTISSLQPEDFATYYCQQHYTTPPTFGQGTKVEIKRTV
overnight      |  2/111| DIDQMTQSPSSLSASVGDRVTITCRASQDVNTAVAWYQQKPGKAPKLLLYSASFLYSGVPSRFSGSRSGTDFTLTLSSLQPENFATYYCQQHYTTPPTFGQGTKVEIKRTV
                         ***                      **                     *                          *
```

Figure S4. Peptide length depending on digestion time. Datasets of four proteases were combined for Supernovo analysis. Peptide length distribution is based on peptides with score >=500. Resulting sequences from Supernovo were aligned to the Herceptin reference sequences to count the number of errors. Every substitution, insertion or deletion was counted as an error as listed before the output sequence; *i.e.* all positions labeled in purple and marked with an asterisks are counted.


Isoleucine 51

Leucine 51

Figure S5. Isoleucine/Leucine assignment at Heavy Chain position 51 of anti-FLAG™-M2. (left panel) Electron density around isoleucine 51 at a contour level of 1.0 RMSD in blue and simulated annealing omit map density of the $C_{\gamma1}$, $C_{\gamma2}$ and C_{δ} atoms of this residue at a contour level of 2.5 R.M.S.D. in green. (right panel) A leucine instead of an isoleucine in this location has a poor fit to both maps.

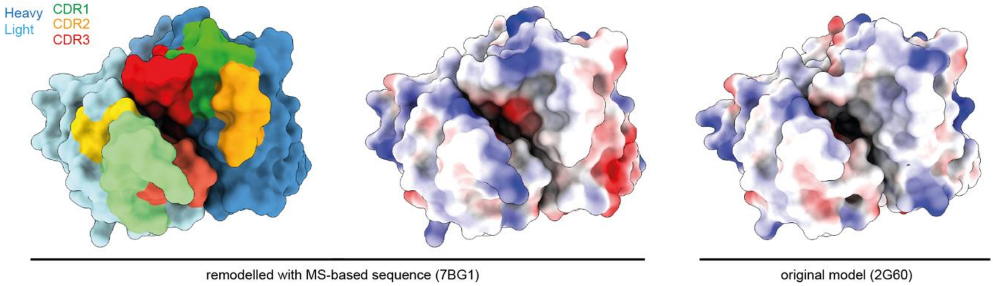


Figure S6. Electrostatic surface potential of the anti-FLAG™-M2 paratope. The revised crystal structure based on the MS-derived sequence (PDB ID: 7BG1) is shown alongside the original model (PDB ID: 2G60). The electrostatic surface was calculated with the default *coulombic* command in ChimeraX.

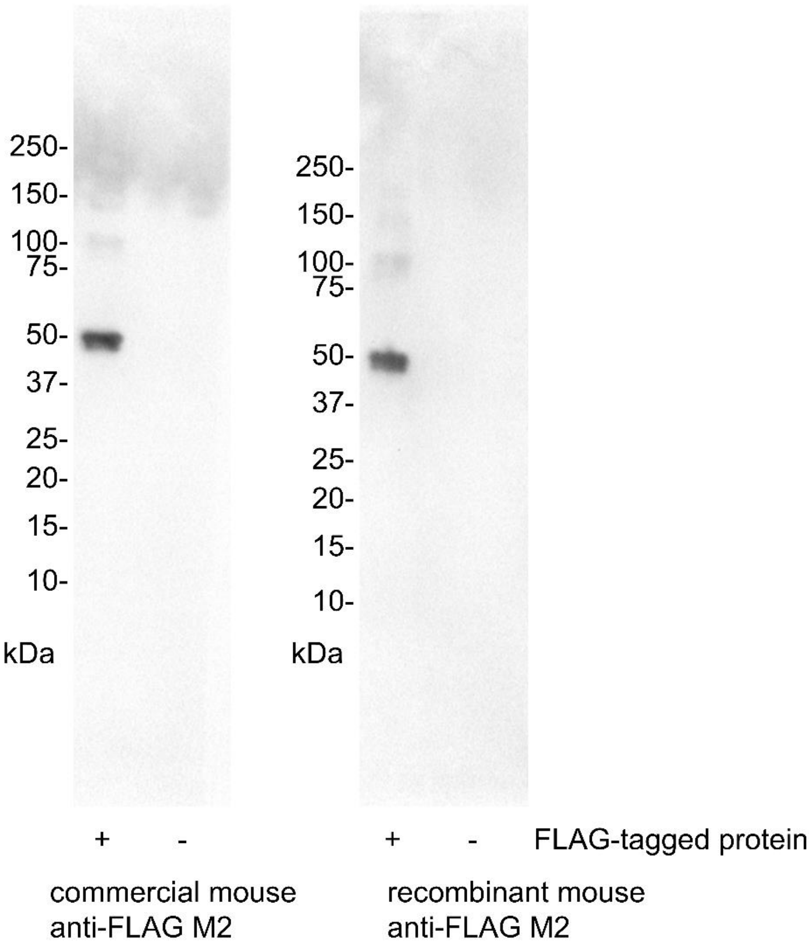


Figure S7. Western blot validation of synthetic recombinant anti-FLAG™-M2 compared to the originally sequenced sample. Same Western blot as shown in Figure 3C, showing complete lanes with marker positions.

References

1. Tonegawa, S., Somatic generation of antibody diversity. *Nature* **1983,** 302, (5909), 575-581.

2. Watson, C. T.; Glanville, J.; Marasco, W. A., The individual and population genetics of antibody immunity. *Trends in immunology* **2017,** 38, (7), 459-470.

3. Carter, P. J.; Lazar, G. A., Next generation antibody drugs: pursuit of the'high-hanging fruit'. *Nature Reviews Drug Discovery* **2018,** 17, (3), 197.

4. Grilo, A. L.; Mantalaris, A., The increasingly human and profitable monoclonal antibody market. *Trends in biotechnology* **2019,** 37, (1), 9-16.

5. Baker, M., Blame it on the antibodies. *Nature* **2015,** 521, (7552), 274.

6. Uhlen, M.; Bandrowski, A.; Carr, S.; Edwards, A.; Ellenberg, J.; Lundberg, E.; Rimm, D. L.; Rodriguez, H.; Hiltke, T.; Snyder, M., A proposal for validation of antibodies. *Nature methods* **2016,** 13, (10), 823-827.

7. Fischer, N. In *Sequencing antibody repertoires: the next generation*, MAbs, 2011; Taylor & Francis: 2011; pp 17-20.

8. Georgiou, G.; Ippolito, G. C.; Beausang, J.; Busse, C. E.; Wardemann, H.; Quake, S. R., The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature biotechnology* **2014,** 32, (2), 158-168.

9. Robinson, W. H., Sequencing the functional antibody repertoire—diagnostic and therapeutic discovery. *Nature Reviews Rheumatology* **2015,** 11, (3), 171.

10. Boutz, D. R.; Horton, A. P.; Wine, Y.; Lavinder, J. J.; Georgiou, G.; Marcotte, E. M., Proteomic identification of monoclonal antibodies from serum. *Analytical chemistry* **2014,** 86, (10), 4758-4766.

11. Castellana, N. E.; McCutcheon, K.; Pham, V. C.; Harden, K.; Nguyen, A.; Young, J.; Adams, C.; Schroeder, K.; Arnott, D.; Bafna, V., Resurrection of a clinical antibody: Template proteogenomic *de novo* proteomic sequencing and reverse engineering of an anti-lymphotoxin-α antibody. *Proteomics* **2011,** 11, (3), 395-405.

12. Chen, J.; Zheng, Q.; Hammers, C. M.; Ellebrecht, C. T.; Mukherjee, E. M.; Tang, H.-Y.; Lin, C.; Yuan, H.; Pan, M.; Langenhan, J., Proteomic analysis of pemphigus autoantibodies indicates a larger, more diverse, and more dynamic repertoire than determined by B cell genetics. *Cell reports* **2017,** 18, (1), 237-247.

13. Cheung, W. C.; Beausoleil, S. A.; Zhang, X.; Sato, S.; Schieferl, S. M.; Wieler, J. S.; Beaudet, J. G.; Ramenani, R. K.; Popova, L.; Comb, M. J., A proteomics approach for the identification and cloning of monoclonal antibodies from serum. *Nature biotechnology* **2012,** 30, (5), 447-452.

14. Guthals, A.; Gan, Y.; Murray, L.; Chen, Y.; Stinson, J.; Nakamura, G.; Lill, J. R.; Sandoval, W.; Bandeira, N., *De novo* MS/MS sequencing of native human antibodies. *Journal of proteome research* **2017,** 16, (1), 45-54.

15. Lee, J.; Boutz, D. R.; Chromikova, V.; Joyce, M. G.; Vollmers, C.; Leung, K.; Horton, A. P.; DeKosky, B. J.; Lee, C.-H.; Lavinder, J. J., Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. *Nature medicine* **2016,** 22, (12), 1456-1464.

16. Lee, J.; Paparoditis, P.; Horton, A. P.; Frühwirth, A.; McDaniel, J. R.; Jung, J.; Boutz, D. R.; Hussein, D. A.; Tanno, Y.; Pappas, L., Persistent antibody clonotypes dominate the serum response to influenza over multiple years and repeated vaccinations. *Cell host & microbe* **2019,** 25, (3), 367-376. e5.

17. Lindesmith, L. C.; McDaniel, J. R.; Changela, A.; Verardi, R.; Kerr, S. A.; Costantini, V.; Brewer-Jensen, P. D.; Mallory, M. L.; Voss, W. N.; Boutz, D. R., Sera antibody repertoire analyses reveal mechanisms of broad and pandemic strain neutralizing responses after human norovirus vaccination. *Immunity* **2019,** 50, (6), 1530-1541. e8.

18. Bandeira, N.; Pham, V.; Pevzner, P.; Arnott, D.; Lill, J. R., Automated *de novo* protein sequencing of monoclonal antibodies. *Nature biotechnology* **2008,** 26, (12), 1336-1338.

19. Rickert, K. W.; Grinberg, L.; Woods, R. M.; Wilson, S.; Bowen, M. A.; Baca, M. In *Combining phage display with de novo protein sequencing for reverse engineering of monoclonal antibodies*, MAbs, 2016; Taylor & Francis: 2016; pp 501-512.

20. Savidor, A.; Barzilay, R.; Elinger, D.; Yarden, Y.; Lindzen, M.; Gabashvili, A.; Tal, O. A.; Levin, Y., Database-Independent Protein Sequencing (DiPS) Enables Full-Length *de novo* Protein and Antibody Sequence Determination. *Molecular & Cellular Proteomics* **2017,** 16, (6), 1151-1161.

21. Sen, K. I.; Tang, W. H.; Nayak, S.; Kil, Y. J.; Bern, M.; Ozoglu, B.; Ueberheide, B.; Davis, D.; Becker, C., Automated antibody *de novo* sequencing and its utility in biopharmaceutical discovery. *Journal of The American Society for Mass Spectrometry* **2017,** 28, (5), 803-810.

22. Sousa, E.; Olland, S.; Shih, H. H.; Marquette, K.; Martone, R.; Lu, Z.; Paulsen, J.; Gill, D.; He, T., Primary sequence determination of a monoclonal antibody against α-synuclein using a novel mass spectrometry-based approach. *International Journal of Mass Spectrometry* **2012,** 312, 61-69.

23. Tran, N. H.; Rahman, M. Z.; He, L.; Xin, L.; Shan, B.; Li, M., Complete *de novo* assembly of monoclonal antibody sequences. *Scientific reports* **2016,** 6, (1), 1-10.

24. Brizzard, B. L.; Chubet, R. G.; Vizard, D., Immunoaffinity purification of FLAG epitope-tagged bacterial alkaline phosphatase using a novel monoclonal antibody and peptide elution. *Biotechniques* **1994,** 16, (4), 730-735.

25. Sigma-Aldrich anti-FLAG-M2 F1804 product page. https://www.sigmaaldrich.com/catalog/product/sigma/f1804?lang=en&region=NL (05-01-2021),

26. Ehrenmann, F.; Kaas, Q.; Lefranc, M.-P., IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic acids research* **2010,** 38, (suppl_1), D301-D307.

27. Ehrenmann, F.; Lefranc, M.-P., IMGT/DomainGapAlign: IMGT standardized analysis of amino acid sequences of variable, constant, and groove domains (IG, TR, MH, IgSF, MhSF). *Cold Spring Harbor Protocols* **2011,** 2011, (6), pdb. prot5636.

28. Roosild, T. P.; Castronovo, S.; Choe, S., Structure of anti-FLAG M2 Fab domain and its use in the stabilization of engineered membrane proteins. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* **2006,** 62, (9), 835-839.

29. Emsley, P.; Cowtan, K., Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D: Biological Crystallography* **2004,** 60, (12), 2126-2132.

30. Afonine, P. V.; Grosse-Kunstleve, R. W.; Echols, N.; Headd, J. J.; Moriarty, N. W.; Mustyakimov, M.; Terwilliger, T. C.; Urzhumtsev, A.; Zwart, P. H.; Adams, P. D., Towards automated crystallographic structure refinement with phenix. refine. *Acta Crystallographica Section D: Biological Crystallography* **2012,** 68, (4), 352-367.

31. Chen, V. B.; Arendall, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C., MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography* **2010,** 66, (1), 12-21.

32. Fuglsang, A., Codon optimizer: a freeware tool for codon optimization. *Protein expression and purification* **2003,** 31, (2), 247-249.

33. García-Nafría, J.; Watson, J. F.; Greger, I. H., IVA cloning: a single-tube universal cloning system exploiting bacterial in vivo assembly. *Scientific reports* **2016,** 6, 27459.

34. Paizs, B.; Suhai, S., Fragmentation pathways of protonated peptides. *Mass spectrometry reviews* **2005,** 24, (4), 508-548.

35. Diedrich, J. K.; Pinto, A. F.; Yates III, J. R., Energy dependence of HCD on peptide fragmentation: stepped collisional energy finds the sweet spot. *Journal of the American Society for Mass Spectrometry* **2013,** 24, (11), 1690-1699.

36. Frese, C. K.; Altelaar, A. M.; van den Toorn, H.; Nolting, D.; Griep-Raming, J.; Heck, A. J.; Mohammed, S., Toward full peptide sequence coverage by dual fragmentation

combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. *Analytical chemistry* **2012,** 84, (22), 9668-9673.

37. Frese, C. K.; Zhou, H.; Taus, T.; Altelaar, A. M.; Mechtler, K.; Heck, A. J.; Mohammed, S., Unambiguous phosphosite localization using electron-transfer/higher-energy collision dissociation (EThcD). *Journal of proteome research* **2013,** 12, (3), 1520-1525.

38. Carter, P.; Presta, L.; Gorman, C. M.; Ridgway, J.; Henner, D.; Wong, W.; Rowland, A. M.; Kotts, C.; Carver, M. E.; Shepard, H. M., Humanization of an anti-p185HER2 antibody for human cancer therapy. *Proceedings of the National Academy of Sciences* **1992,** 89, (10), 4285-4289.

39. Slamon, D. J.; Leyland-Jones, B.; Shak, S.; Fuchs, H.; Paton, V.; Bajamonde, A.; Fleming, T.; Eiermann, W.; Wolter, J.; Pegram, M., Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *New England journal of medicine* **2001,** 344, (11), 783-792.

40. Einhauer, A.; Jungbauer, A., The FLAG™ peptide, a versatile fusion tag for the purification of recombinant proteins. *Journal of biochemical and biophysical methods* **2001,** 49, (1-3), 455-465.

41. Entzminger, K. C.; Hyun, J.-m.; Pantazes, R. J.; Patterson-Orazem, A. C.; Qerqez, A. N.; Frye, Z. P.; Hughes, R. A.; Ellington, A. D.; Lieberman, R. L.; Maranas, C. D., *De novo* design of antibody complementarity determining regions binding a FLAG tetra-peptide. *Scientific reports* **2017,** 7, (1), 1-11.

42. Ikeda, K.; Koga, T.; Sasaki, F.; Ueno, A.; Saeki, K.; Okuno, T.; Yokomizo, T., Generation and characterization of a human-mouse chimeric high-affinity antibody that detects the DYKDDDDK FLAG peptide. *Biochemical and Biophysical Research Communications* **2017,** 486, (4), 1077-1082.

43. Lima, W. C.; Gasteiger, E.; Marcatili, P.; Duek, P.; Bairoch, A.; Cosson, P., The ABCD database: a repository for chemically defined antibodies. *Nucleic acids research* **2020,** 48, (D1), D261-D264.

44. Bondt, A.; Hoek, M.; Tamara, S.; de Graaf, B.; Peng, W.; Schulte, D.; den Boer, M. A.; Greisch, J.-F.; Varkila, M. R.; Snijder, J., Human Plasma IgG1 Repertoires are Simple, Unique, and Dynamic. *SSRN* **2020** https://dx.doi.org/10.2139/ssrn.3749694.

Chapter 3

Template-based assembly of proteomic short reads for *de novo* antibody sequencing and repertoire profiling

Authors:

Douwe Schulte[1], Weiwei Peng[1], Joost Snijder[1]*

* corresponding author: j.snijder@uu.nl

Affiliations:

[1] Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute of Pharmaceutical Sciences, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

Published as:

https://doi.org/10.1021/acs.analchem.2c01300

Abstract:

Antibodies can target a vast molecular diversity of antigens. This is achieved by generating a complementary diversity of antibody sequences though somatic recombination and hypermutation. A full understanding of the antibody repertoire in health and disease therefore requires dedicated *de novo* sequencing methods. Next generation cDNA sequencing methods have laid the foundation of our current understanding of the antibody repertoire, but these methods share one major limitation in that they target the antibody-producing B-cells, rather than the functional secreted product in bodily fluids. Mass spectrometry-based methods offer an opportunity to bridge this gap between antibody repertoire profiling and bulk serological assays, as they can access antibody sequence information straight from the secreted polypeptide products. In a step to meeting the challenge of MS-based antibody sequencing, we present a fast and simple software tool (Stitch) to map proteomic short reads to user-defined templates with dedicated features for both monoclonal antibody sequencing and profiling of polyclonal antibody repertoires. We demonstrate the use of Stitch by fully reconstructing 2 monoclonal antibody sequences with >98% accuracy (including I/L assignment); sequencing a Fab from patient serum isolated by reversed-phase LC fractionation against a high background of homologous antibody sequences; sequencing antibody light chains from urine of multiple-myeloma patients; and profiling the IgG repertoire in sera from patients hospitalized with COVID-19. We demonstrate that Stitch assembles a comprehensive overview of the antibody sequences that are represented in the dataset and provides an important first step towards analyzing polyclonal antibodies and repertoire profiling.

Chapter 3

Introduction

Antibodies bind a wide variety of antigens with high affinity and specificity, playing a major role in the adaptive immune response to infections, but can also target self-antigens to mediate autoimmune diseases [1-4]. Antibodies can mediate immunity by blocking essential steps of a pathogen's replication cycle (*e.g.* receptor binding and cell entry), triggering the complement system, or activating a specific cell-mediated immune response known as antibody-dependent cellular cytotoxicity. Antibodies elicited in response to infection may persist in circulation for several months and regenerate quickly by subsequent exposures to the same antigen through a memory B-cell response [5,6]. All this has made antibodies into popular serological markers of pathogen exposure and vaccine efficacy, therapeutic leads for the treatment of cancer and infectious disease, and invaluable research tools for specific labelling and detection of molecular targets.

The large diversity of antigens that antibodies can target comes from a complementary diversity of available antibody sequences and compositions [2,7-11]. Antibodies of most classes consist of a combination of two unique, paired, homologous polypeptides: the heavy chain and the light chain, each consisting of a series of the characteristic Immunoglobulin (Ig) domains. Both chains can be subdivided into a variable region, involved in antigen binding, and a constant region, which plays a structural role in oligomerization, complement activation and receptor-binding on immune cells. Disulfide bonds covalently link the heavy and light chains, and two copies of this covalent heterodimer are in turn disulfide-linked on the heavy chains to form the characteristic Y-shaped antibody (consisting of two heavy chains and two light chains). Up to four separate gene segments encode each chain by somatic recombination of the Variable, Diversity (only in heavy chain), Joining, and Constant segments, known as V-(D)-J-C recombination. Every unique B-cell clone can draw from many unique alleles for each segment (up to hundreds for the V-segment) creating in the order of 10^5 possible unique V-(D)-J permutations. The number of possible unique pairings between the heavy and light chains further adds to the available variety of the fully assembled antibody.

The V-(D)-J segments collectively make up the variable domain of the heavy and light chains, each containing three so-called complementarity-determining regions (CDRs), which are directly involved in antigen binding and ultimately responsible for binding affinity and specificity [2,9,11]. CDR1 and CDR2 lie fully encoded within the V-segment, while CDR3 lies encoded in the V-(D)-J junction and is therefore inherently more variable. After initial activation of naïve B-cells, each with their specific V-(D)-J recombination, individual clones undergo a process of somatic hypermutation, in which additional sequence variation is introduced in the variable domain of the antibody, especially in the CDRs, resulting in affinity maturation of the coded antibody through a process of natural selection for the strongest antigen binders. This combination of somatic recombination, hypermutation and heavy-light chain pairing thus creates a vast repertoire of mature antibody sequences.

This large sequence diversity within and between individuals requires dedicated *de novo* sequencing of antibodies to uncover the structural basis of antigen binding and to map out the antibody repertoire in both health and disease [12,13]. Established methods for *de novo* antibody sequencing rely on cloning and sequencing of the coding mRNAs from single B-cells, recovering up to hundreds of paired heavy and light chain sequences in a single study. While these methods have laid the foundation for our current understanding of the antibody repertoire, they share one major limitation in that sequencing requires recovery of the antibody-producing B-cells. While immature and memory B-cells do present antibodies on their surface, the major functional contribution of antibodies to adaptive immunity comes from the vast amounts that are secreted in blood and other bodily fluids. In other words, most antibodies are physically disconnected from their producing B-cell and there is no straightforward quantitative relation between serological test results (*e.g.* binding and neutralization titers) and antibody repertoires derived from single B-cell sequencing results. Expansion of memory B-cells into a secreting B-cell population, production and secretion levels of antibodies, and the lifetimes of both B-cells and secreted antibody product in bodily fluid may vary by orders of magnitude between unique B-cell clones. To address this caveat, methods to sequence and profile the functional antibody repertoire on the level of the secreted product are necessary.

Mass spectrometry-based methods are particularly powerful for direct proteomic sequencing and profiling of secreted antibody products. This is illustrated by several recent proteogenomics studies in which targeted single B-cell sequencing data is used to generate a custom database for a conventional proteomics-type LC-MS/MS-based database search to quantitatively profile the abundance of sequenced clones in serum [14-21]. These methods still rely on complementary cDNA sequencing of the antibody producing B-cell, unlike true *de novo* protein sequencing methods. Direct protein sequencing methods have focused especially on *de novo* sequencing of monoclonal antibodies, based on bottom-up analysis of digested peptides [22-31]. With the aid of specialized software packages like DiPS, Supernovo, or ALPS, full and accurate sequences of the heavy and light chains can be reconstructed with the MS/MS spectra of the digested peptides [28,29,31]. The use of multiple complementary proteases and novel hybrid fragmentation techniques provides large benefits in sequence coverage and accuracy in these methods [26]. The obtained sequences are complete and accurate enough to reverse-engineer functional synthetic recombinant antibodies, for instance of monoclonal antibodies from lost hybridoma cell lines. Recently we also demonstrated complete sequencing of a monoclonal antibody isolated from patient serum by reversed-phase LC fractionation and integrated bottom-up and top-down analysis [32]. Plasma proteomics methods to profile polyclonal IgG mixtures and other heterogeneous variant proteins based on *de novo* methods (SpotLight and LAX) have also recently been described [33-36].

Characterization of polyclonal mixtures and a move toward full profiling of the circulating antibody repertoire remain major outstanding challenges for MS-based antibody sequencing. In a step to meeting these challenges, we present a fast and simple software tool (Stitch) to map proteomic short reads to user-defined templates with dedicated features for both monoclonal antibody sequencing and profiling of polyclonal antibody repertoires. We demonstrate the use of Stitch by fully reconstructing 2 monoclonal antibody sequences with >98% accuracy (including I/L assignment); sequencing a Fab from patient serum isolated by reversed-phase LC fractionation against a high background of homologous antibody sequences; sequencing antibody light chains from urine of multiple-myeloma patients; and profiling the IgG repertoire in sera from patients hospitalized with COVID-19.

Methods

Monoclonal antibodies and COVID-19 serum IgG – sample preparation

Herceptin and anti-FLAG-M2 were obtained as described in reference [26], F59 was purified from patient serum as described in reference [32]. Convalescent serum from COVID-19 patients were obtained under the Radboud UMC Biobank protocol; IgG was purified with Protein G affinity resin (Millipore). Samples were denatured in 2% sodium deoxycholate (SDC), 200 mM Tris–HCl, 10 mM Tris(2-carboxyethyl)phosphine (TCEP), pH 8.0 at 95 °C for 10 min, followed by 30 min incubation at 37 °C for reduction. The samples were then alkylated by adding iodoacetic acid to a final concentration of 40 mM and incubated in the dark at room temperature for 45 min. For herceptin and anti-FLAG-M2 a 3 μg sample was then digested by one of the following proteases: trypsin (Promega), chymotrypsin (Roche), lysN (Thermo Fisher Scientific), lysC (FUJIFILM Wako Pure Chemical Corporation), gluC (Roche), aspN (Roche), aLP (Sigma-Aldrich), thermolysin (Promega), and elastase (Sigma-Aldrich) in a 1:50 ratio (w/w) in a total volume of 100 μL of 50 mM ammonium bicarbonate at 37 °C for 4 h. After digestion, SDC was removed by adding 2 μL of formic acid (FA) and centrifugation at 14 000g for 20 min. Following centrifugation, the supernatant containing the peptides was collected for desalting on a 30 μm Oasis HLB 96-well plate (Waters). The F59 monoclonal isolated from patient serum was digested in parallel by four proteases: trypsin, chymotrypsin, thermolysin and pepsin. Digestion with trypsin, chymotrypsin and thermolysin was done with 0.1 μg protease following the SDC protocol described above. For pepsin digestion, a urea buffer was added to a total volume of 80 μL, 2M Urea, 10 mM TCEP. Sample was denatured for 10 min at 95 °C followed by reduction for 20 min at 37 °C. Next, iodoacetic acid was added to a final concentration of 40 mM and incubated in the dark for 45 min at room temperature for alkylation of free cysteines. For pepsin digestion 1 M HCl was added to a final concentration of 0.04 M. Digestion was carried out overnight with 0.1 μg of protease, after which the entire digest was collected for desalting with the Oasis HLB 96-well plate. The Oasis HLB sorbent was activated with 100% acetonitrile and subsequently equilibrated with 10% formic acid in water. Next, peptides were bound to the sorbent, washed twice with 10% formic acid in water, and eluted with 100 μL of 50% acetonitrile/5% formic acid in water (v/v). The eluted peptides were vacuum-dried and reconstituted in 100 μL of 2% FA.

LC-MS/MS

The digested peptides were separated by online reversed phase chromatography on an Agilent 1290 UHPLC coupled to a Thermo Scientific Orbitrap Fusion mass spectrometer. Peptides were separated using a Poroshell 120 EC-C18 2.7-Micron analytical column (ZORBAX Chromatographic Packing, Agilent) and a C18 PepMap 100 trap column (5 mm x 300 µm, 5 µm, Thermo Fisher Scientific). Samples were eluted over a 90 min gradient from 0 to 35% acetonitrile at a flow rate of 0.3 µL/min. Peptides were analyzed with a resolution setting of 60 000 in MS1. MS1 scans were obtained with a standard automatic gain control (AGC) target, a maximum injection time of 50 ms, and a scan range of 350–2000. The precursors were selected with a 3 m/z window and fragmented by stepped high-energy collision dissociation (HCD) as well as electron-transfer high-energy collision dissociation (EThcD). The stepped HCD fragmentation included steps of 25, 35, and 50% normalized collision energies (NCE). EThcD fragmentation was performed with calibrated charge-dependent electron-transfer dissociation (ETD) parameters and 27% NCE supplemental activation. For both fragmentation types, MS2 scans were acquired at a 30 000 resolution, a 4e5 AGC target, a 250 ms maximum injection time, and a scan range of 120–3500.

Peptide sequencing from MS/MS spectra

MS/MS spectra were used to determine *de novo* peptide sequences using PEAKS Studio X (version 10.5). We used a tolerance of 20 ppm and 0.02 Da for MS1 and MS2, respectively. Carboxymethylation was set as fixed modification of cysteine, and variable modification of peptide N-termini and lysine. Oxidation of methionine and tryptophan, pyroglutamic acid modification of N-terminal glutamic acid and glutamine were set as additional variable modifications.

Stitch code and analysis parameters

Stitch was written in C# with compatibility for Windows, Apple and Linux. The source code is available at GitHub (https://github.com/snijderlab/stitch), along with a complete description of all functions in the manual. Stitch is run through a terminal, using a .txt batch file that specifies the input data, run parameters, and filenames with output location. Stitch can read FASTA, PEAKS or Novor.Cloud data as input. The template sequences are derived from IMGT (at https://www.imgt.org/vquest/refseqh.html), grouped by V/J/C segment, and filtered to remove all duplicate sequences at the amino acid level and generate a list of non-redundant unique sequences. These template sequences are included for human, mouse, bovine, dog, and rabbit in the default installation and the retrieval/cleanup procedure for additional species is included. Reads are matched using local Smith Waterman alignment with a user-defined alphabet/scoring matrix; default based on BLOSUM62. Input reads are matched to a template when they exceed the user-defined score cutoff, which is adapted to the square root of the length of the matched template sequence. When this metadata is available for the input reads, the consensus sequences (and sequence logos) in Stitch are weighted by the local and global quality scores scaled from 0-1, with 1 for the best scoring reads/positions, as well as the MS1 peak area scaled logarithmically to a weight between 1 and 2, using 2-1/log10(Area). When specified in the batch file, Stitch can recombine the top-*N* scoring (V/J/C) segments in a user-defined order. A gap (*) can be defined between recombined segments, which will extend the templates with twenty X characters at the junction to look for overhanging reads. This aids reconstruction of CDRH3 at the V-J junction. The potential overlap between these recombined segments is determined within a sliding window of max 40 amino acids, scored using the same scoring matrix/alphabet as the template matching step. If no positive score is found in the sliding window, a single gap is placed at the junction of the extended template sequences. Stitch will then perform a second template matching step on the recombined segments, in which the non-selected templates can be added as decoys. Stitch generates an interactive HTML report, with a summary of the results on the home page, and separate pages with detailed results for every (recombined) template that provides the consensus sequence, sequence logo, depth of coverage, and alignment overview of all reads on the template._All Stitch analysis results of the current work are provided as supplementary data. The batch file parameters used for every analysis are outputted in the results file. Briefly, we typically use PEAKS ALC cutoff of ≥85, local alignment cutoff score of ≥8 and adjust these to the quality and complexity of the input data.

Results

The experimental *de novo* antibody sequence reads obtained from a typical LC-MS/MS experiment are 5-40 amino acids in length. Although these reads are relatively short for completely *de novo* assembly, the rates of somatic hypermutation are typically low enough (1-10%) that the translated germline sequences contained in the IMGT are of sufficient homology to accurately place all peptide reads in the correct framework of the heavy and light chains. Based on this notion, we developed Stitch to perform template-based assembly of antibody-derived *de novo* sequence reads using local Smith-Waterman alignment [37]. Although the program can also perform this task on any user-defined set of templates, using plain FASTA sequences as input, we developed dedicated procedures for both mono- and polyclonal antibody sequencing using *de novo* reads from PEAKS or Novor.Cloud as input [38]. Post-translational modifications can be accommodated in the *de novo* sequence reads and they are handled by scoring the peptide reads using the corresponding unmodified amino acids. With input from PEAKS or Novor.Cloud the program can use metadata of individual reads as filtering criteria and determine weighted consensus sequences from overlapping reads, based on global and local quality scores, as well as MS1 peak area (when available). As output Stitch generates an interactive HTML report that contains a quantitative overview of matched reads, alignment scores and a combined peak area for every template. In addition, it generates the final consensus sequences for all matched templates together with a sequence logo, depth of coverage profiles, and a detailed overview of all assembled reads in the context of their templates (see Figure 1). Finally, the output report also contains a complete overview of all reads assigned to the CDRs.

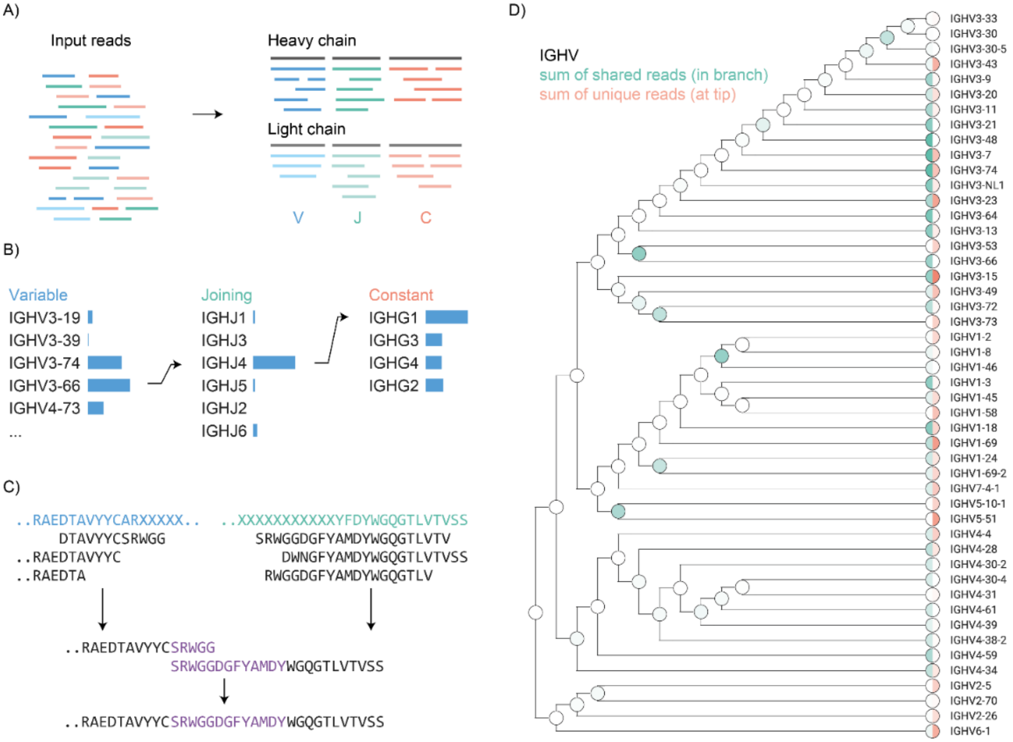
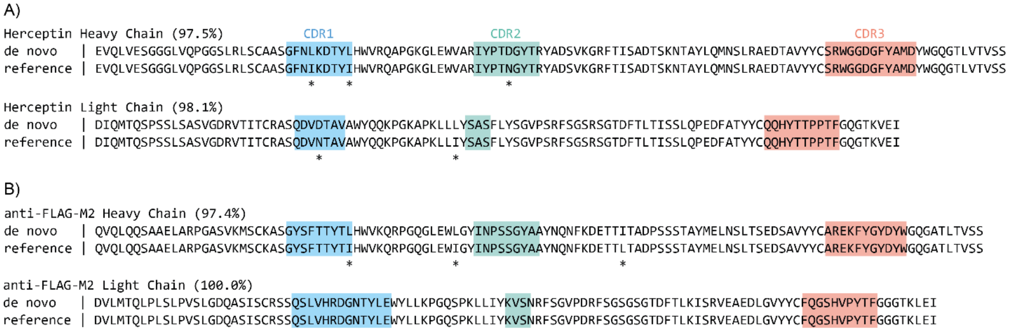


Figure 1. *Schematic overview of Stitch. A) Input reads in PEAKS or FASTA format are matched to user-defined templates for V/J/C segments of the heavy and light chains by local Smith-Waterman alignment. B) For monoclonal antibodies the top-scoring segments can be recombined for a second template matching step on the full heavy and light chain sequences. C) Procedure to reconstruct CDRH3 looks for overlap between the overhanging reads extending the V- and J-segments. D) Shared and unique reads are placed at the corresponding position in a cladogram of homologous template sequences to provide a quantitative overview of the template matching with explicit consideration of ambiguity in the read placement. This example is whole IgG from a COVID-19 hospitalized patient, as further described in Figure 4.*

In its most basic implementation, Stitch can simply match peptide reads to any homologous template in a user-defined database. Peptide reads are placed based on a user-defined cutoff score of the local alignment. When the database contains multiple templates, individual reads may match multiple entries with scores above this cutoff. This scenario is particularly relevant to antibody sequences as the multitude of available V/J/C alleles share a high degree of homology. The program can be set to place reads within all templates above the cutoff score or to place reads only on their single-highest scoring template. With this latter setting, reads with equal scores on multiple templates will be placed at all entries simultaneously. Reads with a single-highest scoring template are thereby defined as 'unique' for the program to track the total 'unique' alignment score and area of every template. Furthermore, Stitch explicitly considers the ambiguity of read placement across multiple homologous template sequences. A multiple sequence alignment is performed on each segment of the user-defined templates to generate a cladogram that represents the homology between the template sequences. Unique reads are placed at the tips of the branches, whereas shared reads are placed at the corresponding branching points of the tree (see Figure 1D). Stitch outputs the consensus sequence of every matched template based on all overlapping reads, accounting for frequency, global quality score and MS peak area with PEAKS data as input. The generated consensus sequence defaults to the template sequence in regions without coverage. Positions corresponding to I/L residues are defaulted to L in PEAKS data, as the two residues have identical masses and are therefore indistinguishable in most MS experiments. The consensus sequences in the output follow the matched template in these instances, changing the position to isoleucine when suggested by the template sequence.

Stitch allows templates to be defined in multiple separate groups, such that for antibody sequences we can sort peptide reads from heavy and light chains, and distinguish peptides from the V-, J- and C-segments of either chain. We have defined separate template databases for IGHV-IGHJ-IGHC, as well as IGLV-IGLJ-IGLC (with all kappa and lambda sequences combined in the same databases). The templates correspond to the germline sequences included in IMGT but filtered to create a reduced and non-redundant set of amino acid sequences (templates for human, mouse, bovine, dog and rabbit antibodies are currently provided and the clean-up procedure to generate the non-redundant databases from additional species is included in the program). Templates for the D-segment are currently not taken from IMGT as they are typically too short and variable for any meaningful read-matching. In addition to the Ig segments, a separate decoy database for common contaminants of cell culture medium, plasma/serum, and proteomics sample preparation can be defined. The output report includes consensus sequences for all matched germline templates with annotation of the CDRs, as well as a quantitative overview of how each germline template is represented in the dataset by number of matched reads, alignment score, and combined peak area for the total set of matched reads, or the set of unique reads. Moreover, the program generates an aligned overview of all reads overlapping the CDRs (grouped by CDRH1, CDRH2, CDRH3, CDRL1, CDRL2, and CDRL3). The resulting report is a comprehensive overview of the antibody sequences that are represented in the dataset and provides an important first step towards analyzing polyclonal antibodies and repertoire profiling.



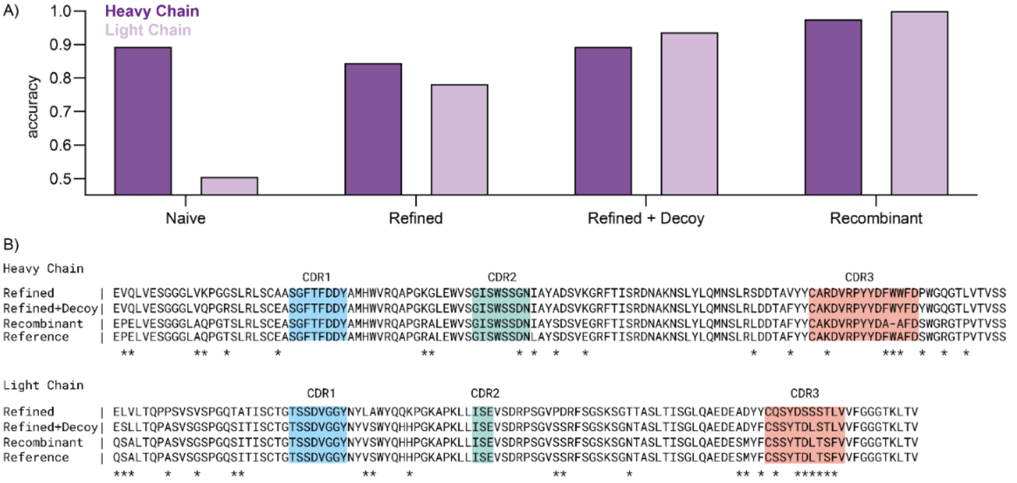
```
A)
Herceptin Heavy Chain (97.5%)           CDR1                CDR2                                      CDR3
de novo   | EVQLVESGGGLVQPGGSLRLSCAASGFNLKDTYLHWVRQAPGKGLEWVARIYPTDGYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYYCSRWGGDGFYAMDYWGQGTLVTVSS
reference | EVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYYCSRWGGDGFYAMDYWGQGTLVTVSS
                                        *    *                    *

