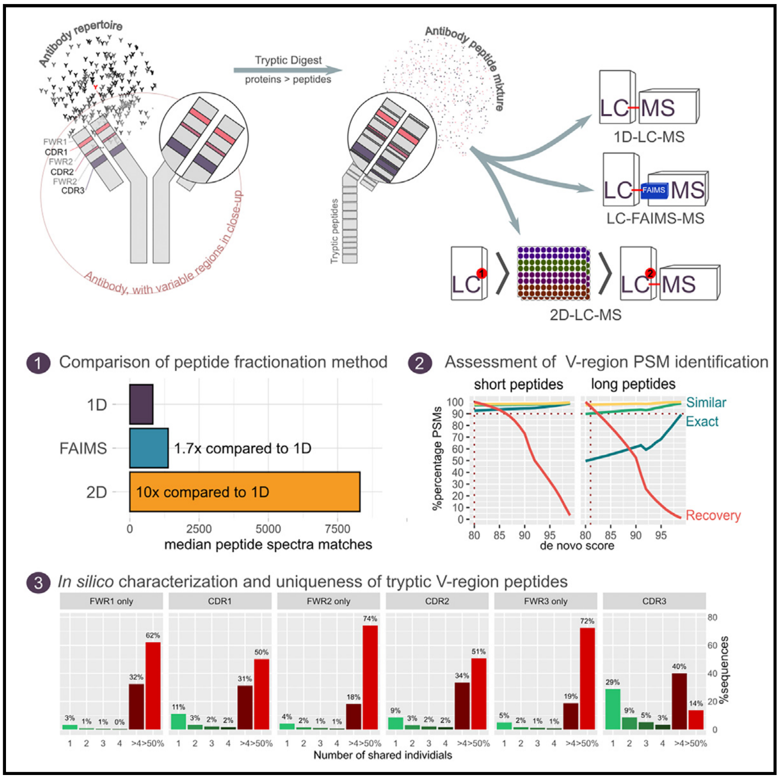
Cell Reports
Methods

Article

Improved detection of tryptic immunoglobulin variable region peptides by chromatographic and gas-phase fractionation techniques

Graphical abstract



Highlights

- 2D LC-MS and FAIMS improve the detection antibody variable region peptides
- *In silico* analysis of Ab sequences allows estimation of peptide uniqueness
- Tryptic CDR and FWR peptides are frequently shared between individuals
- Calculated uniqueness supports selection of the best available antibody surrogate

Authors

Christoph Stingl, Martijn M. VanDuijn, Thomas Dejoie, Peter A.E. Sillevis Smitt, Theo M. Luider

Correspondence

c.stingl@erasmusmc.nl

In brief

The analysis of circulating polyclonal antibodies is a promising area of research with many challenges. Stingl et al. show how multi-dimensional peptide fractionation improves the detection of antibody peptides and how *in silico* analysis provides objective estimates of peptide uniqueness, demonstrating the applicability of bottom-up proteomics for antibody analysis.

Stingl et al., 2024, Cell Reports Methods *4*, 100795
June 17, 2024 © 2024 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.crmeth.2024.100795

Cell Reports Methods

CellPress
OPEN ACCESS

Article

Improved detection of tryptic immunoglobulin variable region peptides by chromatographic and gas-phase fractionation techniques

Christoph Stingl,[1,3,*] Martijn M. VanDuijn,[1] Thomas Dejoie,[2] Peter A.E. Sillevis Smitt,[1] and Theo M. Luider[1]

[1]Clinical and Cancer Proteomics, Department of Neurology, Erasmus MC, 3000 CA Rotterdam, the Netherlands
[2]Laboratoire de Biochimie, Centre Hospitalier Universitaire (CHU), 44000 Nantes, France
[3]Lead contact
*Correspondence: c.stingl@erasmusmc.nl
https://doi.org/10.1016/j.crmeth.2024.100795

MOTIVATION This study was motivated by the question of how to improve the proteomic analysis of circulating antibodies, a class of molecules that we investigate because of their high relevance to disease diagnosis and therapy. Our aim was to determine to what extent methodological developments in plasma proteomics in the field of peptide fractionation are capable of improving the detection of antibody-related peptides by bottom-up proteomics. In particular, the study focused on the detection of antibody variable region peptides.

