Modular Antibody *de novo* Sequence Analysis using Multi-tier LC-MS/MS Data

Bastiaan de Graaf1,2, Douwe Schulte1,2, Max Hoek1,2, Weiwei Peng1,2, Sem Tamara1,2, Joost Snijders1,2, Richard A. Scheltema1,2, Albert J.R. Heck1,2$

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

2Netherlands Proteomics Center, Padualaan 8, 3584 CH Utrecht, The Netherlands;

$Corresponding authors: [a.r.j.heck@uu.nl](mailto:a.r.j.heck@uu.nl)

Keywords: *de novo* sequencing, antibodies, polyclonal, mass spectrometry, middle down

(Make) List of abbreviations

* Ab
* CDR
* Ig
* MS
* BU
* MD
* TD
* PTM
* BCR

# Abstract (246 / 250)

Antibodies (Abs) are an important class of biomolecules that are produced by the immune system to defend humans against infections. Their importance is underlined by their use as therapeutic agents requiring large-scale production as recombinant proteins. The key step to accomplish this is the identification of the amino acid sequence and the PTMs that adorn the protein. Traditionally B-cell sequencing has been used to identify the DNA/RNA sequences that putatively lead to the antibodies of interest. More recently mass spectrometry-based methods (MS) have been used, with the added benefits that they directly target protein sequencings and can provide insights into post translational modifications (PTMs). Both techniques retain their own challenges, and the complete extraction of the amino acid sequence is typically unmanageable. Sequencing is problematic especially when gaps arise in the complementarity determining regions (CDRs) of the antibody that are responsible for the recognition and binding of infectious agents. Sequencing by bottom-up (BU), peptide-centric MS methods is often hampered by sequence homology of Ab sequences and gaps resulting from peptides that either do not ionize well or do not yield informative fragmentation spectra. Top-down (TD), protein-centric MS can retain the link between fragments and precursors, however this field is still rather immature, and *de novo* sequencing by TD-MS remains challenging. Here we describe advances towards the complete sequencing of antibody sequences by using integrative BU and TD approaches and show that the information from these disparate data streams can provide more information and improved sequence coverage. The software is freely available on GitHub.

# Introduction

Antibody molecules, or immunoglobulins (Ig’s), provide one of the cornerstones of the human immune system and are abundantly found in various bodily fluids, such as serum, saliva, milk, the lumen of the gut, and cerebrospinal fluid (Schroeder & Cavacini, 2010). Because of their important role in combatting infectious diseases, immunoglobulins have been intensively studied and have in the last decades taken centre stage for the development of novel therapeutics (Kaplon & Reichert, 2021; Marks & Deane, 2020; Raybould et al., 2020). Recently, antibodies have become the best-selling drugs in the pharmaceutical market, and in 2018, eight of the top ten bestselling drugs worldwide were biologics. New leads for the development of recombinant antibodies as biotherapeutics can be found in various sources, such as immunized animals or recovered patients who carry pathogen neutralizing antibodies (Bornholdt et al., 2016; Corti et al., 2016; Valgardsdottir et al., 2021). The incredible potential for diversity of Ig molecules in the human body, over 1015 theoretical sequences (Briney, Inderbitzin, Joyce, & Burton, 2019; Schroeder Jr., 2006), indicates that each antigen exposure may lead to a unique, personalized antibody response. One way to chart the antibody repertoire is to sequence the B-cell receptor (BCR) of all B cells in that can produce antibodies, but it is thought that only a marginal fraction of all these B cells indeed produce Ig proteins that end up in circulation. Therefore, ideally, one would like to investigate and sequence antibodies directly at the protein level instead of through BCR sequencing (Hom, Tomar, & Tipton, 2022).

Mass spectrometry (MS) has become the method of choice for analysing protein mixtures (Aebersold & Mann, 2016; Altelaar, Munoz, & Heck, 2013), but sequencing polyclonal antibody (Ab) mixtures still poses one of the major remaining challenges (Peng, Pronker, & Snijder, 2021; Sen et al., 2017; Srzentić et al., 2020). Most protein analysis by MS is performed by peptide-centric proteomics, also called shotgun- or bottom-up- (BU) proteomics, where the presence and relative abundance of proteins is inferred from peptides resulting from protein digestion. This approach makes use of a protein sequence database to generate theoretical spectra used to identify the peptide precursors of the obtained fragmentation spectra (Aebersold & Mann, 2003). A sequence database is however not available for the full and wide repertoire of antibodies, as their sequences are the result of the recombination and mutation of several genes, which are encoded for by many different alleles in each person. One option to sequence antibodies by shotgun proteomics makes use of *de novo* sequence analysis, where peptide sequences are directly determined from the fragmentation spectra. The resulting short peptide *reads*, typically 5-25 amino acid residues in length, are assembled into longer *contigs* or even full-length *protein chain* sequences (Guthals, Clauser, & Bandeira, 2012; Sen et al., 2017; Tran et al., 2016). *De novo* sequencing any protein is an impressive feat, but the successful application to large, multimeric targets like antibodies shows how far the field has matured in the previous decade. Another factor that makes read assembly for antibodies particularly difficult is that the sequence of both the light- and heavy chain of an antibody are made up of alternatingly conserved and hypervariable sequence domains (Alberts et al., 2002; Charles A Janeway, Travers, Walport, & Shlomchik, 2001). Collectively, the quality of software platforms for *de novo* sequence analysis of antibodies by MS is steadily increasing. Virtually all published platforms make use of homologous sequence templates (Castellana, Pham, Arnott, Lill, & Bafna, 2010; Schulte, Peng, & Snijder, 2022; Sen et al., 2017; Tran et al., 2016), mostly obtained by comparing the experimental data to an immunogenetic database such as the IMGT (M.-P. Lefranc & Lefranc, 2020; M. P. Lefranc, 2003). The commercially available antibody sequencing platform Supernovo for example takes BU data as an input and returns a full-length sequence, along with the determined germline template sequences. These established solutions in the field are typically limited to monoclonal analysis, although recently an effort to sequence polyclonal mixtures using only BU *de novo* peptides was reported, yielding intriguing results (Schulte et al., 2022).