Herceptin Light Chain (98.1%)
de novo   | DIQMTQSPSSLSASVGDRVTITCRASQDVDTAVAWYQQKPGKAPKLLLYSASFLYSGVPSRFSGSRSGTDFTLTISSLQPEDFATYYCQQHYTTPPTFGQGTKVEI
reference | DIQMTQSPSSLSASVGDRVTITCRASQDVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPSRFSGSRSGTDFTLTISSLQPEDFATYYCQQHYTTPPTFGQGTKVEI
                                         *                 *

B)
anti-FLAG-M2 Heavy Chain (97.4%)
de novo   | QVQLQQSAAELARPGASVKMSCKASGYSFTTYTLHWVKQRPGQGLEWLGYINPSSGYAAYNQNFKDETTITADPSSSTAYMELNSLTSEDSAVYYCAREKFYGYDYWGQGATLTVSS
reference | QVQLQQSAAELARPGASVKMSCKASGYSFTTYTIHWVKQRPGQGLEWIGYINPSSGYAAYNQNFKDETTLTADPSSSTAYMELNSLTSEDSAVYYCAREKFYGYDYWGQGATLTVSS
                                             *             *                     *

anti-FLAG-M2 Light Chain (100.0%)
de novo   | DVLMTQLPLSLPVSLGDQASISCRSSQSLVHRDGNTYLEWYLLKPGQSPKLLIYKVSNRFSGVPDRFSGSGSGTDFTLKISRVEAEDLGVYYCFQGSHVPYTFGGGTKLEI
reference | DVLMTQLPLSLPVSLGDQASISCRSSQSLVHRDGNTYLEWYLLKPGQSPKLLIYKVSNRFSGVPDRFSGSGSGTDFTLKISRVEAEDLGVYYCFQGSHVPYTFGGGTKLEI
```

Figure 2. *Stitch analysis of monoclonal antibodies. A) recombinant purified Herceptin. B) recombinant purified anti-FLAG-M2. CDRs are annotated, sequence conflicts highlighted by an asterisk (*) and the sequence identity listed in parentheses.*

We have also built a dedicated procedure to reconstruct full monoclonal antibody sequences (see Figure 1B-C). Using the template-matching procedure described above, Stitch then selects the top-*N* scoring templates for each segment (with $N = 1$ for monoclonal antibodies) and recombines their consensus sequences into new V-J-C templates. As part of this recombination, CDRH3 is reconstructed by extending the V- and J-segments with the consensus sequences of overhanging reads to fill in the missing D-segment. The program then searches for identical sequences within the V- and J-overhanging regions to find the correct junction between the segments. A single gap is placed at the V-J junction if no overlap between the overhanging sequences can be found. The new recombined templates with reconstructed CDRH3 are then used for a second round of template matching to determine the final consensus sequences of the full heavy and light chains of the antibody. Stitch offers an option to use all non-selected germline templates as decoys in this second step to accommodate sequencing of monoclonal antibodies against a high background of homologous sequences, such as those collected from serum by LC fractionation or to cope with the presence of the multiple light or heavy chains which are often observed in hybridoma-derived antibodies [39].

To demonstrate the use of Stitch, we assembled *de novo* peptide reads to reconstruct the full heavy and light chain sequences of three different monoclonal antibodies. First, we used the human-mouse chimeric therapeutic antibody Herceptin (also known as Trastuzumab). Herceptin is composed of mouse CDR sequences placed within a human IgG1 framework and targets the Her2 receptor in treatment of a variety of cancers [26,40,41]. Second, we reconstructed the sequence of the anti-FLAG-M2 antibody, which is a mouse antibody targeting the DYKDDDDK epitope used to label and purify recombinant proteins [26,42]. Alignment of the assembled output sequences reveals an overall accuracy of 98% and 99% (including I/L assignments) for Herceptin and anti-FLAG-M2, respectively (see Figure 2). A close-up view of the CDRH3 reconstructions demonstrates how the missing D-segment in the heavy chain is obtained through the two-step procedure described above (*i.e.* by extending the V- and J-segments with the consensus sequence of overlapping reads, searching for the V-J junction in the extended templates and performing a second round of template matching on the recombined V-J-C template, see Supplementary Figure S1).

The third monoclonal antibody (F59) represents a more challenging case, as it is a Fab isolated directly from patient serum by reversed-phase LC-fractionation and therefore has to be sequenced against a high background of unrelated antibodies (see Figure 3). The F59 sequence was originally determined by integrated use of both bottom-up and top-down LC-MS/MS data [32]. It consists of an IGHV3-9 heavy chain, coupled to an IGLV2-14 light chain. When we naively provide the input data to Stitch, it initially returns variable domain output sequences of 89% accuracy for the heavy chain, and a mere 50% for the light chain. This rate of errors is caused by the high background of other antibodies in the sample, which results in selection of the ubiquitous IGKV3-20 template for recombination. However, read matching on the C-region clearly point to the presence of a lambda light chain, and within the subset of IGLV templates, IGLV2-14 is indeed assigned the highest score. When we refine the Stitch run to force selection of IGLV2-14 for recombination, the accuracy of the light chain improves to 78%. The remaining errors still stem from the high background of other antibodies, but this can be further reduced by using the non-selected templates as decoys in the final template matching step. With the use of non-selected templates as decoys, the accuracy of the light chain now improves to 94%. By comparison, the sequence of the purified synthetic recombinant antibody can be determined to 98% and 100% for the heavy and light chain, respectively. Of note, in the fractionated serum sample 4/20 of the remaining errors in the heavy and light chains combined occur in CDR3, with all other error occurring outside CDRs in the framework regions. Moreover, most errors are clearly identifiable in the sequence logo from the Stitch analysis (see Supplementary Figure S2), which may be useful to apply manual corrections and refine the sequence with complementary MS data. The F59 Fab from fractionated patient serum can thus be sequenced to >90% accuracy with Stitch, even before input from complementary top-down LC-MS/MS data.



```
Heavy Chain
                                         CDR1                     CDR2                                             CDR3
Refined      | EVQLVESGGGLVKPGGSLRLSCAASGFTFDDYAMHWVRQAPGKGLEWVSGISWSSGNIAYADSVKGRFTISRDNAKNSLYLQMNSLRSDDTAVYYCARDVRPYYDFWWFDPWGQGTLVTVSS
Refined+Decoy| EVQLVESGGGLVQPGRSLRLSCEASGFTFDDYAMHWVRQAPGKGLEWVSGISWSSDNIAYADSVKGRFTISRDNAKNSLYLQMNSLRLDDTAFYYCAKDVRPYYDFWYFDYWGQGTLVTVSS
Recombinant  | EPELVESGGGLAQPGTSLRLSCEASGFTFDDYAMHWVRQAPGRALEWVSGISWSSDNIAYSDSVEGRFTISRDNAKNSLYLQMNSLRLDDTAFYYCAKDVRPYYDA-AFDSWGRGTPVTVSS
Reference    | EPELVESGGGLAQPGTSLRLSCEASGFTFDDYAMHWVRQAPGRALEWVSGISWSSDNLAYSDSVEGRFTISRDNAKNSLYLQMNSLRLDDTAFYYCAKDVRPYYDFWAFDSWGRGTPVTVSS
                **        **  *      *                   **           * *  *   *                      *    *    *       ***  *  *  *

Light Chain
                                         CDR1                  CDR2                                        CDR3
Refined      | ELVLTQPPSVSVSPGQTATISCTGTSSDVGGYNYLAWYQQKPGKAPKLLISEVSDRPSGVPDRFSGSKSGTTASLTISGLQAEDEADYYCQSYDSSSTLVVFGGGTKLTV
Refined+Decoy| ESLLTQPASVSGSPGQSITISCTGTSSDVGGYNYVSWYQHHPGKAPKLLISEVSDRPSGVSSRFSGSKSGNTASLTISGLQAEDEADYFCSSYTDLSTLVVFGGGTKLTV
Recombinant  | QSALTQPASVSGSPGQSITISCTGTSSDVGGYNYVSWYQHHPGKAPKLLISEVSDRPSGVSSRFSGSKSGNTASLTISGLQAEDESMYFCSSYTDLTSFVVFGGGTKLTV
Reference    | QSALTQPASVSGSPGQSITISCTGTSSDVGGYNYVSWYQHHPGKAPKLLISEVSDRPSGVSSRFSGSKSGNTASLTISGLQAEDESMYFCSSYTDLTSFVVFGGGTKLTV
               ***    *   *    **                **    *                   **        *              ** * *  ******
```

***Figure 3.** Stitch analysis of F59 Fab from fractionated serum sample. A) sequence accuracy of the variable domains of heavy and light chain from a naïve Stitch run (Naïve), compared to a run with selected V-segments for recombination (Refined), added use of non-selected templates as decoys in the final template matching step (Refined+Decoy), and the synthetic recombinant antibody run with identical Stitch parameters (Recombinant). B) Sequence*

alignment of the output sequences (Naïve run excluded for low accuracy) with CDRs annotated and sequence conflicts highlighted by an asterisk (*).

In addition to monoclonal antibodies, Stitch can also be used to assemble proteomic short reads of free light chains, such as those observed in multiple myeloma patients. Chamot-Rooke and colleagues recently reported proteomic sequencing of multiple myeloma light chains from patient urine, using an integrated bottom-up and top-down sequencing approach [43]. Here we used the bottom-up *de novo* sequencing reads from that study as a test case to reconstruct the light chains (see Supplementary Figure S3A). This includes one sample consisting of a mixture of two closely related light chains, for which the variable positions are also clearly identifiable in the sequence logo from the Stitch analysis (see Supplementary Figure S3B). The reconstructed sequences are in good agreement with those reported in the original study, with an average accuracy of 98% (ranging from 88% to 100%, including I/L assignments). Notably, in all instances of sequence conflicts, the coverage of the bottom-up data is limited or the corresponding alternative sequence is also present in reads within the dataset (but does not stand out in terms of quality score and MS peak area to dominate the consensus sequence). This further stresses the importance of the (depth of) coverage in the input data and highlights the added benefit of complementary top-down LC-MS/MS data for antibody sequencing.

To demonstrate the use of Stitch for profiling polyclonal antibody mixtures we generated a new dataset of *de novo* peptide reads from human serum. We obtained the total fraction of IgG, isolated by protein G affinity purification, from two individuals hospitalized with COVID-19. The purified IgG fractions were digested in parallel with four different proteases (trypsin, chymotrypsin, elastase and thermolysin), and analyzed by LC-MS/MS with a dual fragmentation scheme using both stepped HCD and EThcD fragmentation, with all obtained *de novo* sequence reads pooled into a single Stitch run. The analysis provides a quantitative overview of the IgG classes, use of kappa vs lambda light chains, and corresponding use of V-alleles across the total IgG repertoire of these patients (see Figure 3). For each of the two patient samples, we mapped 1276 and 1292 reads to IGHC, 697 and 837 reads to IGLC, 513 and 624 reads to IGHV, and 697 and 837 reads to IGKV/IGLV. The profiles of both patients are remarkably similar, dominated by IgG1 with kappa light chains and drawing primarily from IGHV1/3 and IGKV1/3 alleles. Of the matched reads to the IGHV segment, 74 and 89 map to CDRH1, 65 and 92 to CDRH2, and 19 and 23 to CDRH3. Whereas the reads mapping to CDRH1/2 collectively span the full region, the CDRH3 reads are mostly limited to the first conserved AR/K residues following the preceding cysteine, or the conserved parts of the J-segment. Of the matched reads to the IGKV/IGLV segment, 98 and 116 map to CDRL1, 136 and 182 to CDRL2, and 93 and 97 to CDRL3. The CDRL3 reads span a larger region compared to CDRH3, likely because the read assembly does not suffer from the missing D-segment. The Stitch analysis thus provides a quantitative overview of V-gene usage in polyclonal IgG mixtures, obtained straight from human serum samples, covering CDR1 and CDR2, but with notable limitations of CDR3.

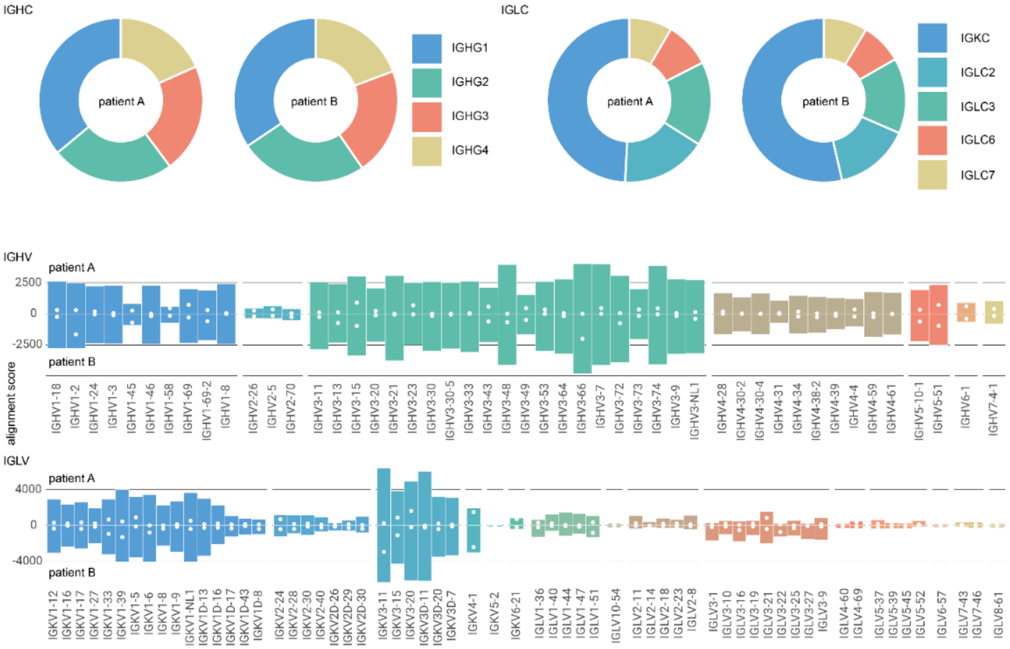


Figure 4. *Repertoire profiling of protein G purified whole IgG from human serum. Profiles of IGHC, IGLC, IGHV and IGLV segments from two hospitalized COVID-19 patients as determined by Stitch. Shown are the total alignment scores of each matched template. The closed white circles in the IGHV/IGLV segments indicate score of the uniquely matched reads.*

Discussion

Stitch provides a quick and accessible way to assemble proteomic short reads against user-defined templates. It enables full reconstruction of monoclonal antibodies and free light chains, as well as profiling of polyclonal antibody mixtures. Stitch was specifically developed to provide better insight into sequence variations of antibodies in complex mixtures. The assumption of a monoclonal antibody is so deeply embedded in the structure of existing software for antibody sequencing that their output provides limited insights into the presence of other (background) sequences and they may even struggle to converge on a consensus sequence for complex mixtures. Compared to other antibody sequencing software, Stitch still requires input from *de novo* peptide sequencing algorithms and is limited to performing a template-based assembly of these input reads. This is an alternative strategy to existing software like pTA, which rather performs the assembly based on overlap between peptide sequences. The template-based assembly of Stitch is on the one hand more dependent on the homology of the input reads to the templates, but on the other hand requires less extensive overlap between peptides.

Given high-quality input reads, Stitch generates accurate consensus sequences, with remaining errors being fundamental to MS-based sequencing. These are errors related to deamidation (N to D) and assignment of isomeric I/L residues. Currently Stitch assigns I/L residues based on the matched template sequence, but this can potentially be further improved by considering experimental information, such as the cleavage specificity of chymotrypsin (cleaves only at L, not I) and use of diagnostic *w*-ions [44-47]. Whereas Stitch already explicitly considers both global and local quality scores of sequence reads, it does not yet provide integrated access to the underlying raw MS/MS data itself, which we aim to implement in the future. It is currently also limited to plain FASTA or PEAKS and Novor.Cloud data as input reads, but we aim to adapt it to data formats from additional *de novo* sequencing software in the future. Current limitations regarding polyclonal antibody profiling will have to be solved with improved experimental approaches: obtaining longer sequence reads will reduce ambiguity in the correlation of sequence variants against the database of homologous templates, and top-down MS/MS of intact Fabs/antibodies or additional cross-linking MS workflows will have to elucidate the heavy-light chain pairings in the antibody mixture.

As illustrated by the sequencing of the F59 clone from fractionated patient serum (Figure 3), as well as the mixture of light chains in the urine of multiple myeloma patients described by Chamot-Rooke and colleagues (Supplementary Figure 3), integrated use of bottom-up and top-down sequencing provides important information to correct, refine and validate *de novo* antibody sequences derived from mass spectrometry, especially from complex mixtures. We have previously described such integrated use of bottom-up and top-down MS to determine the F59 sequence. Future work will focus on better streamlining and automation of the integrated use of both approaches for antibody sequencing. Validation by production and characterization of the synthetic recombinant antibodies, in particular when the target antigen is known, may also become an important addition to the validation stage of MS-based antibody sequencing approaches.

By enabling antibody sequencing and profiling from the purified secreted product, the development of Stitch contributes to an emerging new serology, in which bulk measures of antigen binding and neutralization can be directly related to the composition and sequence of a polyclonal antibody mixture. Direct MS-based sequencing and profiling of secreted antibodies thereby bridges the gap between bulk serological assays and B-cell sequencing approaches. These developments promise to provide a better understanding of antibody-mediated immunity in natural infection, vaccination and autoimmune disorders.

Data and Code availability

The source code of Stitch is available on the Snijderlab GitHub page (https://github.com/snijderlab/stitch). All Stitch HTML results related to this study are provided as supplementary data. The raw data and PEAKS analyses unique to this study have been deposited in the ProteomeXchange Consortium via the PRIDE partner repository with the data set identifier PXD031941. The raw data of the monoclonal antibodies herceptin and anti-FLAG-M2 is available under identifier PXD023419. The raw data and PEAKS analyses of the multiple myeloma light chain dataset of Chamot-Rooke and colleagues is available under identifier PXD025884. The raw data of the serum-derived F59 monoclonal antibody is available at https://doi.org/10.6084/m9.figshare.13194005.

Supplementary Information

Contents

Chapter 3

Legend of Supplementary Data with Stitch HTML reports

CDRH3 reconstructions of Herceptin and anti-FLAG-M2

Details of sequence variations in F59 Fab

Description of multiple myeloma light chain sequences.

Supplementary Data.

All reported Stitch analyses are provided with the complete output reports. The supplementary data can be unzipped to browse the interactive HTML reports. The corresponding analysis parameters are provided under the 'Batch File' menu in each report.

A) Herceptin

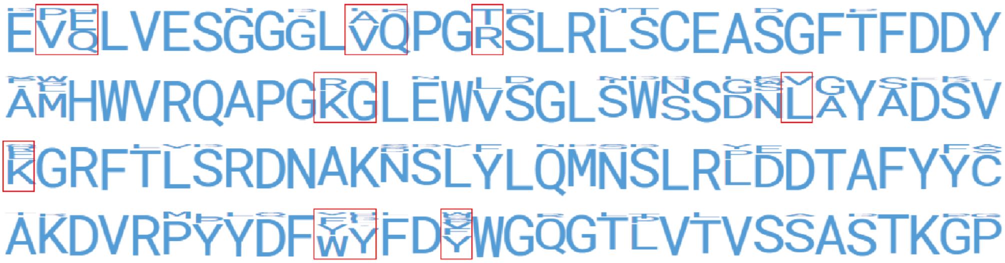
```
template   ...DTAVYYCAR                 YFDYWGQGTLVTVSS...
de novo       DTAVYYCSRWGG     SRWGGDGFYAMDYWGQGTLVTVSS

overlap       DTAVYYCSRWGG
                     SRWGGDGFYAMDYWGQGTLVTVSS

ref        ...DTAVYYCSRWGGDGFYAMDYWGQGTLVTVSS...
```

B) anti-FLAG-M2

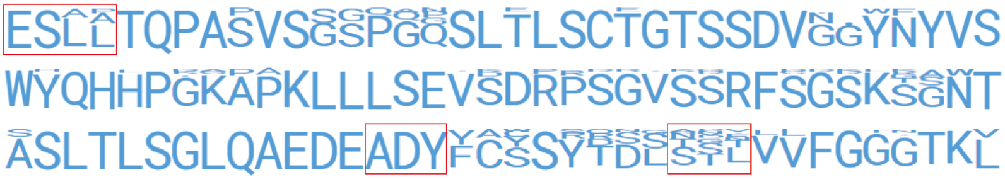
```
template   ...DSAVYYCAR                     YFDYWGQGTTLTVSS...
de novo       DSAVYYCAREKFYGY     VYYCAREKFYGYDYWGQGATLTVSS

overlap       DSAVYYCAREKFYGY
                 VYYCAREKFYGYDYWGQGATLTVSS

ref        ...DSAVYYCAREKFYGYDYWGQGATLTVSS...
```

Supplementary Figure S1. Detailed view of CDRH3 reconstruction by Stitch of the antibodies shown in main text figure 2. Shown are the selected template sequences, aligned reads, found overlap and known reference sequence.

A) F59 Heavy Chain from Refined+Decoy run


EVQLVESGGGLVQPGRSLRLSCEASGFTFDDY
AMHWVRQAPGKGLEWVSGLSWSSDNLAYADSV
KGRFTLSRDNAKNSLYLQMNSLRLDDTAFYYC
AKDVRPYYDFWYFDYWGQGTLVTVSSASTKGP

B) F59 Light Chain from Refined+Decoy run

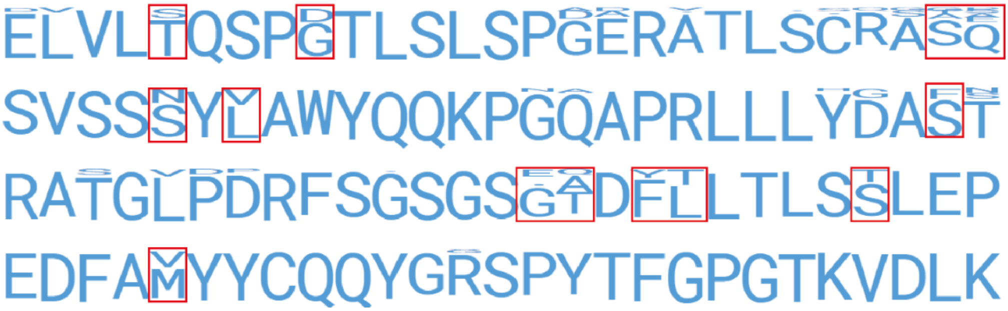

ESLLTQPASVSGSPGQSLTLSCTGTSSDVGGYNYVS
WYQHHPGKAPKLLLSEVSDRPSGVSSRFSGSKSGNT
ASLTLSGLQAEDEADYFCSSYTDLSTLVVFGGGTKL

Supplementary Figure S2. Sequence logos of heavy (A) and light (B) chains of F59 Fab from fractionated serum sample from Refined+Decoy Stitch runs. Sequencing error are highlighted in red boxes.

A)

```
>Patient 1 (94.5% Δ -18.3 Da)
de novo   | ESALTQPRSVSGSPGQSVTISCTGTSSDVGGYNYVSWYQQHPGKAPKLMIYDVTKRPSGVPDRFSGSKSGTTASLTISGLQAEDEADYYCCSYAGLDLFVLFGGTKLTV
reference | GPDLTQPRSVSGSPGQSVTLSCTGTSSDVGGYNYVSWYQQHPGKAPKLMIYDVTKRPSGVPDRFSGSKSGTTASLTISGLQAEDEADYYCCSYAGIDIFVLFGGTKLTV
            ***                *                                                                           * *

>Patient 5 (A: 99.0% Δ -30.4 Da) & (B; 87.9%)
de novo     | EIVLTQSPGTLSLSPGERATLSCRASQSVSSSYLAWYQQKPGQAPRLLIYDASTRATGIPDRFSGSGSGTDFLLTISSLEPEDFAMYYCQQYGRSPYTFGPGTKVDI
reference A | EIVLTQSPGTLSLSPGERATLSCRASQSVSSSYLAWYQQKPGQAPRLLIYDASTRATGIPDRFSGSGSGADFLLTISSLEPEDFAMYYCQQYGRSPYTFGPGTKVDI
reference B | EIVLSQSPDTLSLSPGERATLSCRADKSVSSNYVAWYQQKPGQAPRLLIYDAFTRATGIPDRFSGSGSETDYTLTISTLEPEDFAVYYCQQYGRSPYTFGPGTKVDI
                  *   *                **    * *                  *               ** **    *       *

>Patient 6 (100.0% Δ -0.5 Da)
de novo   | DIQMTQSPSSLSASVGDRVSITCRASESISSYVNWYQQKPGKAPKLLIYTASSLQSGVPPRFSGSASGTDFTLTISSLQPEDFATYYCQQSYSTPITFGQGTRLEI
reference | DIQMTQSPSSLSASVGDRVSITCRASESISSYVNWYQQKPGKAPKLLIYTASSLQSGVPPRFSGSASGTDFTLTISSLQPEDFATYYCQQSYSTPITFGQGTRLEI

>Patient 7 (94.3% Δ 60.8 Da)
de novo   | DIQMTQSPSSLSASVGDRVTITCQASQDIAKYLNWYQQKPGKPPKLLIYDTSNLETGVPSRFSGSGSGTDFTFTINSLQPEDIATYYCQQYDDFPLTFGPGTKVDI
reference | DIQMTQSPSSLSASVGDRVTITCQASQDLAKYLNWYQQKPGKPPKLLIYDTSNLETGVPSRFSNG-GGTDFTFTINSLQPEDLATYYCQQYDDFPLTFGPGTKVDI
                                        *                                   ****              *

>Patient 8 (99.1% Δ -0.5 Da)
de novo   | DIQMTQSPSTLSASVGDRVTITCRASQSISSSLAWYQQKPGKAPKLLIYDASSLETGVPSRFSGSGSGTEFTLSISSLQPDDFATYYCQHYNSYSLTFGQGTKVEI
reference | DIQMTQSPSTLSASVGDRVTITCRASQSLSSSLAWYQQKPGKAPKLLIYDASSLETGVPSRFSGSGSGTEFTLSISSLQPDDFATYYCQHYNSYSLTFGQGTKVEI
                                        *

>Patient 13 (92.8% Δ -12.5 Da)
de novo   | EPALTQPPSVSGAPGQRVTISCTGSSSNIGAGWDVHWYQQLPGTVPKLLIYADRNRPSGVPERFSGSKSGTDAALAIAGLQAEDEADYYCQSYDSALSGFYVFGTGTKVLV
reference | EAPLTQPPSVSGAPGQRVTLSCTGSSSNLGAGWDVHWYQQLPGTVPKLLIYADRNRPSGVPERFSGSKSGTSATVAIAGLQAEDEADYYCQSYDSALSGFYVFGTGTKVIV
            ***                *        *                                          *  *                                  *

>Patient 15 (99.1% Δ -0.8 Da)
de novo   | DIQMTQSPSTLSASVGDAVTITCRASQSINVWLAWYQQKPGKPPKLLIYEASNLESGVPSRFSGSGSGTEFTLTISSLQPDDFATYYCQQYNSYPYTFGQGAKLEI
reference | DIQMTQSPSTLSASVGDAVTITCRASQSLNVWLAWYQQKPGKPPKLLIYEASNLESGVPSRFSGSGSGTEFTLTISSLQPDDFATYYCQQYNSYPYTFGQGAKLEI
                                        *

>Patient 18 (99.1% Δ -0.7 Da)
de novo   | DIQMTQSPSSLSASVGDRVTITCQASRDISNYLNWYQQKPGKAPMLLIYAASNLQTGVPSRFSGSGSGTDFTFTISSLQPEDIATYYCQQYGNLPLTFGGGTKVEI
reference | DIQMTQSPSSLSASVGDRVTITCQASRDISNYLNWYQQKPGKAPMLLIYAASNLQTGVPSRFSGSGSGTDFTFTISSLQPEDLATYYCQQYGNLPLTFGGGTKVEI
                                                                                              *

>Patient 19 (99.1% Δ -0.8 Da)
de novo   | DIQMTQSPSSLSASVGDRVTITCQASQDIGNYLNWYQQKPGKAPRLLIYDASDLEEGVPSRFSGSGSGTDFTFTISSLQPEDFATYYCQQYHTLPPLTFGGGTKVDV
reference | DIQMTQSPSSLSASVGDRVTITCQASQDLGNYLNWYQQKPGKAPRLLIYDASDLEEGVPSRFSGSGSGTDFTFTISSLQPEDFATYYCQQYHTLPPLTFGGGTKVDV
                                        *