SUMMARY

The polyclonal repertoire of circulating antibodies potentially holds valuable information about an individual's humoral immune state. While bottom-up proteomics is well suited for serum proteomics, the vast number of antibodies and dynamic range of serum challenge this analysis. To acquire the serum proteome more comprehensively, we incorporated high-field asymmetric waveform ion-mobility spectrometry (FAIMS) or two-dimensional chromatography into standard trypsin-based bottom-up proteomics. Thereby, the number of variable region (VR)-related spectra increased 1.7-fold with FAIMS and 10-fold with chromatography fractionation. To match antibody VRs to spectra, we combined *de novo* searching and BLAST alignment. Validation of this approach showed that, as peptide length increased, the *de novo* accuracy decreased and BLAST performance increased. Through *in silico* calculations on antibody repository sequences, we determined the uniqueness of tryptic VR peptides and their suitability as antibody surrogate. Approximately one-third of these peptides were unique, and about one-third of all antibodies contained at least one unique peptide.

INTRODUCTION

Circulating antibodies in serum and other body fluids represent a valuable source of information about an individual's immune state in response to infection, cancer, or an autoimmune disease. As a component of the adaptive immune system, B cells possess the capacity to initiate targeted immune responses against pathogens through the production of antibodies that bind specifically with these antigens. Genetic recombination and somatic hypermutation enable the immune system to provide a suitable antibody against a large number of antigens based on a comparatively modest number of germline genes. In theory, this mechanism allows for creating a variety of about 10^{15} distinct antibody clones. However, the actual number of B cell clones within a human is estimated to be around 10^9.[1] Antibodies are secreted by activated plasma cells, and the sequence of the antigen-binding site in the released antibodies corresponds to the B cell receptor (BCR) sequence of the precursor cell.[2] Nonetheless, as only a fraction of all B cells are activated by antigens and subsequently produce antibodies, the immune repertoire of membrane-bound BCR does not completely and sufficiently represent that of the soluble antibodies circulating in the bloodstream.[3] As a result, BCR sequencing alone falls short in providing quantities of circulating antibodies. Thus, additional methods for functional antibody quantification are required.[4]

Over time, the analysis of antibodies has progressed to techniques allowing genetic sequencing of membrane-bound antibodies and circulating antibody repertoires.[3] Currently, the sequencing of BCRs is being accomplished to an ever-increasing extent by high-throughput nucleic acid sequencing techniques.[5] A subset of these techniques is capable of sequencing of antibody repertoires at the single-cell level[6] and preserving the pairing information of heavy and light chains.[7]

While all these techniques are powerful in mapping antibody repertoires up to a depth of millions of sequences, they merely provide information about the antibody-secreting cells, such as peripheral blood mononuclear cells (PBMCs), rather than information about the antibodies that are actually secreted and circulating.[3] Consequently, information about antibody concentrations cannot be acquired, and these techniques cannot be directly applied to cell-free materials, such as serum. More recently, technical improvements of mass spectrometry (MS)-based proteomics have enabled the analysis of secreted and circulating antibodies.[8] This can be achieved on the basis of antibody databases derived from RNA repertoire sequencing. If RNA sequencing is not possible, e.g., in cell-free samples such as serum or plasma, mass spectrometric methods can also facilitate the *de novo* sequencing of antibodies, although this approach currently primarily assesses the most abundant antibodies.[8,9] When analyzing antibodies with bottom-up proteomics, proteins are first cleaved into peptides, usually by enzymes. Hence, these peptides serve as surrogates for the associated proteins, and the presence and quantity of the actual precursors are derived through bioinformatic data analysis. As not all peptide sequences can be clearly linked to their corresponding protein sequences, unambiguous identification and quantification may not always be feasible. This challenge is particularly significant in instances involving protein groups or classes with a high degree of sequence homology, such as antibodies. Nonetheless, bottom-up proteomics is widely used because the technique is highly sensitive, provides accurate quantification, and is easy to perform and to automate. It is, therefore, well suited for large cohort studies or routine measurements.