Recent advances in instrumentation, separation, sample preparation and computational power have empowered the analysis of intact proteins by LC-MS. This enables the simultaneous analysis of an entire protein chain, removing the need for protein inference (Toby, Fornelli, & Kelleher, 2016). This approach, called top-down (TD) MS, is very enticing for *de novo* analysis, as it side-steps the need for assembling peptide sequences into a full protein sequence. While the field has not yet matured to yield spectra that can routinely be used for confident *de novo* sequencing without any additional data, the continuous advances indicate that the future of antibody sequence analysis will surely include using these techniques as a complementary source of information to the more established BU analysis. One particularly striking example of this is the middle-down MD approach used in antibody sequence analysis, which improves sequence coverage and reduces complexity of the spectra by cleaving the constant region of the heavy chain with high specificity (Johansson, Shannon, & Björck, 2008). Reports of sequencing components of polyclonal mixtures are coming out in the form of proof-of-concept studies (Bondt, Hoek, et al., 2021; Dupré et al., 2021), where most of the studies make use of some form of intact protein analysis, pointing towards integrative approaches using multitier MS as the way forward.

Here we describe such an integrated approach that builds further upon the recently reported STITCH algorithm by integrating MD MS data, with the aim to improve antibody sequencing. This workflow sequences a target chain, selected from deconvoluted MS1 spectra of reduced Ab chains, in a modular, three stage process based on germline domains. Each stage deals with increasingly large sequence segments, first sequencing the framework regions (FRs), then CDRs with flanking FRs (FR-CDR-FRs), and ultimately full chain sequences. To demonstrate the performance of this integrated approach, we analysed three samples of various scales of complexity, namely: a purified therapeutic antibody Trastuzumab, both in a monoclonal sample and in a mixture of three monoclonal antibodies, as well as a single abundant clone from the serum IgA1 repertoire of a sepsis patient. We confirmed the effectivity of this workflow, by reconstructing the known sequence of the trastuzumab heavy chain to a high degree in the monoclonal sample and the mixture of three monoclonal antibodies. We then used it to sequence the dominant heavy chain in a polyclonal sample of IgA1 clones from a sepsis patient. We show how integration of MD data can be used to resolve ambiguity in *de novo* sequence predictions, particularly in hypervariable regions, by using MD data to determine the mass of the CDR and using this mass to filter candidate CDR sequences and confirm their pairing to the fragmented precursor chain. We hypothesize that such improvements will be particularly beneficial when analysing polyclonal mixtures of increasing complexity or when lower sample amounts are available. The algorithms supporting the analyses were programmed in the C# programming language and are freely available on Github (XXX).

# Results

Antibody sequencing by any one source of information poses a formidable challenge due to the simultaneously hypervariable yet homologous nature of the target sequences. For example, reference databases are of little to no benefit when sequencing the hypervariable CDRs, which in turn makes assigning bottom-up reads for these regions to templates challenging. While MD MS data is inherently grouped by precursor, fragment coverage is too limited for stand-alone *de novo* sequencing. By using both BU and MD MS data as well as reference sequences from immunological databases, each of these sources of information can complement the other and be used to fill gaps left by the other. By considering sequence domains in isolation, ambiguous residues can be evaluated in their biological context. XNAMEX makes use of MD fragmentation spectra, and the relatively conserved nature of the FRs to determine the size of the CDRs, which is subsequently used as a filter to drastically reduce the number of candidate CDR sequences, while simultaneously confirming their pairing to the fragmented precursor target chain.

Our workflow consists of three stages: first considering only FRs, then FR-CDR-FRs, then full length sequences. Each stage first generates a candidate pool by considering ambiguities left by the previous stage, then evaluates these candidates using the integrated evidence streams, then resolves the ambiguities by discarding candidates that do not have supporting evidence (Figure 3a). By starting with the FRs, which are relatively conserved sequence segments, and resolving ambiguities at this scale before moving to longer, more variable segments by joining adjacent FR candidates into FR-CDR-FR contigs, the size of the search space at each stage is limited to manageable sizes.

To showcase the performance of our workflow, we analysed 3 distinct samples of reduced Ab chains displaying varying levels of complexity: Purified Trastuzumab (referred to as the *monoclonal* sample), an equimolar mixture of 3 monoclonal antibodies including Trastuzumab (referred to as the *mix* sample), and a sample of IgA1 clones from serum of a sepsis patient (referred to as the *polyclonal* sample).