>Patient 20 (99.1% Δ -1.8 Da)
de novo   | DIQMTQSPSTLSTSVGDRVTITCRASQSIRTWLAWYQQKPGKAPKLLIYKASTLETGVPSRFSGSGSGTMFTLTISSLQPEDFATYYCQQYNDYSGTFGQGTKLEI
reference | DIQMTQSPSTLSTSVGDRVTITCRASQSIRTWLAWYQQKPGKAPKLLIYKASTLETGVPSRFSGSGSGTEFTLTISSLQPEDFATYYCQQYNDYSGTFGQGTKLEI
                                                                                 *
```

B)


ELVLTQSPGTLSLSPGERATLSCRASQ
SVSSSYLAWYQQKPGQAPRLLLYDAST
RATGLPDRFSGSGSGTDFLLTLSSLEP
EDFAMYYCQQYGRSPYTFGPGTKVDLK

Supplementary Figure S3. A) Stitch reconstruction of light chains from multiple myeloma patient urine (Chamot-Rooke and colleagues, ref 43). CDRs are annotated, sequence conflicts highlighted by an asterisk (*) and sequence identity in parentheses, along with the mass error of the predicted Stitch sequence compared to the experimentally determined mass of the intact light chain. B) Sequence logo of P5 light chain from Stitch analysis, with highlighted variants in red boxes.

Acknowledgments

We would like to thank Marien de Jonge and Dimitri Diavatopoulos (Radboud UMC) for sharing the COVID-19 patient sera. We would like to thank Keri T. Schmidt, Jade A.C. van der Hout, and Mariëlle E. Floor for extensive testing of Stitch. Further thanks to Bastiaan de Graaf and the rest of our colleagues in the Biomolecular Mass Spectrometry and Proteomics group at Utrecht University for support and helpful discussions. This research was funded by the Dutch Research Council NWO Gravitation 2013 BOO, Institute for Chemical Immunology (ICI; 024.002.009), and the Utrecht Molecular Immunology Hub.

References

(1) Casadevall, A.; Pirofski, L.-a. *Nature immunology* **2012**, *13*, 21-28.
(2) Davies, D. R.; Metzger, H. *Annual review of immunology* **1983**, *1*, 87-115.
(3) Lleo, A.; Invernizzi, P.; Gao, B.; Podda, M.; Gershwin, M. E. *Autoimmunity reviews* **2010**, *9*, A259-A266.
(4) Naparstek, Y.; Plotz, P. H. *Annual review of immunology* **1993**, *11*, 79-104.
(5) McHeyzer-Williams, L. J.; McHeyzer-Williams, M. G. *Annu. Rev. Immunol.* **2005**, *23*, 487-513.
(6) Traggiai, E.; Puzone, R.; Lanzavecchia, A. *Vaccine* **2003**, *21*, S35-S37.
(7) Di Noia, J. M.; Neuberger, M. S. *Annu. Rev. Biochem.* **2007**, *76*, 1-22.
(8) Lefranc, M.-P.; Lefranc, G. *Antibodies* **2019**, *8*, 29.
(9) Tonegawa, S. *Nature* **1983**, *302*, 575-581.
(10) Watson, C. T.; Glanville, J.; Marasco, W. A. *Trends in immunology* **2017**, *38*, 459-470.
(11) Lefranc, M.-P.; Lefranc, G. *The immunoglobulin factsbook*; Academic press, 2001.
(12) Fischer, N. In *MAbs*; Taylor & Francis, 2011, pp 17-20.
(13) Georgiou, G.; Ippolito, G. C.; Beausang, J.; Busse, C. E.; Wardemann, H.; Quake, S. R. *Nature biotechnology* **2014**, *32*, 158-168.
(14) Avram, O.; Kigel, A.; Vaisman-Mentesh, A.; Kligsberg, S.; Rosenstein, S.; Dror, Y.; Pupko, T.; Wine, Y. *PLoS computational biology* **2021**, *17*, e1008607.

(15) Lavinder, J. J.; Horton, A. P.; Georgiou, G.; Ippolito, G. C. *Current opinion in chemical biology* **2015**, *24*, 112-120.

(16) Lee, J.; Boutz, D. R.; Chromikova, V.; Joyce, M. G.; Vollmers, C.; Leung, K.; Horton, A. P.; DeKosky, B. J.; Lee, C.-H.; Lavinder, J. J. *Nature medicine* **2016**, *22*, 1456-1464.

(17) Lee, J.; Paparoditis, P.; Horton, A. P.; Frühwirth, A.; McDaniel, J. R.; Jung, J.; Boutz, D. R.; Hussein, D. A.; Tanno, Y.; Pappas, L. *Cell host & microbe* **2019**, *25*, 367-376. e365.

(18) Lindesmith, L. C.; McDaniel, J. R.; Changela, A.; Verardi, R.; Kerr, S. A.; Costantini, V.; Brewer-Jensen, P. D.; Mallory, M. L.; Voss, W. N.; Boutz, D. R. *Immunity* **2019**, *50*, 1530-1541. e1538.

(19) Boutz, D. R.; Horton, A. P.; Wine, Y.; Lavinder, J. J.; Georgiou, G.; Marcotte, E. M. *Analytical chemistry* **2014**, *86*, 4758-4766.

(20) Chen, J.; Zheng, Q.; Hammers, C. M.; Ellebrecht, C. T.; Mukherjee, E. M.; Tang, H.-Y.; Lin, C.; Yuan, H.; Pan, M.; Langenhan, J. *Cell reports* **2017**, *18*, 237-247.

(21) Fridy, P. C.; Li, Y.; Keegan, S.; Thompson, M. K.; Nudelman, I.; Scheid, J. F.; Oeffinger, M.; Nussenzweig, M. C.; Fenyö, D.; Chait, B. T. *Nature methods* **2014**, *11*, 1253-1260.

(22) Bandeira, N.; Pham, V.; Pevzner, P.; Arnott, D.; Lill, J. R. *Nature biotechnology* **2008**, *26*, 1336-1338.

(23) Castellana, N. E.; McCutcheon, K.; Pham, V. C.; Harden, K.; Nguyen, A.; Young, J.; Adams, C.; Schroeder, K.; Arnott, D.; Bafna, V. *Proteomics* **2011**, *11*, 395-405.

(24) Cheung, W. C.; Beausoleil, S. A.; Zhang, X.; Sato, S.; Schieferl, S. M.; Wieler, J. S.; Beaudet, J. G.; Ramenani, R. K.; Popova, L.; Comb, M. J. *Nature biotechnology* **2012**, *30*, 447-452.

(25) Guthals, A.; Gan, Y.; Murray, L.; Chen, Y.; Stinson, J.; Nakamura, G.; Lill, J. R.; Sandoval, W.; Bandeira, N. *Journal of proteome research* **2017**, *16*, 45-54.

(26) Peng, W.; Pronker, M. F.; Snijder, J. *Journal of Proteome Research* **2021**, *20*, 3559-3566.

(27) Rickert, K. W.; Grinberg, L.; Woods, R. M.; Wilson, S.; Bowen, M. A.; Baca, M. In *MAbs*; Taylor & Francis, 2016, pp 501-512.

(28) Savidor, A.; Barzilay, R.; Elinger, D.; Yarden, Y.; Lindzen, M.; Gabashvili, A.; Tal, O. A.; Levin, Y. *Molecular & Cellular Proteomics* **2017**, *16*, 1151-1161.

(29) Sen, K. I.; Tang, W. H.; Nayak, S.; Kil, Y. J.; Bern, M.; Ozoglu, B.; Ueberheide, B.; Davis, D.; Becker, C. *Journal of The American Society for Mass Spectrometry* **2017**, *28*, 803-810.

(30) Sousa, E.; Olland, S.; Shih, H. H.; Marquette, K.; Martone, R.; Lu, Z.; Paulsen, J.; Gill, D.; He, T. *International Journal of Mass Spectrometry* **2012**, *312*, 61-69.

(31) Tran, N. H.; Rahman, M. Z.; He, L.; Xin, L.; Shan, B.; Li, M. *Scientific reports* **2016**, *6*, 1-10.

(32) Bondt, A.; Hoek, M.; Tamara, S.; de Graaf, B.; Peng, W.; Schulte, D.; van Rijswijck, D. M.; den Boer, M. A.; Greisch, J.-F.; Varkila, M. R. *Cell systems* **2021**, *12*, 1131-1143. e1135.

(33) Coelho, C. H.; Nadakal, S. T.; Hurtado, P. G.; Morrison, R.; Galson, J. D.; Neal, J.; Wu, Y.; King, C. R.; Price, V.; Miura, K. *JCI insight* **2020**, *5*.

(34) Gonzales Hurtado, P. A.; Morrison, R.; Ribeiro, J. M.; Magale, H.; Attaher, O.; Diarra, B. S.; Mahamar, A.; Barry, A.; Dicko, A.; Duffy, P. E. *Journal of proteome research* **2019**, *18*, 3831-3839.

(35) Lundström, S. L.; Heyder, T.; Wiklundh, E.; Zhang, B.; Eklund, A.; Grunewald, J.; Zubarev, R. A. *International journal of molecular sciences* **2019**, *20*, 2157.

(36) Lundström, S. L.; Zhang, B.; Rutishauser, D.; Aarsland, D.; Zubarev, R. A. *Scientific reports* **2017**, *7*, 1-12.

(37) Smith, T. F.; Waterman, M. S. *Journal of molecular biology* **1981**, *147*, 195-197.

(38) Tran, N. H.; Zhang, X.; Xin, L.; Shan, B.; Li, M. *Proceedings of the National Academy of Sciences* **2017**, *114*, 8247-8252.

(39) Bradbury, A. R.; Trinklein, N. D.; Thie, H.; Wilkinson, I. C.; Tandon, A. K.; Anderson, S.; Bladen, C. L.; Jones, B.; Aldred, S. F.; Bestagno, M. In *MAbs*; Taylor & Francis, 2018, pp 539-546.

(40) Carter, P.; Presta, L.; Gorman, C. M.; Ridgway, J.; Henner, D.; Wong, W.; Rowland, A. M.; Kotts, C.; Carver, M. E.; Shepard, H. M. *Proceedings of the National Academy of Sciences* **1992**, *89*, 4285-4289.

(41) Slamon, D. J.; Leyland-Jones, B.; Shak, S.; Fuchs, H.; Paton, V.; Bajamonde, A.; Fleming, T.; Eiermann, W.; Wolter, J.; Pegram, M. *New England journal of medicine* **2001**, *344*, 783-792.

(42) Roosild, T. P.; Castronovo, S.; Choe, S. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* **2006**, *62*, 835-839.

(43) Dupré, M.; Duchateau, M.; Sternke-Hoffmann, R.; Boquoi, A.; Malosse, C.; Fenk, R.; Haas, R.; Buell, A. K.; Rey, M.; Chamot-Rooke, J. *Analytical Chemistry* **2021**, *93*, 10627-10634.

(44) Edwards, H. M.; Wu, H. T.; Julian, R. R.; Jackson, G. P. *Rapid Communications in Mass Spectrometry* **2022**, *36*, e9246.

(45) Johnson, R. S.; Martin, S. A.; Biemann, K.; Stults, J. T.; Watson, J. T. *Analytical chemistry* **1987**, *59*, 2621-2625.
(46) Xiao, Y.; Vecchi, M. M.; Wen, D. *Analytical chemistry* **2016**, *88*, 10757-10766.
(47) Zhokhov, S. S.; Kovalyov, S. V.; Samgina, T. Y.; Lebedev, A. T. *Journal of The American Society for Mass Spectrometry* **2017**, *28*, 1600-1611.

CHAPTER
4

Chapter 4

Reverse engineering the anti-MUC1 hybridoma antibody 139H2 by mass spectrometry-based *de novo* sequencing

Weiwei Peng[1a], Koen C.A.P. Giesbers[2a], Marta Šiborová[1], J. Wouter Beugelink[3], Matti F. Pronker[1], Douwe Schulte[1], John Hilkens[4], Bert J.C. Janssen[3], Karin Strijbis[2*], Joost Snijder[1*]

[1] Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute of Pharmaceutical Sciences, Utrecht University, Padualaan 8, 3584CH Utrecht, The Netherlands

[2] Department of Biomolecular Health Sciences, Division of Infectious Diseases and Immunology, Faculty of Veterinary Medicine, Utrecht University, Utrecht, Netherlands

[3] Structural Biochemistry, Bijvoet Center for Biomolecular Research, Department of Chemistry, Faculty of Science, Utrecht University, Universiteitsweg 99, 3584CG Utrecht, The Netherlands

[4] Division of Molecular Genetics, The Netherlands Cancer Institute, Amsterdam, The Netherlands.

[a] equal contribution

* corresponding authors: k.strijbis@uu.nl, j.snijder@uu.nl

Abstract:

Mucin 1 (MUC1) is a transmembrane mucin expressed at the apical surface of epithelial cells at mucosal surfaces. MUC1 has a barrier function against bacterial invasion and is well known for its aberrant expression and glycosylation in adenocarcinomas. The MUC1 extracellular domain contains a variable number of tandem repeats (VNTR) of 20 amino acids, which are heavily *O*-linked glycosylated. Monoclonal antibodies against the MUC1 VNTR are powerful research tools with applications in the diagnosis and treatment of MUC1-expressing cancers. Here we report direct mass spectrometry-based sequencing of anti-MUC1 hybridoma-derived 139H2 IgG, enabling reverse engineering of the functional recombinant monoclonal antibody. The crystal structure of the 139H2 Fab fragment in complex with the MUC1 epitope was solved, revealing the molecular basis for 139H2 binding specificity to MUC1 and its tolerance to *O*-glycosylation of the VNTR. The available sequence of 139H2 will allow further development of MUC1-related diagnostics, targeting and treatment strategies.

Introduction

The mucin MUC1 is a transmembrane glycoprotein expressed by epithelial cells at different mucosal surfaces including breast tissue, the airways and gastrointestinal tract. The full-length MUC1 protein extends 200-500 nm from the apical surface of epithelial cells and is therefore an important component of the glycocalyx[1,2]. At the mucosal surface, MUC1 has an essential barrier function against bacterial and viral invasion[3,4] but it can also be used as entry receptor by pathogenic *Salmonella* species [5]. Using knockout mice, it was demonstrated that MUC1 has anti-inflammatory functions[6–8]. However, MUC1 is most well-known for its aberrant expression and glycosylation in different types of adenocarcinomas[9].

The full-length MUC1 heterodimer consists of an extracellular domain with a variable number of tandem repeats (VNTR) of 20 amino acids, which are heavily *O*-linked glycosylated, a non-covalently attached SEA domain, a transmembrane domain, and a cytoplasmic tail with signaling capacity (see Figure 1). The VNTR region consists of repeats of 20 amino acids with the sequence GSTAPPAHGVTSAPDTRPAP[10,11]. Each repeat contains five serine and threonine residues that can be *O*-linked glycosylated and experiments with synthetic MUC1 fragments demonstrated a high glycosylation occupancy at these residues[12]. In healthy tissue, the *O*-glycans on the MUC1 VNTR predominantly consist of elongated core 2 structures, while it remains restricted to predominant core 1 structures in many cancerous cells[13,14].

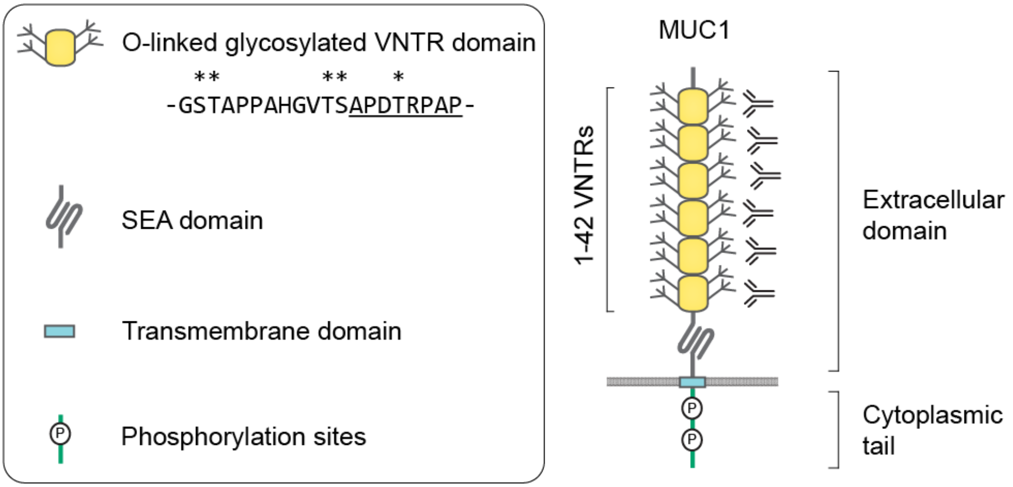


Figure 1. Schematic overview of MUC1 domain structure. VNTR: Variable Number of Tandem Repeats. SEA: domain name from initial identification in a sperm protein, enterokinase, and agrin.

The overexpression and altered glycosylation of MUC1 in cancerous cells makes it a potentially viable candidate target for cancer immunotherapy. In addition, MUC1 could be an interesting target for therapeutic strategies that require delivery to the (healthy) mucosal surface. Monoclonal antibodies against the MUC1 VNTR can be powerful tools because of their multiplicity of binding and possible applications in the diagnosis and treatment of MUC1-expressing cancers. Since the late 1980's, several monoclonal antibodies against MUC1 have been described and explored for the diagnosis and treatment of MUC1 overexpressing cancers[15,16]. Peptide mapping experiments have revealed that many such monoclonal antibodies target a similar region within the VNTR of MUC1, resulting in the definition of an immunodominant peptide corresponding to the subsequence APDTRPAP[17]. One such antibody is 139H2, a hybridoma monoclonal antibody that was raised against human breast cancer plasma membranes[15,16]. In different studies, 139H2 has been applied for the diagnostics of MUC1-overexpressing cancers and radioimmunotherapy[15,16,18]. In addition, the antibody is also widely applied as a research tool in Western blot, ELISA, immunohistochemistry and immunofluorescence microscopy to study MUC1 biology[16,19,20]. To make this antibody available for general use, we set out to determine its sequence based on the available hybridoma-derived product. Recently we have reported a method to reverse engineer monoclonal antibodies by determining the sequence directly from the purified protein product based on liquid chromatography coupled to mass spectrometry (LC-MS), using a bottom-up proteomics approach[21–24]. Here we applied this method to obtain the full sequence of 139H2. The sequence was successfully validated by comparing the performance of the reverse engineered 139H2 and its Fab fragment to the hybridoma-derived product in Western blot and immunofluorescence microscopy. Reverse engineering 139H2 enabled us to characterize binding to the immunodominant peptide epitope within the MUC1 VNTR by surface plasmon resonance (SPR) and map out the epitope by solving a crystal structure of the 139H2 Fab fragment in complex with the APDTRPAP peptide. These analyses reveal the molecular basis for 139H2 binding to MUC1 and illustrate a remarkable diversity of binding modes to the immunodominant epitope in comparison to other reported structures of anti-MUC1 monoclonals targeting the VNTR.

Methods

Purification of 139H2 from hybridoma cultures supernatant:

The 139H2 in hybridoma culture supernatant was a kind gift from John Hilkins from The Netherlands Cancer Institute (NKI). The 139H2 was purified with Protein G Sepharose 4 Fast Flow beads (Merck), washed with PBS, eluted with 0.2 mM Glycine-buffer pH 2.5, neutralized with 1 M Tris-HCL pH 8 and dialyzed against PBS with Pierce Protein Concentrators PES, 30 kDa MWCO.

Bottom-up proteomics

in-solution digestion:

139H2 was denatured in 2% sodium deoxycholate (SDC), 200 mM Tris-HCl, and 10 mM Tris(2-carboxyethyl)phosphine (TCEP), pH 8.0 at 95 °C for 10 min, followed by 30 min incubation at 37 °C for reduction. The samples were then alkylated by adding iodoacetic acid to a final concentration of 40 mM and incubated in the dark at room temperature for 45 min. 3 µg sample was then digested by one of the following proteases: trypsin (Promega) and elastase (Sigma-Aldrich) in a 1:50 ratio (w/w) in a total volume of 100 µL of 50 mM ammonium bicarbonate at 37 °C for 4 h. After digestion, SDC was removed by adding 2 µL of formic acid (FA) and centrifuged at 14000× g for 20 min. Following centrifugation, the supernatant containing the peptides was collected for desalting on a 30 µm Oasis HLB 96-well plate (Waters). The Oasis HLB sorbent was activated with 100% acetonitrile and subsequently equilibrated with 10% formic acid in water. Next, peptides were bound to the sorbent, washed twice with 10% formic acid in water, and eluted with 100 µL of 50% acetonitrile/5% formic acid in water (v/v).

in-gel digestion:

The hybridoma 139H2 was loaded on a 4%-12% Bis-Tris precast gel (Bio-Rad) in non-reducing conditions and run at 120 V in 3-Morpholinopropane-1-sulfonic acid (MOPS) buffer (Bio-Rad). Bands were visualized with Imperial Protein Stain (Thermo Fisher

Scientific), and the size of the fragments evaluated by running a protein standard ladder (Bio-Rad). The Fab bands were cut and reduced by 10 mM TCEP at 37 °C, then alkylated in 40 mM IAA at RT in the dark, followed by alkylation in 40 mM IAA at RT in the dark. The Fab bands were digested by chymotrypsin and thermolysin at 37 °C overnight in 50 mM ammonium bicarbonate buffer. The peptides were extracted with two steps incubation at RT in 50% ACN, and 0.01% TFA, and then 100% ACN respectively.

LC-MS/MS:

The peptides obtained by in-solution and in-gel digestion were vacuum-dried and reconstituted in 100 μL of 2% FA. The digested peptides were separated by online reversed-phase chromatography on an Agilent 1290 Ultra-high performance LC (UHPLC) or Dionex UltiMate 3000 (Thermo Fisher Scientific) coupled to a Thermo Scientific Orbitrap Fusion mass spectrometer. Peptides were separated using a Poroshell 120 EC-C18 2.7-Micron analytical column (ZORBAX Chromatographic Packing, Agilent) and a C18 PepMap 100 trap column (5 mm × 300, 5 μm, Thermo Fisher Scientific). Samples were eluted over a 90 min gradient from 0 to 35% acetonitrile at a flow rate of 0.3 μL/min. Peptides were analyzed with a resolution setting of 60 000 in MS1. MS1 scans were obtained with a standard automatic gain control (AGC) target, a maximum injection time of 50 ms, and a scan range of 350–2000. The precursors were selected with a 3 m/z window and fragmented by stepped high-energy collision dissociation (HCD) as well as electron-transfer high-energy collision dissociation (EThcD). The stepped HCD fragmentation included steps of 25, 35, and 50% normalized collision energies (NCE). EThcD fragmentation was performed with calibrated charge-dependent electron-transfer dissociation (ETD) parameters and 27% NCE supplemental activation. For both fragmentation types, MS2 scans were acquired at a 30 000 resolution, a 4e5 AGC target, a 250 ms maximum injection time, and a scan range of 120–3500.

peptide sequencing from MS/MS Spectra:

MS/MS spectra were used to determine *de novo* peptide sequences using PEAKS Studio X (version 10.6)[25,26]. We used a tolerance of 20 ppm and 0.02 Da for MS1 and MS2, respectively. Carboxymethylation was set as fixed modification of cysteine and variable modification of peptide N-termini and lysine. Oxidation of methionine and tryptophan and pyroglutamic acid modification of N-terminal glutamic acid and glutamine were set as additional variable modifications. The CSV file containing all the *de novo* sequenced peptide was exported for further analysis.

template-based assembly via Stitch:

Stitch (nightly version 1.4.0+802a5ba) was used for the template-based assembly[27]. The mouse antibody database from IMGT was used as template[28]. The cutoff score for the *de novo* sequenced peptide was set as 90 and the cutoff score for the template matching was set as 10. All the peptides supporting the sequences were examined manually.

Cloning and Expression of recombinant 139H2 IgG and Fab:

To recombinantly express full-length anti-MUC1 antibodies, the proteomic sequences of both the light and heavy chains were reverse-translated and codon-optimized for expression in human cells using the Thermo Fisher webtool (https://www.thermofisher.com/order/gene-design/index.html). For the linker and Fc region of the heavy chain, the standard mouse Ig γ-1 (IGHG1) amino acid sequence (Uniprot P01868.1) was used. An N-terminal secretion signal peptide derived from human IgG light chain (MEAPAQLLFLLLLWLPDTTG) was added to the N-termini of both heavy and light chains. BamHI and NotI restriction sites were added to the 5′ and 3′ ends of the coding regions, respectively. Only for the light chain, a double stop codon was introduced at the 3′ site before the NotI restriction site. The coding regions were subcloned using BamHI and NotI restriction-ligation into a pRK5 expression vector with a C-terminal octahistidine tag between the NotI site and a double stop codon 3′ of the insert, so that only the heavy chain has a C-terminal AAAHHHHHHHH sequence for nickel-affinity purification (the triple alanine resulting from the NotI site). After the sequence was validated by Sanger Sequencing, the HC/LC were mixed in a 1:1 DNA ratio and expressed in HEK293 cells by the ImmunoPrecise Antibodies (Europe) B.V company. After expression the culture supernatant of the cells was harvested and purified using a prepacked HisTrap excel column (Cytiva), following standard protocols. (see Supplementary Figure S2))

To recombinantly express anti-MUC1 Fab the coding regions of HC variable region were subcloned using AgeI and NheI restriction-ligation into a pRK5 expression vector. The subcloned region contains the mouse Ig γ-1 (IGHG1) Fab constant region with a C-terminal octahistidine tag followed by a double stop codon 3′ of the insert, so that only the heavy chain has a C-terminal AAAHHHHHHHH sequence for nickel-affinity purification (the triple alanine resulting from the NotI site). After the sequence was validated by Sanger Sequencing the HC/LC were mixed in a 1:1 (*m*/*m*) DNA ratio and expressed in HEK293 cells by the ImmunoPrecise Antibodies (Europe) B.V company. After expression the culture supernatant was loaded onto a 5 ml HisTrap excel column (Cytiva) using peristatic pump. Column was reconnected to the ÄktaGo system (Cytiva) for column wash (50 mM Tris at pH=8, 150 mM NaCl) and step elution (50 mM Tris at pH=8, 150 mM NaCl, 300 mM imidazole). Fraction from the peak corresponding to the Fab were concentrated using Amicon Ultra-15 (Millipore) and further purified by size-exclusion chromatography using Superdex 200 Increase 10/300 GL (Cytiva) in buffer 50 mM Tris (pH=8), 150 mM NaCl.

Mammalian cell lines and culture conditions:

The human gastrointestinal epithelial cell lines HT29-MTX[29] and HT29-MTX ΔMUC1[5] were cultured in 25 cm^2 flasks in Dulbecco's modified Eagle's medium (DMEM) containing 10% fetal calf serum (FCS) at 37 °C in 10% CO^2.

Western blot:

HT29-MTX and HT29-MTX ΔMUC1 lysates were prepared form cells grown to full confluency for 7 days in a 6-well plate. Cells were harvested by scraping and lysed with lysis buffer (10% SDS in PBS with 1× Halt Protease Inhibitor Cocktail). Concentration was measured by BCA-assay, 5× Laemmli buffer was added and sample was boiled for 15 min at 95 °C. A mucin-SDS gel was made according to Li et al.[5]; 40 µg of protein was added to each well and run in Boric acid-Tris buffer (192 mM Boric acid, Merck; 1 mM EDTA, Merck; 0.1% SDS, to pH 7.6 with Tris) at 25 mA for 1.5 h. Proteins were transferred to a polyvinylidene fluoride (PVDF) membrane using wet transfer for 3 h at 90 V/4 °C in transfer buffer (25 mM Tris; 192 mM glycine, Merck; 20% methanol, Merck). Afterwards, membranes were blocked with 5% BSA in TSMT (20 mM Tris; 150 mM NaCl, Merck; 1 mM $CaCl_2$ (Sigma); 2 mM $MgCl_2$, Merck; adjusted to pH 7 with HCl; 0.1% Tween 20 (Sigma)) overnight at 4 °C. The following day, membranes were washed with TSMT and incubated with 139H2 Wildtype, Synthetic or FAB antibodies (1:1000) in TSMT containing 1% BSA for 1h at RT. Membranes were washed again with TSMT and incubated with α-mouse IgG secondary antibody (A2304, Sigma) diluted 1:8000 in TSMT with 1% BSA for 1 h at RT, washed with TSMT followed by TSM. For detection of actin, cell lysates were loaded onto a 10% SDS-PAGE gel, transferred to PVDF membranes and incubated with α-Actin antibody (1:2,000; bs-0061R, Bioss) and α-rabbit IgG (1: 10,000; A4914, Sigma). Blots were developed with the Clarity Western ECL kit (Bio-Rad) and imaged in a Gel-Doc system (Bio-Rad).

Western blot of MUC1 reporter constructs:

Four MUC1 reporter constructs, expressed in engineered HEK293 cells, were a kind gift from Chistian Büll of the Copenhagen Center for Glycomics. Each reporter construct in 1× PBS was boiled in 5× laemmli buffer. 10 ng/25 ng of each construct was loaded per well on a 10% bis-acrylamide SDS gel for the 139H2/6× His-tag blots respectively. Samples were run in 1× Novex Tris-Glycine SDS Running Buffer (Thermo Fisher Scientific) for 1.5 h at 120 V. Proteins were transferred to a 0.2 μm Trans-Blot PVDF membrane (Bio-Rad) and transferred at 1.3 A/25 V for 7 min using the Trans-Blot Turbo system (Bio-Rad). Afterwards, membranes were blocked with 5% BSA in TSMT (20 mM Tris; 150 mM NaCl, Merck; 1 mM $CaCl_2$, Sigma; 2 mM $MgCl_2$, Merck; adjusted to pH 7 with HCl; 0.1% Tween 20, Sigma) overnight at 4 °C. The following day, membranes were washed with TSMT and incubated with 139H2 Wiltype, Synthetic antibody (1:1000) or HisProbe-HRP Conjugate (15165,Thermo Fisher Scientific,1:5000) in TSMT containing 1% BSA for 1h at RT. The 6× His-tag blots were washed with TMST and TSM and developed with the Clarity Western ECL kit (Bio-Rad) and imaged in a Gel-Doc system (Bio-Rad). The 139H2 membranes were washed again with TSMT and incubated with α-mouse IgG secondary antibody (A2304, Sigma) diluted 1:8000 in TSMT with 1% BSA for 1 h at RT, washed with TSMT followed by TSM and developed.

Confocal microscopy:

HT29-MTX and HT29-MTX ΔMUC1 cells were grown for 7 days to reach a confluent monolayer on cover slips (8 mm diameter#1.5) in 24-well plates. Cells were washed with Dulbecco's Phosphate Buffered Saline (DPBS, D8537) and fixed with 4% paraformaldehyde in PBS (Affimetrix) for 30 min at RT. Fixation was stopped by adding

50 mM NH_4Cl in PBS for 15 min. Cells were washed 2 times and permeabilized in binding buffer (0.1% saponin (Sigma) and 0.2% BSA (Sigma) in DPBS) for 30 min. Coverslips were incubated with 139H2 Wildtype, Synthetic of FAB at 1:100 dilution for 1h, washed 3× with binding buffer, incubated with Alexa Fluor-488-conjugated α-mouse IgG secondary antibodies (1:200; A11029, ThermoFisher) and DAPI at 2 μg/ml (D21490, Invitrogen) for 1 h. Coverslips were washed 3× with DPBS, desalted in MiliQ, dried and embedded in Prolong Diamond mounting solution (ThermoFisher) and allowed to harden. Images were collected on a Leica SPE-II confocal microscope with a 63× objective (NA 1.3, HCX PLANAPO oil). Controlled by Leica LAS AF software with default settings to detect DAPI, Alexa488, Alexa568 and Alexa647. Axial series were collected with step sizes of 0.29 μm.

Surface Plasmon Resonance:

N-terminally biotinylated synthetic MUC1 peptide with the sequence biotin-GGS-APDTRPAPG was ordered from Genscript. This was dissolved in PBS and printed on a planar streptavidin-coated SPR chip (P-Strep, SSens B.V.) using a continuous flow microfluidics spotter (Wasatch), flowing for 1 hour at RT, after which it was washed with SPR buffer (150 mM NaCl, 25 mM 4-(2-hydroxyethyl)-1-piperazineethanesulphonic acid (HEPES) with 0.005% Tween 20) for 15 min and quenched with biotin solution (10 mM biotin in SPR buffer). SPR experiments were performed using an IBIS-MX96 system (IBIS technologies) with SPR buffer as the running buffer. Dilution series of 2× steps of the full recombinant 139H2 or Fab were prepared, starting from a 10.0 μM stock for full IgG and a 7.88 μM stock for the Fab, diluting with SPR buffer. 20 dilution steps (including the stock) were used for the full IgG, and 10 dilutions were used for the Fab. SPR experiments were performed as a kinetic titration without regenerating in between association/dissociation cycles, with 30 min association and 10 min dissociation time for the full IgG and 6 min association and 4 min dissociation for the Fab. Binding affinity was determined by fitting data at binding equilibrium to a 2-site binding model for the full IgG and a 1-site (Langmuir) binding model for the Fab, using Scrubber 2.0 (Biologic software) and Graphpad Prism 5 (Graphpad software, Inc.).

Crystallization and data collection:

Sitting-drop vapor diffusion crystallization trials were set up at 20 °C by mixing 150 nl of complex with 150 nl of reservoir solution. The complex sample consisted of purified 139H2 Fab and MUC1 epitope peptide (APDTRPAPG; GeneScript) in a 1:2.5 molar ratio, at a total concentration of 3.8 mg/mL in a buffer of 50 mM trisaminomethane at pH 8.0 and 150 mM NaCl. The diffracting crystals grew in a condition of 0.2 M NaCl, 0.1 M sodium phosphocitrate, and 20% w/v Polyethylene glycol (PEG) 8000 used as reservoir solution. A 3:1 mixture of reservoir solution and glycerol was added as cryo-protectant to the crystals before plunge freezing them in liquid nitrogen. Datasets were collected at 100 K at Diamond Light Source beamline I24, equipped with an Eiger 9M detector (Dectris), at a wavelength of 0.6199 Å.

Structure determination and refinement:

Collected datasets were integrated using the xia2.multiplex pipeline[30], and the three best datasets were subsequently merged and scaled in AIMLESS to a maximum resolution of 2.5 Å. Resolution limit cut off was determined based on mean intensity correlation coefficient of half-data sets, $CC_{1/2}$. An initial model of 139H2 Fab was generated using ColabFold[31]. The variable region and constant region were placed in subsequent PHASER[32], the short linkers between the two regions were built manually and the CDRs were adjusted in COOT[33]. Clear density for the MUC1 peptide was present in the Fo-Fc map, and the peptide was built manually in COOT. The structure was refined by iterative rounds of manual model building in COOT and refinement in REFMAC5[34]. The final model was assessed using MolProbity[35]. All programs were used as implemented in CCP4i2 version 1.1.0[36].

Results

De novo sequencing by bottom-up mass spectrometry

The goal of our study was to obtain the sequence of the full length 139H2 IgG antibody using a bottom-up proteomics approach. As a starting point, we used 139H2 IgG hybridoma supernatant and purified the antibody using protein G affinity resin. The purified IgG was digested with a panel of 4 proteases in parallel (trypsin, chymotrypsin, α-lytic protease, and thermolysin) to generate overlapping peptides for the LC-MS/MS analysis, using a hybrid fragmentation scheme with stepped high-energy collision dissociation (sHCD) and electron-transfer high energy collision dissociation (EThcD) on all peptide precursors. The peptide sequences were predicted from the MS/MS spectra using PEAKS and assembled into the full-length heavy and light chain sequences using the in-house developed software Stitch. This resulted in the identification of a mouse IgG1 antibody with an IGHV1-53 heavy chain paired with an IGKV8-30 light chain (the full sequence is provided in the Supplementary Information). The depth of coverage for the complementarity determining regions (CDRs) varies from around 10 to 100, indicating a high sequence accuracy (see Supplementary Figure S1). Examples of MS/MS spectra supporting the CDRs of both heavy chain and light chain are shown in Figure 2. Comparison to the inferred germline precursors indicate a typical moderate level of somatic hypermutation (3% in the light chain; 10% in the heavy chain), with some notable mutations in the framework regions, also directly flanking CDRH2.