The proteomics analyses of antibodies depends also on the availability of the antibody sequences information. If the expected antibody sequences have already been determined, a database-driven approach can be used, where the acquired mass spectra are matched against database-derived spectra. This approach is primarily carried out to ascertain the presence and quantify the abundance of antibodies. Chueng et al. demonstrated an analysis of circulating polyclonal antibodies by combining next-generation sequencing, to generate a sequence database, with MS-based proteomics for data acquisition.[10] Lindesmith and co-workers used that approach to determine the quantitative response of antibody repertoire related to human norovirus vaccination. This information provided the basis to select antibody clonotypes for epitope and structural analysis, which in turn led to the discovery of a neutralizing antibody.[11] Other researchers, along with our group, have applied this approach to quantify M-protein—a patient-specific antibody with characteristically high abundance in multiple myeloma (MM) patients—to monitor disease progression and treatment.[12–14] If the proteomics analysis is applied on antibodies with unknown sequences and no database can be generated, *de novo* sequencing becomes necessary. In this approach, the amino acid sequence of the variable region (VR) is directly derived from the acquired spectra. Peng and co-workers showed sequencing of the VR of a monoclonal antibody with 99% accuracy.[15] McDonald and co-workers demonstrated a workflow for sequencing M-protein light and heavy chains directly from MM patient samples, without relying on BCR sequencing.[16] Bondt et al. combined (protein-centric) top-down and (peptide-centric) bottom-up experiments to profile the immunoglobulin (Ig) G1 repertoire in serum. This method allowed for the quantification and sequencing of individual Ig clones.[9] Most recently, Peng and co-workers used a multi-enzyme bottom-up proteomics approach to sequence the light and heavy chain of an M-protein. They validated the accuracy of the sequence through top-down proteomics.[17]

The minimally invasive nature of blood sampling and the routine collection of serum and plasma samples in hospitals make peripheral blood samples a common choice for clinical studies. Nevertheless, the proteome analysis of serum samples is challenging because of the substantial dynamic range of the serum proteome. Igs, as a collective group, and specifically IgG as the most abundant class, constitute a quantitatively significant proportion of the overall plasma protein content. However, the concentration of individual antibodies can vary over several orders of magnitude (ranging from 10 g/L to below 10 μg/L). According to calculations conducted by Lavinder et al., the polyclonal antibody concentration in serum required for antigen elimination is estimated to be approximately 10 μg/L.[3] This estimation further indicates that the concentration of individual antibodies is roughly a million times lower in concentration than that of highly abundant plasma proteins. Techniques to overcome this problem in MS-based proteomics are depletion of high-abundant (and middle-abundant) proteins,[18,19] affinity enrichment strategies in cases where the analysis is focused on a small and specific set of proteins,[20] or multi-dimensional fractionation techniques.[21–23] In the latter techniques, peptides undergo separation through a sequence of at least two orthogonal chromatographic separation techniques. As a result of enhanced chromatographic separation, low-abundant peptides are better separated from high-abundant peptides, which in turn reduces suppression and improves detection of low-abundant peptides. Improvements of sensitivity can further be attributed to the fact that an increasing number of fractionations enables a higher input amount of sample material for analysis. Furthermore, peptides can also be separated using ion mobility in the gas phase of the mass spectrometer. In ion mobility, the spatial shape of the peptide ion is decisive for separation, which is defined by length, side chains, and charge of the amino acid chain. Ion-mobility fractionation is conducted in the gas phase, hence after chromatographic separation and ionization. Consequently, the same limitations concerning the amount of sample that can be loaded and suppression during ionization apply as for measurements without ion mobility. Improvements of detection due to the inclusion of ion mobility can thus be attributed to the reduction in the complexity of the peptide mixture and the better separation from background ions. Several studies have shown that employing high-field asymmetric waveform ion-mobility spectrometry (FAIMS) has resulted in heightened sensitivity, improved signal-to-noise ratios in peptide quantification, and an increased number of peptide and protein identifications.[24–27]

In this study, we aimed to improve the detection of VR peptides using gas-phase ion-mobility (FAIMS) and preparative high-pH liquid chromatography (LC) peptide fractionation (2D).