## Target mass determination and sample characterization using MD MS

To characterize the complexity of our samples and determine the precursor masses of our target chains, we collected MD LC-MS/MS data on our samples and deconvoluted the MS1 spectra (Figure 1a-c). For the monoclonal sample, as expected, 2 highly abundant peaks were extracted (originating from the separated heavy and light chains), accounting for over half of the total deconvoluted intensity. When adjacent peaks in both mass and retention time (±50 Da and 1 minute; from here on referred to as misidentifications) are considered, this increases to over 90% with the remaining masses consisting of *background* peaks of less than 5% relative abundance (Figure 1a). For the mix sample 6 abundant peaks were extracted. The abundance of the target chains and misidentifications made up ~33%, as expected based on the equimolar mixing ratio. The other clones make up a total of 50% of intensity and ~20% of intensity is background (Figure 1b). Lastly, for the polyclonal sample, the target clone and misidentifications make up less than 20% of intensity, with more than half of the intensity representing other clones over 5% relative abundance and ~25% background (Figure 1c). Highlighting the challenges posed for deconvoluting the acquired top-down spectra, we observe that the used deconvolution software cannot deconvolute these spectra into accurate masses (Table S1), as peaks are incorrectly grouped and averaged. To obtain the best exact mass, we averaged the MS1 spectra recorded over the elution window of each target chain and deconvoluted the averaged spectra. This improved the mass assignments to within 30 ppm accuracy for the trastuzumab heavy chain in the monoclonal and mix sample and yielded a target precursor mass of 24811.17 Da for the polyclonal sample. The MS2 fragmentation spectra for the elution windows where our precursor was selected for fragmentation were averaged and deconvoluted to yield a fragmentation spectrum for each target chain, with 919, 265 and 469 deconvoluted peaks for the monoclonal, mix and polyclonal sample respectively.

## **Using multi-enzyme shotgun proteomics data for de novo sequencing**

Each sample was also digested using 4 proteases in parallel, and LC-MS/MS data was collected for these digestions. The resulting BU spectra were submitted for *de novo* peptide identification through PEAKS (Ma et al., 2003), yielding a total (i.e., cumulatively from all protease treatments) of 14000, 27421 and 35003 *de novo* peptide reads for the monoclonal, mix and polyclonal sample respectively (supplementary table S2).

To confirm the effectivity of our workflow we first reconstructed the known sequence of the Trastuzumab heavy chain using the monoclonal and mix sample, before moving on to the polyclonal sample. To this end, the peptide reads for these samples were each submitted to the de novo peptide assembly tool STITCH, using settings derived from those reported by Schulte *et al*. in their seminal paper (Schulte et al., 2022). The resulting output for the monoclonalsample, was nearly perfect (Figure 2a). However, such superb performance could not be retained in the more complex mix sample, where the resulting consensus sequence contained 4 erroneous residue predictions in the FR2 (Figure 2c). We observed that these errors were the result of low peptide coverage, caused by assigning reads to the wrong templates, thereby splitting reads that belonged to the same chain. Furthermore, the unassisted germline recombination by STITCH failed to select our V-region for recombination, as it was not the highest scoring V-region in the mix sample.

To deal with these issues, we ran STITCH again with refined templates (*i.e.* the consensus sequence as output by our initial STITCH run, or *template selection* run, rather than the germline sequence) and a lower score cut-off for the input reads (50 instead of 85). Additionally, to ensure recombination of our target V-region, we manually defined which V-region templates should be recombined by STITCH, by providing a number of refined templates equal to the number of abundant clones present in the MD data (Figure 1a and 1b). For the monoclonal sample, the best scoring V-region, IGHV3-66, was selected. For the mix sample, 3 V-region templates were selected: the highest unique score (IGHV4-39), the highest score (IGHV4-30-4), and the highest score in a different family (IGHV3-66) (Supplemental materials). This additional STITCH run, or *definitive* run, proved a major improvement for the mix sample as it improved the depth of coverage 2-28 and raised depth of coverage above our dynamic cut-off (the depth of coverage at Cys104, Figure S3) for 13 out of 21 previously lacking positions (Figure 2d) without being detrimental to the performance of the monoclonal sample (Figure 2b). However, ambiguity remained in the consensus sequence for the mix sample, highlighting the challenges in sequencing antibody mixtures using solely bottom-up data.

.

## **Integrating multiple evidence streams – Recombinant benchmarks**

### **Framework region sequencing**

Using the residue frequency tables (figure S3) from both STITCH runs, as well as a residue frequency table generated from the IMGT database, FR candidate sequences were generated by to expand ambiguous residues into sequence candidates (figure S3). This yielded between 1 and 756 candidates per target FR (Supplementary Table S4) and included the correct candidate for all *benchmark samples* (i.e., the monoclonal and mix sample). These candidates were evaluated against BU and MD evidence and ranked by the summed rank of the resulting scores. For BU scoring, the Multiscore was used, which represents the depth coverage of exact sequence matches longer than 6 residues, weighted by match length. For MD scoring, the Spectrumscore was used, which represents the overlap between theoretical fragments of the sequence and peaks in the MD fragmentation spectrum (Supplementary Table S5). The Spectrumscore is obtained using the *sliding window* scoring algorithm, which also outputs the optimal prefix- and suffix- mass of a given contig (Figure S2). Candidates missing highly conserved residues (Cys23, Cys104), and terminal segment (i.e., FR1 and FR4) candidates with highly unlikely prefix- or suffix- masses were removed in a first pass filtering step. This reduced the pools to a maximum of 90 candidates.

We further filtered the candidate pools to a maximum of 40 candidates (Supplementary Table S4) without eliminating any correct candidates through manual inspection of the scores. For the monoclonal sample, the correct FR1 candidate was ranked #1 (50.1#1, 51.2k#1) with a large discrepancy between scores (supplementary figure X). As FR2-4 only had one candidate each, no selection was needed. However, it was encouraging to see that the sliding window algorithm was able to correctly determine the prefix- masses for these contigs, with a maximum error of 18 ppm.

The candidate pools for the mix sample could be filtered down from 240, 756, 5 and 4 candidates to 40, 7, 1 and 2 candidates for FR1-4 respectively (table S4). For FR1, we reject 200 candidates in the first pass, leaving 40 candidates. No further filtering was possible, as the fragment and read coverages are too low for confident filtering (max 2 fragments, and no read coverage past Cys23). The FR2 candidates have an extremely high degree of overlapping scores due to low read coverage of the n-terminal ambiguous residues (figure 2b) and a near total overlap of theoretical fragments for these candidates. We rejected the lower Spectrumscores (105.79 vs 121.247) as they represent the same fragment matches but without a matched c2 fragment. This reduced the candidate pool from 756 to 84 candidates. Subsequent filtering using the Multiscore (rejecting 8.9k vs 9.4k) left only 7 candidates, representing a single remaining ambiguous N-terminal residue. For FR3 only 1 out of 5 candidates had the highly conserved Cys104, so we rejected all other candidates. For FR4 we reject all candidates not starting with the conserved Trp118 but consider difference in Multiscore for the remaining 2 candidates too small too reject either, doubly so because BU evidence can be misleading due to the presence of other clones in the sample.