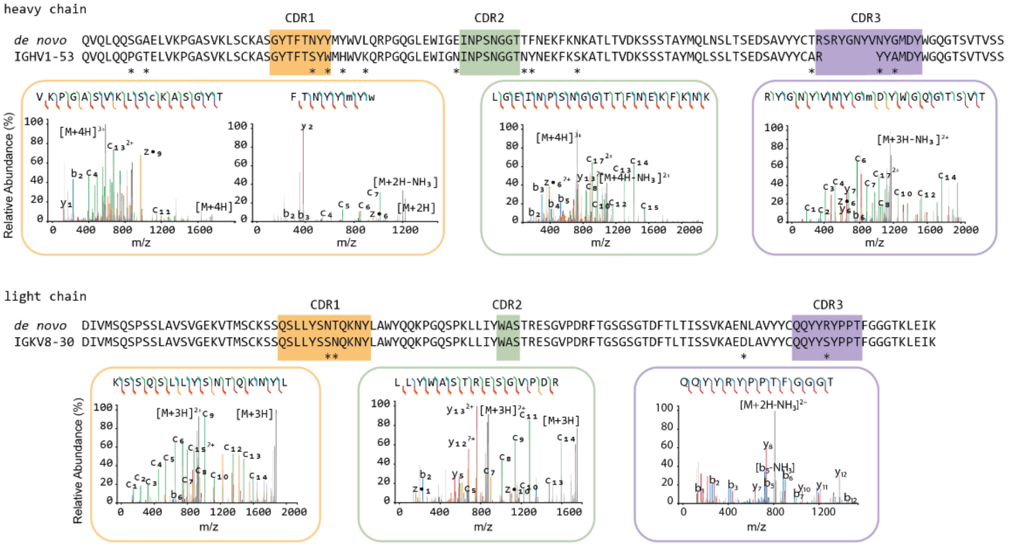


Figure 2. De novo sequencing of the hybridoma 139H2 based on bottom-up proteomics. The variable region alignment to the inferred germline sequence is shown for both heavy and light chains. Positions with putative somatic hypermutation are highlighted with asterisks (*). The MS/MS spectra supporting the CDR regions are shown beneath the sequence alignment, b/y ions are indicated in blue and red, while c/z ions are indicated in green and yellow.

Validation of the experimentally determined 139H2 sequence

The experimentally determined sequences of the 139H2 variable domains were codon optimized for mammalian expression and subcloned into expression vectors with the mouse IgG1 heavy chain (with an 8xHis-tag) and the kappa light chain backbones (see Supplementary Information for the full amino acid sequences). Co-transfection of the two plasmids in HEK293 cells yielded ca. 10 mg from a 1 L culture following His-trap purification (see Supplementary Figure S2). Additionally, the Fragment antigen-binding (Fab) region was expressed to study the monovalent binding to MUC1. The recombinant 139H2 and Fab were then compared with the hybridoma-derived 139H2 in Western blot and confocal immunofluorescence microscopy.

To investigate the specificity of the recombinant 139H2 antibody for MUC1, we performed immunoblot analysis on lysates of the methotrexate-adapted human colon cancer cell line HT29-MTX, known for its high MUC1 expression, and a MUC1 knockout of the same cell line that was previously described (see Figure 2)[5]. The original hybridoma-derived 139H2 recognizes one predominant band at an estimated molecular weight of 600 kDa, corresponding to full length MUC1, and this band is absent in lysates of the MUC1-knockout cells. The recombinant 139H2 showed the same binding pattern. In confocal immunofluorescence microscopy, original hybridoma-derived 139H2 stains MUC1 at the apical surface in a confluent culture of HT29-MTX, and this signal is reduced to background in the MUC1-knockout cell line. A similar staining is observed with the recombinant 139H2. Western blot and immunofluorescence microscopy using the monovalent Fab fragment also showed specific binding to MUC1 in the wild type background but with reduced avidity compared to the full bivalent IgG molecule. These results confirm that the reverse engineered 139H2 antibody is functional and recognizes the full length MUC1 glycoprotein at the apical surface of intestinal epithelial cells.

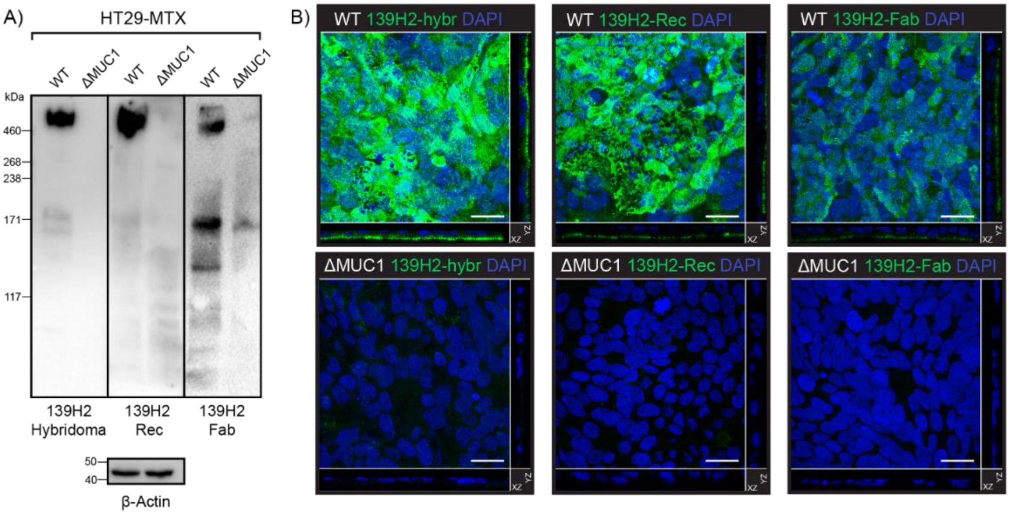


Figure 3. Validation of synthetic recombinant 139H2 following the mass spectrometry-derived sequence. (A) Immunoblot analysis of lysates of intestinal epithelial HT29-MTX and HT29-MTX ΔMUC1 cells with the original hybridoma-derived 139H2 IgG antibody and synthetic recombinant 139H2. (B) Immunofluorescence confocal microscopy imaging of confluent HT29-MTX and HT29-MTX ΔMUC1 monolayers. Cells were stained for nuclei (DAPI, blue) and MUC1 (139H2, green). The signal of the 139H2 Fab was enhanced to compensate for the expected low signal/binding. White scale bars represent 20 µm.

Epitope mapping of 139H2

Using the reverse engineered 139H2 product, we next characterized binding to the immunodominant epitope APDTRPAPG within the MUC1 VNTR. Binding to the synthetic peptide, including an N-terminal biotin and short peptide linker for immobilization to the SPR substrate (*i.e.* biotin-GGS-APDTRPAPG), was determined by SPR. Binding of the full IgG was characterized by a high and low affinity phase with dissociation constants of 17×10^{-9} M and 43×10^{-7} M, respectively (Figure S3). We interpret this biphasic binding as an avidity-enhanced bivalent mode (both Fab arms engaged with epitope, high affinity), and a monovalent mode (single Fab arm, low affinity) of binding, respectively. In line with this interpretation, binding to a recombinant monovalent 139H2 Fab yielded a dissociation constant of 45×10^{-7} M, similar to the low affinity binding phase of the full IgG.

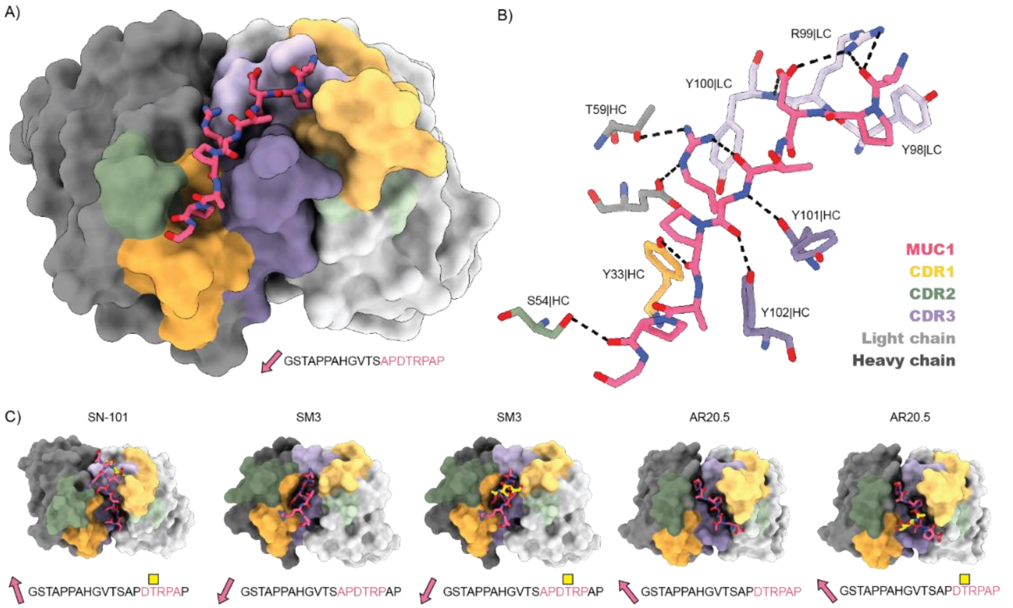


Figure 4. Structure of 139H2 Fab in complex with MUC1 peptide. (A) Surface representation of *the Fab with CDRs highlighted in colours and MUC1 peptide shown as a model. N- to C-terminus direction of MUC1 peptides is shown as a pink arrow. (B) Interactions of interactions between 139H2 Fab and MUC1 peptide. (C) Comparison with previously reported structures of monoclonal anti-MUC1 antibodies targeting the VNTR. Glycosylated residues of the epitope are depicted by yellow square above.*

To better understand the molecular basis of 139H2 binding to the immunodominant epitope within the VNTR we determined a crystal structure of the Fab fragment in complex with the synthetic APDTRPAPG peptide (without N-terminal biotin or peptide linker). Crystals diffracted to a resolution of 2.5 Å and a structure was solved using molecular replacement with a ColabFold model of the 139H2 Fab. This also revealed clear density for the peptide epitope in contact with the CDRs of 139H2 (see Supplementary Table S1 and Supplementary Figure S4).

The APDTRPAPG peptide binds diagonally across the cleft between the heavy and light chains, making direct contact with all CDRs, except CDRL2 (see Figure 4 and Supplementary Table S2). Contact points between the peptide and the 139H2 Fab include hydrogen bonds with the peptide backbone at 6 out of 8 positions. Both the aspartic acid and arginine residues within the epitope make salt bridges with side chains from 139H2. While D3 interacts with R99 within CDRL1, R5 interacts with E50 and T59 near CDRH2, in addition to a stacking interaction with Y100 in CDRL3. Neither residue E50 nor T59 in 139H2 is formally part of CDRH2, though both residues directly flank the loop. Previous studies on the binding specificity of 139H2 have shown that R5 of the epitope is crucial for 139H2 binding. The crystal structure reported here shows that interactions with R5 are mediated by residues in 139H2 that are formally part of the framework regions of the heavy chain, but both mutated compared to the inferred germline precursors (see Figure 2). Two additional framework mutations in the heavy chain, *i.e.* Y35 and T97, appear indirectly involved in MUC1 binding by positioning CDRH3 through hydrogen bonds with N106 and the backbone of Y111, respectively (see Supplementary Figure S5). Finally, the T4 residue of the APDTRPAPG epitope is a known glycosylation site, although 139H2 binding is reported to be unaffected by the presence of a single *O*-linked GalNAc at this position[14,37]. The crystal structure reported here shows the T4 side chain to be pointing outwards from the 139H2 paratope with no indication of potential clashes that would preclude binding of the epitope with glycosylated APDTRPAPG at the T4 position. In line with this previous report and our own structural data, we also found that 139H2 binds equally well to MUC1 reporter constructs with different types of *O*-linked glycans (Supplementary Figure S6).

Comparison with previously reported structures of monoclonal anti-MUC1 antibodies targeting the VNTR reveal a striking diversity in the modes of binding (a full overview of reported structures is listed in Supplementary Table S3)[38–48]. Monoclonal antibodies 14A, 16A, and 5E5 all target a different region within the VNTR. While monoclonal antibodies SM3, SN101, and AR20.5 all bind to the same immunodominant epitope of the VNTR as 139H2, the peptide is either shifted or oriented in the opposite direction relative to the cleft between the heavy and light chains. For SN101 and AR20.5, the peptide runs across this cleft in the opposite direction compared to 139H2. In SM3 the peptide is oriented in a similar direction but shifted by approximately 2 residues such that both D3 and R5 are contacting different CDRs. In contrast to 139H2, each of the monoclonals compared above bind stronger to the glycosylated epitope. In the case of AR20.5 and SN101 this specificity can be explained by direct contacts made between the glycan and CDRs of the antibody. However, for SM3 the orientation of the glycosylated T4 residue is more similar to 139H2.

In SM3 the GalNAc residue makes an additional hydrogen bond with a tyrosine in CDRL1. A similar interaction is predicted for 139H2, albeit through a different group of the GalNAc residue (see Supplementary Figure S7).

Discussion

Our study demonstrates how direct mass spectrometry-based protein sequencing enables the reconstruction of antibodies from hybridoma supernatants. In addition to recovering such precious resources for research and therapeutic applications, it also contributes to open and reproducible science by making the sequences of crucial monoclonal antibody reagents more readily available and accessible. Poorly defined (monoclonal) antibody products have notoriously been a challenge to reproducibility in life science research and the present work shows that MS-based sequencing can offer helpful improvements in this regard[49,50].

The reverse-engineered anti-MUC1 monoclonal antibody 139H2 reported here is suitable for Western blotting and immunofluorescence microscopy and is likely suitable for other applications in FACS sorting of MUC1 positive cells, immunohistochemistry and ELISA, as demonstrated for the original hybridoma-derived product[16,19,20]. We show that 139H2 binds the immunodominant epitope of the VNTR in a unique way compared to previously described monoclonal antibodies against MUC1. Because of its previously reported glycan-independent binding, which we supported in this study by the determined structure in complex with the epitope, the 139H2 antibody is an important tool for current and future MUC1 research.

Acknowledgements

This research was funded by the Dutch Research Council NWO Gravitation 2013 BOO, Institute for Chemical Immunology (ICI; 024.002.009) to J.S. KS and KCAPG are supported by the European Research Council under the European Union’s Horizon 2020 research and innovation program (ERC-2019-STG-852452). The authors would like to thank Diamond Light Source for beamtime (proposal mx25413), and the staff of beamline I24 for assistance with data collection.

Data Availability

The raw LC-MS/MS files and analyses have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD043489. Coordinates and structure factors for 139H2 bound to the MUC1 epitope peptide have been deposited to the Protein Data Bank with accession code 8P6I.

Supplementary Information:

>139H2 Heavy Chain

QVQLQQSGAELVKPGASVKLSCKASGYTFTNYYMYWVLQRPGQGLEWIGEINPSNGGTTFNEK
FKNKATLTVDKSSSTAYMQLNSLTSEDSAVYYCTRSRYGNYVNYGMDYWGQGTSVTVSSASTT
PPSVYPLAPGSAAQTNSMVTLGCLVKGYFPEPVTVTWNSGSLSSGVHTFPAVLQSDLYTLSSSV
TVPSSPRPSETVTCNVAHPASSTKVDKKIVPRDCGCKPCICTVPEVSSVFIFPPKPKDVLTITLTP
KVTCVVVDISKDDPEVQFSWFVDDVEVHTAQTQPREEQFNSTFRSVSELPIMHQDWLNGKEF
KCRVNSAAFPAPIEKTISKTKGRPKAPQVYTIPPPKEQMAKDKVSLTCMITDFFPEDITVEWQW
NGQPAENYKNTQPIMNTNGSYFVYSKLNVQKSNWEAGNTFTCSVLHEGLHNHHTEKSLSHSP
GK

>139H2 Light Chain

DIVMSQSPSSLAVSVGEKVTMSCKSSQSLLYSNTQKNYLAWYQQKPGQSPKLLIYWASTRESGV
PDRFTGSGSGTDFTLTISSVKAENLAVYYCQQYYRYPPTFGGGTKLEIRRADAAPTVSIFPPSSEQ
LTSGGASVVCFLNNFYPKDINVKWKIDGSERQNGVLNSWTDQDSKDSTYSMSSTLTLTKDEYE
RHNSYTCEATHKTSTSPIVKSFNRNEC

Supplementary Table S1. Details of X-ray data collection, processing, and structure refinement.

PDB accession code	**8P6I**
Data collection and processing	
Space group	*C*2
a, b, c (Å)	211.8, 42.8, 129.1
α, β, γ (deg)	90.0, 122.3, 90.0
Wavelength (Å)	0.6199
Resolution (Å)	56.05-2.50 (2.60-2.50)*
Rmerge	0.387 (1.813)
Rpim	0.093 (0.673)
CC(1/2)	0.988 (0.689)
No. of observations	630424 (58514)
No. unique	34554 (3836)
Mean I/σ (I)	9.0 (0.7)
Completeness (%)	100.0 (100.0)
Redundancy	18.2 (15.2)
Structure Refinement	
Rwork/Rfree	0.202/0.253
Model composition	
Non-hydrogen atoms	6719
Protein residues	854
Water	229
B factors (Å^2)	
Protein	33.6
R.m.s. deviations	
Bond lengths (Å)	0.0086
Bond angles (°)	1.53
Validation	
MolProbity score	1.53
Clash score	2.27
Poor rotamers (%)	3.12
Ramachandran plot	
Favoured (%)	97.19
Allowed (%)	2.69
Outliers (%)	0.12

*Values in parenthesis are for the highest resolution shell.

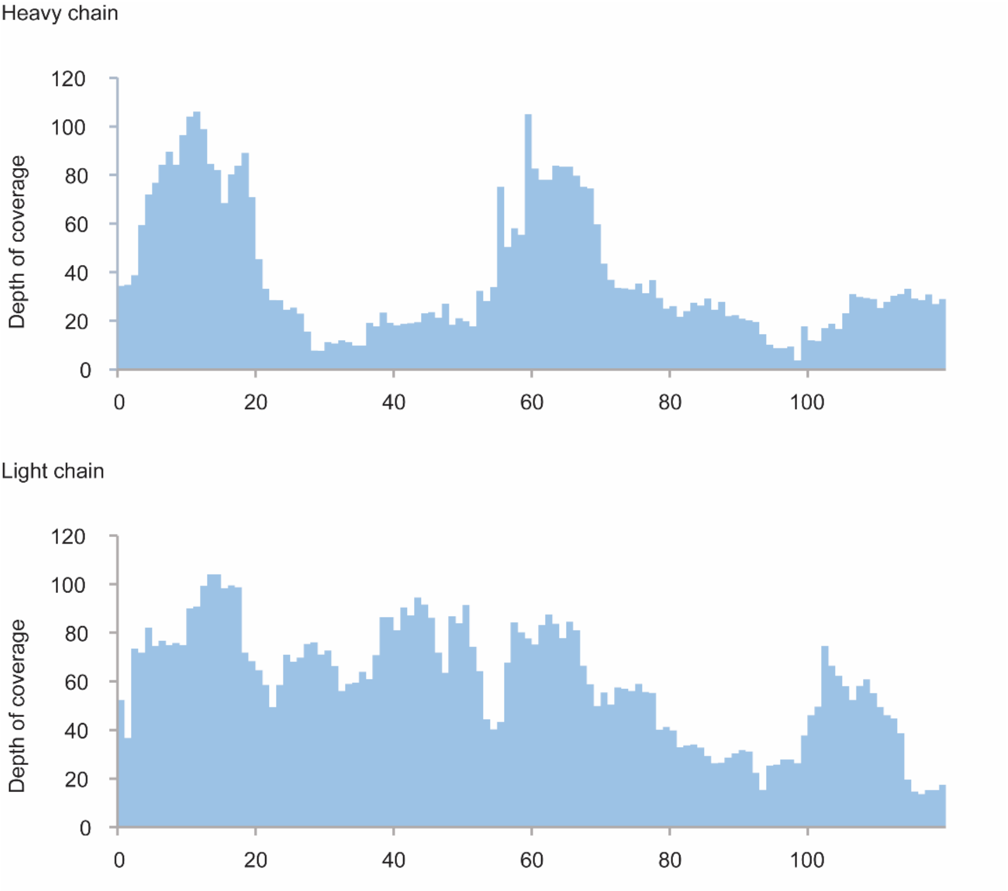
Supplementary Table S2. Overview of 139H2-MUC1 epitope contacts observed in crystal structure.

MUC1 peptide	group	139H2	group	interaction
Ala1	backbone	Arg99|LC	sidechain	hydrogen bond
Pro2	sidechain	Tyr98|LC	sidechain	stacking
Asp3	backbone	Tyr98|LC	backbone	hydrogen bond
	sidechain	Arg99|HC	sidechain	salt bridge
	sidechain	Tyr100|LC	backbone	hydrogen bond
Thr4	-	-	-	-
Arg5	backbone	Tyr101|HC	backbone	hydrogen bond
	backbone	Tyr102|HC	sidechain	hydrogen bond
	sidechain	Glu50|HC	sidechain	salt bridge
	sidechain	Thr59|HC	sidechain	hydrogen bond
	sidechain	Tyr100|LC	sidechain	stacking
Pro6	backbone	Tyr33|HC	sidechain	hydrogen bond
Ala7	-	-	-	-
Pro8	backbone	Ser54|HC	sidechain	hydrogen bond
Gly9	-	-	-	-

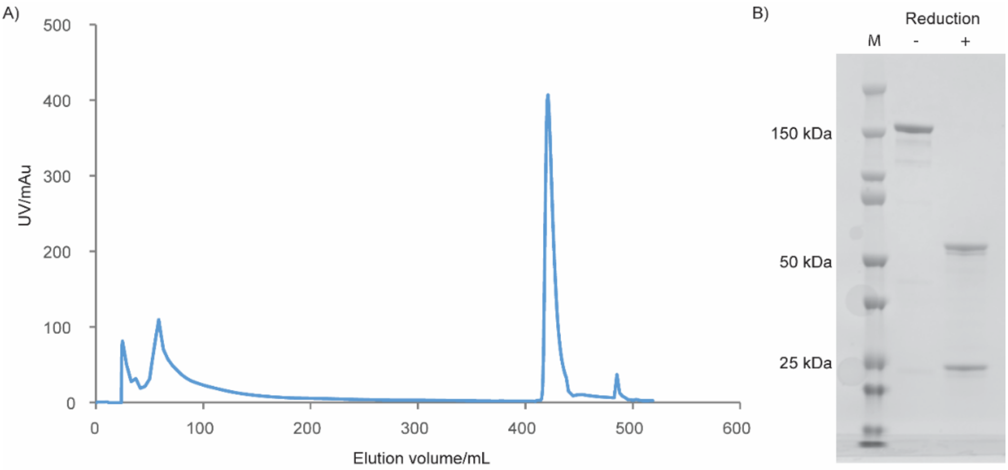
Supplementary Table S3. Overview of reported MUC1-Fab structures. Structures in yellow contain O-glycopeptide, green unglycosylated peptide. The final column highlights the bound peptides in the context of MUC1's repeat region.

PDB ID	**mAb**	**peptide**	**remarks**	**ref**	**VNTR (2x repeat GSTAPPAHGVTSAPDTRPAP)**
7VAZ	14A	RPAPGS(GalNAc)TAPPAHG	higher affinity for glycopeptide	https://doi.org/10.1101/2022.07.24.501275	GSTAPPAHGVTSAPDT**RPAPGSTAPPAHG**VTSAPDTRPAP
7V8Q	14A	RPAPGST(GalNAc)APPAHG	higher affinity for glycopeptide	https://doi.org/10.1101/2022.07.24.501275	GSTAPPAHGVTSAPDT**RPAPGSTAPPAHG**VTSAPDTRPAP
7VAC	14A	RPAPGS(GalNAc)T(GalNAc)APPAHG	higher affinity for glycopeptide	https://doi.org/10.1101/2022.07.24.501275	GSTAPPAHGVTSAPDT**RPAPGSTAPPAHG**VTSAPDTRPAP
7V4W	16A	RPAPGSTAPPAHG	higher affinity for glycopeptide	https://doi.org/10.1101/2022.07.24.501275	GSTAPPAHGVTSAPDT**RPAPGSTAPPAHG**VTSAPDTRPAP
7V64	16A	RPAPGST(GalNAc)APPAHG	higher affinity for glycopeptide	https://doi.org/10.1101/2022.07.24.501275	GSTAPPAHGVTSAPDT**RPAPGSTAPPAHG**VTSAPDTRPAP
7V7K	16A	RPAPGS(GalNAc)T(GalNAc)APPAHG	higher affinity for glycopeptide	https://doi.org/10.1101/2022.07.24.501275	GSTAPPAHGVTSAPDT**RPAPGSTAPPAHG**VTSAPDTRPAP
6TNP	5E5	APGST(GalNAc)AP	higher affinity for glycopeptide	https://doi.org/10.1039/D0CC06349E	GSTAPPAHGVTSAPDTRP**APGSTAP**PAHGVTSAPDTRPAP
6KX1	SN-101	VTSAPDT(GalNAc)RPAPGSTA	higher affinity for glycopeptide	https://doi.org/10.1039/D0SC00317D	GSTAPPAHG**VTSAPDTRPAPGSTA**PPAHGVTSAPDTRPAP
5T6P	AR20.5	APDTRPAP	higher affinity for glycopeptide	https://doi.org/10.1093/glycob/cww131	GSTAPPAHGVTS**APDTRPAP**GSTAPPAHGVTSAPDTRPAP
5T78	AR20.5	APDT(GalNAc)RPAP	higher affinity for glycopeptide	https://doi.org/10.1093/glycob/cww131	GSTAPPAHGVTS**APDTRPAP**GSTAPPAHGVTSAPDTRPAP
5A2J	SM3	APDTRP	higher affinity for glycopeptide	https://doi.org/10.1002/anie.201502813	GSTAPPAHGVTS**APDTRP**APGSTAPPAHGVTSAPDTRPAP
5A2K	SM3	APDT(GalNAc)RP	higher affinity for glycopeptide	https://doi.org/10.1002/anie.201502813	GSTAPPAHGVTS**APDTRP**APGSTAPPAHGVTSAPDTRPAP
5A2I	SM3	APDS(GalNAc)RP	higher affinity for glycopeptide	https://doi.org/10.1002/anie.201502813	GSTAPPAHGVTS**APDTRP**APGSTAPPAHGVTSAPDTRPAP
5A2L	SM3	APDC(GalNAc)RP	higher affinity for glycopeptide	https://doi.org/10.1002/anie.201502813	GSTAPPAHGVTS**APDTRP**APGSTAPPAHGVTSAPDTRPAP
5N7B	SM3	APDT(GalNAc)RP	higher affinity for glycopeptide; substitution in glycosidic linkage	https://doi.org/10.1021/jacs.8b13503	GSTAPPAHGVTS**APDTRP**APGSTAPPAHGVTSAPDTRPAP
6FRJ	SM3	APDT(GalNAc)RP	higher affinity for glycopeptide; substitution in glycosidic linkage	https://doi.org/10.1021/jacs.8b13503	GSTAPPAHGVTS**APDTRP**APGSTAPPAHGVTSAPDTRPAP
6TGG	SM3	APDT(GalNAc)RP	higher affinity for glycopeptide; with iminosugar	https://doi.org/10.1039/C9SC06334J	GSTAPPAHGVTS**APDTRP**APGSTAPPAHGVTSAPDTRPAP

6FZR	SM3	APDT(GalNAc)RP	higher affinity for glycopeptide; with fluorinated glycan	https://doi.org/10.1021/jacs.8b04801	GSTAPPAHGVTS**APDTRP**APGSTAPPAHGVTSAPDTRPAP
6FZQ	SM3	APDT(GalNAc)RP	higher affinity for glycopeptide; with fluorinated glycan	https://doi.org/10.1021/jacs.8b04801	GSTAPPAHGVTS**APDTRP**APGSTAPPAHGVTSAPDTRPAP
5FXC	SM3	APDT(GalNAc)RP	higher affinity for glycopeptide; glycan connected with linker	https://doi.org/10.1021/acs.joc.6b00833	GSTAPPAHGVTS**APDTRP**APGSTAPPAHGVTSAPDTRPAP
5OWP	SM3	GVTSA(2fP)DT(GalNAc)RPAP	higher affinity for glycopeptide; with fluorinated proline	https://doi.org/10.1021/jacs.7b09447	GSTAPPAH**GVTSAPDTRPAP**GSTAPPAHGVTSAPDTRPAP
1SM3	SM3	TSAPDTRPAPGST	higher affinity for glycopeptide	https://doi.org/10.1006/jmbi.1998.2209	GSTAPPAHGV**TSAPDTRPAPGST**APPAHGVTSAPDTRPAP
8P6I	139H2	APDTRPAPG		current work	GSTAPPAHGVTS**APDTRPAPG**STAPPAHGVTSAPDTRPAP

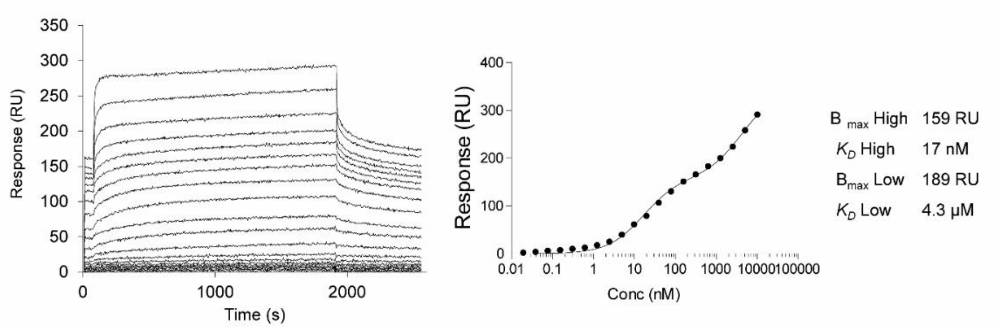


Supplementary Figure S1. Depth of coverage (total number of overlapping peptides mapped per position) for the variable domains of the 139H2 heavy and light chains.

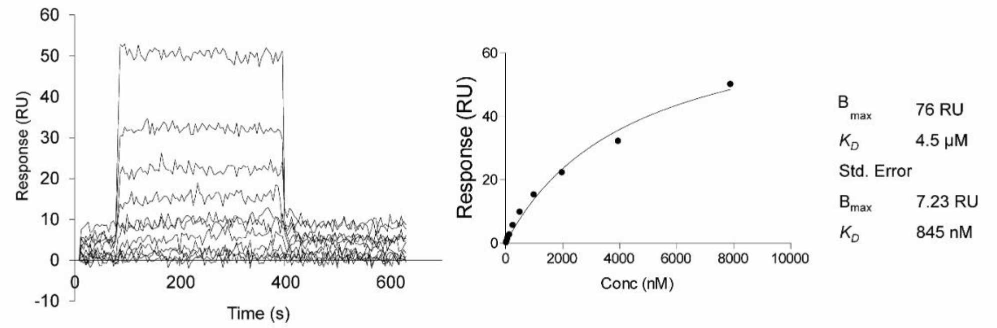


Supplementary Figure S2. Purification of recombinant 139H2. A) Elution profile across the imidazole gradient from the His-Trap purification. B) SDS-PAGE of the purified IgG product under reducing/non-reducing conditions.

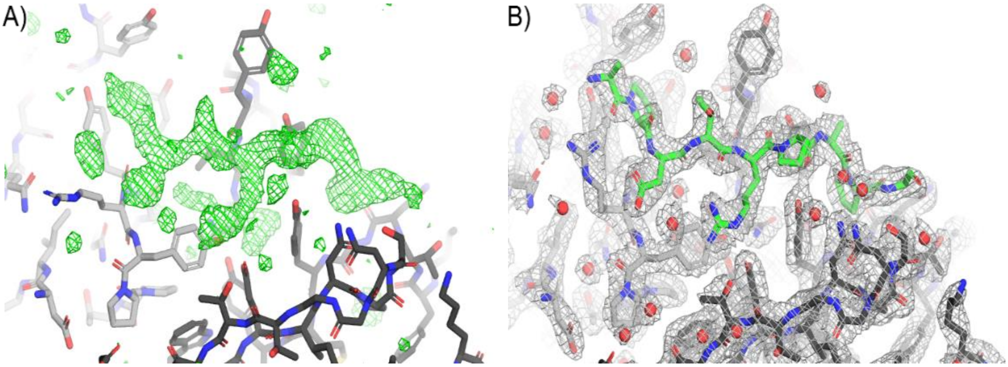
A) 139H2 IgG



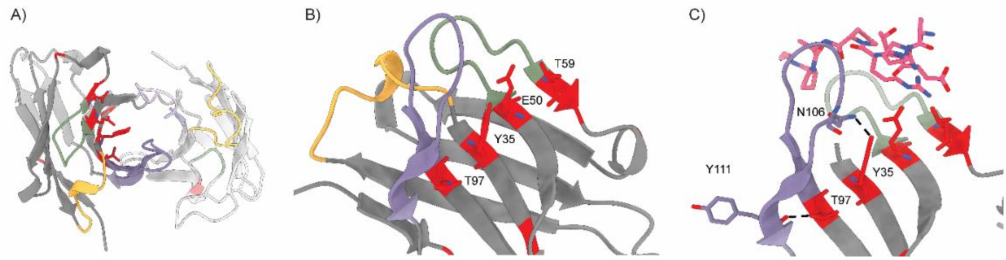
B) 139H2 Fab



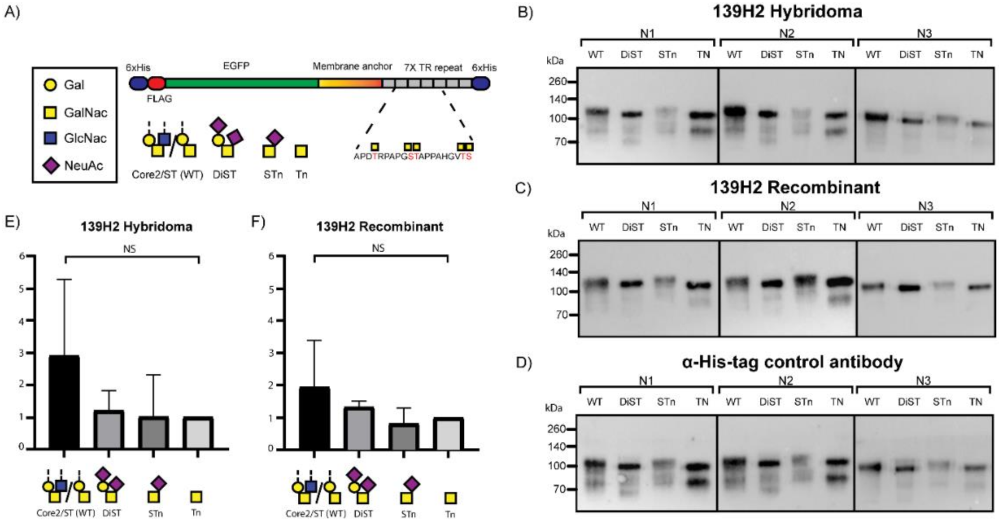
Supplementary Figure S3. Surface plasmon resonance quantification of binding affinity for the recombinant full-length 139H2 and its Fab to surface-immobilized MUC1 peptide. a) Binding of 139H2 full IgG to surface-immobilized biotinylated MUC1 peptide, analyzed by SPR. Equilibrium binding was fitted using a two-site binding model (right panel), with the respective K_D and B_{max} values for the high- and low-affinity binding indicated. b) Binding of 139H2 Fab to surface-immobilized biotinylated MUC1 peptide, analyzed by surface plasmon resonance SPR. Equilibrium binding was fitted using a one-site binding model (Langmuir isotherm, right panel), with K_D and B_{max} value indicated.



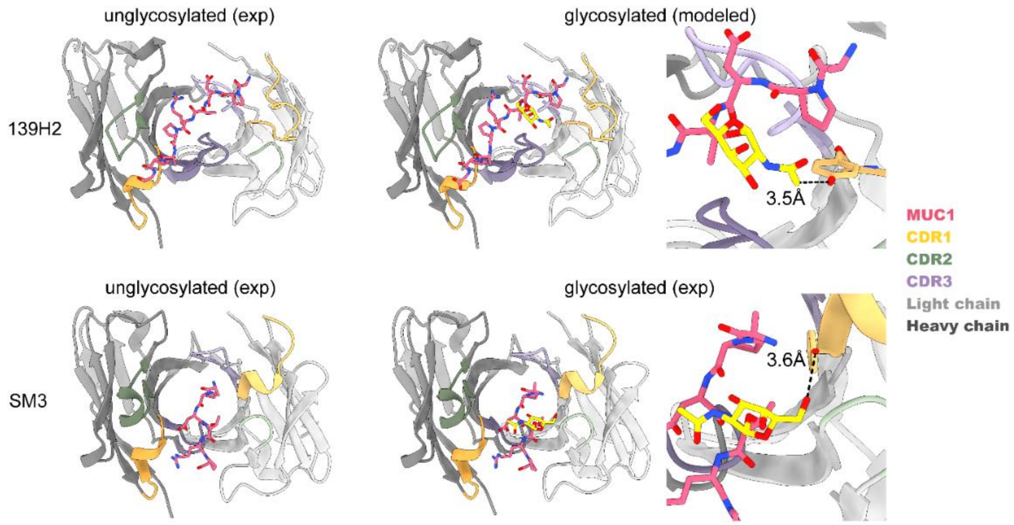
Supplementary Figure S4. Electron density for the MUC1 peptide bound to 139H2 Fab. (A) Positive F_o-F_c omit density plotted at 3.0σ (green mesh), after correcting CDRs and refinement in REFMAC but excluding placement of the MUC1 peptide, shows well-resolved additional density at the peptide binding site. (B) The $2F_o$-F_c density plotted at 1.0σ (grey mesh) of the final refined model including waters (red spheres) shows a good fit for the modelled MUC1 peptide (green sticks). In both panels the 139H2 Fab molecule is shown in stick representation, with the light chain coloured in light grey, and the heavy chain coloured in dark grey, as in Figure 3.



Supplementary Figure S5. Sites of somatic hypermutation in 139H2 framework. (A) Structure of 139H2 FAB with somatic hypermutations highlighted in red. (B) Hypermutations in heavy chain are organized in a stripe across the beta-sheet, side chains are oriented to the center of beta-barrel formed by heavy and light chains. (C) Interaction of Y35 and T97 with N106 and Y111, respectively, tilt CDR3 loop into the position where it can interact with MUC1.



Supplementary Figure S6. Binding of 139H2 to MUC1 reporter constructs with different O-linked glycosylation. A) Schematic representation of the MUC1 Fragments used, adopted from Nason/Büll et al. 2021. The four fragments used contain 7 transmembrane repeats (TR) of MUC1 with 5 O-glycosylation sites with WT Core2/ST (WT)/ DiST/ STn or Tn glycan structures. Fragments were a kind gift from Christian Büll. Fig. 1B-D) Western blots against the MUC1 WT/DiST/STn/Tn fragments with 139H2 Hybridoma-derived antibody (B) (N=3), 139H2 Synthetic Recombinant antibody (C) (N=3) and a α-His-tag antibody control (D) (N=3). Fig. 1E) Western blot band intensities analyzed with Image Lab 6.0 software. Calculated intensity ratios were made relative to the intensity of MUC1-Tn. No significant difference in binding of 139H2 Hybridoma-derived or 139H2 Recombinant Synthetic was observed compared to the 6 α-His-tag antibody control.



Supplementary Figure S7. Comparison of 139H2 and SM3 binding to MUC1. In SM3 the GalNAc residue makes an additional hydrogen bond with a tyrosine in CDRL1, similar interaction between T4 and GalNAc is predicted to be present also in 139H2.

Reference

(1) Bramwell, M. E.; Wiseman, G.; Shotton, D. M. Electron-Microscopic Studies of the ca Antigen, Epitectin. *Journal of Cell Science* **1986**, *86* (1), 249–261. https://doi.org/10.1242/jcs.86.1.249.

(2) Role of the Glycocalyx in Regulating Access of Microparticles to Apical Plasma Membranes of Intestinal Epithelial Cells: Implications for Microbial Attachment and Oral Vaccine Targeting. *J Exp Med* **1996**, *184* (3), 1045–1059.

(3) Lindén, S. K.; Sheng, Y. H.; Every, A. L.; Miles, K. M.; Skoog, E. C.; Florin, T. H. J.; Sutton, P.; McGuckin, M. A. MUC1 Limits Helicobacter Pylori Infection Both by Steric Hindrance and by Acting as a Releasable Decoy. *PLOS Pathogens* **2009**, *5* (10), e1000617. https://doi.org/10.1371/journal.ppat.1000617.

(4) McAuley, J. L.; Linden, S. K.; Png, C. W.; King, R. M.; Pennington, H. L.; Gendler, S. J.; Florin, T. H.; Hill, G. R.; Korolik, V.; McGuckin, M. A. MUC1 Cell Surface Mucin Is a Critical Element of the Mucosal Barrier to Infection. *J Clin Invest* **2007**, *117* (8), 2313–2324. https://doi.org/10.1172/JCI26705.

(5) Li, X.; Bleumink-Pluym, N. M. C.; Luijkx, Y. M. C. A.; Wubbolts, R. W.; Putten, J. P. M. van; Strijbis, K. MUC1 Is a Receptor for the Salmonella SiiE Adhesin That Enables Apical Invasion into Enterocytes. *PLOS Pathogens* **2019**, *15* (2), e1007566. https://doi.org/10.1371/journal.ppat.1007566.

(6) Kato, K.; Lillehoj, E. P.; Lu, W.; Kim, K. C. MUC1: The First Respiratory Mucin with an Anti-Inflammatory Function. *J Clin Med* **2017**, *6* (12), 110. https://doi.org/10.3390/jcm6120110.

(7) Lu, W.; Hisatsune, A.; Koga, T.; Kato, K.; Kuwahara, I.; Lillehoj, E. P.; Chen, W.; Cross, A. S.; Gendler, S. J.; Gewirtz, A. T.; Kim, K. C. Cutting Edge: Enhanced Pulmonary Clearance of Pseudomonas Aeruginosa by Muc1 Knockout Mice1. *The Journal of Immunology* **2006**, *176* (7), 3890–3894. https://doi.org/10.4049/jimmunol.176.7.3890.

(8) Sheng, Y. H.; Triyana, S.; Wang, R.; Das, I.; Gerloff, K.; Florin, T. H.; Sutton, P.; McGuckin, M. A. MUC1 and MUC13 Differentially Regulate Epithelial Inflammation in Response to Inflammatory and Infectious Stimuli. *Mucosal Immunology* **2013**, *6* (3), 557–568. https://doi.org/10.1038/mi.2012.98.

(9) Taylor-Papadimitriou, J.; Burchell, J.; Miles, D. W.; Dalziel, M. MUC1 and Cancer. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1999**, *1455* (2), 301–313. https://doi.org/10.1016/S0925-4439(99)00055-1.

(10) *Characterization and Molecular Cloning of a Novel MUC1 Protein, Devoid of Tandem Repeats, Expressed in Human Breast Cancer Tissue - Zrihan-Licht - 1994 - European Journal of Biochemistry - Wiley Online Library*. https://febs.onlinelibrary.wiley.com/doi/full/10.1111/j.1432-1033.1994.00787.x?sid=nlm%3Apubmed (accessed 2023-05-25).

(11) Gendler, S.; Taylor-Papadimitriou, J.; Duhig, T.; Rothbard, J.; Burchell, J. A Highly Immunogenic Region of a Human Polymorphic Epithelial Mucin Expressed by Carcinomas Is Made up of Tandem Repeats. *Journal of Biological Chemistry* **1988**, *263* (26), 12820–12823. https://doi.org/10.1016/S0021-9258(18)37632-4.

(12) Nason, R.; Büll, C.; Konstantinidi, A.; Sun, L.; Ye, Z.; Halim, A.; Du, W.; Sørensen, D. M.; Durbesson, F.; Furukawa, S.; Mandel, U.; Joshi, H. J.; Dworkin, L. A.; Hansen, L.; David, L.; Iverson, T. M.; Bensing, B. A.; Sullam, P. M.; Varki, A.; Vries, E. de; de Haan, C. A. M.; Vincentelli, R.; Henrissat, B.; Vakhrushev, S. Y.; Clausen, H.; Narimatsu, Y. Display of the Human Mucinome with Defined O-Glycans by Gene Engineered Cells. *Nat Commun* **2021**, *12* (1), 4070. https://doi.org/10.1038/s41467-021-24366-4.

(13) Yang, J.-M.; Byrd, J. C.; Siddiki, B. B.; Chung, Y.-S.; Okuno, M.; Sowa, M.; Kim, Y. S.; Matta, K. L.; Brockhausen, I. Alterations of O-Glycan Biosynthesis in Human Colon Cancer Tissues. *Glycobiology* **1994**, *4* (6), 873–884. https://doi.org/10.1093/glycob/4.6.873.

(14) Lloyd, K. O.; Burchell, J.; Kudryashov, V.; Yin, B. W. T.; Taylor-Papadimitriou, J. Comparison of O-Linked Carbohydrate Chains in MUC-1 Mucin from Normal Breast Epithelial Cell Lines and Breast Carcinoma Cell Lines:: DEMONSTRATION OF SIMPLER AND FEWER GLYCAN CHAINS IN TUMOR CELLS*. *Journal of Biological Chemistry* **1996**, *271* (52), 33325–33334. https://doi.org/10.1074/jbc.271.52.33325.

(15) Hilkens, J.; Buijs, F.; Hilgers, J.; Hageman, Ph.; Calafat, J.; Sonnenberg, A.; Van Vlak, M. D. Monoclonal Antibodies against Human Milk-Fat Globule Membranes Detecting Differentiation Antigens of the Mammary Gland and Its Tumors. *International Journal of Cancer* **1984**, *34* (2), 197–206. https://doi.org/10.1002/ijc.2910340210.

(16) Hilkens, J.; Kroezen, V.; Buijs, F.; Hilgers, J.; van Vliet, M.; de Voogd, W.; Bonfrer, J.; Bruning, P. F. MAM-6, a Carcinoma Associated Marker: Preliminary Characterisation and Detection in Sera of Breast Cancer Patients. In *Monoclonal Antibodies and Breast Cancer: Proceedings of the International Workshop on Monoclonal Antibodies and Breast Cancer San Francisco, California — November 8–9, 1984*; Ceriani, R. L., Ed.; Developments in Oncology; Springer US: Boston, MA, 1985; pp 28–42. https://doi.org/10.1007/978-1-4613-2617-5_3.

(17) Ryuko, K.; Schol, D. J.; Snijdewint, F. G. M.; von Mensdorff-Pouilly, S.; Poort-Keesom, R. J. J.; Karuntu-Wanamarta, Y. A.; Verstraeten, R. A.; Miyazaki, K.; Kenemans, P.; Hilgers, J. Characterization of a New MUC1 Monoclonal Antibody (VU-2-G7) Directed to the Glycosylated PDTR Sequence of MUC1. *Tumor Biology* **2000**, *21* (4), 197–210. https://doi.org/10.1159/000030126.

(18) Molthoff, C. F. M.; Pinedo, H. M.; Schlüper, H. M. M.; Rutgers, D. H.; Boven, E. Comparison Of131I-Labelled Anti-Episialin 139H2 with Cisplatin, Cyclophosphamide or External-Beam Radiation for Anti-Tumor Efficacy in Human Ovarian Cancer Xenografts. *Int. J. Cancer* **1992**, *51* (1), 108–115. https://doi.org/10.1002/ijc.2910510120.

(19) Molthoff, C. F. M.; Calame, J. J.; Pinedo, H. M.; Boven, E. Human Ovarian Cancer Xenografts in Nude Mice: Characterization and Analysis of Antigen Expression. *International Journal of Cancer* **1991**, *47* (1), 72–79. https://doi.org/10.1002/ijc.2910470114.

(20) Tashiro, Y.; Yonezawa, S.; Kim, Y. S.; Sato, E. Immunohistochemical Study of Mucin Carbohydrates and Core Proteins in Human Ovarian Tumors. *Human Pathology* **1994**, *25* (4), 364–372. https://doi.org/10.1016/0046-8177(94)90144-9.

(21) Peng, W.; Pronker, M. F.; Snijder, J. Mass Spectrometry-Based *De novo* Sequencing of Monoclonal Antibodies Using Multiple Proteases and a Dual Fragmentation Scheme. *J. Proteome Res.* **2021**, *20* (7), 3559–3566. https://doi.org/10.1021/acs.jproteome.1c00169.

(22) Bondt, A.; Hoek, M.; Tamara, S.; de Graaf, B.; Peng, W.; Schulte, D.; van Rijswijck, D. M. H.; den Boer, M. A.; Greisch, J.-F.; Varkila, M. R. J.; Snijder, J.; Cremer, O. L.; Bonten, M. J. M.; Heck, A. J. R. Human Plasma IgG1 Repertoires Are Simple, Unique, and Dynamic. *Cell Systems* **2021**, *12* (12), 1131-1143.e5. https://doi.org/10.1016/j.cels.2021.08.008.

(23) Peng, W.; Boer, M. A. den; Tamara, S.; Mokiem, N. J.; Lans, S. P. A. van der; Schulte, D.; Haas, P.-J.; Minnema, M. C.; Rooijakkers, S. H. M.; Zuilen, A. D. van; Heck, A. J. R.; Snijder, J. Direct Mass Spectrometry-Based Detection and Antibody Sequencing of Monoclonal Gammopathy of Undetermined Significance from Patient Serum – a Case Study. bioRxiv May 24, 2023, p 2023.05.22.541697. https://doi.org/10.1101/2023.05.22.541697.

(24) Bondt, A.; Hoek, M.; Dingess, K.; Tamara, S.; Graaf, B. de; Peng, W.; Boer, M. A. den; Damen, M.; Zwart, C.; Barendregt, A.; Rijswijck, D. M. H. van; Grobben, M.; Tejjani, K.; Rijswijk, J. van; Völlmy, F.; Snijder, J.; Fortini, F.; Papi, A.; Volta, C. A.; Campo, G.; Contoli, M.; Gils, M. J. van; Spadaro, S.; Rizzo, P.; Heck, A. J. R. No Patient Is the Same; Lessons Learned from Antibody Repertoire Profiling in Hospitalized Severe COVID-19 Patients. medRxiv December 26, 2022, p 2022.12.23.22283896. https://doi.org/10.1101/2022.12.23.22283896.

(25) Tran, N. H.; Zhang, X.; Xin, L.; Shan, B.; Li, M. *De novo* Peptide Sequencing by Deep Learning. *Proceedings of the National Academy of Sciences* **2017**, *114* (31), 8247–8252. https://doi.org/10.1073/pnas.1705691114.

(26) Tran, N. H.; Qiao, R.; Xin, L.; Chen, X.; Liu, C.; Zhang, X.; Shan, B.; Ghodsi, A.; Li, M. Deep Learning Enables *de novo* Peptide Sequencing from Data-Independent-Acquisition Mass Spectrometry. *Nat Methods* **2019**, *16* (1), 63–66. https://doi.org/10.1038/s41592-018-0260-3.

(27) Schulte, D.; Peng, W.; Snijder, J. Template-Based Assembly of Proteomic Short Reads For *De novo* Antibody Sequencing and Repertoire Profiling. *Anal. Chem.* **2022**, *94* (29), 10391–10399. https://doi.org/10.1021/acs.analchem.2c01300.

(28) Lefranc, M.-P. IMGT® Databases, Web Resources and Tools for Immunoglobulin and T Cell Receptor Sequence Analysis, Http://Imgt.Cines.Fr. *Leukemia* **2003**, *17* (1), 260–266. https://doi.org/10.1038/sj.leu.2402637.

(29) *Differential expression of the human mucin genes MUC1 to MUC5 in relation to growth and differentiation of different mucus-secreting HT-29 cell subpopulations | Journal of Cell Science | The Company of Biologists*. https://journals.biologists.com/jcs/article/106/3/771/23672/Differential-expression-of-the-human-mucin-genes (accessed 2023-05-25).

(30) Gildea, R. J.; Beilsten-Edmands, J.; Axford, D.; Horrell, S.; Aller, P.; Sandy, J.; Sanchez-Weatherby, J.; Owen, C. D.; Lukacik, P.; Strain-Damerell, C.; Owen, R. L.; Walsh, M. A.; Winter, G. Xia2.Multiplex: A Multi-Crystal Data-Analysis Pipeline. *Acta Cryst D* **2022**, *78* (6), 752–769. https://doi.org/10.1107/S2059798322004399.

(31) Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: Making Protein Folding Accessible to All. *Nat Methods* **2022**, *19* (6), 679–682. https://doi.org/10.1038/s41592-022-01488-1.