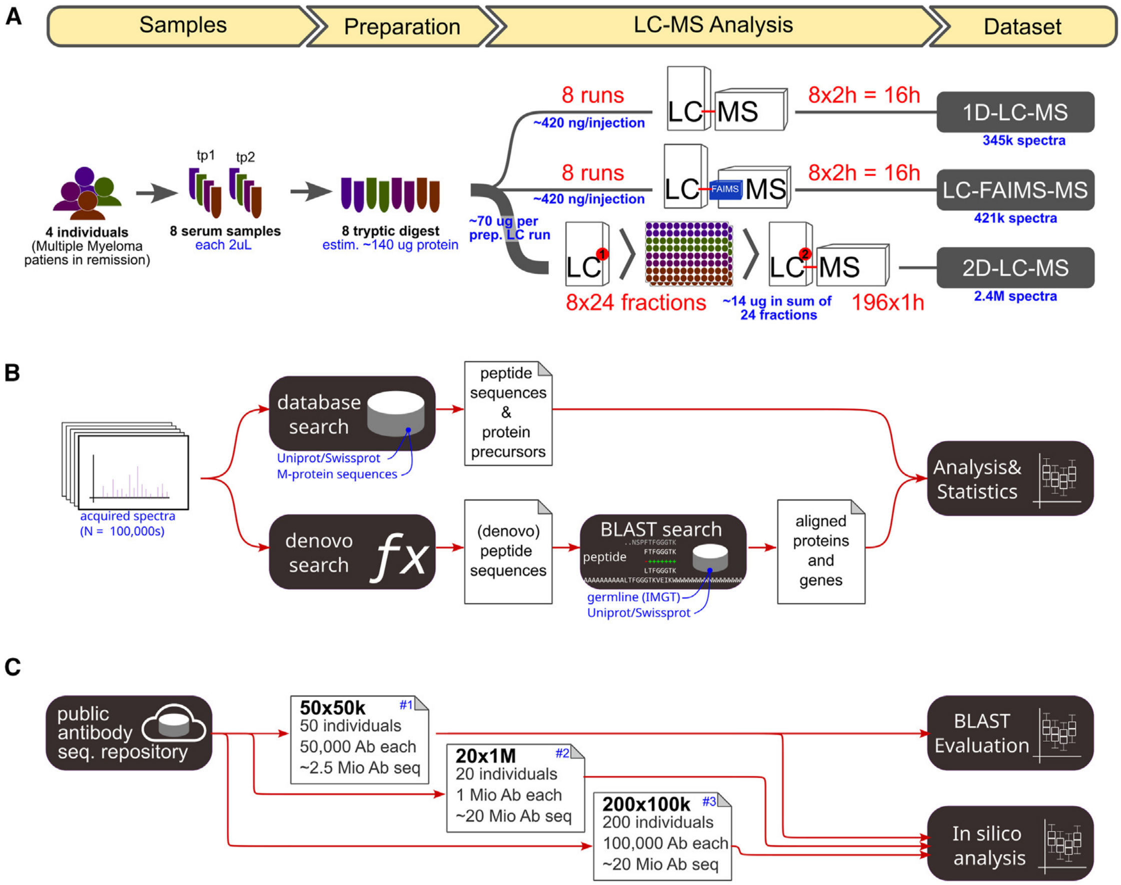


Figure 1. Experimental overview

(A) Flowchart of the sample preparation. Samples were analyzed using three different methods: 1D, FAIMS, and 2D. The number of runs and the LC-MS acquisition time are indicated in red for each method.

(B) Flow chart of data analysis.

(C) *In silico* analysis was conducted of three datasets (50 × 50,000, 20 × 1 million, and 200 × 100,000) obtained from a public repository (OAS).