### **Complementarity determining region sequencing**

Using the selected FR candidates, MD fragmentation data, and bottom-up reads, a CDR pool was generated for all target CDRs by recombining adjacent FR candidates (e.g., 2 FR1 and 3 FR2 candidates yield 6 permutations of FR1-FR2 pairs, or *permutations)*. These permutations used to find CDR candidates (mass-matching, bridging sequences between the FRs), generated from overhanging BU reads using an algorithm outlined in Figure S6. Briefly, reads containing the 3 CDR-flanking residues are taken for a given pair of FRs. The N-terminal (e.g., extending from FR1 towards FR2) overhanging reads are then combined with the C-terminal (e.g., extending from FR2 towards FR1) overhanging reads to yield CDR-candidates. The mass of the CDR (i.e. the mass gap between the adjacent FRs) is then calculated using the FR candidates and the sliding window algorithm(Figure S2) and used to filter the CDR candidates, retaining only those matching the mass within a 5 Da tolerance. ~~These CDR candidates are scored using the Spectrum- and Peaks- score (table S5) and ranked by the summed rank of these scores.~~ However, as CDRs are difficult to score in isolation, the top 10 CDR candidates for each FR-candidate pair are then rescored as FR-CDR-FR contigs. The complete list of FR-CDR-FR candidates (i.e., up to 10 candidates per FR-pair) are then scored and ranked in the same way as the FR candidates.

For all benchmark samples, the generated FR-CDR-FR candidates contained the right candidate for CDR1-3. For the monoclonal sample, 10 FR-CDR-FR candidates were generated for CDR1-3 (30 candidates in total, table S4), and for each CDR the correct candidate was ranked #1 (CDR1 (255#1, 137k#1), CDR2 (508#1, 56k#2) and CDR3 (1561#2, 122.5k#1)), with enough score discrepancy to reject the remaining candidates (supplementary figure X).

For the mix sample 1106, 49 and 20 FR-CDR-FR candidates were generated for the CDR1-3 respectively (table S4). Despite the much larger starting pools, the correct candidate was ranked #1 for CDR1 (142.8#2, 29.7k#1) and CDR2 (257.2#1, 39.8k#1), with enough score discrepancy to reject the remaining candidates (supplementary figure, figure 3c). CDR1 selection also filtered out the remaining incorrect FR1 and FR2 candidates, which left only 7 FR-CDR-FR candidates for CDR2 as the rest did not contain the right FR2. The correct FR-CDR-FR candidate for CDR3 could was ranked #14 and the pool contained 3 near isobaric candidates, similar to those found in the monoclonal sample (SR***WNDG***FYAMDY, SR***WGGDG***FYAMDY, SR***DNWG***FYAMDY). However, due to far lower fragment coverage (scores 280-282) there was a far lower spread of the Spectrumscores compared to the monoclonal CDR3 FR-CDR-FR candidates (1053-1601). The Multiscores were distributed in two clusters based on which FR4 candidate was included: ~221k (WGQGT..) and ~244k (WGQGS..). Upon manual inspection, we observed that the incorrect FR4 candidate, as well as incorrect CDR candidates, had a large number of short reads traversing the CDR as supporting evidence, while the correct candidate had fewer but longer reads (average ~25 vs average ~12). However, we could not discriminate between the 3 isobaric candidates based on the experimental data, leaving 3 candidates for the CDR3.

~~The former is the correct candidate, but the latter is the result of peptides from another component of the mixture with the YYCAREGHTAAPFDYWGQGSL motif. When excluded, we still do not get the right candidate as a top scorer, as part of the inserted CDR3 matches the CDR3 from this other component. Only upon close inspection, we could observe a complete break in coverage between the AAPFDY/WGQGS motifs, thereby identifying the correct answer. This was particularly challenging because the different components of the mixture were present in a similar concentration.~~

### **Full chain sequencing**

We then expanded the scope of sequencing to the entire target chain to verify the compatibility of the selected FR-CDR-FR candidates and evaluate the resulting chains in their entirety and confirm that they indeed represented the best prediction for the given evidence. To achieve this, we recombined all remaining FR candidates for FR1-4 and transformed the permutations into full length chain candidates by joining the FR candidates with CDR candidates in the same way as before (Figure S6). The resulting chain candidates that deviated more than 5 Da from the precursor mass in the MD data were discarded and the remaining pool was scored and ranked same as before. This yielded 930 chain candidates for our monoclonal sample and 616 for our mix sample.

Chain candidates resulting from combination of 3 selected FR-CDR-FR candidates from the previous stage were considered *primary chain candidates*. These primary chain candidates were all found for both benchmark samples, meaning all selected FR-CDR-FR candidates were recombined into 1 and 3 primarychain candidates respectively. This confirmed the compatibility of the selected FR-CDR-FR candidates, in that they could be combined into a single sequence prediction matching the precursor mass.