(32) McCoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J. Phaser Crystallographic Software. *J Appl Cryst* **2007**, *40* (4), 658–674. https://doi.org/10.1107/S0021889807021206.

(33) Emsley, P.; Cowtan, K. Coot: Model-Building Tools for Molecular Graphics. *Acta Cryst D* **2004**, *60* (12), 2126–2132. https://doi.org/10.1107/S0907444904019158.

(34) Kovalevskiy, O.; Nicholls, R. A.; Long, F.; Carlon, A.; Murshudov, G. N. Overview of Refinement Procedures within REFMAC5: Utilizing Data from Different Sources. *Acta Cryst D* **2018**, *74* (3), 215–227. https://doi.org/10.1107/S2059798318000979.

(35) Williams, C. J.; Headd, J. J.; Moriarty, N. W.; Prisant, M. G.; Videau, L. L.; Deis, L. N.; Verma, V.; Keedy, D. A.; Hintze, B. J.; Chen, V. B.; Jain, S.; Lewis, S. M.; Arendall III, W. B.; Snoeyink, J.; Adams, P. D.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. MolProbity: More and Better Reference Data for Improved All-Atom Structure Validation. *Protein Science* **2018**, *27* (1), 293–315. https://doi.org/10.1002/pro.3330.

(36) Potterton, L.; Agirre, J.; Ballard, C.; Cowtan, K.; Dodson, E.; Evans, P. R.; Jenkins, H. T.; Keegan, R.; Krissinel, E.; Stevenson, K.; Lebedev, A.; McNicholas, S. J.; Nicholls, R. A.; Noble, M.; Pannu, N. S.; Roth, C.; Sheldrick, G.; Skubak, P.; Turkenburg, J.; Uski, V.; von Delft, F.; Waterman, D.; Wilson, K.; Winn, M.; Wojdyr, M. CCP4i2: The New Graphical User Interface to the CCP4 Program Suite. *Acta Cryst D* **2018**, *74* (2), 68–84. https://doi.org/10.1107/S2059798317016035.

(37) Burchell, J.; Taylor-Papadimitriou, J. Effect of Modification of Carbohydrate Side Chains on the Reactivity of Antibodies with Core-Protein Epitopes of the MUC1 Gene Product. *Epithelial Cell Biology* **1993**, *2* (4), 155–162.

(38) Han, Y.; Niu, J.; Pan, D.; Feng, C.; Song, K.; Meng, B.; Westerlind, U.; Zhang, Y.; Liu, H.; Xu, L.; Zhou, D. Structural Basis of a High-Affinity Antibody Binding to Glycoprotein Region with Consecutive Glycosylation Sites. bioRxiv July 24, 2022, p 2022.07.24.501275. https://doi.org/10.1101/2022.07.24.501275.

(39) Kinoshita, N.; Ohno, M.; Nishiura, T.; Fujii, S.; Nishikawa, A.; Kawakami, Y.; Uozumi, N.; Taniguchi, N. Glycosylation at the Fab Portion of Myeloma Immunoglobulin G and Increased Fucosylated Biantennary Sugar Chains: Structural Analysis by High-Performance Liquid Chromatography and Antibody-Lectin Enzyme Immunoassay Using Lens Culinaris Agglutinin1. *Cancer Research* **1991**, *51* (21), 5888–5892.

(40) Wakui, H.; Tanaka, Y.; Ose, T.; Matsumoto, I.; Kato, K.; Min, Y.; Tachibana, T.; Sato, M.; Naruchi, K.; Martin, F. G.; Hinou, H.; Nishimura, S.-I. A Straightforward Approach to Antibodies Recognising Cancer Specific Glycopeptidic Neoepitopes. *Chem. Sci.* **2020**, *11* (19), 4999–5006. https://doi.org/10.1039/D0SC00317D.

(41) Movahedin, M.; Brooks, T. M.; Supekar, N. T.; Gokanapudi, N.; Boons, G.-J.; Brooks, C. L. Glycosylation of MUC1 Influences the Binding of a Therapeutic Antibody by Altering the Conformational Equilibrium of the Antigen. *Glycobiology* **2017**, *27* (7), 677–687. https://doi.org/10.1093/glycob/cww131.

(42) Martínez-Sáez, N.; Castro-López, J.; Valero-González, J.; Madariaga, D.; Compañón, I.; Somovilla, V. J.; Salvadó, M.; Asensio, J. L.; Jiménez-Barbero, J.; Avenoza, A.; Busto, J. H.; Bernardes, G. J. L.; Peregrina, J. M.; Hurtado-Guerrero, R.; Corzana, F. Deciphering the Non-Equivalence of Serine and Threonine O-Glycosylation Points: Implications for Molecular Recognition of the Tn Antigen by an Anti-MUC1 Antibody. *Angewandte Chemie International Edition* **2015**, *54* (34), 9830–9834. https://doi.org/10.1002/anie.201502813.

(43) Compañón, I.; Guerreiro, A.; Mangini, V.; Castro-López, J.; Escudero-Casao, M.; Avenoza, A.; Busto, J. H.; Castillón, S.; Jiménez-Barbero, J.; Asensio, J. L.; Jiménez-Osés, G.; Boutureira, O.; Peregrina, J. M.; Hurtado-Guerrero, R.; Fiammengo, R.; Bernardes, G. J. L.; Corzana, F. Structure-Based Design of Potent Tumor-Associated Antigens: Modulation of Peptide Presentation by Single-Atom O/S or O/Se Substitutions at the Glycosidic Linkage. *J. Am. Chem. Soc.* **2019**, *141* (9), 4063–4072. https://doi.org/10.1021/jacs.8b13503.

(44) Bermejo, I. A.; Navo, C. D.; Castro-López, J.; Guerreiro, A.; Jiménez-Moreno, E.; Fernández, E. M. S.; García-Martín, F.; Hinou, H.; Nishimura, S.-I.; Fernández, J. M. G.; Mellet, C. O.; Avenoza, A.; Busto, J. H.; Bernardes, G. J. L.; Hurtado-Guerrero, R.; Peregrina, J. M.; Corzana, F. Synthesis, Conformational Analysis and in Vivo Assays of an Anti-Cancer Vaccine That Features an Unnatural Antigen Based on an Sp2-Iminosugar Fragment. *Chem. Sci.* **2020**, *11* (15), 3996–4006. https://doi.org/10.1039/C9SC06334J.

(45) Bermejo, I. A.; Usabiaga, I.; Compañón, I.; Castro-López, J.; Insausti, A.; Fernández, J. A.; Avenoza, A.; Busto, J. H.; Jiménez-Barbero, J.; Asensio, J. L.; Peregrina, J. M.; Jiménez-Osés, G.; Hurtado-Guerrero, R.; Cocinero, E. J.; Corzana, F. Water Sculpts the

Distinctive Shapes and Dynamics of the Tumor-Associated Carbohydrate Tn Antigens: Implications for Their Molecular Recognition. *J. Am. Chem. Soc.* **2018**, *140* (31), 9952–9960. https://doi.org/10.1021/jacs.8b04801.

(46) Rojas-Ocáriz, V.; Compañón, I.; Aydillo, C.; Castro-Loṕez, J.; Jiménez-Barbero, J.; Hurtado-Guerrero, R.; Avenoza, A.; Zurbano, M. M.; Peregrina, J. M.; Busto, J. H.; Corzana, F. Design of α-S-Neoglycopeptides Derived from MUC1 with a Flexible and Solvent-Exposed Sugar Moiety. *J. Org. Chem.* **2016**, *81* (14), 5929–5941. https://doi.org/10.1021/acs.joc.6b00833.

(47) Somovilla, V. J.; Bermejo, I. A.; Albuquerque, I. S.; Martínez-Sáez, N.; Castro-López, J.; García-Martín, F.; Compañón, I.; Hinou, H.; Nishimura, S.-I.; Jiménez-Barbero, J.; Asensio, J. L.; Avenoza, A.; Busto, J. H.; Hurtado-Guerrero, R.; Peregrina, J. M.; Bernardes, G. J. L.; Corzana, F. The Use of Fluoroproline in MUC1 Antigen Enables Efficient Detection of Antibodies in Patients with Prostate Cancer. *J. Am. Chem. Soc.* **2017**, *139* (50), 18255–18261. https://doi.org/10.1021/jacs.7b09447.

(48) Dokurno, P.; Bates, P. A.; Band, H. A.; Stewart, L. M. D.; Lally, J. M.; Burchell, J. M.; Taylor-Papadimitriou, J.; Snary, D.; Sternberg, M. J. E.; Freemont, P. S. Crystal Structure at 1.95 å Resolution of the Breast Tumour-Specific Antibody SM3 Complexed with Its Peptide Epitope Reveals Novel Hypervariable Loop Recognition1 1Edited by R. Huber. *Journal of Molecular Biology* **1998**, *284* (3), 713–728. https://doi.org/10.1006/jmbi.1998.2209.

(49) Baker, M. Reproducibility Crisis: Blame It on the Antibodies. *Nature* **2015**, *521* (7552), 274–276. https://doi.org/10.1038/521274a.

(50) Uhlen, M.; Bandrowski, A.; Carr, S.; Edwards, A.; Ellenberg, J.; Lundberg, E.; Rimm, D. L.; Rodriguez, H.; Hiltke, T.; Snyder, M.; Yamamoto, T. A Proposal for Validation of Antibodies. *Nat Methods* **2016**, *13* (10), 823–827. https://doi.org/10.1038/nmeth.3995.

Chapter 5

Structural basis for postfusion-specific binding to Respiratory Syncytial Virus F protein by the antigenic site-I antibody 131-2a.

Weiwei Peng[1a], Marta Šiborová[1a], Xuesheng Wu[2], Wenjuan Du[2], Douwe Schulte[1], Matti F. Pronker[1], Cornelis A. M. de Haan[2], Joost Snijder[1*]

[1] Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute of Pharmaceutical Sciences, Utrecht University, Padualaan 8, 3584CH Utrecht, The Netherlands

[2] Virology Group, Division of Infectious Diseases and Immunology, Department of Biomolecular Health Sciences, Faculty of Veterinary Medicine, Utrecht University, Utrecht, Yalelaan 1, 3584CL, the Netherlands

[a] equal contribution

* corresponding author: j.snijder@uu.nl

Abstract

The Respiratory Syncytial Virus (RSV) Fusion (F) protein is a major target of antiviral antibodies following natural infection and vaccination and responsible for mediating fusion between the viral envelope and the host membrane. The fusion process is driven by a large-scale conformational change in F, switching irreversibly from the metastable prefusion state to the stable postfusion conformation. Previous research has identified six distinct antigenic sites in RSV-F, termed sites Ø, I, II, III, IV, and V. Of these, only antigenic site I is fully specific to the postfusion conformation of F. A monoclonal antibody 131-2a that targets postfusion F specifically has been widely employed as a research tool to probe for postfusion F and to define antigenic site I in serological studies, yet the sequence and epitope of the antibody remained unknown. Here we use mass spectrometry-based *de novo* sequencing of 131-2a to reverse engineer a recombinant product and study the epitope to define antigenic site I with molecular detail. The experimentally determined sequence was validated by comparison of the reverse-engineered 131-2a and the sequenced input material in Western blot and ELISA. We used reverse engineered 131-2a to investigate the epitope by single particle cryo electron microscopy, revealing the molecular basis for the antibody binding to the antigenic site I of the postfusion RSV-F.

Chapter 5

Introduction

Respiratory Syncytial Virus (RSV) is the second major cause for hospital admissions of young infants globally, following malaria. It is estimated that RSV is responsible for approximately 60,000 childhood deaths each year, especially in low-resource settings[1,2]. RSV is a Pneumovirus, member of the family of Paramyxoviruses. It consists of a negative sense, single stranded RNA genome enveloped in a lipid bilayer containing three virally encoded envelope proteins: the Small Hydrophobic Protein (SHP), Fusion protein (F), and Glycoprotein (G)[3]. Both G and F are involved in host cell attachment and receptor binding and as the name suggests, F also mediates cell entry by fusing the viral envelope with the host membrane, thereby delivering the genetic material of the virus into its host cell.[4] The antibodies that target G and F play a pivotal role in the antiviral immune response. This is especially true for the F protein, making it the focus of vaccines and monoclonal antibody therapies currently under clinical development[5–7].

F is a trimeric class I viral fusion protein that exists in a metastable prefusion conformation. Its conformational change into the stable postfusion conformation drives fusion of the viral envelope with the host membrane. Owing to the drastic conformational changes between the pre- and postfusion states of F, each conformation presents distinct epitopes. Previous studies have described this complex antigenic landscape by defining six antigenic sites: Ø, I, II, III, IV, and V (see Figure 1)[8]. Antigenic sites Ø and V are specific to the prefusion state of F, and while antibodies directed against antigenic site III can bind F in both conformations, they also bind stronger to the prefusion conformation. Antigenic sites II and IV are shared between pre- and postfusion states, but antibodies directed to antigenic site I are fully specific to the postfusion conformation[9].

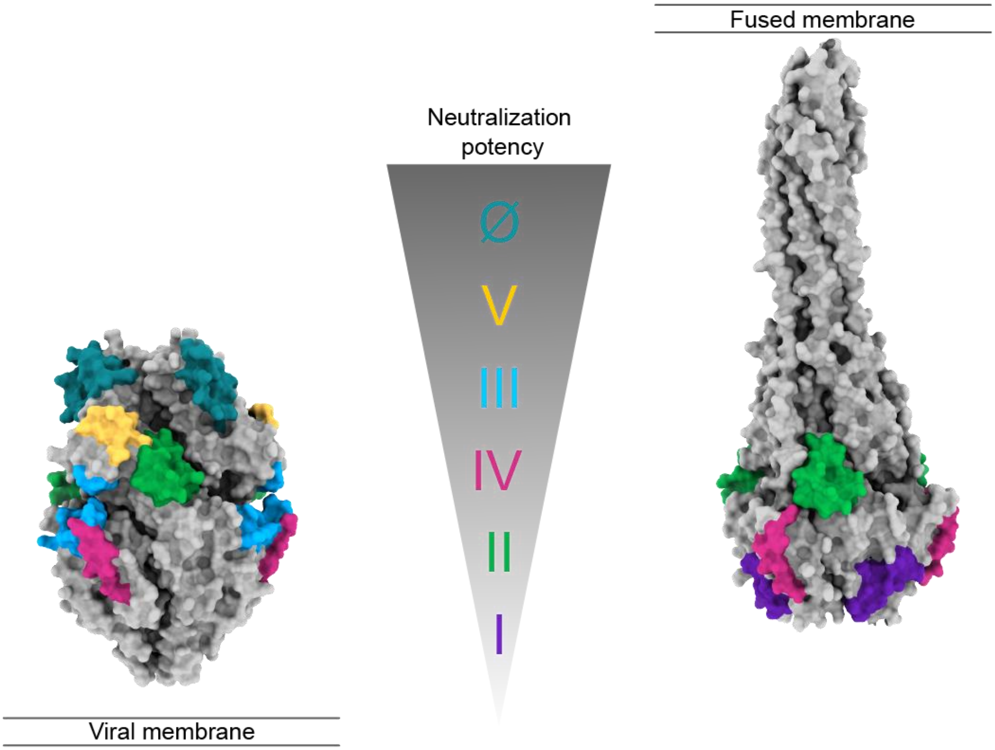


Figure 1. RSV F protein antigenic sites: Antigenic sites Ø (petrol) and V (yellow) are specific to the prefusion state; antigenic II (green) and IV (magenta) are present in both pre- and postfusion F; antigenic site III (blue) show preference to the prefusion state; antigenic site I (purple) is specific to postfusion F.

Even in newly produced virions from an infected host cell, F is present on the envelope in a complex mixture of pre- and postfusion states. Meanwhile, neutralizing activity in immune sera can be largely attributed to prefusion F-specific antibodies. The postfusion F on virions has been speculated to act as a decoy to the immune system, directing it towards non-neutralizing epitopes at the expense of effectively neutralizing epitopes present in prefusion F. The antigenic site-I directed antibodies thereby constitute an important, possibly counterproductive component of the antibody response to natural RSV infection or vaccination.

Serological studies have used monoclonal antibody standards to probe the antigenic site-specific antibody response in human subjects. This is typically done in competition binding experiments of the monoclonal antibody standards with polyclonal sera to pre- and postfusion F[10]. The mouse monoclonal antibody 131-2a has become the canonical antigenic site I-~~-~~defining antibody in these studies and is also widely used to specifically detect postfusion F in viral cultures and vaccine preparations by ELISA, Western blot, immunofluorescence microscopy, and cell sorting experiments. While 131-2a was discovered in the early 1980's and has since been widely used in RSV studies, its sequence is not publicly available[11]. Moreover, the epitope of 131-2a on postfusion F has not been studied at a detailed structural level, so the basis of its specificity to postfusion F has not been conclusively established. We have recently developed a mass spectrometry-based workflow to sequence antibodies straight from the purified protein, which we apply here to reverse engineer a functional recombinant 131-2a monoclonal antibody[12,13]. This enabled detailed structural studies of the 131-2a interaction with F by single particle cryo electron microscopy (cryo-EM), revealing the molecular basis for its binding specificity to the postfusion conformation.

Methods

Expression and Purification of RSV-F Proteins

Design, expression and purification of RSV pre-fusion F (DSCav1-T4fd, Genbank JX015498) has been described previously[14,15]. Briefly, cDNA encoding pre-fusion RSV F (DSCav1-T4fd) was cloned into the pCD5 expression vector in frame with the CD5 signal peptide coding sequence, followed by sequences encoding a C-terminal T4 fibritin trimerization motif, thrombin site, and Strep-tagII (IBA, Germany). Pre-fusion F was expressed transiently in HEK-293T cells [ATCC, CRL-11268] and secreted protein was purified from culture supernatants using Strep-tactin Sepharose beads (IBA) following the manufacturer's protocol and as described previously[15,16]. RSV post-fusion F corresponds to Flys-GCN described previously except that the strep tag was not present in the current protein and that it was expressed in CHO cells[16]. This construct was generated by cloning cDNAs encoding the RSV F ectodomain (amino acids 26 to 515, Genbank: JX015498.1) in frame with a CD5 signal peptide-encoding sequence and followed by sequences coding for the GCN4 isoleucine zipper trimerization motif and a a LysM peptidoglycan binding domain[17,18]. In addition, arginines in the two furin-cleavage sites were substituted by lysines. This protein was previously used in a clinical trial[19].

MS-based sequencing of 131-2a

Sample preparation – 24 µg of 131-2a (MAB8599P-K, Sigma) was denatured and reduced in 2% sodium deoxycholate (SDC), 200 mM Tris-HCl, 10 mM tris(2-carboxyethyl)phosphine (TCEP), 40 mM iodoacetic acid, pH 8.5 at 95°C for 10 min, followed with 20 min incubation at room temperature in the dark for alkylation. 3 µg Sample was then digested by one of the following proteases trypsin, chymotrypsin, α-lytic protease, thermolysin, elastase, gluC, lysC and lysN in a 1:50 ratio (w:w) in a total volume of 100 uL of 50 mM ammonium bicarbonate at 37°C overnight. After digestion, SDC was removed by adding 2 uL formic acid (FA) and centrifugation at 14000 g for 20 min. Following centrifugation, the supernatant containing the peptides was collected for desalting on a 30 µm Oasis HLB 96-well plate (Waters). The Oasis HLB sorbent was activated with 100% acetonitrile and subsequently equilibrated with 10% formic acid in water. Next, peptides were bound to the sorbent, washed twice with 10% formic acid in water and eluted with 100 µL of 50% acetonitrile/5% formic acid in water (v/v). The eluted peptides were vacuum-dried and reconstituted in 100 µL 2% FA.

Mass Spectrometry –The digested peptides (single injection of 0.2 ug) were separated by online reversed phase chromatography on an Agilent 1290 UHPLC (column packed with Poroshell 120 EC C18; dimensions 50 cm x 75 µm, 2.7 µm, Agilent Technologies) coupled to a Thermo Scientific Orbitrap Fusion mass spectrometer. Samples were eluted over a 90 min gradient from 0% to 35% acetonitrile at a flow rate of 0.3 µL/min. Peptides were analyzed with a resolution setting of 60000 in MS1. MS1 scans were obtained with standard AGC target, maximum injection time of 50 ms, and scan range 350-2000. The precursors were selected with a 3 m/z window and fragmented by stepped HCD as well as EThcD. The stepped HCD fragmentation included steps of 25%, 35% and 50% NCE. EThcD fragmentation was performed with calibrated charge-dependent ETD parameters and 27% NCE supplemental activation. For both fragmentation types, ms2 scan were acquired at 30000 resolution, 800% Normalized AGC target, 250 ms maximum injection time, scan range 120-3500.

Data analysis – MS/MS spectra were used to determine *de novo* peptide sequences using PEAKS Studio X (version 10.6)[20]. We used a tolerance of 20 ppm and 0.02 Da for MS1 and MS2, respectively. Carboxymethylation was set as fixed modification of cysteine and variable modification of peptide N-termini and lysine. Oxidation of methionine and tryptophan and pyroglutamic acid modification of N-terminal glutamic acid and glutamine were set as additional variable modifications. The CSV file containing all the *de novo* sequenced peptide was exported for further analysis. Stitch (nightly version 1.4.0+6dbb8b7) was used for the template-based assembly[21]. The mouse antibody database from IMGT was used as templates[22,23]. The cutoff score for the *de novo* sequenced peptide was set as 85 and the cutoff score for the template matching was set as 10. All the peptides supporting the sequences were examined manually.

Reverse engineering 131-2a

Cloning and Expression of recombinant 131-2a

To recombinantly express full-length 131-2a, the proteomic sequences of both the light and heavy chains were reverse-translated and codon-optimized for expression in human cells using the Thermo Fisher webtool (https://www.thermofisher.com/order/gene-design/index.html). For the linker and Fc region of the heavy chain, the standard mouse IGG2A amino acid sequence (IMGT database) was used. An N-terminal secretion signal peptide derived from human IgG light chain (MEAPAQLLFLLLLWLPDTTG) was added to the N-termini of both heavy and light chains. BamHI and NotI restriction sites were added to the 5′ and 3′ ends of the coding regions, respectively. Only for the light chain, a double stop codon was introduced at the 3′ site before the NotI restriction site. The coding regions were subcloned using BamHI and NotI restriction-ligation into a pRK5 expression vector with a C-terminal octahistidine tag between the NotI site and a double stop codon 3′ of the insert, so that only the heavy chain has a C-terminal AAAHHHHHHHH sequence for nickel-affinity purification (the triple alanine resulting from the NotI site). After the sequence was validated by Sanger Sequencing, the HC/LC were mixed in a 1:1 DNA ratio and expressed in HEK293 cells by the ImmunoPrecise Antibodies (Europe) B.V company. After expression the culture supernatant of the cells was harvested and purified using a HisPur Ni-NTA Resin (Thermos Scientific). After purification, 131-2a was buffer exchanged and concentrated into PBS by Amicon Ultra Filter.

131-2a Fab generation

The full 131-2a IgG was digested by immobilized papain (ThermoFisher) in digestion buffer (0.22 mM Cysteine HCl, 20 mM phosphate buffer, pH=7) at 37 degree, 1000 rpm shaking for 5 hours. After digestion, Fc was removed by incubation with protein A agarose Resin (ThermoFisher) at room temperature for 15 minutes. The 131-2a Fab was further purified by size-exclusion chromatography using Superdex 200 Increase 10/300 GL (Cytiva) in PBS buffer.

Validation of 131-2a product

Western blot

Binding of 131-2a was analyzed by western blot assay utilizing both prefusion and postfusion proteins mentioned above[14–16]. Briefly, 0.25 μg of either post- or pre-fusion RSV F proteins were mixed with native protein buffer (Bio-Rad, 1610738) and then loaded into 7% polyacrylamide gel devoid of SDS, accompanied by protein standards (Bio-Rad, 1610375). The proteins were transferred to a cellulose nitrate membrane using a Trans-Blot Turbo Transfer System (Bio-Rad). Following this, the membrane was subjected to blocking utilizing 3% BSA alongside 0.1% Tween 20. HRP-conjugated rabbit anti-mouse IgG (Dako, P0260) was used at 1:5,000 for detection of 131-2a antibody. Visualization was performed using BeyoECL Moon (Beyotime). The Western blots were scanned by using an imaging system (Odyssey).

ELISA

Nunc MaxiSorp ELISA plates (Thermo Fisher Scientific) were icoated with 50 ng of RSV F and incubated overnight at 4 °C, followed by three washing steps with phosphate-buffered saline (PBS) containing 0.05% Tween 20. Plates were blocked with 2% bovine serum albumin (BSA; Fitzgerald) in PBS with 0.1% Tween 20 at 4 °C overnight. Subsequently, 131-2a (home-made or commercial (MAB8599P-K, Sigma)) antibodies were allowed to bind the plates at 3-fold serial dilutions, starting at 1 μg/ml diluted in PBS containing 2% BSA and 0.1% Tween 20, at RT for 1 hour. After washing, plates were incubated with 1:1000 diluted horseradish peroxidase (HRP)–conjugated rabbit anti-mouse IgG (Dako, P0260) for 1 hour at RT. HRP reactivity with tetramethylbenzidine substrate (BioFX) was measured with an ELISA plate reader at 450nM (EL-808, BioTek).

CryoEM sample preparation, data collection, motion correction, and CTF estimation

RSV postF and 131-2a Fabs were mixed in 3:4 molar ratio and incubated 30 minutes on ice. Sample was diluted to 0.2 mg/ml in PBS and was pipetted onto a holey carbon-coated copper grid (R1.2/1.3, mesh 200; Quantifoil), blotted and vitrified by plunging into liquid ethane using an FEI Vitrobot Mark IV (Thermo Fisher Scientific). The vitrified sample was transferred to a Titan Krios electron microscope (Thermo Fisher Scientific) operated under cryogenic conditions and at an acceleration voltage of 300 kV.

Micrographs were collected at a magnification of 105,000x on a K3 direct electron detector in counted super-resolution mode, resulting in a calibrated pixel size of 0.418 Å/pix. Imaging was done under low-dose conditions (total dose 50 $e^-/Å^2$) and defocus values ranging from −0.8 to −2.0 µm. The 2.52 second exposure was fractionated into 50 frames and saved as a tiff. Automated data acquisitions were performed using the software EPU with AFIS (Thermo Fisher Scientific).

The movies were motion-corrected globally and locally (5 × 5 patches) using the software MotionCor2[24] and saved as dose-weighted micrographs. Defocus values were estimated from aligned non-dose-weighted micrographs using the program CTFFINF4[25].

Cryo-EM reconstruction of postF in complex with 131-2a Fab

Particle picking was done in two iterations, first particles were picked using Topaz embedded in Relion-4.0.1. and subjected to 2D classification[26,27]. Selected 2D classes were used as a reference for reference based particle picking in Relion. Extracted particles were then transferred to Cryosparc v4.2.1.[28]. Several rounds of 2D classification were performed to select intact postF particles. *Ab initio* reconstruction with four requested classes resulted into the densities of postF with none, one, two, and three Fabs attached. Global 3D refinements were performed. Particle stacks were sorted for one, two and three Fab-bound complexes by iterative heterogeneous refinements using *ab initio* reconstructions as initial models. All particle stacks were separately reconstructed using homogenous, local and ctf refinements.

Cryo-EM structure determination and refinement

131-2a Fab Alphafold2 model and crystal structure of RSV postF (PDB:3RKI) was fitted into their densities as rigid bodies using Chimera and then iteratively refined in real space using the program PHENIX real_space_refine.py and corrected manually in COOT and

ISOLDE[9,29–33]. The quality of the structures and their fit to the cryo-EM density maps was assessed using comprehensive validation in the program PHENIX.

Result

De novo sequencing by bottom-up mass spectrometry

The purified mouse monoclonal antibody 131-2a was sequenced by mass spectrometry, using a bottom-up proteomics approach. The antibody was digested with a panel of 8 proteases in parallel (trypsin, chymotrypsin, α-lytic protease, thermolysin, elastase, gluC, lysC and lysN) to generate overlapping peptides for the LC-MS/MS analysis, using an in-solution digestion protocol. Peptides were sequenced from MS/MS spectra, following a hybrid fragmentation scheme with both stepped high-energy collision dissociation (sHCD) and electron-transfer high energy collision dissociation (EThcD) on all peptide precursors. The peptide sequences were predicted from the MS/MS spectra using PEAKS and assembled into the full length heavy and light chain sequences using the in-house developed software Stitch. This resulted in the identification of a mouse IgG2a antibody with an IGHV1S29 heavy chain, paired with an IGKV3-2 light chain (the full sequence is provided in the Supplementary Information). The depth of coverage for the complementarity determining regions (CDRs) varies from around 10 to 200, indicating a high sequence accuracy (see Supplementary Figure S1). Examples of MS/MS spectra supporting the CDRs of both heavy chain and light chain are shown in Figure 2. Both heavy and light chains exhibit a typical, moderate degree of somatic hypermutation of an estimated 12% and 5%, respectively. This includes several inferred mutations in the framework regions of the chains.

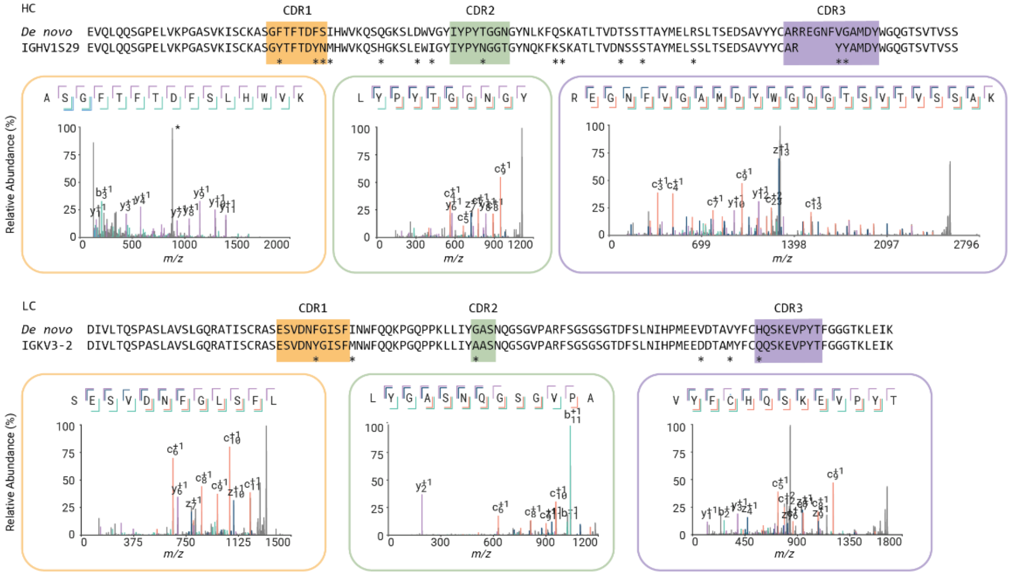


Figure 2. De novo sequencing of the commercial 131-2a based on bottom-up proteomics. The variable region alignment to the inferred germline sequence is shown for both heavy and light chains. Positions with putative somatic hypermutation are highlighted with asterisks (*). The MS/MS spectra supporting the CDR regions are shown beneath the sequence alignment, b/y ions are indicated in green and purple, while c/z ions are indicated in orange and dark blue.

Reverse engineering a functional 131-2a monoclonal antibody

The experimentally determined sequences of 131-2a were reverse translated to DNA with codon optimization for expression in HEK293 cells. The synthetic DNA for the variable domains was inserted into pRK5 plasmids containing the mouse IgG2a backbone with a C-terminal 8xHis-tag for purification for the heavy chain, and the mouse Ig Kappa backbone for the light chain. Plasmids were co-transfected in HEK293E cells, with the recombinant 131-2a yielding 95 mg from a 1 L culture following his-tag purification (see Supplementary Figure S2). The reverse engineered 131-2a was then compared with input material for sequencing in Western blot and enzyme-linked immunosorbent assay (ELISA). As shown in Figure 3, the reverse engineered 131-2a binds specifically to postfusion F, in a manner that is indistinguishable from the input material. This demonstrates that the mass spectrometry derived sequence yielded a functionally equivalent antibody product.

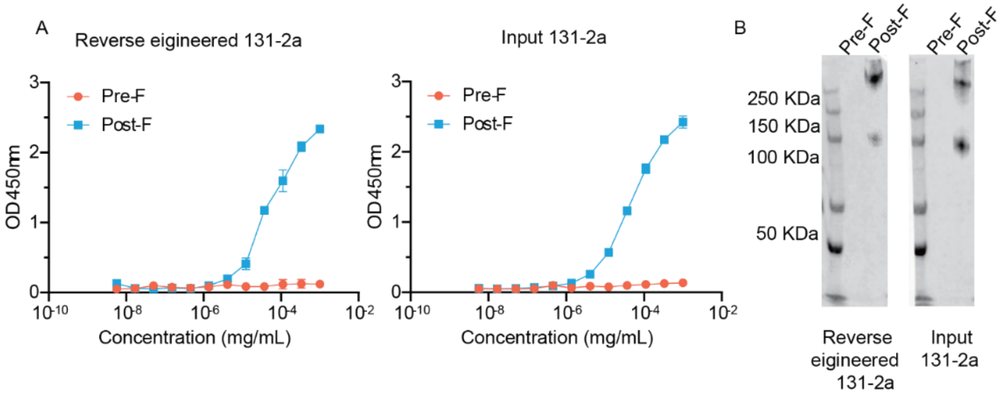


Figure 3. Validation of recombinant 131-2a expressed with mass spectrometry-derived sequence A) Elisa analysis of pre/post- F with the commercial 131-2a and the recombinant 131-2a; B) Western blot analysis of pre/post- F with the commercial 131-2a and the recombinant 131-2a

Epitope mapping of 131-2a

Despite its use as an antibody standard to define antigenic site I and specifically detect the F protein in a postfusion state, the molecular basis of 131-2a's postfusion specificity is not well understood. Reverse engineering 131-2a enabled us to map the epitope in greater detail using single particle cryoEM. The antigen-binding fragment (Fab) of 131-2a was purified by SEC following papain cleavage and added to postfusion F to form the complex. These F:131-2a complexes were deposited on holey carbon grids for vitrification and imaging without further purification. Several rounds of 2D and 3D classification recovered a clearly identifiable postfusion F head domain, while the stalk remained largely unresolved. The imaged particles consist of a mixture of four F:131-2a stoichiometries, containing 3:0, 3:1, 3:2, and 3:3 subunits of each component. The 3:1 complex was most populated in this dataset and refined to a final resolution of 3.7 Å. There were no distinguishable differences in Fab binding between the 3:1 or 3:2 and 3:3 complexes, indicating that the 3:1 stoichiometry we focused on is simply the result of the specific mixing ratio of the postfusion F and 131-2a Fab in this preparation.

The epitope of 131-2a consists mainly of a region spanning the C382-C393 loop in the F1 subunit. This loop is contacted by both CDRL2 and CDRH3, with the latter making additional contacts with the N-terminal region of the F2 subunit. CDRH1 and CDRL1 make additional contacts with the alpha helix and beta strand flanking the C382-C393 loop, respectively. CDRL2 makes additional contacts with the C322-C333 loop in the F1 subunit, which is disordered in previously published postfusion F structures, as well as in unbound F subunits in the present reconstructions[34]. The conformation of the 131-2a epitope is similar in prefusion F. However, in prefusion F access by 131-2a is blocked by the a10 helix of the F1 subunit. This helix refolds to the stalk in the postfusion conformation[14].

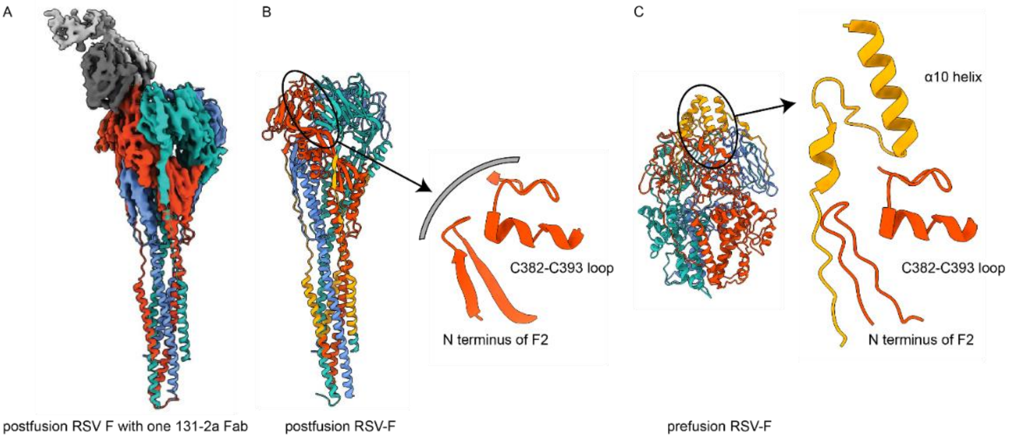


Figure 4: Panel A illustrates the Cryo-EM reconstruction of RSV postF with bound 131-2a Fab, Panel B illustrates the previously published unbound postfusion F structure (PDB: 3RRR), the C382-C393 loop is highlighted in the zoom in. Panel C illustrates the prefusion F structure (PBD: 4MMU). The α10 helix that hindered the binding and C382-C393 loop are highlighted in the zoom in. The 131-2a Fab is shown in grey, the C-terminus of F1 subunit is shown in yellow.

Conclusions

In this work, the application of the bottom-up proteomics based mass spectrometry method allows us to derive the full sequence from the anti-RSV post fusion F mAb 131-2a. This sequence information enabled the reverse engineering of a functional recombinant 131-2a antibody, which was demonstrated to possess equivalent binding specificity to RSV postfusion F when compared to the commercial 131-2a, as confirmed by Western blot and ELISA