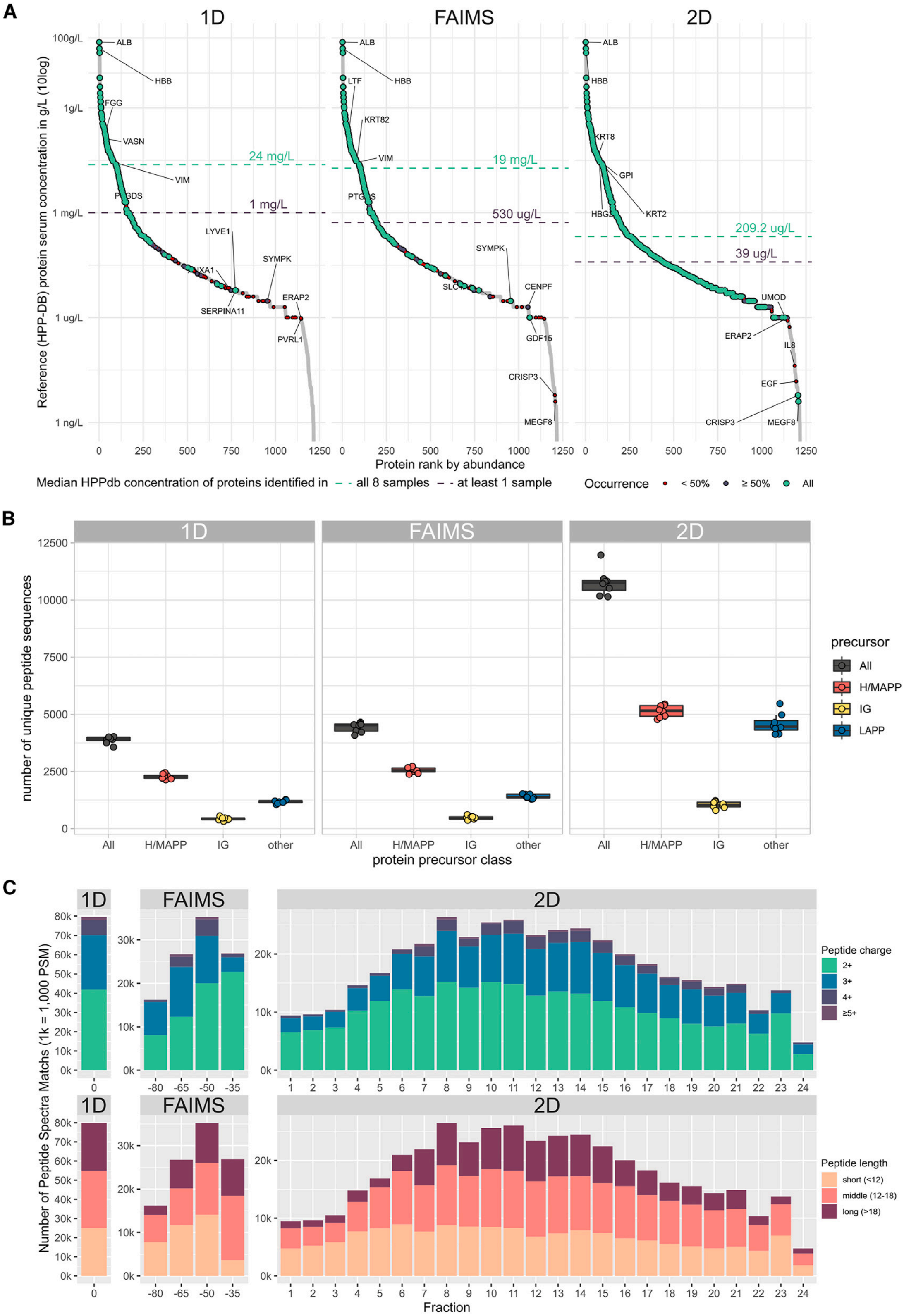
The study was conducted on eight serum samples from four MM patients, collected during their treatment or remission phases. The M-protein amino acid sequences were used as reference sequence, but otherwise the immune repertoire was unknown. To effectively analyze VR peptides with *a priori* unknown sequences, we combined and evaluated two data-processing steps: firstly, the *de novo* sequencing of acquired spectra, and, secondly, the alignment of these sequences with Ig germline sequences.

Furthermore, we conducted an analysis to determine the presence of unique VR peptides suitable as proxies for specific antibody proteoforms. Through *in silico* analysis based on datasets with up to ∼20 million Ig sequences, sourced from the Observed Antibody Space repository,[28,29] we built a reference database of tryptic VR peptides. This reference base was then used to compare characteristics of the peptides acquired through experimentation and those generated *in silico*. This approach provided an objective and systematic estimate of the potentialities and constraints associated with analyzing the circulating antibody repertoire via a trypsin-based bottom-up proteomic approach.