To ensure that these primary chain candidates indeed represented the best prediction, all chain candidates were scored and ranked in the same way as the FR candidates (figure 3c, supplementary figure X). The primary chain candidate for the monoclonal sample was ranked #1, despite not having the highest Multiscore (267k vs 270k) or Spectrumscore (1815 vs 1818). The primary chain candidates for the mix sample were ranked #3-5, with the correct sequence at #5. The #1 and #2 candidates have the same CDR3s as we saw in the CDR3 sequencing stage and can be rejected on the same bases (shorter CDR3 reads), and the isobaric CDR3s still cannot be confidently ranked as the scores are too close together, with Multiscores of 255.7k-255.8k and Spectrumscores of 426.1-427.2. The large difference in fragment coverage between the monoclonal and mix sample (265 peaks vs 919 peaks) may have added to the difficulty of resolving the final ambiguities (table S1), and is highlighted by the large difference in spectrumscores between the correct candidates (426.1 vs 1815.3).

## **Performance on polyclonal samples**

We then moved on to sequencing the most abundant heavy chain in a polyclonal sample, following the same steps outlined above. The template V-region with the highest cumulative area under the curve from the STITCH *template selection* run, IGHV3-33, was selected as our sequencing target (Figure 4a). Deconvoluted fragmentation spectra for several elution windows with were scored against the IgA1 C-region sequence to identify the fragmented heavy chains. The most abundant heavy chain spectra were scored against the consensus FR sequences for the selected template using the sliding window algorithm, identifying the selected elution window as the elution window for our target clone. For the *definitive* run, refined versions for all V-regions were provided as templates to provide STITCH with accurate decoy sequences, as this was a polyclonal sample containing many clones. The resulting frequency tables were expanded into FR candidates same as before (figure S3).

This yielded 8, 2, 384 and 64 candidates for the FR1-4 respectively, which after scoring and filtering was reduced to 2, 1, 3 and 4 candidates (Figure 4b, Supplementary Table S4). For FR1, the #1 ranked candidate had both the top Spectrumscore (10,6) and the top Multiscore (35k), however, due to the polyclonal nature of the sample we included an additional isobaric candidate with an N-terminal pyro-Q despite a lower Multiscore (34k). For FR2, the #1 Spectrumscore was much higher than the #2 (155 vs 105), so we the latter candidate despite a substantially lower Multiscore (12.5k vs 21.9k), as we believe this Multiscore was caused by reads from another clone in the sample. For FR3 we rejected all but the top scoring candidates with respect to Spectrumscore (163.1), as the fragment coverage was high. However, low fragment coverage at the C-terminal residues of these candidates led us to include the top 3 Multiscore candidates (27.7k-30.4k). FR4 showed a similar situation, with a Spectrumscore of 123.6, leading us to include the top 4 candidates in terms of Multiscore here (308.4k-309.9k).

Using these FR-candidates, 20, 30 and 120 FR-CDR-FR candidates were generated for CDR1-3 respectively. Spectrum- and Multi- scores were unambiguous for both CDR1 and CDR2, identifying the CDR1 as being GLTFSTYD (117.7#1, 57.3k#1), and CDR2 as LWNDGYNK (377.1#1 and 50.9k#1). The selected FR-CDR-FR candidates for CDR2 for also enabled us to reject 2 out of 3 remaining FR3 candidates. This left 40 FR-CDR-FR candidates for CDR3. From these we selected the top ranked selected 2 isobaric FR-CDR-FR candidates (..LGQRRPL.. and ..GLQRRLP..) with the top Multiscores (346.2k#1-346.4k#2) and high Spectrumscores (370.7#2) (Figure 4c).

Recombining the remaining FR candidates into chain candidates yielded 975 chain candidates and confirmed the 2 primary chain candidates matched the precursor mass and were ranked #1 (422.9#4, 411.3k#1) and #2 (422.9#4, 411.1k#2) (supplementary Figure X). These primary candidates had the same ambiguities: 2 isobaric CDR3 candidate sequences (Figure 4d). Close inspection of these 2 candidates’ Multiscore coverage revealed a break in CDR3 coverage for one of the candidates (..GL/QRRLP.. ). Strong support for the LP motif in the CDR3 led us to revisit the de novo reads manually, where we found several reads located entirely in the CDR3, suggesting the CDR3 sequence to be LGQRRLP, a sequence not present in any bridging or overhanging CDR3 reads (Figure 4e). Rescoring this sequence against the experimental evidence indeed reveals an increased Multiscore, from 411.3k to 411.7k.

# Discussion

With this work we show that integration of bottom-up and top/middle-down data is essential to achieve a higher fidelity for extraction of the sequences of antibodies. To provide a solid basis with the *de* novo peptide data, we utilize STITCH (Schulte et al., 2022) although this step still results in errors and leads to many hypotheses. To correct the errors and filter the options, top-down fragmentation data is used. Although the data for even the most abundant clone in a mixture is far from complete, we show that it can be used as a potent filter and even to assist with filling gaps in the sequence. We have used the presented workflow to simultaneously sequence light and heavy chains, but for the sake of brevity have omitted these efforts in this manuscript. As we analyse one chain at a time, there is little difference between the analysis of light and heavy chains aside from differences arising from the quality of the data or the complexity of the target. Light chains are less complex owing to a lower degree of somatic hypermutation and the lack of a D-segment. Unsurprisingly therefore, these targets performed equally well or better than their heavy chain counterparts.