Moreover, the study employed single particle cryo-electron microscopy (cryo-EM) to map the epitope of 131-2a in detail. It revealed that the epitope primarily encompasses a region spanning the C382-C393 loop in the F1 subunit and the N-terminus of the F2 subunit. While the conformation of the epitope remains similar in the prefusion conformation of F, access is hindered by the a10 helix of the F1 subunit, which refolds to the stalk region in the postfusion conformation. The postF specificity of 131-2a can thus be explained by negative selection of binding in the prefusion conformation. *De novo* sequencing of 131-2a by mass spectrometry enabled an in-depth molecular characterization of antigenic site I of the RSV-F protein, shedding new light on decades of serological studies characterizing the antibody response to RSV infection and vaccination.

Acknowledgements

This research was funded by the Dutch Research Council NWO Gravitation 2013 BOO, Institute for Chemical Immunology (ICI; 024.002.009) to J.S. We additionally acknowledge support from NWO, funding the MS facilities through the X-omics Road Map program (project 184.034.019). [NeCEN] [Utrecht cryoEM]

Supplementary information to:

Structural basis for postfusion-specific binding to Respiratory Syncytial Virus F protein
by the antigenic site-I antibody 131-2a.

>Heavy chain_131-2a

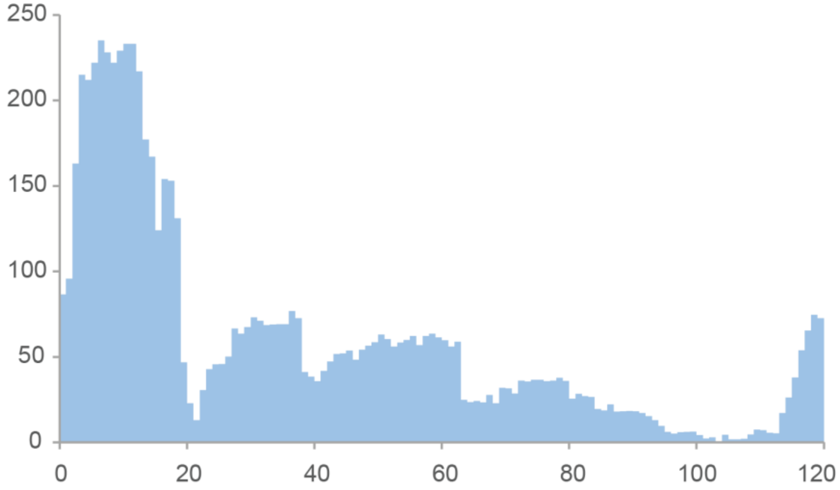
EVQLQQSGPELVKPGASVKISCKASGFTFTDFSIHWVKQSQGKSLDWVGYIYPYTGGNGYNLKF
QSKATLTVDTSSTTAYMELRSLTSEDSAVYYCARREGNFVGAMDYWGQGTSVTVSSAKTTAPS
VYPLAPVCGDTTGSSVTLGCLVKGYFPEPVTLTWNSGSLSSGVHTFPAVLQSDLYTLSSSVTVTS
STWPSQSITCNVAHPASSTKVDKKIEPRGPTIKPCPPCKCPAPNLLGGPSVFIFPPKIKDVLMISL
SPIVTCVVVDVSEDDPDVQISWFVNNVEVHTAQTQTHREDYNSTLRVVSALPIQHQDWMSGK
EFKCKVNNKDLPAPIERTISKPKGSVRAPQVYVLPPPEEEMTKKQVTLTCMVTDFMPEDIYVE
WTNNGKTELNYKNTEPVLDSDGSYFMYSKLRVEKKNWVERNSYSCSVVHEGLHNHHTTKSFS
RTPGK

>Light chain_131-2a

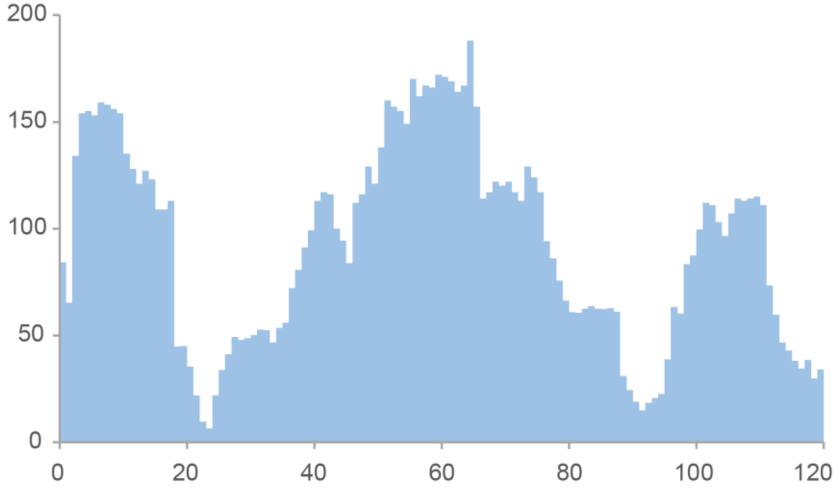
DIVLTQSPASLAVSLGQRATISCRASESVDNFGISFINWFQQKPGQPPKLLIYGASNQGSGVPARF
SGSGSGTDFSLNIHPMEEVDTAVYFCHQSKEVPYTFGGGTKLEIKRADAAPTVSIFPPSSEQLTS
GGASVVCFLNNFYPKDINVKWKIDGSERQNGVLNSWTDQDSKDSTYSMSSTLTLTKDEYERHN
SYTCEATHKTSTSPIVKSFNRNCCSNTTGSKT

Coverage map

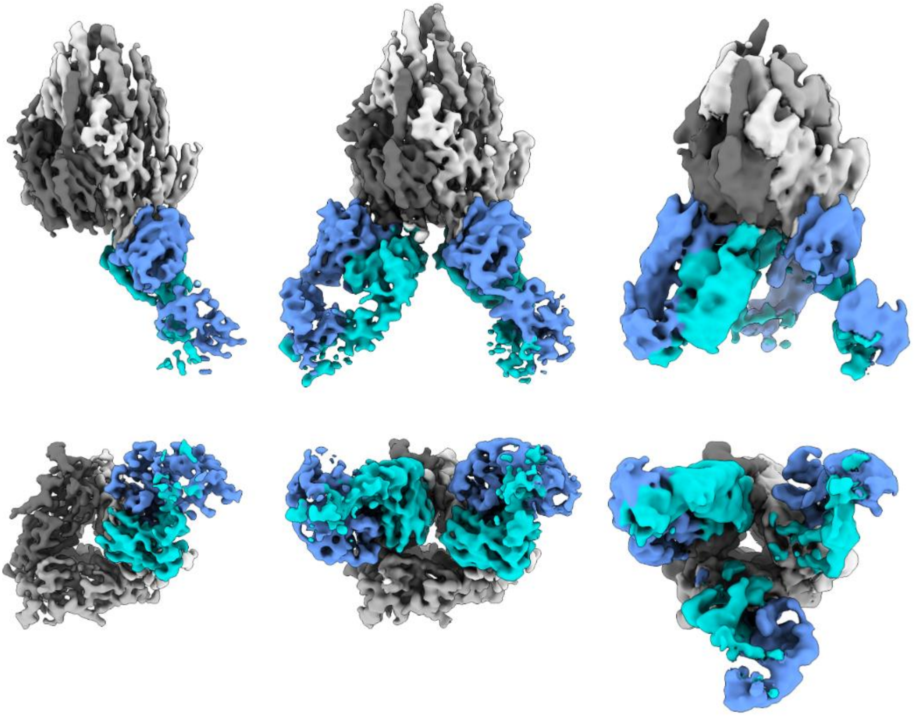
Heavy chain



Light chain



Supplementary Figure S1. Depth of coverage (total number of overlapping peptides mapped per position) for the variable domains of the 131-2a heavy and light chains.



Supplementary Figure S2. Cryo-EM reconstructions of RSV postF (grey) with one, two, and three 131-2a Fabs attached.

References

(1) Cohen, C.; Zar, H. J. Deaths from RSV in Young Infants—the Hidden Community Burden. *The Lancet Global Health* **2022**, *10* (2), e169–e170. https://doi.org/10.1016/S2214-109X(21)00558-1.

(2) Reichert, H.; Suh, M.; Jiang, X.; Movva, N.; Bylsma, L. C.; Fryzek, J. P.; Nelson, C. B. Mortality Associated With Respiratory Syncytial Virus, Bronchiolitis, and Influenza Among Infants in the United States: A Birth Cohort Study From 1999 to 2018. *The Journal of Infectious Diseases* **2022**, *226* (Supplement_2), S246–S254. https://doi.org/10.1093/infdis/jiac127.

(3) Bergeron, H. C.; Tripp, R. A. Immunopathology of RSV: An Updated Review. *Viruses* **2021**, *13* (12), 2478. https://doi.org/10.3390/v13122478.

(4) Battles, M. B.; McLellan, J. S. Respiratory Syncytial Virus Entry and How to Block It. *Nat Rev Microbiol* **2019**, *17* (4), 233–245. https://doi.org/10.1038/s41579-019-0149-x.

(5) Higgins, D.; Trujillo, C.; Keech, C. Advances in RSV Vaccine Research and Development – A Global Agenda. *Vaccine* **2016**, *34* (26), 2870–2875. https://doi.org/10.1016/j.vaccine.2016.03.109.

(6) Haynes, L. M. Progress and Challenges in RSV Prophylaxis and Vaccine Development. *The Journal of Infectious Diseases* **2013**, *208* (suppl_3), S177–S183. https://doi.org/10.1093/infdis/jit512.

(7) Shaw, C. A.; Ciarlet, M.; Cooper, B. W.; Dionigi, L.; Keith, P.; O'Brien, K. B.; Rafie-Kolpin, M.; Dormitzer, P. R. The Path to an RSV Vaccine. *Current Opinion in Virology* **2013**, *3* (3), 332–342. https://doi.org/10.1016/j.coviro.2013.05.003.

(8) López, J. A.; Bustos, R.; Örvell, C.; Berois, M.; Arbiza, J.; García-Barreno, B.; Melero, J. A. Antigenic Structure of Human Respiratory Syncytial Virus Fusion Glycoprotein. *Journal of Virology* **1998**, *72* (8), 6922–6928. https://doi.org/10.1128/jvi.72.8.6922-6928.1998.

(9) Mousa, J. J.; Sauer, M. F.; Sevy, A. M.; Finn, J. A.; Bates, J. T.; Alvarado, G.; King, H. G.; Loerinc, L. B.; Fong, R. H.; Doranz, B. J.; Correia, B. E.; Kalyuzhniy, O.; Wen, X.; Jardetzky, T. S.; Schief, W. R.; Ohi, M. D.; Meiler, J.; Crowe, J. E. Structural Basis for Nonneutralizing Antibody Competition at Antigenic Site II of the Respiratory Syncytial Virus Fusion Protein. *Proceedings of the National Academy of Sciences* **2016**, *113* (44), E6849–E6858. https://doi.org/10.1073/pnas.1609449113.

(10) Anderson, L. J.; Hierholzer, J. C.; Tsou, C.; Hendry, R. M.; Fernie, B. F.; Stone, Y.; McIntosh, K. Antigenic Characterization of Respiratory Syncytial Virus Strains with Monoclonal Antibodies. *The Journal of Infectious Diseases* **1985**, *151* (4), 626–633. https://doi.org/10.1093/infdis/151.4.626.

(11) Fernie, B. F.; Cote, P. J.; Gerin, J. L. Classification of Hybridomas to Respiratory Syncytial Virus Glycoproteins. *Proceedings of the Society for Experimental Biology and Medicine* **1982**, *171* (3), 266–271. https://doi.org/10.3181/00379727-171-41509.

(12) Peng, W.; Pronker, M. F.; Snijder, J. Mass Spectrometry-Based De Novo Sequencing of Monoclonal Antibodies Using Multiple Proteases and a Dual Fragmentation Scheme. *J. Proteome Res.* **2021**, *20* (7), 3559–3566. https://doi.org/10.1021/acs.jproteome.1c00169.

(13) Peng, W.; den Boer, M. A.; Tamara, S.; Mokiem, N. J.; van der Lans, S. P. A.; Bondt, A.; Schulte, D.; Haas, P.-J.; Minnema, M. C.; Rooijakkers, S. H. M.; van Zuilen, A. D.; Heck, A. J. R.; Snijder, J. Direct Mass Spectrometry-Based Detection and Antibody Sequencing of Monoclonal Gammopathy of Undetermined Significance from Patient Serum: A Case Study. *J. Proteome Res.* **2023**. https://doi.org/10.1021/acs.jproteome.3c00330.

(14) *Structure-Based Design of a Fusion Glycoprotein Vaccine for Respiratory Syncytial Virus | Science*. https://www.science.org/doi/10.1126/science.1243283 (accessed 2023-08-29).

(15) *Characterization of Epitope-Specific Anti-Respiratory Syncytial Virus (Anti-RSV) Antibody Responses after Natural Infection and after Vaccination with Formalin-Inactivated RSV | Journal of Virology*. https://journals.asm.org/doi/10.1128/jvi.00235-16?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub++0pubmed (accessed 2023-08-29).

(16) Rigter, A.; Widjaja, I.; Versantvoort, H.; Coenjaerts, F. E. J.; Roosmalen, M. van; Leenhouts, K.; Rottier, P. J. M.; Haijema, B. J.; Haan, C. A. M. de. A Protective and Safe Intranasal RSV Vaccine Based on a Recombinant Prefusion-Like Form of the F Protein Bound to Bacterium-Like Particles. *PLOS ONE* **2013**, *8* (8), e71072. https://doi.org/10.1371/journal.pone.0071072.

(17) Harbury, P. B.; Zhang, T.; Kim, P. S.; Alber, T. A Switch Between Two-, Three-, and Four-Stranded Coiled Coils in GCN4 Leucine Zipper Mutants. *Science* **1993**, *262* (5138), 1401–1407. https://doi.org/10.1126/science.8248779.

(18) van Roosmalen, M. L.; Kanninga, R.; El Khattabi, M.; Neef, J.; Audouy, S.; Bosma, T.; Kuipers, A.; Post, E.; Steen, A.; Kok, J.; Buist, G.; Kuipers, O. P.; Robillard, G.; Leenhouts, K. Mucosal Vaccine Delivery of Antigens Tightly Bound to an Adjuvant Particle Made from Food-Grade Bacteria. *Methods* **2006**, *38* (2), 144–149. https://doi.org/10.1016/j.ymeth.2005.09.015.

(19) Ascough, S.; Vlachantoni, I.; Kalyan, M.; Haijema, B.-J.; Wallin-Weber, S.; Dijkstra-Tiekstra, M.; Ahmed, M. S.; van Roosmalen, M.; Grimaldi, R.; Zhang, Q.; Leenhouts, K.; Openshaw, P. J.; Chiu, C. Local and Systemic Immunity against Respiratory Syncytial Virus Induced by a Novel Intranasal Vaccine. A Randomized, Double-Blind, Placebo-Controlled Clinical Trial. *Am J Respir Crit Care Med* **2019**, *200* (4), 481–492. https://doi.org/10.1164/rccm.201810-1921OC.

(20) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. PEAKS: Powerful Software for Peptide de Novo Sequencing by Tandem Mass Spectrometry. *Rapid Communications in Mass Spectrometry* **2003**, *17* (20), 2337–2342. https://doi.org/10.1002/rcm.1196.

(21) Schulte, D.; Peng, W.; Snijder, J. Template-Based Assembly of Proteomic Short Reads For De Novo Antibody Sequencing and Repertoire Profiling. *Anal. Chem.* **2022**, *94* (29), 10391–10399. https://doi.org/10.1021/acs.analchem.2c01300.

(22) Lefranc, M.-P. IMGT® Databases, Web Resources and Tools for Immunoglobulin and T Cell Receptor Sequence Analysis, Http://Imgt.Cines.Fr. *Leukemia* **2003**, *17* (1), 260–266. https://doi.org/10.1038/sj.leu.2402637.

(23) Ehrenmann, F.; Kaas, Q.; Lefranc, M.-P. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: A Database and a Tool for Immunoglobulins or Antibodies, T Cell Receptors, MHC, IgSF and MhcSF. *Nucleic Acids Research* **2010**, *38* (suppl_1), D301–D307. https://doi.org/10.1093/nar/gkp946.

(24) Zheng, S. Q.; Palovcak, E.; Armache, J.-P.; Verba, K. A.; Cheng, Y.; Agard, D. A. MotionCor2: Anisotropic Correction of Beam-Induced Motion for Improved Cryo-Electron Microscopy. *Nat Methods* **2017**, *14* (4), 331–332. https://doi.org/10.1038/nmeth.4193.

(25) Rohou, A.; Grigorieff, N. CTFFIND4: Fast and Accurate Defocus Estimation from Electron Micrographs. *J Struct Biol* **2015**, *192* (2), 216–221. https://doi.org/10.1016/j.jsb.2015.08.008.

(26) Bepler, T.; Morin, A.; Rapp, M.; Brasch, J.; Shapiro, L.; Noble, A. J.; Berger, B. Positive-Unlabeled Convolutional Neural Networks for Particle Picking in Cryo-Electron Micrographs. *Nat Methods* **2019**, *16* (11), 1153–1160. https://doi.org/10.1038/s41592-019-0575-8.

(27) *A Bayesian approach to single-particle electron cryo-tomography in RELION-4.0 | eLife*. https://elifesciences.org/articles/83724 (accessed 2024-01-11).

(28) Punjani, A.; Rubinstein, J. L.; Fleet, D. J.; Brubaker, M. A. cryoSPARC: Algorithms for Rapid Unsupervised Cryo-EM Structure Determination. *Nat Methods* **2017**, *14* (3), 290–296. https://doi.org/10.1038/nmeth.4169.

(29) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(30) *UCSF ChimeraX: Meeting modern challenges in visualization and analysis - Goddard - 2018 - Protein Science - Wiley Online Library*. https://onlinelibrary.wiley.com/doi/10.1002/pro.3235 (accessed 2024-01-11).

(31) Afonine, P. V.; Poon, B. K.; Read, R. J.; Sobolev, O. V.; Terwilliger, T. C.; Urzhumtsev, A.; Adams, P. D. Real-Space Refinement in PHENIX for Cryo-EM and Crystallography. *Acta Crystallogr D Struct Biol* **2018**, *74* (Pt 6), 531–544. https://doi.org/10.1107/S2059798318006551.

(32) Emsley, P.; Cowtan, K. Coot: Model-Building Tools for Molecular Graphics. *Acta Crystallogr D Biol Crystallogr* **2004**, *60* (Pt 12 Pt 1), 2126–2132. https://doi.org/10.1107/S0907444904019158.

(33) Croll, T. I. ISOLDE: A Physically Realistic Environment for Model Building into Low-Resolution Electron-Density Maps. *Acta Crystallogr D Struct Biol* **2018**, *74* (Pt 6), 519–530. https://doi.org/10.1107/S2059798318002425.

(34) McLellan, J. S.; Yang, Y.; Graham, B. S.; Kwong, P. D. Structure of Respiratory Syncytial Virus Fusion Glycoprotein in the Postfusion Conformation Reveals

Preservation of Neutralizing Epitopes. *Journal of Virology* **2011**, *85* (15), 7788–7796. https://doi.org/10.1128/jvi.00555-11.

Chapter 6

Direct mass spectrometry-based detection and antibody sequencing of Monoclonal Gammopathy of Undetermined Significance from patient serum – a case study.

Weiwei Peng[1], Maurits A. den Boer[1], Sem Tamara[1], Nadia J. Mokiem[1], Sjors P.A. van der Lans[2], Albert Bondt[1], Douwe Schulte[1], Pieter-Jan Haas[2], Monique C. Minnema[3], Suzan H.M. Rooijakkers[2], Arjan D. van Zuilen[4], Albert J.R. Heck[1], Joost Snijder[1*]

[1] Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute of Pharmaceutical Sciences, Utrecht University, Padualaan 8, 3584, CH, Utrecht, The Netherlands
[2] Medical Microbiology, University Medical Center Utrecht, Utrecht University, Heidelberglaan 100, 3584CX Utrecht, The Netherlands
[3] Department of Hematology, University Medical Center Utrecht, Utrecht University, Heidelberglaan 100, 3584CX Utrecht, the Netherlands
[4] Department of Nephrology and Hypertension, University Medical Center Utrecht, Utrecht University, Heidelberglaan 100, 3584CX Utrecht, the Netherlands

* corresponding author: j.snijder@uu.nl

Published as:

https://doi.org/10.1021/acs.jproteome.3c00330

Abstract:

Monoclonal gammopathy of undetermined significance (MGUS) is a plasma cell disorder, characterized by the presence of a predominant monoclonal antibody (*i.e.,* M-protein) in serum, without clinical symptoms. Here we present a case study in which we detect MGUS by liquid-chromatography coupled with mass spectrometry (LC-MS) profiling of IgG1 in human serum. We detected a Fab-glycosylated M-protein and determined the full heavy and light chain sequences by bottom-up proteomics techniques using multiple proteases, further validated by top-down LC-MS. Moreover, the composition and location of the Fab-glycan could be determined in CDR1 of the heavy chain. The outlined approach adds to an expanding mass spectrometry-based toolkit to characterize monoclonal gammopathies such as MGUS and multiple myeloma, with fine molecular detail. The ability to detect monoclonal gammopathies and determine M-protein sequences straight from blood samples by mass spectrometry provides new opportunities to understand the molecular mechanisms of such diseases.

Keywords:

M-protein; monoclonal gammopathy; glycosylation; antibody; *de novo* sequencing; middle-down

Introduction

Monoclonal gammopathy of undetermined significance (MGUS) is a plasma cell disorder, characterized by the presence of a predominant monoclonal antibody (*i.e.,* M-protein) in patient serum. [1] MGUS is a preclinical stage of Multiple Myeloma (MM), with an estimated annual risk of 1% to progress to MM. [2] The most common antibody isotype in MGUS patients is IgG, which is a heterodimer consisting of two identical pairs of heavy chains (HC) and light chains (LC).[3,4] All IgG share one conserved N-linked glycosylation site on each copy of the HC in the Fc region.[5] However, the M proteins present in both MGUS and MM patients have been reported to have a high frequency of unusual additional glycosylation in the Fab region, present in the variable domains of either the light or heavy chains. [6,7]

We recently developed methods for direct mass spectrometry-based repertoire profiling and sequencing of IgG1 from human serum.[8,9] In this method IgG is affinity purified from serum samples, followed by selective digestion and release of the IgG1 Fab portion.[10] Subsequent analysis of the released Fabs by reversed phase liquid chromatography, coupled with mass spectrometry (LC-MS), establishes a highly resolved map of antibody clones separated by mass and retention time, spanning at least 2 orders of magnitude in abundance. By spiking in monoclonal antibodies at known concentrations, also absolute concentrations of endogenous clones can be estimated by normalizing their signal intensities. Typically, we detect a few hundred of the most abundant clones, together making up 50-90% of the total subclass concentration. The most abundant IgG1 clones are generally in the order of 0.01-0.1 mg/mL, with sometimes outliers of up to about 1 mg/mL in hospitalized patients in critical condition.[11,12]

During screening of serum IgG1 repertoires of a new cohort of donors by LC-MS we observed a donor whose repertoire was exceptionally dominated by a single IgG1 clone exhibiting a very high concentration of approximately 10 mg/mL in serum. This single clone contributes approximately 98% to the total amount of all IgG1 molecules in the serum of this patient. Subsequent clinical tests confirmed diagnosis of MGUS. Our MS data also indicated that this MGUS M-protein harbored abundant Fab glycosylation. Combining the above IgG1 profiling method with bottom-up proteomics-based sequencing, we were able to recover the full sequence of the antibody heavy and light chains. The Fab glycosylation could be traced back to a specific residue in CDR1 of the heavy chain. The attached N-glycan structures could be assigned and quantified based on the intact Fab MS spectra and tandem MS spectra of the corresponding glycopeptides. This case illustrates how integrated bottom-up and top-down proteomics can be used to detect MGUS and other monoclonal gammopathies, sequence the associated monoclonal antibody against a background of serum IgG1, even when the M-protein is Fab-glycosylated, and that the composition of this Fab-glycosylation can be determined and localized from the same sampled material.

Method

Cohort and Trial information

In the period of 2015-2019 patients who underwent kidney transplantation were asked to participate in a biobank to evaluate immunological developments after kidney transplantation. All participants provided written informed consent to collect clinical data and serum samples pre-transplantation and at month 1, 3, 6, and 12 post-transplantation. The study was approved by the local Biobank Research Ethics Committee (protocol 15-019). Serum samples of patients with a recorded bacterial infection after kidney transplantation were identified and analyzed to evaluate the immunological response to such an infection. In one of the patients the samples showed a few extremely abundant clones.

IgG purification and Fab production

IgG1 clonal profiling was performed based on a method previously described by Bondt et al. [8] Two internal reference mAbs (trastuzumab and alemtuzumab) were spiked into serum samples of 10 µL to a final concentration of 20 µg/mL (200 ng), after which IgG was captured using 10 µL CaptureSelect FcXL affinity matrix (20 µL slurry, Thermo Fisher) in a spin column. After binding for 60 min on a shaker at 750 rpm and room temperature, columns were washed in four sequential rounds by adding 150 µL PBS and removing the liquid by centrifugation for 1 min at 500 g. IgG1 Fab molecules were released on through on-bead proteolytic digestion using 100 U IgdE (FabALACTICA; Genovis) in 50 µL 150 mM sodium phosphate pH 7.0 overnight on a shaker at 750 rpm and 37 °C. Liquid containing free IgG1 Fabs was captured through centrifugation for 1 min at 1,000 g.

Analysis was performed by reversed-phase LC-MS using a Vanquish Flex UHPLC system (Thermo Fisher) coupled to an Orbitrap Exploris 480 instrument (Thermo Scientific. Chromatographic separation was performed on a 1 x 150 mm MAbPac column at 80 °C and using a flow rate of 150 µL/min. Mobile phase A consisted of MilliQ water with 0.1% formic acid, mobile phase B of acetonitrile with 0.1% formic acid. Samples were run starting with a 10%-25% B ramp with the spray voltage turned off for 2 minutes to wash away salts. This was followed by a 54 min linear gradient of 25%-40% B, a 95% B wash and re-equilibration at 10% B. The mass spectrometer was operated at low pressure setting in Intact Protein mode at a set resolution of 7,500 at 200 *m/z.* For every scan, 5 µscans were acquired with an *m/z* range of 500 - 4,000 using an AGC target of 300% with a maximum injection time of 50 ms. The RF lens was set to 40% and a source fragmentation energy of 15 V was used. Raw data were processed by sliding window deconvolution using the ReSpect algorithm in BioPharma Finder v3.2 (Thermo Fisher). Further analysis was performed using an in-house python library described by Bondt et al. Components with masses between 45,000 and 53,000 Da, most intense charge states above m/z 1,000, and a Score of over 40 were considered valid Fab identifications.

Bottom-up proteomics

In-gel digestion – Fab (3 µg/lane) was loaded on a 4%-12% Bis-Tris precast gel (Bio-rad) in non-reducing conditions and run at 120 V in 3-Morpholinopropane-1-sulfonic acid (MOPS) buffer (Bio-rad). Bands were visualized with Imperial Protein Stain (Thermo Fisher Scientific), and the size of the fragments evaluated by running a protein standard ladder (Bio-rad). The Fab bands were cut and reduced by 10 mM TCEP at 37°C, then alkylated in 40 mM IAA at RT in the dark, followed by alkylation in 40 mM IAA at RT in the dark. The Fab bands were digested by trypsin, chymotrypsin, thermolysin, and alpha lytic protease at 37 °C overnight in 50 mM ammonium bicarbonate buffer. The peptides were extracted with two steps incubation at RT in 50% ACN, and 0.01% TFA, and then 100% ACN respectively. The peptides were dried in speed-vac. To obtain the sequence of the glycosylated Fab, the N-linked glycan was removed by PNGaseF at 37 °C overnight then in gel digested as described above.

Mass Spectrometry – The digested peptides were separated by online reversed phase chromatography on an Dionex UltiMate 3000 (Thermo Fisher Scientific) (column packed with Poroshell 120 EC C18; dimensions 50 cm × 75 µm, 2.7 µm, Agilent Technologies) coupled to a Thermo Scientific Orbitrap Fusion mass spectrometer or Thermo Scientific Orbitrap Fusion LUMOS mass spectrometer. Samples were eluted over a 90 min gradient from 0 to 35% acetonitrile at a flow rate of 0.3 µL/min. Peptides were analyzed with a resolution setting of 60 000 in MS1. MS1 scans were obtained with a standard automatic gain control (AGC) target, a maximum injection time of 50 ms, and a scan range of 350–2000. The precursors were selected with a 3 m/z window and fragmented by stepped high-energy collision dissociation (HCD) and electron-transfer higher-energy collision dissociation (EThcD). The stepped HCD fragmentation included steps of 25, 35, and 50% normalized collision energies (NCE). EThcD fragmentation was performed with calibrated charge-dependent electron-transfer dissociation (ETD) parameters and 27% NCE supplemental activation. For both fragmentation types, MS2 scans were acquired at a 30 000 resolution, a 4e5 AGC target, a 250 ms maximum injection time, and a scan range of 120–3500.

Peptide Sequencing from MS/MS Spectra – MS/MS spectra were used to determine *de novo* peptide sequences using PEAKS Studio X (version 10.6).[13,14] We used a tolerance of 20 ppm and 0.02 Da for MS1 and 0.02 Da for MS2, respectively. Carboxymethylation was set as fixed modification of cysteine and variable modification of peptide N-termini and lysine. Oxidation of methionine and tryptophan, pyroglutamic acid modification of N-terminal glutamic acid, and glutamine were set as additional variable modifications. The CSV file containing all the *de novo* sequenced peptides was exported for further analysis.

Template-based assembly via Stitch – Stitch[15] (1.1.2) was used for the template-based assembly. The human antibody database from IMGT was used as template. The cutoff score for the *de novo* sequenced peptide was set as 90/70 and the cutoff score for the template matching was set as 10. All the peptides supporting the sequences were examined manually. The ions for annotation of the CDR regions were exported and visualized by Interactive Peptide Spectral Annotator.[16]

Glycoproteomics data analysis – Chymotryptic digested peptides were used to search for site specific glycosylation via Byonic (v5.0.3).[17] The *de novo* obtained sequences were selected as protein database. Four missed cleavages were permitted using C-terminal cleavage at WFMLY for chymotrypsin. Carboxymethylation of cysteine was set as fixed modification, oxidation of methionine/tryptophan as variable rare 1, Gln to pyro-Glu and Glu to pyro-glu on the N-temmius of protein as rare 1, and N-glycan modifications were set as variable rare 1. The N-glycan 132 human database from Byonic was applied in the search. All reported glycopeptides in the Byonic result files were manually inspected for quality of fragment assignments.

Native MS

To remove the N-linked glycan on the fab, the samples was incubated with 1% Rapigest (Waters Corporation, USA) for 3 min at 90 °C.the PNGaseF was added to the sample and incubated at 50 °C for 10 min. Both the native fab and the deglycosylated fab were buffer exchanged into 150 mM ammonium acetate (pH 7.5) using Amicon 10 kDa MWCO centrifugal filters (Merck Millipore). The samples were loaded into gold-coated borosilicate capillaries (in-house prepared) and analyzed on an ultra-high mass range (UHMR) Q-Exactive Orbitrap (Thermo Fisher Scientific, Bremen, Germany). The mass spectra were obtained in positive mode with an ESI voltage of 1.3 kV. The maximum injection was set at 100 ms and the HCDenergy was set at 100 V. The used resolution was 12500 at 400 m/z. The S-Lens level was set at 200. UniDec was used to generating the charge-deconvoluted spectrum. [18]

MD proteomics

The reduced Fab was freshly prepared by incubating with TCEP at 60°C for 30 min before injecting to MS. Around 1 μg sample was used for a single measurement. Reduced Fab was measured by LC-MS/MS. Samples were loaded on a Thermo Scientific Vanquish Flex UHPLC instrument, equipped with a 1 mm x 150 mm MAbPac RP analytical column, directly coupled to an Orbitrap Fusion Lumos Tribrid (Thermo Scientific, San Jose, CA, USA). The samples were eluted over 22 min at a 150 μL/min flow rate. Gradient elution was achieved by using two mobile phases A (0.1% HCOOH in Milli-Q) and B (0.1% HCOOH in CH3CN) and ramping up B from 10 to 25% over one minute, from 25 to 40% over 14 min, and from 40 to 95% over one minute. MS data were collected with the instrument operating in Intact Protein and Low Pressure mode. Spray voltage was set at 3.5 kV, capillary temperature 350 °C, probe heater temperature 100 °C, sheath gas flow 15, auxiliary gas flow 5, and source-induced dissociation was set at 15 V. Separate Fab chains were analyzed with a resolution setting of 120,000. MS1 scans were acquired in a range of 500-3,000 Th with the 250% AGC target and a maximum injection time set to either 50 ms for the 7,500 resolution or 250 ms for the 120,000 resolution. In MS1, 2 µscans were recorded for the 7,500 resolution and 5 µscans for the 120,000 resolution per scan. Data-dependent mode was defined by the number of scans: single scan for intact Fabs and two scans for separate Fab chains. MS/MS scans were acquired with a resolution of 120,000, a maximum injection time of 500 ms, a 1,000% AGC target, and 5 µscans averaged and recorded per scan for the separate Fab chains. The EThcD active was set at true. The ions of interest were mass-selected by quadrupole in a 4 Th isolation window and accumulated to the AGC target prior to fragmentation. MS/MS spectra were used to validate the sequences using LC-MS Spectator (Version 1.1.8313.28552) and ProSight Lite (1.4.8).[19,20] In LC-MS Spectator, we used a tolerance of 10 ppm for MS1 and 20 ppm for MS2, respectively and applied the S/N threshold filtering (1.5). All the annotated ions were exported and visualized in ProSight Lite.

Structural model of glycosylated MGUS Fab

The variable domain of the MGUS Fab was modelled using the ABodyBuilder2 webserver from the SAbPred suite. The predominant HexNAc(5)Hex(5)Fuc(1)NeuAc(2) glycoform was modelled as diantennary, bisected complex glycan with core fucosylation using the GLYCAM glycoprotein builder webserver. Figures were rendered in ChimeraX.

Results

Observation of a Fab-glycosylated IgG1 M protein

When analyzing the serum IgG1 clonal repertoire of a patient that had undergone a recent kidney transplant, we unexpectedly encountered an atypical antibody profile. This patient was part of a longitudinal study cohort who were observed after kidney transplantation for immunological monitoring. Per protocol in these patients, serum samples were obtained at different time points, starting from moments before surgery (t = 0 days) with follow-up samplings for over a year which were stored in a biobank. Samples of patients who developed designated bacterial infections in follow up were drawn from the biobank and to investigate how the antibody repertoire responded to the surgery and subsequent infections, we applied our LC-MS IgG1 profiling approach to these longitudinal serum samples.

Strikingly, the IgG1 repertoire of this particular kidney transplant patient was very different from that of other donors, being dominated by seemingly a few extremely abundant clones that also exhibit relatively high masses (Figure 1A). This pattern remained unchanged before or after kidney transplantation or even after an episode of sepsis with a *Klebsiella* species (Supplementary Figure S1).

Because these antibodies were detected at relatively high masses, we hypothesized that the unusual antibody profile of the patient resulted from a single IgG1 clone of very high concentration that may carry Fab-glycosylation. Based on experimental data from our lab and theoretical calculations using sequences from the ImMunoGeneTics (IMGT) database,[21] the bulk of IgG1 Fabs have backbone masses of roughly 46-50 kDa. The abundant Fabs that we detected, however, have higher masses of more than 50 kDa. This strongly suggests that they are modified by N-glycosylation, which would contribute roughly 2 kDa to their mass. Combining the signal intensities of the multiple putative glycoforms, the total concentration of this clone is approximately 10 mg/mL at t = 0, remaining high throughout the longitudinal follow up (see Supplementary Figure S1). Compared to the total IgG1 concentration in human plasma, approximately 8 mg/mL on average, this is extremely high, prompting additional clinical tests. The patient was tested for an M protein using serum immunofixation, which confirmed the IgG kappa M protein, as well as serum electrophoresis, which could not quantify the M protein due to low amount. In addition, free light chains (FLC) were determined (Binding Site®), demonstrating an elevated FLC kappa of 61.52 mg/L, and a FLC ratio of 3.92 (normal range 0.26-1.65). No bone marrow biopsy was done and a diagnosis of monoclonal gammopathy of undetermined significance (MGUS) was made.

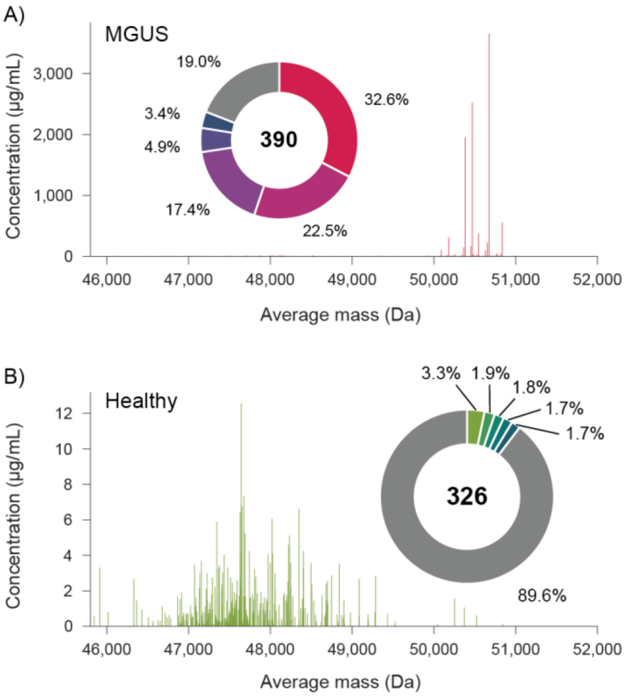


Figure 1. Detection of MGUS by LC-MS based IgG1 Fab-profiling. A) Serum IgG1 Fab profile of human subject with putative MGUS. B) Illustrative IgG1 Fab profile from serum of a healthy human donor. Note the difference in complexity, average mass, and concentration of detected IgG1 Fabs, hinting at the presence of a Fab-glycosylated M-protein in A). Pie charts show the relative abundance (%) of the top-5 most abundant Fab species.

Direct MS-based sequencing and glycan localization of the serum-derived MGUS clone