RESULTS

Mapping of the common serum proteome

We digested 2 μL of eight serum samples taken at two time points from four patients and measured from each a volume corresponding to 6 nL of initial serum (estimated 420 ng of proteins) by direct LC-MS (1D) and LC-FAIMS-MS (FAIMS) using a 90-min LC gradient in a method with approximately 120-min total run time. Further, half of the digest (1 μL, corresponding to estimated 70 μg of protein) was fractioned by high-pH reversed-phase chromatography into 24 fractions. Subsequently, 20% of each fraction was measured individually using LC-MS with a 30-min gradient (2D). This method had an overall run time of approximately 60 min. Hence, for 2D, the total sample amount that could be analyzed was roughly 30× larger than that for 1D, and data acquisition took around 12× longer (Figure 1). A total of 345,000, 421,000 and 2.4 million MS/MS spectra were acquired in the 1D, FAIMS, and 2D datasets, respectively. These acquisitions resulted in the identification of 296, 310, and 890 proteins through an MS/MS database search on the common human proteome, with the thresholds



(legend on next page)

of a false discovery rate (FDR) of 1% maximum and at least two peptides per protein.

All subsequent analyses were carried out on the basis of peptide-spectra-match (PSM)-centric PEAKS X peptide reports. This involved a database search with an FDR threshold of maximum 1% and a PEAKS X *de novo* score threshold of 80. To show the covered depth of the serum proteome, the corresponding precursor proteins were assigned to the expected serum concentrations according to the information of the Human Plasma Protein Database (HPPdb).[30,31] In total, 275, 303, and 729 proteins could be matched to corresponding database entries for datasets 1D, FAIMS, and 2D, respectively. The median HPPdb concentration of all proteins was similar between 1D and FAIMS measurements (1 and 0.53 mg/L, respectively), and approximately 25-fold lower in 2D samples (39 μg/L). Notably, the median HPPdb concentrations of proteins present across all samples dropped from 24 and 19 mg/L in 1D and FAIMS, respectively, to 209 μg/L in 2D samples. This observation indicated a greater than 100-fold sensitivity increase in measuring reproducibly low-abundant serum proteins through the 2D-LC-MS approach (Figure 2A). FAIMS resulted in a median increase of 1.14 times (+14%) more uniquely identified peptides. Furthermore, the application of chromatographic fractionation (2D) increased the overall number of peptide identifications by about 2.7-fold (+171%). The count of PSM increased by FAIMS 1.3-fold and by 2D 5.3-fold. We further categorized the hits from the database search based on their precursors. These were grouped into peptides derived from high- or medium-abundance plasma proteins (H/MAPP), Igs (primarily germline sequences present in the Uniprot/Swissprot database), and low-abundance plasma proteins (LAPPs). The proportions of the peptide groups (with H/MAPP accounting for 68%–69%, Ig for 12%, and LAPP for 17%–18%) remained consistent regardless of the fractionation technique used. Consequently, improvements in the number of identifications attributable to LAPPs stemmed from the overall increased number of acquired spectra, rather than a selective enrichment process (Figure 2B; Table S1). The average sequence coverage of the specific M-proteins used as reference antibodies did not show an increase with the utilization of FAIMS (gain: 1.02, $p = 0.741$). However, a relative increase of 40% ($p = 0.0067$) was observed when employing the 2D technique (Figure S1; Table S2). The distribution of PSMs from the database search across FAIMS compensation voltage (CV) fractions demonstrated that most PSMs were identified within FAIMS fraction −50 V. For the 2D method, the majority of peptide identifications were concentrated in the mid-range and particularly in the second quarter of the 2D fractionation. For the FAIMS approach, the numbers and proportions of higher-charged (>2+) and shorter peptides (<12 aa) were lowest in fraction −35 V and increased progressively with a rise in negative CV. For the 2D approach, peptides in the earlier fractions (first quarter) were primarily of low charge (2+) and short length (<12 aa), while later fractions showed a more equal distribution. Interestingly, longer peptides (>18 aa) were most frequently found in the center of the fractionation runs, with their frequency decreasing in later eluting fractions (Figure 2C).