The samples used in this study still represent an ideal case for sequencing. Monoclonal antibodies, and mixtures thereof are of course less complex than those extracted from biological samples. While moving to more complex samples will surely present new challenges, it has been shown that circulating antibody repertoires are, more often than previously thought, dominated by a limited number of clones(Bondt, Dingess, et al., 2021; Bondt, Hoek, et al., 2021). We are therefore optimistic that the presented approach will be applicable to a significant fraction of polyclonal samples and can be extended to the cases where it falls short. The manual interpretation at various points in this workflow significantly limits the throughput. While the main goal of the presented work has been to define a broadly applicable protocol for polyclonal antibody sequencing, we have not yet been able to analyse a large enough number of targets to define robust score cut-offs for several decision points. The integration of multiple data sources, as well as the diversity of the analysed samples (polyclonal, complex, *etc*), targets (light or heavy chain, dominant clones, isotypes), regions (FR1-4, CDR1-3) and segments (FRs, FR-CDR-FR, chain), makes this an even bigger challenge. Current sequencing efforts require high levels of expertise, effort, and modern instrumentation. As such, the requirement that users manually interpret analysis results throughout the analysis is not a deal-breaker especially when considering that one or a few clones dominate the sample. As the field matures, a point will be reached where scoring functions and corresponding cut-offs can be defined. This will automate an ever-increasing portion of this work, eventually leading to a high throughput, fully automated method.

Acknowledgements -R.A.S. and A.J.R.H acknowledge funding through the European Union Horizon 2020 program INFRAIA project Epic-XS (Project 823839). Additionally, B.d.G. and A.J.R.H. acknowledge the Dutch Research Council (NWO) supporting this work via project number 15575.

**Contributions –** AJRH conceived the study, BDG performed the data analysis, wrote the required code and prepared the manuscript, RAS prepared the manuscript, DS prepared the manuscript, ST recorded and analyzed all middle down proteomics data, MH and WP recorded and analysed all shot-gun proteomics data.

Data availability –

# Materials and Methods

## Plasma IgG purification and Fab generation

The IgA/IgG purification and generation of Fabs was performed basically as previously published (Bondt, Dingess, Hoek, van Rijswijck, & Heck, 2021; Bondt, Hoek, et al., 2021). Mobicol spin filters were assembled according to the manufacturer’s instructions and placed in 2 mL Eppendorf tubes. Then 20 µL FcXL affinity matrix slurry was added to the spin filter, followed by three washing steps with 150 µL PBS, in which the liquid was removed by centrifugation for 1 min at 1000 × *g*. Two additional washing steps with 150 µL were performed. The samples were then incubated, under shaking conditions for one hour. Next, the flow through was collected and the affinity matrix with bound Igs was washed four times with 150 µL PBS. Finally, IgG samples were digested overnight using 50 µL PBS containing 100 U of the IgdE protease (FabALACTICA®, Genovis, Llund, Sweden) on a thermal shaker at 37 °C. IgA samples were digested overnight with the O-glycopeptidase from Akkermansia muciniphila, OgpA (OpeRATOR®, Genovis, Llund, Sweden). Digestion was performed using 40 U SialEXO (a sialidase cocktail to remove sialic acids from the O-glycans) and 40 U of OgpA enzyme and incubation overnight, both digestions were performed overnight in an Eppendorf thermal shaker (Eppendorf, The Netherlands) at 37 °C. Next, 10 µL Ni-NTA beads were added to bind and remove the His-tagged protease, where after the samples were incubated for an additional 30 minutes. The flow through of this incubation contained the Fab fragments.

## Protein-centric (middle-down) LC-MS/MS

All Fab samples were denatured and reduced in 10 mM tris(2-carboxyethyl)phosphine (TCEP) at 60 °C for 30 min prior to LC-MS/MS analysis. For each LC-MS/MS experiment 2-5 µg of sample was injected. Reversed-phase liquid chromatography was performed by using a Thermo Scientific Vanquish Flex UHPLC instrument, equipped with a 1 mm x 150 mm MAbPac analytical column and directly coupled to an Orbitrap Fusion Lumos Tribrid (Thermo Fisher Scientific, Bremen, Germany). The column preheater, as well as the analytical column chamber, were heated to 80 °C during chromatographic separation.

The monoclonal and mix samples were separated over 27 min at a flow rate of 250 µL/min. The polyclonal sample in 22 min at a flow rate of 150 µL/min. Gradient elution was achieved by using two mobile phases A (0.1% HCOOH in Milli-Q HOH) and B (0.1% HCOOH in CH3CN) and ramping up B from 10 to 25% over six and one minute respectively, 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. The spray voltage was set at 3.3 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. The electrospray voltage was only turned on after 5 and 2 min respectively to prevent the salts in the sample from entering the MS.

The separated Fab chains were analysed with a resolution setting of 120,000 (@ 200 *m/z*) in MS1, which allows for more accurate mass detection of smaller proteins (< 30 kDa). MS1 scans were acquired in a range of 500-3000 Th and 600-2000 Th respectively, with the default and 250% AGC target and a maximum 246 and 500 ms injection time respectively. In MS1, 2 and 5 µscans were recorded per scan, respectively. Data-dependent mode was defined as two scans. In both cases, MS/MS scans were acquired with a resolution of 120k, a maximum injection time of default and 500 ms, a default and 10,000% AGC target, and default µscans averaged and recorded per scan respectively. The ions of interest were mass-selected by quadrupole in a 10, 2 and 4 Th isolation window, for the experiments on the monoclonal, antibody mixture and polyclonal mixture, respectively, and accumulated to the AGC target prior to fragmentation. Electron-transfer dissociation (ETD) was performed using the following settings: 16 ms reaction time, a maximum injection time of 200 ms, and an AGC target of 1e6 for the ETD reagent. For data-dependent MS/MS acquisition, the intensity threshold was set to 5e4. MS/MS scans were recorded in the range of *m/*z = 350-5000 Th using high mass range quadrupole isolation.