We have recently demonstrated the direct MS-based sequencing of both recombinant and serum-derived antibodies, using bottom-up proteomics methods.[8,9,22,23] Owing to the high abundance of the MGUS M-protein in the serum samples of this donor, we were able to sequence the full antibody without further fractionation. First, we used an in-gel digestion protocol, using in parallel four proteases of complementary specificity, to obtain overlapping peptides for *de novo* sequencing by LC-MS/MS analysis. We identified the M-protein as consisting of an IGHV4-4 heavy chain, coupled with an IGKV3-20 light chain. Notably, this experiment was performed with intact N-glycosylation and resulted in a lack of coverage in CDR1 of the heavy chain. The germline sequence of IGHV4-4 CDRH1 contains 5 serine residues and a single asparagine, priming it to obtain an N-glycosylation sequon by as many as 6 independent substitutions. These observations pointed to CDRH1 as a likely region to contain the putative Fab-glycosylation.

Digestion with PNGase F results in the removal of the N-glycan and converts the glycan-linked asparagine to an aspartic acid residue.[24] We digested the serum sample with PNGase F, followed by proteolysis with chymotrypsin and thermolysin in parallel. This recovered the previously missing CDRH1 sequence, containing a clear DSS motif, which would have corresponded with an NSS glycosylation sequon in the antibody prior to PNGase F digestion. Using the experimentally determined sequence, we then performed a glycoproteomics database search including common human N-glycans and were able to detect a predominant HexNAc(5)Hex(5)Fuc(1)NeuAc(2) glycan at the identified NSS sequon in CDRH1 (see Supplementary Figure S2).

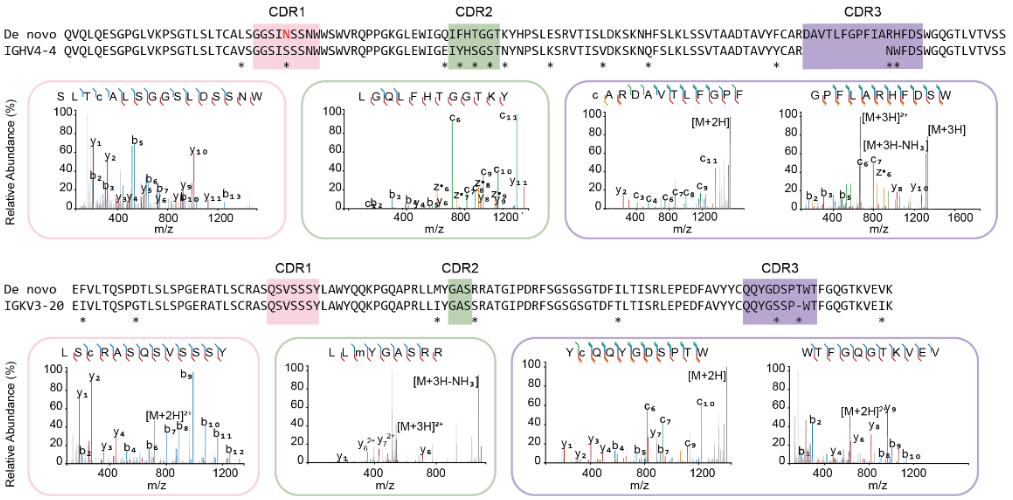


Figure 2. De novo sequencing of the MGUS Fab by bottom-up proteomics. The variable region alignment to the inferred germline sequence is shown for both heavy and light chains. Positions with putative somatic hypermutation are highlighted with asterisks (*). The MS/MS spectra supporting the annotation of the CDRs are shown beneath the sequence

alignment. b/y ions are indicated in blue and red, while c/z ions are indicated in green and yellow.

Validation of the sequence and Fab glycosylation by middle-down LC-MS/MS

To validate our bottom-up *de novo* sequencing result, we further analyzed the MGUS Fab by native MS and middle-down LC-MS/MS. The intact mass profile of the Fab is shown in Figure 3A. The observed masses are consistent with the determined sequence, considering a pyroglutamic acid modification of the heavy chain N-terminus, and reveals heterogeneous glycosylation with a predominant HexNAc(5)Hex(5)Fuc(1)NeuAc(2) glycan, as also observed by bottom-up LC-MS/MS (see Supplementary Table S1). A predicted structural model of the variable domain shows this glycan protruding outwards from CDRH1, leaving the other CDR loops exposed (see Figure 3B). The observed Fab glycosylation pattern follows a similar trend as reported by Bondt *et al.* in that it is enriched in galactosylation, sialylation, and bisection compared to Fc glycosylation at the conserved N297 site (see Figure 3C; a full overview of glycoforms of the MGUS Fab is provided in Supplementary Table S1).[25]

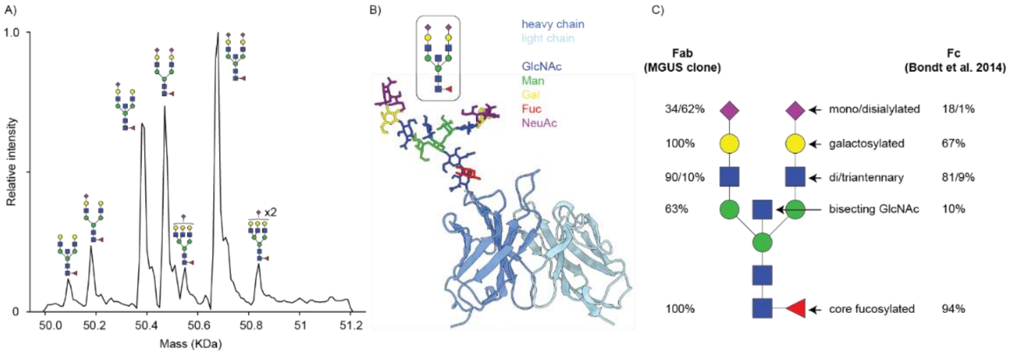
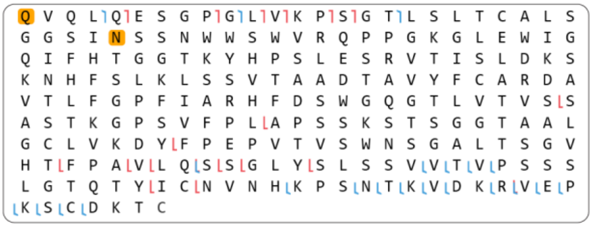


Figure 3. N-Glycosylation of MGUS Fab. A) Intact mass profile from native MS; peaks are annotated according to the assigned glycan structure. B) ABodyBuilder2 structural model prediction of the MGUS Fab variable domain with HexNAc(5)Hex(5)Fuc(1)NeuAc(2) grafted on CDRH1 using GLYCAM. C) Glycosylation profile of MGUS Fab compared to typical Fc glycosylation at N297, according to Bondt et al. 2014[25]

We performed middle-down fragmentation of the reduced Fab by EThcD to confirm the sequence determined by bottom-up proteomics. This resulted in a coverage of 25.7% for the Fd and 60.5% for the LC (see Figure 4 and Supplementary Figure S3). The obtained sequence coverage is in line with a recent interlaboratory middle-down EThcD fragmentation benchmark on known monoclonal antibody standards, where an average coverage of 20-25% was obtained on both Fd and LC[26]. The coverage we obtain on the Fd of this M-protein is similar, while the coverage of the LC is substantially higher. We attribute the relatively lower coverage of the Fd to the presence of the N-glycan in CDRH1, which is likely to fragment during EThcD, producing more complex spectra, with lower signal for individual fragments. Furthermore, glycan fragmentation is not yet implemented in currently available peak matching algorithms for middle-down LC-MS/MS. In line with this explanation, none of the observed fragment ions for the Fd supersedes the position of the N-glycan in either the b/c or y/z series. Nonetheless, the intact masses and middle-down fragmentation patterns support the M-protein sequence determined by bottom-up proteomics.

Heavy Chain (coverage 26%)


Q V Q L Q E S G P G L V K P S G T L S L T C A L S
G G S I N S S N W W S W V R Q P P G K G L E W I G
Q I F H T G G T K Y H P S L E S R V T I S L D K S
K N H F S L K L S S V T A A D T A V Y F C A R D A
V T L F G P F I A R H F D S W G Q G T L V T V S S
A S T K G P S V F P L A P S S K S T S G G T A A L
G C L V K D Y F P E P V T V S W N S G A L T S G V
H T F P A V L Q S S G L Y S L S S V V T V P S S S
L G T Q T Y I C N V N H K P S N T K V D K R V E P
K S C D K T C

Light Chain (coverage 61%)

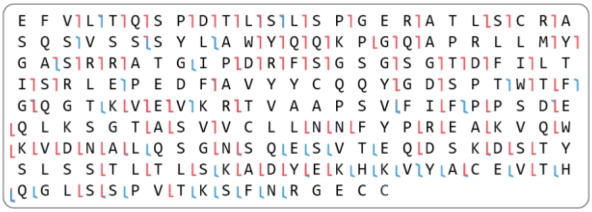

E F V L T Q S P D T L S L S P G E R A T L S C R A
S Q S V S S S Y L A W Y Q Q K P G Q A P R L L M Y
G A S R R A T G I P D R F S G S G S G T D F I L T
I S R L E P E D F A V Y Y C Q Q Y G D S P T W T F
G Q G T K V E V K R T V A A P S V F I F P P S D E
Q L K S G T A S V V C L L N N F Y P R E A K V Q W
K V D N A L Q S G N S Q E S V T E Q D S K D S T Y
S L S S T L T L S K A D Y E K H K V Y A C E V T H
Q G L S S P V T K S F N R G E C C

Figure 4. Top-down LC-MS/MS data on the of Fab-glycosylated M-protein. Shown are the heavy and light chain sequences, position supported by b/c or y/z ions from EThcD fragmentation are indicated in red and blue, respectively.

Conclusion

Here we demonstrate that LC-MS based IgG1 profiling of patient serum can lead to the detection of an M-protein, which can be related to diseases such as MGUS or Multiple Myeloma. The associated sequence of the M-protein can be fully derived by mass spectrometry. In this case, mass spectrometry also revealed the presence, location and composition of Fab glycosylation in the heavy chain of the M-protein. While the Fab-profiling workflow presented here is limited to the IgG1 subclass, and the *de novo* sequencing and glycan profiling methods are best suited for research applications, we believe that a robust implementation of Fab profiling across all IgG subclasses by LC-MS holds promise for the clinical detection of M-proteins and diagnostics of monoclonal gammopathies. The outlined approach in this case study adds to an expanding mass spectrometry-based toolkit to characterize monoclonal gammopathies such as MGUS and MM with fine molecular detail.[27–32] The ability to detect monoclonal gammopathies and determine M-protein sequences straight from peripheral blood samples by mass spectrometry provides opportunities to understand the molecular mechanisms of these diseases.

Acknowledgements

This research was funded by the Dutch Research Council NWO Gravitation 2013 BOO, Institute for Chemical Immunology (ICI; 024.002.009) to J.S and through the NWO TTW-NACTAR Grant #16442 (to A.J.R.H. and S.H.M.R.). We additionally acknowledge support from NWO, funding the MS facilities through the X-omics Road Map program (project 184.034.019). We would like to thank the Utrecht University (UU) Electron Microscopy Centre and at the Netherlands Centre for Electron Microscopy (NeCEN) for the EM imaging. We would also like to thank Bert Jan Haijema for providing information of the postfusion F protein used in this study.

Data availability

The raw LC-MS/MS files and analyses have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD042266.

Supporting Information:

Contents:

- Supplementary Table S1. Intact mass analysis of MGUS Fab and derivatives.
- Supplementary Figure S1. Longitudinal IgG1 Fab profile of the MGUS patient.
- Supplementary Figure S2. Tandem MS spectrum of N-glycopeptide in CDRH1 of M-protein.

- Supplementary Figure S3. Middle-down LC-MS/MS spectra of M-protein Fd and LC.

Supplementary Table S1. Intact mass analysis of MGUS Fab and derivatives. Glycoforms are denoted with: N-acetyl hexosamine (N), hexose (H), fucose (F), and N-acetyl neuraminic acid (S).

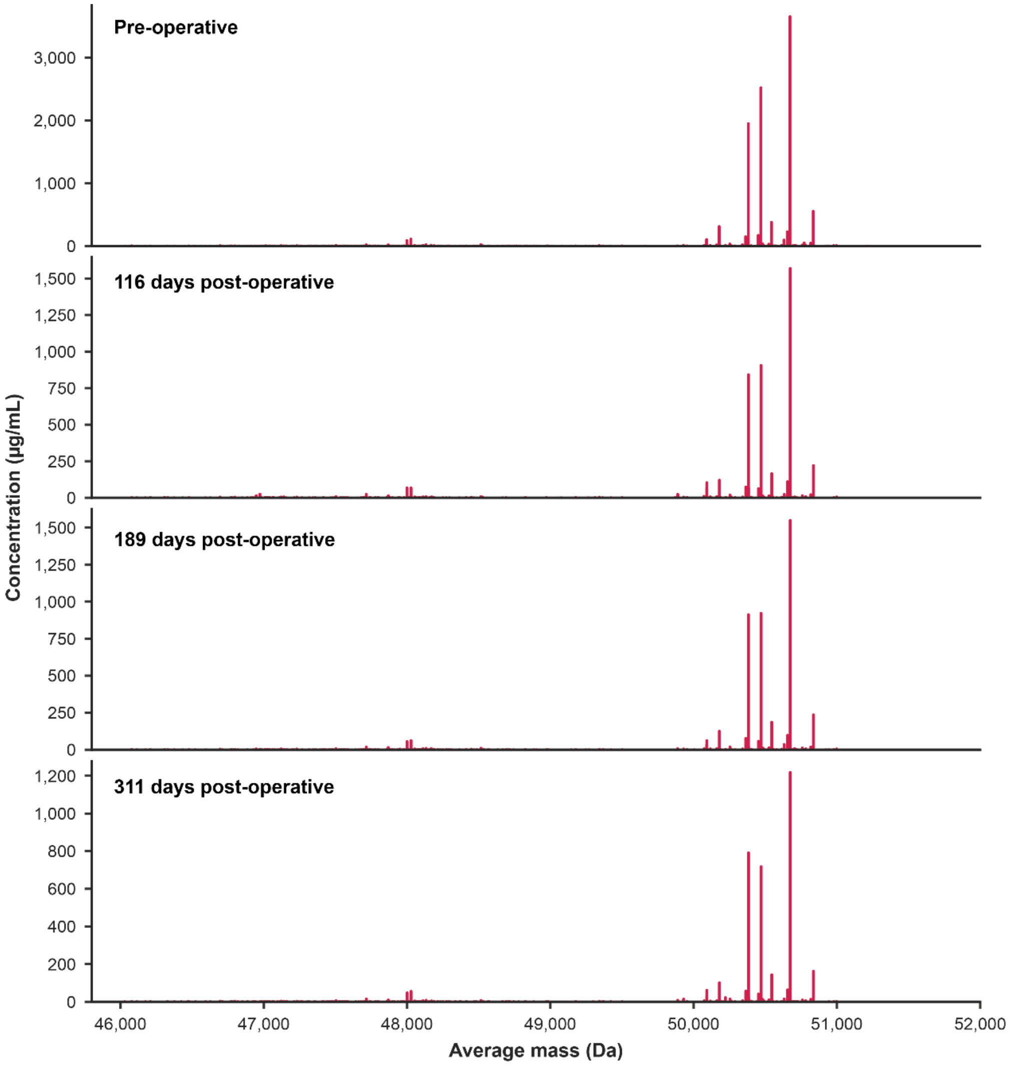
Sample	**glycoform**	**mass exp. (Da)**	**SD (Da)**[c]	**mass theor. (Da)**[d]	**Δmass (Da)**	**relative intensity (%)**
intact Fab[a]	N5H5F1	50093.96	0.40	50090.79	3.17	3.63
	N4H5F1S1	50181.62	0.16	50178.85	2.77	7.52
	N5H5F1S1	50384.96	0.21	50382.05	2.91	22.04
	N4H5F1S2	50472.65	0.02	50470.11	2.54	23.78
	N5H6F1S1	50549.69	0.04	50544.19	5.50	4.93
	N5H5F1S2	50675.96	0.13	50673.30	2.66	32.57
	N5H6F1S2	50837.75	0.17	50835.45	2.30	5.54
Fab + PNGase F[a]	-	48120.43	0.12	48117.67	2.76	-
Fab + TCEP HC[b]	N5H5F1S2	26952.69	0.37	26951.09	1.60	-
Fab + TCEP LC[b]	-	23729.78	0.37	23731.31	-1.53	-

[a] *native MS*

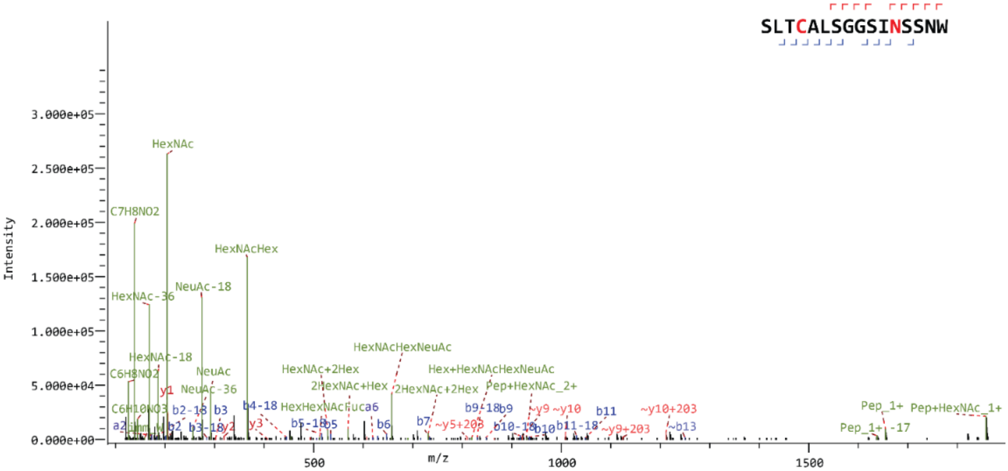
[b] *LC-MS*

[c] *average and standard deviations are calculated across the series of charge states for native MS, across 4 replicate measurements for LC-MS.*

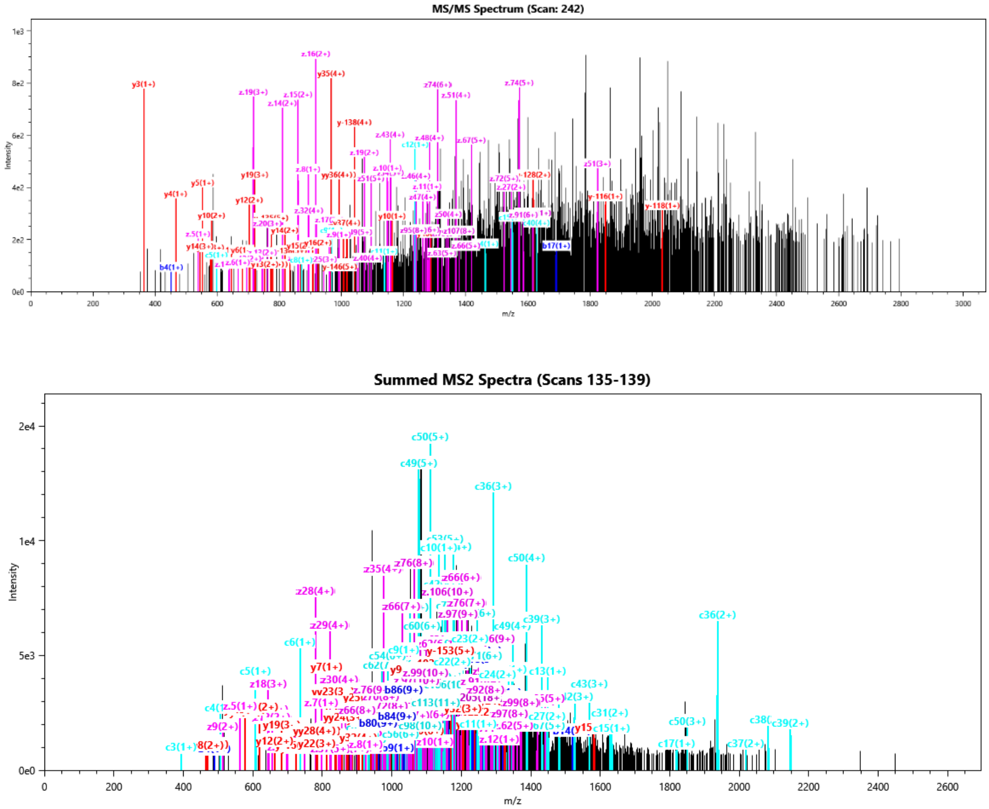
[d] *average theoretical mass considering disulfide bond formation and pyroglytamic acid conversion of the heavy chain N-terminus.*



Supplementary Figure S1. Longitudinal IgG1 Fab profile of the MGUS patient.



Supplementary Figure S2. Tandem MS spectrum of N-glycopeptide in CDRH1 of M-protein. Shown is the HexNAc(5)Hex(5)Fuc(1)NeuAc(2) glycoform.



Supplementary Figure S3. Top-down LC-MS/MS spectra of M-protein Fab Heavy Chain (top) and Light Chain.

References

(1) Glavey, S. V.; Leung, N. Monoclonal Gammopathy: The Good, the Bad and the Ugly. *Blood Rev* **2016**, *30* (3), 223–231. https://doi.org/10.1016/j.blre.2015.12.001.

(2) Dhodapkar, M. V. MGUS to Myeloma: A Mysterious Gammopathy of Underexplored Significance. *Blood* **2016**, *128* (23), 2599–2606. https://doi.org/10.1182/blood-2016-09-692954.

(3) Dasari, S.; Kohlhagen, M. C.; Dispenzieri, A.; Willrich, M. A. V.; Snyder, M. R.; Kourelis, T. V.; Lust, J. A.; Mills, J. R.; Kyle, R. A.; Murray, D. L. Detection of Plasma Cell Disorders by Mass Spectrometry: A Comprehensive Review of 19,523 Cases. *Mayo Clinic Proceedings* **2022**, *97* (2), 294–307. https://doi.org/10.1016/j.mayocp.2021.07.024.

(4) *B-Cell Development and the Antibody Response - ClinicalKey*. https://www.clinicalkey.com/#!/content/book/3-s2.0-B978070207844600009X (accessed 2023-02-21).

(5) Huhn, C.; Selman, M. H. J.; Ruhaak, L. R.; Deelder, A. M.; Wuhrer, M. IgG Glycosylation Analysis. *PROTEOMICS* **2009**, *9* (4), 882–913. https://doi.org/10.1002/pmic.200800715.

(6) Mittermayr, S.; Lê, G. N.; Clarke, C.; Millán Martín, S.; Larkin, A.-M.; O'Gorman, P.; Bones, J. Polyclonal Immunoglobulin G *N* -Glycosylation in the Pathogenesis of Plasma Cell Disorders. *J. Proteome Res.* **2017**, *16* (2), 748–762. https://doi.org/10.1021/acs.jproteome.6b00768.

(7) Kinoshita, N.; Ohno, M.; Nishiura, T.; Fujii, S.; Nishikawa, A.; Kawakami, Y.; Uozumi, N.; Taniguchi, N. Glycosylation at the Fab Portion of Myeloma Immunoglobulin G and Increased Fucosylated Biantennary Sugar Chains: Structural Analysis by High-Performance Liquid Chromatography and Antibody-Lectin Enzyme Immunoassay Using Lens Culinaris Agglutinin1. *Cancer Research* **1991**, *51* (21), 5888–5892.

(8) Bondt, A.; Hoek, M.; Tamara, S.; de Graaf, B.; Peng, W.; Schulte, D.; van Rijswijck, D. M. H.; den Boer, M. A.; Greisch, J.-F.; Varkila, M. R. J.; Snijder, J.; Cremer, O. L.; Bonten, M. J. M.; Heck, A. J. R. Human Plasma IgG1 Repertoires Are Simple, Unique, and Dynamic. *Cell Systems* **2021**, *12* (12), 1131-1143.e5. https://doi.org/10.1016/j.cels.2021.08.008.

(9) Bondt, A.; Hoek, M.; Dingess, K.; Tamara, S.; Graaf, B. de; Peng, W.; Boer, M. A. den; Damen, M.; Zwart, C.; Barendregt, A.; Rijswijck, D. M. H. van; Grobben, M.; Tejjani, K.; Rijswijk, J. van; Völlmy, F.; Snijder, J.; Fortini, F.; Papi, A.; Volta, C. A.; Campo, G.; Contoli, M.; Gils, M. J. van; Spadaro, S.; Rizzo, P.; Heck, A. J. R. No Patient Is the Same; Lessons Learned from Antibody Repertoire Profiling in Hospitalized Severe COVID-19 Patients. medRxiv December 26, 2022, p 2022.12.23.22283896. https://doi.org/10.1101/2022.12.23.22283896.

(10) Spoerry, C.; Hessle, P.; Lewis, M. J.; Paton, L.; Woof, J. M.; Pawel-Rammingen, U. von. Novel IgG-Degrading Enzymes of the IgdE Protease Family Link Substrate Specificity

to Host Tropism of Streptococcus Species. *PLOS ONE* **2016**, *11* (10), e0164809. https://doi.org/10.1371/journal.pone.0164809.

(11) Cassidy, J.; Nordby, G. Human Serum Immunoglobulin Concentrations: Prevalence of Immunoglobulin Deficiencies+. *Journal of Allergy and Clinical Immunology* **1975**, *55* (1), 35–48. https://doi.org/10.1016/S0091-6749(75)80006-6.

(12) Gonzalez-Quintela, A.; Alende, R.; Gude, F.; Campos, J.; Rey, J.; Meijide, L. M.; Fernandez-Merino, C.; Vidal, C. Serum Levels of Immunoglobulins (IgG, IgA, IgM) in a General Adult Population and Their Relationship with Alcohol Consumption, Smoking and Common Metabolic Abnormalities. *Clinical and Experimental Immunology* **2008**, *151* (1), 42–50. https://doi.org/10.1111/j.1365-2249.2007.03545.x.

(13) Tran, N. H.; Zhang, X.; Xin, L.; Shan, B.; Li, M. *De novo* Peptide Sequencing by Deep Learning. *Proceedings of the National Academy of Sciences* **2017**, *114* (31), 8247–8252. https://doi.org/10.1073/pnas.1705691114.

(14) Tran, N. H.; Qiao, R.; Xin, L.; Chen, X.; Liu, C.; Zhang, X.; Shan, B.; Ghodsi, A.; Li, M. Deep Learning Enables *de novo* Peptide Sequencing from Data-Independent-Acquisition Mass Spectrometry. *Nat Methods* **2019**, *16* (1), 63–66. https://doi.org/10.1038/s41592-018-0260-3.

(15) Schulte, D.; Peng, W.; Snijder, J. Template-Based Assembly of Proteomic Short Reads For *De novo* Antibody Sequencing and Repertoire Profiling. *Anal. Chem.* **2022**, *94* (29), 10391–10399. https://doi.org/10.1021/acs.analchem.2c01300.

(16) Brademan, D. R.; Riley, N. M.; Kwiecien, N. W.; Coon, J. J. Interactive Peptide Spectral Annotator: A Versatile Web-Based Tool for Proteomic Applications. *Mol Cell Proteomics* **2019**, *18* (8 Suppl 1), S193–S201. https://doi.org/10.1074/mcp.TIR118.001209.

(17) Bern, M.; Kil, Y. J.; Becker, C. Byonic: Advanced Peptide and Protein Identification Software. *Current Protocols in Bioinformatics* **2012**, *40* (1), 13.20.1-13.20.14. https://doi.org/10.1002/0471250953.bi1320s40.

(18) Marty, M. T.; Baldwin, A. J.; Marklund, E. G.; Hochberg, G. K. A.; Benesch, J. L. P.; Robinson, C. V. Bayesian Deconvolution of Mass and Ion Mobility Spectra: From Binary Interactions to Polydisperse Ensembles. *Anal. Chem.* **2015**, *87* (8), 4370–4376. https://doi.org/10.1021/acs.analchem.5b00140.

(19) Fellers, R. T.; Greer, J. B.; Early, B. P.; Yu, X.; LeDuc, R. D.; Kelleher, N. L.; Thomas, P. M. ProSight Lite: Graphical Software to Analyze Top-down Mass Spectrometry Data. *PROTEOMICS* **2015**, *15* (7), 1235–1238. https://doi.org/10.1002/pmic.201400313.

(20) Park, J.; Piehowski, P. D.; Wilkins, C.; Zhou, M.; Mendoza, J.; Fujimoto, G. M.; Gibbons, B. C.; Shaw, J. B.; Shen, Y.; Shukla, A. K.; Moore, R. J.; Liu, T.; Petyuk, V. A.; Tolic, N.; Pasa-Tolic, L.; Smith, R. D.; Payne, S. H.; Kim, S. Informed-Proteomics: Open Source

Software Package for Top-down Proteomics. *Nat Methods* **2017**, *14* (9), 909–914. https://doi.org/10.1038/nmeth.4388.

(21) Lefranc, M.-P. IMGT® Databases, Web Resources and Tools for Immunoglobulin and T Cell Receptor Sequence Analysis, Http://Imgt.Cines.Fr. *Leukemia* **2003**, *17* (1), 260–266. https://doi.org/10.1038/sj.leu.2402637.

(22) Peng, W.; Pronker, M. F.; Snijder, J. Mass Spectrometry-Based *De novo* Sequencing of Monoclonal Antibodies Using Multiple Proteases and a Dual Fragmentation Scheme. *J. Proteome Res.* **2021**, *20* (7), 3559–3566. https://doi.org/10.1021/acs.jproteome.1c00169.

(23) Peng, W.; Giesbers, K. C. A. P.; Siborova, M.; Beugelink, J. W.; Pronker, M. F.; Schulte, D.; Hilkens, J.; Janssen, B. J. C.; Strijbis, K.; Snijder, J. Reverse Engineering the Anti-MUC1 Hybridoma Antibody 139H2 by Mass Spectrometry-Based *de novo* Sequencing. bioRxiv July 5, 2023, p 2023.07.05.547778. https://doi.org/10.1101/2023.07.05.547778.

(24) Mann, A. C.; Self, C. H.; Turner, G. A. A General Method for the Complete Deglycosylation of a Wide Variety of Serum Glycoproteins Using Peptide-N-Glycosidase-F. *Glycosylation & Disease* **1994**, *1* (4), 253–261. https://doi.org/10.1007/BF00919333.

(25) Bondt, A.; Rombouts, Y.; Selman, M. H. J.; Hensbergen, P. J.; Reiding, K. R.; Hazes, J. M. W.; Dolhain, R. J. E. M.; Wuhrer, M. Immunoglobulin G (IgG) Fab Glycosylation Analysis Using a New Mass Spectrometric High-Throughput Profiling Method Reveals Pregnancy-Associated Changes *. *Molecular & Cellular Proteomics* **2014**, *13* (11), 3029–3039. https://doi.org/10.1074/mcp.M114.039537.

(26) Srzentić, K.; Fornelli, L.; Tsybin, Y. O.; Loo, J. A.; Seckler, H.; Agar, J. N.; Anderson, L. C.; Bai, D. L.; Beck, A.; Brodbelt, J. S.; van der Burgt, Y. E. M.; Chamot-Rooke, J.; Chatterjee, S.; Chen, Y.; Clarke, D. J.; Danis, P. O.; Diedrich, J. K.; D'Ippolito, R. A.; Dupré, M.; Gasilova, N.; Ge, Y.; Goo, Y. A.; Goodlett, D. R.; Greer, S.; Haselmann, K. F.; He, L.; Hendrickson, C. L.; Hinkle, J. D.; Holt, M. V.; Hughes, S.; Hunt, D. F.; Kelleher, N. L.; Kozhinov, A. N.; Lin, Z.; Malosse, C.; Marshall, A. G.; Menin, L.; Millikin, R. J.; Nagornov, K. O.; Nicolardi, S.; Paša-Tolić, L.; Pengelley, S.; Quebbemann, N. R.; Resemann, A.; Sandoval, W.; Sarin, R.; Schmitt, N. D.; Shabanowitz, J.; Shaw, J. B.; Shortreed, M. R.; Smith, L. M.; Sobott, F.; Suckau, D.; Toby, T.; Weisbrod, C. R.; Wildburger, N. C.; Yates, J. R. I.; Yoon, S. H.; Young, N. L.; Zhou, M. Interlaboratory Study for Characterizing Monoclonal Antibodies by Top-Down and Middle-Down Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2020**, *31* (9), 1783–1802. https://doi.org/10.1021/jasms.0c00036.

(27) Barnidge, D. R.; Dasari, S.; Botz, C. M.; Murray, D. H.; Snyder, M. R.; Katzmann, J. A.; Dispenzieri, A.; Murray, D. L. Using Mass Spectrometry to Monitor Monoclonal Immunoglobulins in Patients with a Monoclonal Gammopathy. *J. Proteome Res.* **2014**, *13* (3), 1419–1427. https://doi.org/10.1021/pr400985k.

(28) Deighan, W. I.; Winton, V. J.; Melani, R. D.; Anderson, L. C.; McGee, J. P.; Schachner, L. F.; Barnidge, D.; Murray, D.; Alexander, H. D.; Gibson, D. S.; Deery, M. J.; McNicholl, F. P.; McLaughlin, J.; Kelleher, N. L.; Thomas, P. M. Development of Novel Methods for Non-Canonical Myeloma Protein Analysis with an Innovative Adaptation of Immunofixation Electrophoresis, Native Top-down Mass Spectrometry, and Middle-down *de novo* Sequencing. *Clinical Chemistry and Laboratory Medicine (CCLM)* **2021**, *59* (4), 653–661. https://doi.org/10.1515/cclm-2020-1072.

(29) McDonald, Z.; Taylor, P.; Liyasova, M.; Liu, Q.; Ma, B. Mass Spectrometry Provides a Highly Sensitive Noninvasive Means of Sequencing and Tracking M-Protein in the Blood of Multiple Myeloma Patients. *J. Proteome Res.* **2021**, *20* (8), 4176–4185. https://doi.org/10.1021/acs.jproteome.0c01022.

(30) Dupré, M.; Duchateau, M.; Sternke-Hoffmann, R.; Boquoi, A.; Malosse, C.; Fenk, R.; Haas, R.; Buell, A. K.; Rey, M.; Chamot-Rooke, J. *De novo* Sequencing of Antibody Light Chain Proteoforms from Patients with Multiple Myeloma. *Anal. Chem.* **2021**, *93* (30), 10627–10634. https://doi.org/10.1021/acs.analchem.1c01955.

(31) Noori, S.; Zajec, M.; Russcher, H.; Tintu, A. N.; Broijl, A.; Jacobs, J. F. M.; Luider, T. M.; de Rijke, Y. B.; vanDuijn, M. M. Retrospective Longitudinal Monitoring of Multiple Myeloma Patients by Mass Spectrometry Using Archived Serum Protein Electrophoresis Gels and *De novo* Sequence Analysis. *Hemasphere* **2022**, *6* (8), e758. https://doi.org/10.1097/HS9.0000000000000758.

(32) Noori, S.; Wijnands, C.; Langerhorst, P.; Bonifay, V.; Stingl, C.; Touzeau, C.; Corre, J.; Perrot, A.; Moreau, P.; Caillon, H.; Luider, T. M.; Dejoie, T.; Jacobs, J. F. M.; van Duijn, M. M. Dynamic Monitoring of Myeloma Minimal Residual Disease with Targeted Mass Spectrometry. *Blood Cancer J.* **2023**, *13* (1), 1–3. https://doi.org/10.1038/s41408-023-00803-z.

CHAPTER 7

Summary And Outlook

Chapter 7

Summary

Although the importance of antibodies in immunity was proposed more than one hundred years ago[1], comprehensively studying and characterizing antibodies remains a challenge. This thesis demonstrates how mass spectrometry (MS) can be utilized to address this challenge through *de novo* antibody sequencing on the protein level.

In Chapter 1, I reviewed the current state of mass spectrometry-based *de novo* antibody sequencing. Beginning with an overview of B cell development and antibody diversity, we acknowledged the great diversity of antibodies that safeguard our body against numerous harmful antigens, while also acknowledging the difficulties this poses for their characterization. Recent advancements include partial *de novo* methods like Ig-seq, which merges single B cell sequencing with bottom-up proteomics to semi-quantitatively explore serological antibody repertoires. Then I described the complete *de novo* antibody sequencing based on a bottom-up proteomics approach, based on LC-MS/MS, which doesn't require the DNA database and better represents the antibody repertoire on the protein level. I describe a range of peptide *de novo* sequencing tools, discussing their strengths and limitations. Lastly, I review *de novo* sequencing based on the middle-down approach as a complementary method to bottom-up, enabling comprehensive antibody repertoire profiling and protein-level sequence insights.

In Chapter 2, we introduced a bottom-up approach employing multiple proteases and a dual fragmentation scheme for monoclonal antibody *de novo* sequencing. We first examined the efficacy of this method on a monoclonal antibody Herceptin (sequence is known), showcasing a sequence coverage and accuracy of 100% and 99% in variable region of both heavy chain and light chain. The high accuracy enables us to test this method on another monoclonal antibody anti-FLAG-M2 whose sequence was unknown. Furthermore the sequence was successfully validated by expression of this monoclonal antibody in HEK293 cells and comparing its performance in western blot with the input antibody. Moreover, the sequence accuracy was confirmed by remodeling the published crystal structure of the anti-FLAG-M2 Fab.

However the *de novo* sequencing software "Supernovo" that we used in Chapter 2 was limited to monoclonal antibodies. Thereby, in Chapter 3 we introduced a new tool "Stitch" that uses template-based assembly of the peptide reads for *de novo* antibody sequencing and repertoire profiling. Stitch supports multiple peptide *de novo* sequencing tools, e.g. PEAKS, MaxNovo, Casanovo, and Novor. Stitch provides the chance to semi-quantitatively study the V- and C- gene usage and profile the serological antibody repertoire.

In Chapter 4, we applied the bottom-up approach that was introduced in Chapter 2 on an anti-MUC1 antibody 139H2 and obtained the sequence in *de novo* fashion. 139H2 was developed in the last century for the diagnosis and treatment of MUC1 overexpressing cancer. In addition, it has been widely used in the lab as a research tool for western blot and immunofluorescence microscopy owing to its binding ability to the tandem repeats region (VNTR) on the extracellular domain of MUC1, however, its sequence remained unknown. We used the software PEAKS to *de novo* sequence peptides and Stitch to assemble the peptide reads into a full sequence. The sequence was successfully validated by comparing the performance of the reverse engineered 139H2 and its Fab fragment to the hybridoma-derived product in Western blot and immunofluorescence microscopy. The sequence enables us to further characterize the binding to the VNTR peptide epitope by surface plasmon resonance (SPR) and solve the crystal structure of the 139H2 Fab fragment in complex with MUC1 VNTR peptide.

In Chapter 5, the bottom-up approach we discussed above was employed to an anti-respiratory syncytial virus (RSV) antibody 131-2a, which targets the envelope glycoprotein F, responsible for host cell binding and entry, in its postfusion conformation.post fusion (F) antibody 131-2a. F plays an important role in host cell attachment, receptor binding and mediating cell entry. 131-2a has been developed in the 1980s and widely used as research tool due to its high binding affinity and great specificity to RSV-F in the postfusion conformation. However, prior to our study, its precise amino acid sequence remained undisclosed. After recombinant expression of 131-2a in HEK293 cells, we validate its sequence accuracy by comparing its performance with the input 131-2a in western blot and ELISA. The precise sequence allows us to further characterize its binding epitope to RSV-post fusion protein by single particle cryo-EM, revealing the molecular basis for 131-2a's specificity to the postfusion conformation of F.

Though chapters 2 to 5 underscored the power of the bottom-up approach for monoclonal antibody *de novo* sequencing, it faces challenges when handling more complex samples. Chapter 6 introduced a hybrid bottom-up and middle-down approach to profile the IgG1 antibody repertoire from a monoclonal gammopathy of undetermined significance (MGUS) patient, successfully obtaining the full sequence from a most abundant clone that has Fab glycosylation. Native MS characterized its Fab glycosylation profile, providing insights into aberrant glycosylation patterns, comparing with the normal Fc glycosylation profile.

Outlook

Since the first monoclonal antibody was generated by hybridoma technology in 1975, more than 100 antibody-derived therapeutics have been approved by FDA and deployed in the treatment for various diseases: infectious disease, cancer, autoimmune disorders, and more[2,3]. The great potential for antibodies to cure or treat a wide range of diseases creates a large demand for the continual development of novel antibodies. Nowadays, antibody development predominantly relies on the single-B cell sequencing technology, wherein antigen-specific B cells are derived from patient or other host and the cDNA was sequenced by next-generation sequencing technology. However, a significant proportion of B cells reside in bone marrow and spleen, which are normally not accessible. The B cells in sera only represent a minute fraction of the overall B cell repertoire. To better understand the antibody repertoire as secreted glycoproteins in bodily fluids, in direct relation to serological studies probing antibody binding and neutralization titers, and potentially discover superior therapeutic antibody candidates, it is imperative to comprehensively elucidate the antibody repertoire straight at the protein level. Furthermore, confirmation of the antibody's amino acid sequences is crucial for antibody production quality control. The omission of this critical step often leads to reproducibility crisis in research[4]. The work in this thesis described how bottom-up proteomics based *de novo* sequencing method derived accurate amino acid sequence of monoclonal antibodies and the sequence information is subsequently employed for characterization, such as epitope mapping.

While the bottom-up approach stands as a powerful tool for *de novo* antibody sequencing of monoclonal antibodies, it exhibits limitations when confronted with complex polyclonal antibodies originating from sera. The peptides generated by a conventional bottom-up approach typically consist of around 15-20 amino acids. However, due to the extensive heterogeneity in the complementarity-determining regions (CDR) regions and high level of conservation in the framework, assembling these short peptides into complete antibody sequences for thousands of variants becomes nearly impossible. In chapter 6 of this thesis, we illustrate how application of complementary bottom-up and middle-down methods can be used to derive a full sequence of the most abundant clone of a MGUS patient. Nonetheless, this represents an uncharacteristically simple antibody profile, and comprehending the full antibody repertoire with a deeper coverage of paired heavy-light chain antibody sequences in a polyclonal mixture remains a formidable task.

Extended bottom-up approach (proteases, fragmentation)

A more favorable scenario for bottom-up proteomics sequencing of antibodies would utilize some proteases that yield longer peptides of approximately 50+ amino acids, encompassing at least two CDRs. This strategy would ensure that the overlapping portions of these longer peptides provide sufficient information for accurate sequence assembly. Notably, Laskay *et al* have published that aspartic acid protease 9 (Sap9) is capable of yielding peptides containing complete variable regions or two CDR regions[5,6]. However, the Sap9 has so far only been applied on monoclonal antibody analysis. Cotham *et al* also demonstrated that using LysC with shortened digestion time (2h) can generate full length CDR3 peptides[7], while most peptides still fail to cover at least 2 CDR regions, thereby not fully solving the problem for the polyclonal antibody sequencing challenge. Hence, thorough investigation of proteases and digestion conditions is warranted to establish an extended bottom-up approach. The long peptides also lead to new challenges for LC/MS-MS analysis: the peptides are separated depending on their hydrophobicity in the C18 column that is normally applied in a bottom-up approach. But the selectivity of the long peptides on this material is not very optimal. Columns with C4 or C8 stationary phases have shown superior selectivity for this special purpose compared to the C18 column that is normally used in traditional bottom-up proteomics methods. Addressing the fragmentation of these lengthy peptides is also crucial. Fragmentation schemes, such as activated ion electron transfer dissociation (AI-ETD), electron capture dissociation (ECD), electron activated dissociation (EAD),electron transfer high-energy collision dissociation (EThcD), and Ultraviolet Photodissociation (UVPD), frequently employed for intact protein fragmentation, can be effectively adapted for the extended bottom-up here.

Middle-down approach