Assignment of spectra to variable antibody regions

In addition to the MS/MS database searches, we conducted a *de novo* search using the spectra of the three datasets. All further analyses included *de novo* hits with a PEAKS X *de novo* score >80 (most stringent export setting in *de novo* software PEAKS X). As the *de novo* search method does not directly provide an FDR estimate comparable to the database search, we indirectly assessed the *de novo* search by comparing PSMs of both methods on a confident set of reference spectra. We defined confident spectra as those that, on the basis of the database search, met the two criteria of having an FDR < 1% and being assigned to a high- or middle-abundant plasma protein. The PSMs identified in the two searches were compared on three different levels: (1) exact sequence correspondence, (2) ≥90% sequence similarity (by Levenshtein distance relative to sequence length), or (3) recovery of the initial gene identifier by a subsequent BLAST search of the *de novo* hit. Sequencing of peptides was consistent to a degree of >90% for short peptides (7–10 aa) above *de novo* score 80 in every of the three aspects (exact, similarity, alignment). With increasing peptide length, the frequency of deviating results increased, most distinctly in the context of exact sequence comparison. For instance, regarding peptides 10–12 aa long acquired by the FAIMS method, the average exact correspondence was ∼82% at *de novo* score level 80 and the threshold of 90% exact correspondence was first exceeded above *de novo* score level 95. In contrast, the assessment based on similarity relative to sequence length was more robust and demonstrated an average true-positive rate of over 90%, even for long peptides (16–39 aa) (Figure S2). For further downstream analysis, we applied a peptide-length-dependent *de novo* score threshold, resulting in an estimated 90% average sequence conformity.(Table S3).

BLAST alignment was evaluated using tryptic peptide sequences computed on the basis of 2.5 million antibody sequences downloaded from the Observed Antibody Space (OAS) antibody sequence repository (50 × 50,000 OAS set). BLAST alignment was further carried out with four different

Figure 2. Results of MS/MS database search (plasma proteome mapping)

(A) Waterfall plots of the human plasma proteome, with protein abundances (obtained from the human plasma proteome database) plotted on a logarithmic scale against protein abundance rank.

(B) Boxplots of the number of identified peptides, including all peptides (dark gray), peptides derived from high- and middle-abundant plasma proteins (H/MAPP, in red), peptides originating from antibodies (IG, in yellow), and peptides from low-abundant plasma proteins (LAPPs, in blue). The lower and upper hinges correspond to the first and third quartiles, and the whisker extends from the hinge to the lowest and largest value, respectively, no further than 1.5 times the interquartile range.

(C) Distribution of the number of peptide-spectra matches (PSMs) identified in 1D, across four fractions of FAIMS, and within 24 fractions from 2D measurements, categorized by peptide charge and peptide length. PSM counts represent the sum across all eight samples.

composition-based matrices available in NCBI BLAST.[32–35] The evaluation of the BLAST search was carried out by means of the recovery of the initial VR (complementarity-determining regions [CDRs], or framework regions [FWR]) through BLAST search of tryptic peptides. Performing the BLAST search on VR tryptic peptide sequences without using a composition-based matrix (BLAST search parameter) yielded the highest recovery rates. Consequently, this setting was adopted for BLAST searches on the acquired data. Peptides with lengths of 6 aa or less were not efficiently aligned by BLAST. Low recovery rates were also observed for peptides shorter than 10 aa with already a few mutations (Figure S3A). This indicates further that long peptides are required for solid mapping of the hyper-mutated CDR regions, first and foremost CDR3. Less than 40% of all CDR3 peptides with a length $\leq$10 aa were successfully assigned by a BLAST search. In contrast, CDR3 peptides with lengths intervals of $10 < L \leq 20$ and $20 < L \leq 35$ could be aligned successfully in 91.9% and >99.8%, respectively. Alignment was more efficient for peptides originating from other, less frequently mutated regions, having recovery rates of 85.7% for peptides with lengths $\leq$10 aa and 99.0% for peptides longer than 10 aa. Peptides longer than 10 aa also constituted the majority ($10 < L \leq 20$, 42.3%; and $20 < L \leq 35$, 25.4%) of all peptides. (Table S4; Figure S3B).

In summary, when utilizing *de novo* search, a high level of precision in relation to the database search was evident for shorter peptides. However, as peptide length increased, the effect of missing and unrecognized fragment ions most likely contributed to increasingly inaccurate sequencing outcomes. In contrast, a subsequent BLAST search resulted in generally low recovery rates for shorter peptides and high recovery rates (97.7%) for peptides longer than 10 aa. Fortunately, the majority of tryptic antibody VR peptides are expected to exceed 10 aa in length.