## Peptide-centric (bottom-up) LC-MS/MS

All purified Fab antibody fragments were dried under vacuum and resuspended in a 50 mM aqueous ammonium bicarbonate buffer. Samples were equally split for subsequent digestion, with each of the four proteases used.  For digestion with trypsin, chymotrypsin, elastase and thermolysin, a sodium deoxycholate (SDC) buffer was added to a total volume of 80 µL, 200 mM Tris pH 8.5, 10 mM TCEP, 2% (w/v) SDC final concentration. For digestion with pepsin, a urea buffer was added to a total volume of 80 µL, 2M urea, 10 mM TCEP. Samples were 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. Then for trypsin, chymotrypsin and thermolysin, 50 mM ammonium bicarbonate buffer was added to a total volume of 100 µL. For pepsin 1 M HCl was added to a final concentration of 0.04 M. A total of 0.1 µg of each protease was added and incubated for 4 hours at 37 °C. After digestion 2 µL formic acid was added to precipitate the SDC. SDC was removed by centrifugation for 20 min at max speed (20817 × *g*) after which the supernatant was moved to a new tube. The final samples were desalted by Oasis HLB (Oasis).  Sorbent was wetted using 2x 200 µL ACN, followed by equilibration with 2x 200 µL water/10% formic acid.  The sample was loaded and washed with 2x 200 µL Mili Q water/10% formic acid.  Finally, the sample was eluted using 2x 50 µL water/50% ACN/10% formic acid and dried down by vacuum centrifuge. Prior to MS analysis samples were reconstituted in 2% FA.

## **IMGT database**

The full IMGT database was used as a source of homologous germline sequences. This database was filtered by excluding non-human entries, entries with identical sequences, and partial or non-functional entries (supplementary materials). Furthermore, sequences including wildcards and non-canonical AAs were excluded. Constant regions were cleaved to match the Fab fragments produced by the IgdE and OgpA enzymes.

## Top-down spectral processing

Following the protein-centric (middle-down) LC-MS/MS on the Fab fragments,MS1 features were retrieved from the generated RAW files using BioPharmaFinder 3.2 (Thermo Scientific). Deconvolution was performed using the ReSpect algorithm, deconvoluting averaged scans over a selected RT window where the target clone eluted (Table S1). The output mass range for the fragment ions was set at 10 to 40 kDa. Charge states between 10 and 50 were included with a minimum of 6 and 10 adjacent charges for the low and high model mass respectively. No relative abundance or score threshold was used. The target mass was set to 25 kDa, the number of peak models to 1, with a shape of 2 and 2 (left/right). The peak detection minimum significance measure was set to 1 standard deviation and the peak detection quality measured was set to 95%.  The MS2 spectra over the selected retention time were deconvoluted to yield their protonated monoisotopic fragment masses using the Freestyles Xtract algorithm. The minimum charge was set to 1, the maximum charge was set to 50, no thresholds were set for the minimum number of detected charges and the relative abundance.

Throughout the manuscript, we make use of an error tolerant sliding window algorithm to optimize contig placement against a MD fragmentation spectrum (Figure S4). This algorithm slides a set of theoretical fragments along a provided m/z range, incrementing the fragment masses with a set increment (default 0.01 Th). To limit processing time, peaks in the spectra are binned and the number of non-empty bins are counted for each offset. The top scoring offsets (default: 100) are then scored by a more refined scoring function (Olsen & Mann, 2004) and the best scoring offset is returned. This enables error-tolerant scoring of (sub)sequences (used for scoring imperfect contigs), but also determination their prefix- and suffix- masses (used for CDR gap size calculation) and offset from previous positions (used during manual inspection to find modifications/mutations/sequencing errors around the termini).

## Bottom-up *de novo* sequencing

Bottom-up MS/MS spectra were processed with the PEAKS-AB *de novo* sequencing suite. Default settings were used unless explicitly mentioned. Variable modifications were set to: Deamidation(N/Q), pyro-Glu from E, pyro-Glu from Q, oxidation (H/W), oxidation (M). Max 4 variable PTMs per peptide, max 5 peptides reported per spectrum, 0.02 fragment mass error tolerance, 20 ppm parent mass tolerance, fixed modification: Carboxymethyl. The resulting *de novo* predictions (referred to as *reads* throughout the manuscript), were inserted into the proteomic short read assembly tool STITCH for two subsequent runs, (parameter files in supplemental information) to yield a frequency table and sequence template for each target chain.

The sequence template is internally numbered according to the IMGT numbering convention. The de novo reads are numbered by aligning them to the sequence template using the smith waterman algorithm with a custom scoring matrix (supplementary materials) and copying the numbering.

# References

Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, *422*(6928), 198–207. https://doi.org/10.1038/nature01511

Aebersold, R., & Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. *Nature*, *537*(7620), 347–355. https://doi.org/10.1038/nature19949

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). The Generation of Antibody Diversity. In *Molecular Biology of the Cell* (4th ed.). Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK26860/

Altelaar, A. F. M., Munoz, J., & Heck, A. J. R. (2013). Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature Reviews Genetics*, *14*(1), 35–48. https://doi.org/10.1038/nrg3356

Bondt, A., Dingess, K. A., Hoek, M., van Rijswijck, D. M. H., & Heck, A. J. R. (2021). A Direct MS-Based Approach to Profile Human Milk Secretory Immunoglobulin A (IgA1) Reveals Donor-Specific Clonal Repertoires With High Longitudinal Stability. *Frontiers in Immunology*, *12*. https://doi.org/10.3389/FIMMU.2021.789748

Bondt, A., Hoek, M., Tamara, S., de Graaf, B., Peng, W., Schulte, D., … Heck, A. J. R. (2021). Human plasma IgG1 repertoires are simple, unique, and dynamic. *Cell Systems*, *12*(12), 1131-1143.e5. https://doi.org/10.1016/j.cels.2021.08.008

Bornholdt, Z. A., Turner, H. L., Murin, C. D., Li, W., Sok, D., Souders, C. A., … Walker, L. M. (2016). Isolation of potent neutralizing antibodies from a survivor of the 2014 Ebola virus outbreak. *Science*, *351*(6277), 1078–1083. https://doi.org/10.1126/science.aad5788