In the middle-down approach, the Fc region of the antibody is cleaved by proteases to release the antigen binding fragment, reducing the molecular weight and the complexity of the target analytes. However, so far there is no protease available that can generate Fab for the whole range of human IgG subclasses, let alone across all classes of immunoglobulins. IgdE , for instance, cleaves above the hinge region to generate Fab fragment but is limited to human IgG1[8,9]. IdeS, on the other hand, works across all the human IgG subclasses, but yields Fab2 upon cleavage below the hinge region[10,11]. The Fab2 is twice big as the Fab, presenting challenges in subsequent separation on the LC and fragmentation with MS-MS. Papain, a widely used enzyme for Fab generation from various species, lacks specificity and generates multiple products with different molecular weights from a single monoclonal antibody. However, a unique molecular weight is considered as a unique clone in middle-down MS methods. This heterogeneity in product generation renders papain unsuitable for MS-based methods. While EThcD, ETD, ECD, and UVPD display promise in middle-down approaches, the sequence coverage remains suboptimal[7,12–14]. Achieving higher coverage is imperative to support peptides assembly from bottom-up approach and validating the sequence, ideally aiming for 100% coverageThe inherent similarity in antibody sequences complicates the separation of these products generated from polyclonal samples. It has often been considered that the longer column length can give better separation. Though lots of middle-down analysis are using 50 cm column for separation of Fab, recent evidence suggests that effective column lengths differ greatly between peptides and larger proteins. 1-2 cm column lengths are appropriate for moderately sized molecule (5-10 kDa) and even shorter columns would be adequate for large molecular (100-150 kDa) like Fab2. Consequently, column length optimization is essential for Fab/Fab2 profiling[15].

De novo **sequencing tool**

As reviewed in chapter 1, each *de novo* sequencing tools have its own benefits and drawbacks. Despite the assembly capabilities of Stitch, manual examination of spectra supporting CDR regions remains necessary now, even for monoclonal antibodies[16]. For high-throughput *de novo* antibody sequencing, automation with precise outputs is ideal. Many peptide sequencing tools have constraints on peptide length; for instance, PEAKS outputs peptides shorter than 40 amino acids. But for polyclonal antibody *de novo* sequencing, long peptides that can cover at least two CDR regions, which are normally longer than 50-60 amino acids, are desired. A significant enhancement of *de novo* sequencing tools needs to be made to surpass these peptide length restrictions, catering to extended bottom-up data. As discussed above, EThcD/UVPD are powerful fragmentation scheme for long peptide discovery, however, now PEAKS and Novor are the only two software supporting EThcD spectra, while PEAKS is the unique one that is able to *de novo* solve UVPD spectra[17–20].

There are some algorithms developed in the past few years for the intact protein data analysis, however none of them can do it in a *de novo* way. Presently, no software seamlessly integrates bottom-up and middle-down data for comprehensive *de novo* sequencing. Manually assembling sequences and validating them with middle-down data prove time-intensive and subjective. In an ideal scenario, a novel algorithm would automate the consideration of both bottom-up and middle-down data, yielding complete sequences. Beyond sequence information, this innovative tool could quantitatively profile entire IgG subclasses from both bottom-up and middle-down data, presenting a remarkable advancement in antibody sequencing methodologies.

All things considered, MS stands a robust and indispensable tool to study antibody sequence on the protein level offering valuable insights into the identification of potential novel antibody therapeutic candidates. Ongoing advancements in both bottom-up and middle-down approaches, coupled with the refinement of *de novo* sequencing methodologies, hold the promise of elucidating the complete antibody repertoire from sera samples and enable the derivation of at least the top ten most abundant antibody sequences, thereby facilitating a deeper exploration of related diseases. Furthermore, antibody sequence information assumes a pivotal role in quality control measures, safeguarding against research reproducibility issues. MS is poised to significantly expand its influence in antibody discovery, promising substantial contributions to the recovery of numerous patients coping with diverse diseases.

Reference

(1) Behring, N.; Kitasato, N. Ueber das Zustandekommen der Diphtherie-Immunität und der Tetanus-Immunität bei Thieren. *Dtsch Med Wochenschr* **1890**, *16* (49), 1113–1114. https://doi.org/10.1055/s-0029-1207589.

(2) Köhler, G.; Milstein, C. Continuous Cultures of Fused Cells Secreting Antibody of Predefined Specificity. *Nature* **1975**, *256* (5517), 495–497. https://doi.org/10.1038/256495a0.

(3) Mullard, A. FDA Approves 100th Monoclonal Antibody Product. *Nature Reviews Drug Discovery* **2021**, *20* (7), 491–495. https://doi.org/10.1038/d41573-021-00079-7.

(4) Baker, M. Reproducibility Crisis: Blame It on the Antibodies. *Nature* **2015**, *521* (7552), 274–276. https://doi.org/10.1038/521274a.

(5) Laskay, Ü. A.; Srzentić, K.; Monod, M.; Tsybin, Y. O. Extended Bottom-up Proteomics with Secreted Aspartic Protease Sap9. *Journal of Proteomics* **2014**, *110*, 20–31. https://doi.org/10.1016/j.jprot.2014.07.035.

(6) Srzentić, K.; Fornelli, L.; Laskay, Ü. A.; Monod, M.; Beck, A.; Ayoub, D.; Tsybin, Y. O. Advantages of Extended Bottom-Up Proteomics Using Sap9 for Analysis of Monoclonal Antibodies. *Anal. Chem.* **2014**, *86* (19), 9945–9953. https://doi.org/10.1021/ac502766n.

(7) Cotham, V. C.; Horton, A. P.; Lee, J.; Georgiou, G.; Brodbelt, J. S. Middle-Down 193-Nm Ultraviolet Photodissociation for Unambiguous Antibody Identification and Its Implications for Immunoproteomic Analysis. *Anal. Chem.* **2017**, *89* (12), 6498–6504. https://doi.org/10.1021/acs.analchem.7b00564.

(8) Spoerry, C.; Hessle, P.; Lewis, M. J.; Paton, L.; Woof, J. M.; Pawel-Rammingen, U. von. Novel IgG-Degrading Enzymes of the IgdE Protease Family Link Substrate Specificity to Host Tropism of Streptococcus Species. *PLOS ONE* **2016**, *11* (10), e0164809. https://doi.org/10.1371/journal.pone.0164809.

(9) Deveuve, Q.; Lajoie, L.; Barrault, B.; Thibault, G. The Proteolytic Cleavage of Therapeutic Monoclonal Antibody Hinge Region: More Than a Matter of Subclass. *Frontiers in Immunology* **2020**, *11*.

(10) Johansson, B. P.; Shannon, O.; Björck, L. IdeS: A Bacterial Proteolytic Enzyme with Therapeutic Potential. *PLoS One* **2008**, *3* (2), e1692. https://doi.org/10.1371/journal.pone.0001692.

(11) Rosenstein, S.; Vaisman-Mentesh, A.; Levy, L.; Kigel, A.; Dror, Y.; Wine, Y. Production of F(Ab′)2 from Monoclonal and Polyclonal Antibodies. *Current Protocols in Molecular Biology* **2020**, *131* (1), e119. https://doi.org/10.1002/cpmb.119.

(12) Fornelli, L.; Ayoub, D.; Aizikov, K.; Beck, A.; Tsybin, Y. O. Middle-Down Analysis of Monoclonal Antibodies with Electron Transfer Dissociation Orbitrap Fourier Transform Mass Spectrometry. *Anal. Chem.* **2014**, *86* (6), 3005–3012. https://doi.org/10.1021/ac4036857.

(13) Fornelli, L.; Srzentić, K.; Huguet, R.; Mullen, C.; Sharma, S.; Zabrouskov, V.; Fellers, R. T.; Durbin, K. R.; Compton, P. D.; Kelleher, N. L. Accurate Sequence Analysis of a Monoclonal Antibody by Top-Down and Middle-Down Orbitrap Mass Spectrometry

Applying Multiple Ion Activation Techniques. *Anal. Chem.* **2018**, *90* (14), 8421–8429. https://doi.org/10.1021/acs.analchem.8b00984.

(14) Pandeswari, P. B.; Sabareesh, V. Middle-down Approach: A Choice to Sequence and Characterize Proteins/Proteomes by Mass Spectrometry. *RSC Adv.* **2018**, *9* (1), 313–344. https://doi.org/10.1039/C8RA07200K.

(15) Fekete, S.; Lauber, M. Studying Effective Column Lengths in Liquid Chromatography of Large Biomolecules. *Journal of Chromatography A* **2023**, *1692*, 463848. https://doi.org/10.1016/j.chroma.2023.463848.

(16) Schulte, D.; Peng, W.; Snijder, J. Template-Based Assembly of Proteomic Short Reads For *De novo* Antibody Sequencing and Repertoire Profiling. *Anal. Chem.* **2022**, *94* (29), 10391–10399. https://doi.org/10.1021/acs.analchem.2c01300.

(17) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. PEAKS: Powerful Software for Peptide *de novo* Sequencing by Tandem Mass Spectrometry. *Rapid Communications in Mass Spectrometry* **2003**, *17* (20), 2337–2342. https://doi.org/10.1002/rcm.1196.

(18) Ma, B. Novor: Real-Time Peptide *de novo* Sequencing Software. *J. Am. Soc. Mass Spectrom.* **2015**, *26* (11), 1885–1894. https://doi.org/10.1007/s13361-015-1204-0.

(19) *de novo Peptide Sequencing | LC-MS/MS Software*. Bioinformatics Solutions Inc. https://www.bioinfor.com/de-novo-sequencing/ (accessed 2023-08-16).

(20) *novor.cloud*. https://app.novor.cloud/request (accessed 2023-08-16).

Chapter 7

samenvatting (Nederlands)

Hoewel het belang van antilichamen in de immuniteit al meer dan honderd jaar geleden werd voorgesteld, blijft het uitgebreid bestuderen en karakteriseren van antilichamen een uitdaging. Dit proefschrift laat zien hoe massaspectrometrie (MS) kan worden gebruikt om deze uitdaging aan te pakken door middel van de novo antilichaam sequentie bepaling op eiwitniveau.

In Hoofdstuk 1 besprak ik de huidige stand van zaken op het gebied van op massaspectrometrie gebaseerde de novo antilichaamsequencing. Beginnend met een overzicht van de ontwikkeling van B-cellen en de diversiteit van antilichamen, erkenden we de grote diversiteit aan antilichamen die ons lichaam beschermen tegen talrijke schadelijke antigenen, terwijl we ook de moeilijkheden erkenden die dit met zich meebrengt voor hun karakterisering. Recente ontwikkelingen omvatten gedeeltelijke de novo-methoden zoals Ig-seq, die sequencing van enkele B-cellen combineert met bottom-up proteomics om semi-kwantitatief serologische antilichaamrepertoires te onderzoeken. Vervolgens beschreef ik de volledige de novo-sequentiebepaling van antilichamen op basis van een bottom-up proteomics-benadering, gebaseerd op LC-MS/MS, waarvoor de DNA-database niet nodig is en die het antilichaamrepertoire op eiwitniveau beter vertegenwoordigt. Ik beschrijf een reeks nieuwe peptide-sequencing-instrumenten, waarbij ik hun sterke en zwakke punten bespreek. Ten slotte bespreek ik de novo sequencing op basis van de middle-down benadering als een complementaire methode voor bottom-up, waardoor uitgebreide profilering van antilichaamrepertoire en sequentie-inzichten op eiwitniveau mogelijk worden.

In Hoofdstuk 2 introduceerden we een bottom-up benadering waarbij gebruik werd gemaakt van meerdere proteasen en een duaal fragmentatieschema voor de novo sequencing van monoklonale antilichamen. We onderzochten eerst de werkzaamheid van deze methode op een monoklonaal antilichaam Herceptin (sequentie is bekend), waarbij een sequentiedekking en nauwkeurigheid van 100% en 99% in het variabele gebied van zowel de zware keten als de lichte keten werd aangetoond. Dankzij de hoge nauwkeurigheid kunnen we deze methode testen op een ander monoklonaal antilichaam anti-FLAG-M2 waarvan de sequentie onbekend was. Bovendien werd de sequentie met succes gevalideerd door expressie van dit monoklonale antilichaam in HEK293-cellen en door de prestaties ervan in Western blot te vergelijken met het ingevoerde antilichaam. Bovendien werd de sequentienauwkeurigheid bevestigd door het hermodelleren van de gepubliceerde kristalstructuur van het anti-FLAG-M2 Fab.

De de novo sequencing-software “Supernovo” die we in Hoofdstuk 2 gebruikten, was echter beperkt tot monoklonale antilichamen. Daarom introduceerden we in Hoofdstuk 3 een nieuwe tool “Stitch” die gebruik maakt van op templates gebaseerde assemblage van de peptide-reads voor de novo antilichaamsequencing en repertoireprofilering. Stitch ondersteunt meerdere peptide de novo sequencing-tools, b.v. PEAKS, MaxNovo, Casanovo en Novor. Stitch biedt de kans om het gebruik van V- en C-genen semi-kwantitatief te bestuderen en het serologische antilichaamrepertoire te profileren.

In Hoofdstuk 4 hebben we de bottom-up benadering toegepast die in Hoofdstuk 2 werd geïntroduceerd op een anti-MUC1 antilichaam 139H2 en de sequentie op de novo manier verkregen. 139H2 werd in de vorige eeuw ontwikkeld voor de diagnose en behandeling van kanker die MUC1 tot overexpressie brengt. Bovendien wordt het in het laboratorium op grote schaal gebruikt als onderzoeksinstrument voor western blot- en immunofluorescentiemicroscopie vanwege het bindingsvermogen ervan aan het tandem repeats-gebied (VNTR) op het extracellulaire domein van MUC1, maar de sequentie ervan bleef onbekend. We gebruikten de software PEAKS om peptiden de nieuwe sequentie te geven en Stitch om de peptide-reads samen te stellen tot een volledige sequentie. De sequentie werd met succes gevalideerd door de prestaties van het reverse-engineerde 139H2 en het Fab-fragment ervan te vergelijken met het van hybridoma afgeleide product in Western blot- en immunofluorescentiemicroscopie. De sequentie stelt ons in staat om de binding aan het VNTR-peptide-epitoop verder te karakteriseren door oppervlakteplasmonresonantie (SPR) en de kristalstructuur van het 139H2 Fab-fragment in complex met MUC1 VNTR-peptide op te lossen.

In Hoofdstuk 5 werd de bottom-up benadering die we hierboven bespraken toegepast op een anti-respiratoir syncytieel virus (RSV) antilichaam 131-2a, dat zich richt op het envelopglycoproteïne F, verantwoordelijk voor de binding en toegang tot de gastheercel, in zijn postfusieconformatie.post fusie (F) antilichaam 131-2a. F speelt een belangrijke rol bij de hechting van gastheercellen, receptorbinding en het bemiddelen in celinvoer. 131-2a is in de jaren tachtig ontwikkeld en wordt op grote schaal gebruikt als onderzoeksinstrument vanwege de hoge bindingsaffiniteit en grote specificiteit voor RSV-F in de postfusieconformatie. Voorafgaand aan onze studie bleef de precieze aminozuursequentie echter onbekend. Na recombinante expressie van 131-2a in HEK293-cellen valideren we de sequentienauwkeurigheid door de prestaties ervan te vergelijken met de input 131-2a in Western blot en ELISA. De precieze sequentie stelt ons in staat om zijn bindende epitoop aan RSV-postfusie-eiwit verder te karakteriseren door cryo-EM van een enkel deeltje, waardoor de moleculaire basis wordt onthuld voor de specificiteit van 131-2a voor de postfusieconformatie van F.

Hoewel de hoofdstukken 2 tot en met 5 de kracht van de bottom-up benadering voor de novo sequencing van monoklonale antilichamen onderstreepten, wordt deze geconfronteerd met uitdagingen bij het omgaan met complexere monsters. Hoofdstuk 6 introduceerde een hybride bottom-up en middle-down benadering om het IgG1-antilichaamrepertoire van een monoklonale gammopathie van onbepaalde significantie (MGUS) patiënt te profileren, waarbij met succes de volledige sequentie werd verkregen van een meest voorkomende kloon die Fab-glycosylatie heeft. Native MS karakteriseerde het Fab-glycosylatieprofiel en verschafte inzicht in afwijkende glycosylatiepatronen, vergeleken met het normale Fc-glycosylatieprofiel.

Chapter 7

总结（中文）

尽管在一百多年前科学家们就已经意识到了抗体对于免疫的重要性，然而抗体的全面表征仍是一个挑战。为了解决这一难题，本论文展示了如何利用质谱进行蛋白质水平的抗体从头测序。

本文第一章回顾了基于质谱的抗体从头测序的发展现状。抗体的多样性可以保护我们的身体免受多种抗原的侵害，然而这种极高的复杂度也带来了对抗体表征的极大挑战。现在的研究方法（包括 Ig-seq 在内的部分从头测序）都是将单 B 细胞测序与自下而上的蛋白质组学相结合，从而对血清抗体库进行半定量分析。我在此后介绍了基于自下而上的蛋白质组学方法的完整从头测序，与部分从头测序方法相比，此方法不依靠 DNA 数据库，并且可以更客观的在蛋白质水平上表征抗体。我还介绍了一系列的肽段从头测序软件，客观地描述了它们的优缺点。最后我还介绍了从中而下的蛋白质组学的方法，很好的弥补了从上而下的蛋白质组学的不足之处。将这两种方法结合可以全面的进行抗体库分析以及抗体的表征。

本文第二章介绍了一种自下而上的方法，并采用多种蛋白酶和双重碎裂方案进行单克隆抗体的全新测序。本章首先将此方法对已知序列的单抗 Herceptin 进行了测试，结果显示重链和轻链可变区的覆盖率和准确率高达 100%和 99%。本章接下来又将此方法在未知序列的单抗 anti-flag M2 上进行了检测。根据实验得到的序列，本章将 anti-flag M2 在 HEK293 细胞中表达并将其性能在西方印迹中与原始的 anti-flag M2 单抗进行了比较，成功验证了序列的准确性。此外，通过重新构建 anti-FLAG-M2 Fab 的已发表晶体结构，确认了序列的准确性。

然而，第二章中使用的抗体从头测序软件“Supernovo”仅限于单克隆抗体。因此，第三章引入了一个新工具“Stitch”，它使用基于模板的肽段读取组装进行抗体测序和库分析。Stitch 支持多个肽段从头测序工具，如 PEAKS、MaxNovo、Casanovo 和 Novor。Stitch 提供了半定量研究 V 和 C 基因使用以及分析血清抗体库的机会。

第四章应用了在第二章中介绍的自下而上的方法对抗-MUC1 抗体 139H2 进行从头测序。139H2 是在上个世纪为诊断和治疗 MUC1 过度表达的癌症而开发的。此外，由于其对 MUC1 细胞外区域的串联重复区（VNTR）的结合能力，被广泛应用于西方印迹

和免疫荧光显微镜等方法中。然而，139H2 的序列仍然未被解锁。第四章使用了软件 PEAKS 对肽段进行从头测序，并使用 Stitch 将肽段读取组装成完整的序列。通过将在 HEK293 细胞中表达得到的 139H2 及其 Fab 片段的性能与原始单抗在西方印记和免疫荧光显微镜方法中进行比较，序列得到了成功验证。精确的序列确保了我们能够通过表面等离子共振（SPR）进一步表征对 VNTR 肽段抗原的结合，并解析与 MUC1 VNTR 肽段形成复合物的 139H2 Fab 片段的晶体结构。

第五章采用了上文讨论的自下而上的方法对反呼吸道合胞病毒（RSV）抗体 131-2a 进行了测序，该抗体以其后融合构象中的信使 F 包膜糖蛋白为靶点。F 在宿主细胞附着、受体结合和介导细胞进入中发挥着重要作用。 131-2a 在 1980 年代被开发，并因其对 RSV-F 在后融合构象中的高结合亲和力和极高的特异性而广泛用作研究工具。然而，在我们的研究之前，其精确的氨基酸序列仍然未知。在 HEK293 细胞中重新表达 131-2a 后，我们通过将其在西方印迹和 ELISA 中的性能与原始 131-2a 进行比较来验证其序列准确性。精确的序列使我们能够通过单颗粒冷冻电子显微镜对其与 RSV 后融合蛋白的结合抗原表位进行进一步表征，揭示了 131-2a 对 F 后融合构象的特异性的分子基础。

尽管第 2 章到第 5 章强调了自下而上方法在单克隆抗体全新测序中的强大能力，但在处理更复杂的样本时仍然面临挑战。第 6 章引入了一种混合自下而上和自中间而下的方法，用于分析单克隆临床不确定意义的浆细胞病（MGUS）患者的 IgG1 抗体库，成功地解析了该病人血浆 IgG1 库中丰度最高且有 Fab 糖基化修饰的抗体序列。非变性质谱表征了其 Fab 糖基化的多样性，深入了解了异常糖基化模式，并与正常 Fc 糖基化谱进行比较。

Chapter 7

Zusammenfassung (Deutsch)

Obwohl die Bedeutung von Antikörpern für die Immunität bereits vor mehr als hundert Jahren dargelegt wurde, bleibt die umfassende Erforschung und Charakterisierung von Antikörpern eine Herausforderung. Diese Dissertation zeigt, wie die Massenspektrometrie (MS) genutzt werden kann, um diese Herausforderung durch die *de novo* Sequenzierung von Antikörpern auf Proteinebene zu bewältigen.

Im ersten Kapitel habe ich den aktuellen Stand der *de novo* Antikörpersequenzierung mittels Massenspektrometrie besprochen. Dabei begann ich mit einer Übersicht über die Entwicklung von B-Zellen und die Vielfalt von Antikörpern. Wir erkannten die große Diversität von Antikörpern, die unseren Körper vor zahlreichen schädlichen Antigenen schützen, und gleichzeitig die Schwierigkeiten bei ihrer Charakterisierung. Zu den neuesten Fortschritten gehören teilweise *de novo* Methoden wie Ig-seq, die die Einzelzell-Sequenzierung mit bottom up Proteomik kombinieren, um serologische Antikörperrepertoires semi-quantitativ zu erforschen. Dann beschrieb ich die vollständige *de novo* Antikörpersequenzierung auf der Grundlage eines bottom up Proteomik-Ansatzes, basierend auf LC-MS/MS, der keine DNA-Datenbank erfordert und das Antikörperrepertoire auf Proteinebene besser repräsentiert. Ich stellte eine Reihe von Peptid *de novo* Sequenzierungswerkzeugen vor und diskutierte ihre Stärken und Limitationen. Schließlich überprüfte ich die de-novo-Sequenzierung auf der Grundlage des Middle-down-Ansatzes als ergänzende Methode zu bottom-up, um eine umfassende Profilierung des Antikörperrepertoires und Einblicke in die Sequenz auf Proteinebene zu ermöglichen.

Im zweiten Kapitel führten wir einen bottom-up-Ansatz ein, der mehrere Proteasen und ein duales Fragmentierungsschema für die de-novo-Sequenzierung von monoklonalen Antikörpern verwendet. Zunächst untersuchten wir die Wirksamkeit dieser Methode an dem monoklonalen Antikörper Herceptin (Sequenz bekannt) und zeigten eine Sequenzabdeckung und Genauigkeit von 100% bzw. 99% in der variablen Region sowohl der schweren als auch der leichten Kette. Die hohe Genauigkeit ermöglichte es uns, diese Methode auf den monoklonalen Antikörper anti-FLAG-M2 anzuwenden, dessen Sequenz unbekannt war. Die Sequenz wurde erfolgreich validiert, indem dieser monoklonale Antikörper in HEK293-Zellen exprimiert und seine Leistung im Western Blot mit dem

Ausgangsantikörper verglichen wurde. Die Sequenzgenauigkeit wurde auch durch die Neumodellierung der veröffentlichten Kristallstruktur des anti-FLAG-M2 Fab bestätigt.

Allerdings war die de-novo-Sequenzierungssoftware "Supernovo", die im zweiten Kapitel verwendet wurde, auf monoklonale Antikörper beschränkt. Daher führten wir im dritten Kapitel ein neues Werkzeug "Stitch" ein, das die templatebasierte Montage der Peptiden für die de-novo-Antikörpersequenzierung und Profilierung des Repertoires verwendet. Stitch unterstützt mehrere Bioinformatische Algorthitmen wie PEAKS, MaxNovo, Casanovo und Novor. Stitch bietet die Möglichkeit, die V- und C-Gennutzung semiquantitativ zu untersuchen und das serologische Antikörperrepertoire zu profilieren.

Im vierten Kapitel haben wir den im zweiten Kapitel vorgestellten bottom-up-Ansatz auf einen Anti-MUC1-Antikörper 139H2 angewendet und die Sequenz de-novo erhalten. 139H2 wurde im letzten Jahrhundert für die Diagnose und Behandlung von MUC1-überexprimierenden Krebsarten entwickelt und wird aufgrund seiner Bindungsfähigkeit zum tandemrepetierenden Bereich (VNTR) auf der extrazellulären Domäne von MUC1 auch als Forschungswerkzeug in Laboren weit verbreitet eingesetzt. Die Sequenz wurde erfolgreich validiert, indem die Leistung des reverse konstruierten 139H2 und seines Fab-Fragments mit dem hybriderzeugten Produkt im Western Blot und in der Immunfluoreszenzmikroskopie verglichen wurde. Die Sequenz ermöglichte es uns, die Bindung an das VNTR-Peptid-Epitop durch Oberflächenplasmonenresonanz (SPR) zu charakterisieren und die Kristallstruktur des 139H2 Fab-Fragments im Komplex mit dem MUC1 VNTR-Peptid zu lösen.

Im fünften Kapitel wurde der zuvor diskutierte bottom-up-Ansatz auf einen Anti-respiratorischen Synzytialvirus (RSV)-Antikörper 131-2a angewendet, der das F-Glykoprotein in seiner postfusionären Konformation anvisiert. F spielt eine wichtige Rolle bei der Anheftung an Wirtszellen, der Rezeptorbindung und der Vermittlung des Zelleintritts. 131-2a wurde in den 1980er Jahren entwickelt und aufgrund seiner hohen Bindungsaffinität und großen Spezifität für RSV-F in der postfusionären Konformation weit verbreitet als Forschungswerkzeug eingesetzt. Vor unserer Studie blieb jedoch seine genaue Aminosäuresequenz unbekannt. Nach der rekombinanten Expression von 131-2a in HEK293-Zellen validierten wir seine Sequenzgenauigkeit, indem wir seine Leistung im Western Blot und ELISA mit dem ursprünglichen 131-2a verglichen. Die genaue Sequenz ermöglichte es uns, das Bindungsepitop an das RSV-Postfusion-Protein durch Einzelpartikel-Kryoelektronenmikroskopie zu charakterisieren und die molekulare Grundlage für die Spezifität von 131-2a für die postfusionäre Konformation von F aufzudecken.

Acknowledgment

First of all, I'd like to thank my supervisor **Joost** for providing me with the opportunity to work on this fascinating and thrilling project. Particularly, as I came here with limited knowledge of mass spectrometry, you were always patient in guiding me through the intricacies of both laboratory work and data analysis. Under your mentorship, I have not only gained a profound understanding of mass spec but also learned how to become an independent researcher. I also want to thank **Albert Heck** for the advanced mass spec lab, where I had the privilege of undertaking numerous captivating projects and collaborating with exceptional scientists worldwide.

As a freshman of mass spec, I owe a debt of gratitude to the senior lab members and technicians who generously shared their knowledge. Especially **Riccardo**. We were team member for Orbi12 and then later Fusion. I can't forget all the time we spent together on fixing up fusion. You were always friendly and patient when I came to "Riccado, fusion is down again!" Though you personally barely used fusion. Of course I appreciate everyone from fusion team, particularly **Joshua**! I am also really grateful for all the help from **Arjan**. Every time when there was problem of mass spec, you were always there and helped me! Thank you so much. Beside Mass spec, I couldn't collect so much cool data without great maintenance of the wet lab. I want to thank **Mirjam** for all your help and effort. I really appreciate your extra taking care of my XS gloves during the difficult covid time.

I love our great IgG team, couldn't get the achievement without all your help, especially **Douwe**. You are not only my best team worker but also best Dutch friend here. We worked through all these four years together. You trust my data and I trust your code. We talked so much about our interest in science and future plan. I appreciate your help solving my "Dutch problem" and introducing Dutch culture to me. I enjoyed the time cooking and baking with you and I prepared a secret present for you and Cynthia! I hope we can still work together though I am leaving. I am really looking forward to participating your defense as your paranymph if I could make it back! Also want to thank **Sem** and **Bastiaan**. I have learned lots of MD knowledge from you! Of course **Albert Bondt, Maurits and Danique**, thanks for your help for the fab profiling work and loads of great suggestion. **Rien**, I am not forgetting you. Though we haven't worked together for long time, I like your enthusiasm about science. Thanks for your support in the lab and your great idea about

the de novo algorithm. We will keep contact, work together and maybe enjoy coffee together.

Beside antibody, I have also conducted lots of glycoprojects as “hobby”, thank you **Karli, Joshua** and **Weiwei Wang** for your input and great suggestions! **Kat**! Thanks for your help on cross-linking but of course as a friend! I enjoyed the trip with you in Germany and our great moment in Asian restaurants!

I would like to thank everyone from **Snijder lab**, especially **Matti** and **Marta**. **Matti** you have helped me so much with expression and purification, I am always impressed by your great knowledge of structure biology.

Beside the lab work, I also want to appreciate all the help from **Corine**, thank you for making everything so smoothly for me and your patience for all the questions!

I also want to thank all the Chinese colleagues, especially **Yang, Weiwei Wang**. We had so many great trips in Europe together and I really enjoyed the time with you. I also want to express my thank to **Lai, Yuxian** and **博会姐**, for your help in both work and life! **Yang** we have spent three years as officemate, friend and now you are my paranymph. I am looking forward to being your paranymph in your defense! See you then, my 大宝儿!

I couldn’t forget all the support and remote company I received during my whole PhD, particularly in covid time. A big thanks for all my friends in China, Germany, France and the Netherlands! **远方**, we had so many calls with 7 hours’ time difference. Best friends forever! **王珺姐姐**， thanks for the great time in France, which really shined the darkness in the lockdown period. Ich danke auch meiner deutschen Mama, **Johanna**! Ich weiß noch nicht ob du an meine Verteidigung teilnehmen kannst, aber ich danke dir für alle schöne Zeit in Buggingen und Gott segnen euch. I received many great ideas for designing and printing of this thesis from **Maurits**, **Douwe** and **Inge**! Thank you for your nice suggestion and tips!

Last but not the least, I appreciate the company and help of everyone from **Biomass Group** and the support of all my collaborators. Thank you all!

Now I am going to write in Chinese because this is for my loving mommy.

亲爱的**老妈**，终于我要博士毕业啦！你也终于在我的软磨硬泡之下同意来参加我的毕业答辩！多么不容易，我来欧洲八年多了，第一次你同意来看我！我在欧洲的每一天离不开你的支持，无论是学习工作还是生活，你永远是我的后背我的依靠！接下来新的旅途，你也要陪伴我一起！也谢谢**薛阿姨**，说服老妈一起来参加的我的答辩，谢谢你在我不在时帮助我家的点点滴滴！

最后写给**小多**，我们彼此的情谊无需多言。谢谢你的支持与陪伴，无论前路如何。Austin 见！

Chapter 7

Curriculum vitae

My name is Weiwei Peng, and I was born on the 12th of July in Wuhan, China. I pursued my undergraduate studies in chemical biology at Wuhan University from 2010 to 2014. Following the completion of my bachelor's degree, I participated in a winter program at the University Freiburg, where I cultivated a keen interest in studying in Germany. Subsequently, I dedicated nearly one year to learning the German language. Following my preparatory efforts, I commenced my master program in Biotechnology at TU Dresden in 2016.

In April 2018, I relocated to Magdeburg to undertake my master's internship, titled "Synthesis of lipid-linked oligosaccharide for in vitro N-glycosylation," within the research group led by Thomas Rexer at the Max Planck Institute. During this period, I acquired substantial knowledge in the fields of glycobiology and analytical chemistry.

After completing my master's degree, I came to Utrecht, the Netherlands, where I commenced my Ph.D. study at Utrecht University in September 2019. Under the supervision of Joost Snijder and Albert Heck, my doctoral research focused on *de novo* antibody sequencing based on mass spectrometry. The findings from this research are presented in this thesis.

List of Publications

1. **Peng W**, Giesbers K, Šiborová, Beugelink M.J.W, Pronker MF, Schulte D, Hilkens J, Janssen B.J.C, Strijbis K, Snijder J, Reverse engineering the anti-MUC1 hybridoma antibody 139H2 by mass spectrometry-based de novo sequencing. bioRxiv 2023.07.05.547778; https://doi.org/10.1101/2023.07.05.547778 (Under review)

2. Bondt A, Hoek M, Dingess K, Tamara S, de Graaf B, **Peng W**, den Boer MA, Damen M, Zwart C, Barendregt A, van Rijswijck DMH, Schulte D, Grobben M, Tejjani K, van Rijswijk J, Völlmy F, Snijder J, Fortini F, Papi A, Volta CA, Campo G, Contoli M, van Gils MJ, Spadaro S, Rizzo P, Heck AJR. Into the Dark Serum Proteome: Personalized Features of IgG1 and IgA1 Repertoires in Severe COVID-19 Patients. Mol Cell Proteomics. 2023 Dec 6;23(1):100690. https://doi: 10.1016/j.mcpro.2023.100690.

3. **Peng W**, Rayaprolu V, Parvate AD, Pronker MF, Hui S, Parekh D, Shaffer K, Yu X, Saphire EO, Snijder J. Glycan shield of the ebolavirus envelope glycoprotein GP. Commun Biol. 2022 Aug 4;5(1):785. https://doi.org/10.1038/s42003-022-03767-1

4. Schulte D, **Peng W**, Snijder J. Template-Based Assembly of Proteomic Short Reads For De Novo Antibody Sequencing and Repertoire Profiling. Anal Chem. 2022 Jul 26;94(29):10391-10399. https://doi.org/10.1021/acs.analchem.2c01300

5. Bondt A, Hoek M, Tamara S, de Graaf B, **Peng W**, Schulte D, van Rijswijck DMH, den Boer MA, Greisch JF, Varkila MRJ, Snijder J, Cremer OL, Bonten MJM, Heck AJR. Human plasma IgG1 repertoires are simple, unique, and dynamic. Cell Syst. 2021 Dec 15;12(12):1131-1143.e5. https://doi.org/10.1016/j.cels.2021.08.008

6. **Peng W**, Pronker MF, Snijder J. Mass Spectrometry-Based De Novo Sequencing of Monoclonal Antibodies Using Multiple Proteases and a Dual Fragmentation Scheme. J Proteome Res. 2021 Jul 2;20(7):3559-3566. https://doi.org/10.1021/acs.jproteome.1c00169

7. Olmedillas, E., Mann, C. J., **Peng, W**., Wang, Y.-T., Avalos, R. D., Bedinger, D., Valentine, K., Shafee, N., Schendel, S. L., Yuan, M., Lang, G., Rouet, R., Christ, D., Jiang, W., Wilson, I. A., Germann, T., Shresta, S., Snijder, J., & Saphire, E. O. (2021). Structure-based design of a highly stable, covalently-linked SARS-CoV-2 spike trimer with improved structural properties and immunogenicity. BioRxiv, 2021.05.06.441046. https://doi.org/10.1101/2021.05.06.441046