Briney, B., Inderbitzin, A., Joyce, C., & Burton, D. R. (2019). Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature*, *566*(7744), 393–397. https://doi.org/10.1038/s41586-019-0879-y

Castellana, N. E., Pham, V., Arnott, D., Lill, J. R., & Bafna, V. (2010). Template Proteogenomics: Sequencing Whole Proteins Using an Imperfect Database. *Molecular & Cellular Proteomics*, *9*(6), 1260–1270. https://doi.org/10.1074/mcp.M900504-MCP200

Charles A Janeway, J., Travers, P., Walport, M., & Shlomchik, M. J. (2001). The generation of diversity in immunoglobulins. In *Immunobiology: The Immune System in Health and Disease* (5th ed.). Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK27140/

Corti, D., Misasi, J., Mulangu, S., Stanley, D. A., Kanekiyo, M., Wollen, S., … Sullivan, N. J. (2016). Protective monotherapy against lethal Ebola virus infection by a potently neutralizing antibody. *Science (New York, N.Y.)*, *351*(6279), 1339–1342. https://doi.org/10.1126/science.aad5224

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

Guthals, A., Clauser, K. R., & Bandeira, N. (2012). Shotgun Protein Sequencing with Meta-contig Assembly. *Molecular & Cellular Proteomics*, *11*(10), 1084–1096. https://doi.org/10.1074/mcp.M111.015768

Hom, J. R., Tomar, D., & Tipton, C. M. (2022). Exploring the Diversity of the B-Cell Receptor Repertoire Through High-Throughput Sequencing. In J. Rast & K. Buckley (Eds.), *Methods in Molecular Biology: Immune Receptors* (Vol. 2421, pp. 231–241). https://doi.org/10.1007/978-1-0716-1944-5\_16

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

Kaplon, H., & Reichert, J. M. (2021). Antibodies to watch in 2021. *MAbs*, *13*(1). https://doi.org/10.1080/19420862.2020.1860476

Lefranc, M.-P., & Lefranc, G. (2020). Immunoglobulins or Antibodies: IMGT® Bridging Genes, Structures and Functions. *Biomedicines*, *8*(9), 319. https://doi.org/10.3390/biomedicines8090319

Lefranc, M. P. (2003, January 1). IMGT, the international ImMunoGeneTics database®. *Nucleic Acids Research*, Vol. 31, pp. 307–310. https://doi.org/10.1093/nar/gkg085

Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., & Lajoie, G. (2003). PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, *17*(20), 2337–2342. https://doi.org/10.1002/rcm.1196

Marks, C., & Deane, C. M. (2020). How repertoire data are changing antibody science. *The Journal of Biological Chemistry*, *295*(29), 9823–9837. https://doi.org/10.1074/jbc.rev120.010181

Olsen, J. V, & Mann, M. (2004). *Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation*. *101*. Retrieved from www.pnas.orgcgidoi10.1073pnas.0405549101

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

Raybould, M. I. J., Marks, C., Lewis, A. P., Shi, J., Bujotzek, A., Taddese, B., & Deane, C. M. (2020). Thera-SAbDab: the Therapeutic Structural Antibody Database. *Nucleic Acids Research*, *48*(D1), D383–D388. https://doi.org/10.1093/nar/gkz827

Schroeder, H. W., & Cavacini, L. (2010). Structure and function of immunoglobulins. *Journal of Allergy and Clinical Immunology*, *125*(2), S41–S52. https://doi.org/10.1016/j.jaci.2009.09.046

Schroeder Jr., H. W. (2006). Similarity and divergence in the development and expression of the mouse and human antibody repertoires. *Dev Comp Immunol*, *30*(1–2), 119–135. https://doi.org/10.1016/j.dci.2005.06.006

Schulte, D., Peng, W., & Snijder, J. (2022). Template-Based Assembly of Proteomic Short Reads For De Novo Antibody Sequencing and Repertoire Profiling. *Analytical Chemistry*, *94*(29), 10391–10399. https://doi.org/10.1021/ACS.ANALCHEM.2C01300

Sen, K. I., Tang, W. H., Nayak, S., Kil, Y. J., Bern, M., Ozoglu, B., … Becker, C. (2017). Automated Antibody De Novo Sequencing and Its Utility in Biopharmaceutical Discovery. *Journal of the American Society for Mass Spectrometry*, *28*(5), 803–810. https://doi.org/10.1007/s13361-016-1580-0

Srzentić, K., Fornelli, L., Tsybin, Y. O., Loo, J. A., Seckler, H., Agar, J. N., … Zhou, M. (2020). Interlaboratory Study for Characterizing Monoclonal Antibodies by Top-Down and Middle-Down Mass Spectrometry. *Journal of the American Society for Mass Spectrometry*, *31*(9), 1783–1802. https://doi.org/10.1021/jasms.0c00036

Toby, T. K., Fornelli, L., & Kelleher, N. L. (2016). Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annual Review of Analytical Chemistry*, *9*(1), 499–519. https://doi.org/10.1146/annurev-anchem-071015-041550

Tran, N. H., Rahman, M. Z., He, L., Xin, L., Shan, B., & Li, M. (2016). Complete de Novo Assembly of Monoclonal Antibody Sequences. *Scientific Reports*, *6*(31730), 1–10. https://doi.org/10.1038/srep31730

Valgardsdottir, R., Cattaneo, I., Napolitano, G., Raglio, A., Spinelli, O., Salmoiraghi, S., … Golay, J. (2021). Identification of Human SARS-CoV-2 Monoclonal Antibodies from Convalescent Patients Using EBV Immortalization. *Antibodies*, *10*(3), 26. https://doi.org/10.3390/antib10030026