Effect of VR peptide mapping by peptide fractionation

The number of *de novo* search identifications ranged in the three datasets from 80,000 to 502,000 PSMs (1D, 80,000; FAIMS, 122,000; and 2D, 502,000 PSM). Compared to data of the 1D (15,000) datasets, the number of PSMs aligned to antibody genes was 1.6-fold higher in FAIMS (24,000) and 8-fold higher in 2D (117,000). The number of PSMs linked to a VR was 1.7-fold higher in FAIMS (11,500) and 10-fold higher in 2D (69,000) compared to PSMs acquired with the 1D (6,900) method. Hence, the proportion of antibody-related PSMs was similar in the 1D (18.9%) and FAIMS (19.3%) datasets, and about 1.2-fold higher in the 2D (23.3%) dataset. Similar to this, the proportion of VR-related PSMs was also comparable in the 1D (8.6%) and FAIMS (9.4%) datasets and 1.6-fold higher (13.7%) in data of the 2D approach (Table S5). Consequently, the proportion of VR PSMs relative to all antibody PSMs increased through the use of fractionation and was 46% in 1D, 49% in FAIMS, and 59% in 2D data (Figure 3 and Table S5). Using FAIMS, a significant portion of antibody and VR peptides were identified in both relative and absolute quantities within the fraction with the lowest negative CV (−35 V). This voltage setting generally corresponds to longer peptides with low charge states. Regarding the 2D approach, while the majority of the *de novo* PSMs were retrieved from the midsection of the preparative LC run (around fraction 8–16), the fractions collected at the end of this section (fractions 14–16) contained the highest proportions and absolute counts of PSMs associated with antibodies and VRs (Figure 3). This observation indicates that these peptides possess a slightly above-average hydrophobicity compared to peptides originating from the common proteome.

The *de novo* PSMs were annotated more specifically according to their CDR or FWR, and the improvements by applying FAIMS or 2D fractionation were determined relative to the 1D method. Using FAIMS, the increase ranged between approximately 2-fold for FWR1 (gain = 1.91, p = 0.005), FWR2 (gain = 1.97, p = 0.023), and CDR2 (gain = 1.93, p = 0.015) to 1.36-fold for CDR1 (gain = 1.36, p = 0.05). The counts of FWR3 and CDR3 peptides had not increased significantly, and notably the CDR3 PSMs count remained relatively low. Overall, the improvements achieved through FAIMS when compared to the 1D approach (1.65-fold, p = 0.019) fell within the range of the enhancements observed from the database search (FAIMS: 1.32×). In contrast, when employing 2D LC-MS, the number of PSMs related to VRs had increased more than 10-fold (11.43×). Notably for the CDR, there was an even more substantial increase, with over 1,000 PSM/sample associated with CDR3 (13.1×) being identified. (Figure 4A; Table 1)

Next, we analyzed the distribution of the VR peptide in the different fractions. With the exception of CDR1—and in contrast to the PSM of the common proteome—PSMs related with VR showed the strongest enrichment in the −35-V FAIMS fraction. This was particularly evident with respect to FWR1. CDR1 peptides were predominantly identified in the FAIMS CV fraction of −50 V. In contrast to the 2D LC-MS analysis of the common proteome (Figure 2C), VR peptides displayed a marginal shift to later 2D fractions and were highest around fractions 14–16. This trend was not observed for peptides linked to FWR1 and FWR3 (Figure S4). The variations in the count of *de novo* PSMs derived from VRs were considerably more pronounced than those for the common proteome identified through the database search. Comparison of *de novo* spectra counts and the measured concentration of residual M-protein revealed that the diversity of different antibody sequences correlated inversely with the residual M-protein concentration. Consequently, the residual concentration of M-protein accounts for a significant part of the variance observed in the sequenced variable peptides (Figure 4B).

The alignment was further extended to heavy chain constant regions in order to assess the distribution of Ig classes (IgG, IgA, IgM, and IgE found) and to light-chain constant regions, in order to determine the counts of kappa and lambda chains. On the basis of the heavy-chain constant-region peptides, IgG—expected to be the most abundant class—was by far the most frequently found Ig class, followed by IgA and IgM (both about 10-times fewer PSMs than IgG). IgE was found only sporadically, and IgD not at all. Kappa-chain PSMs were identified at roughly twice the frequency of lambda chains, and a similar count of VR PSMs for both light and heavy chains was observed. Although the overall PSM counts varied between the three methods applied, the relative distribution of the various region and class assignments remained unaltered (Figure 4C). Similar to the findings regarding the variability in the

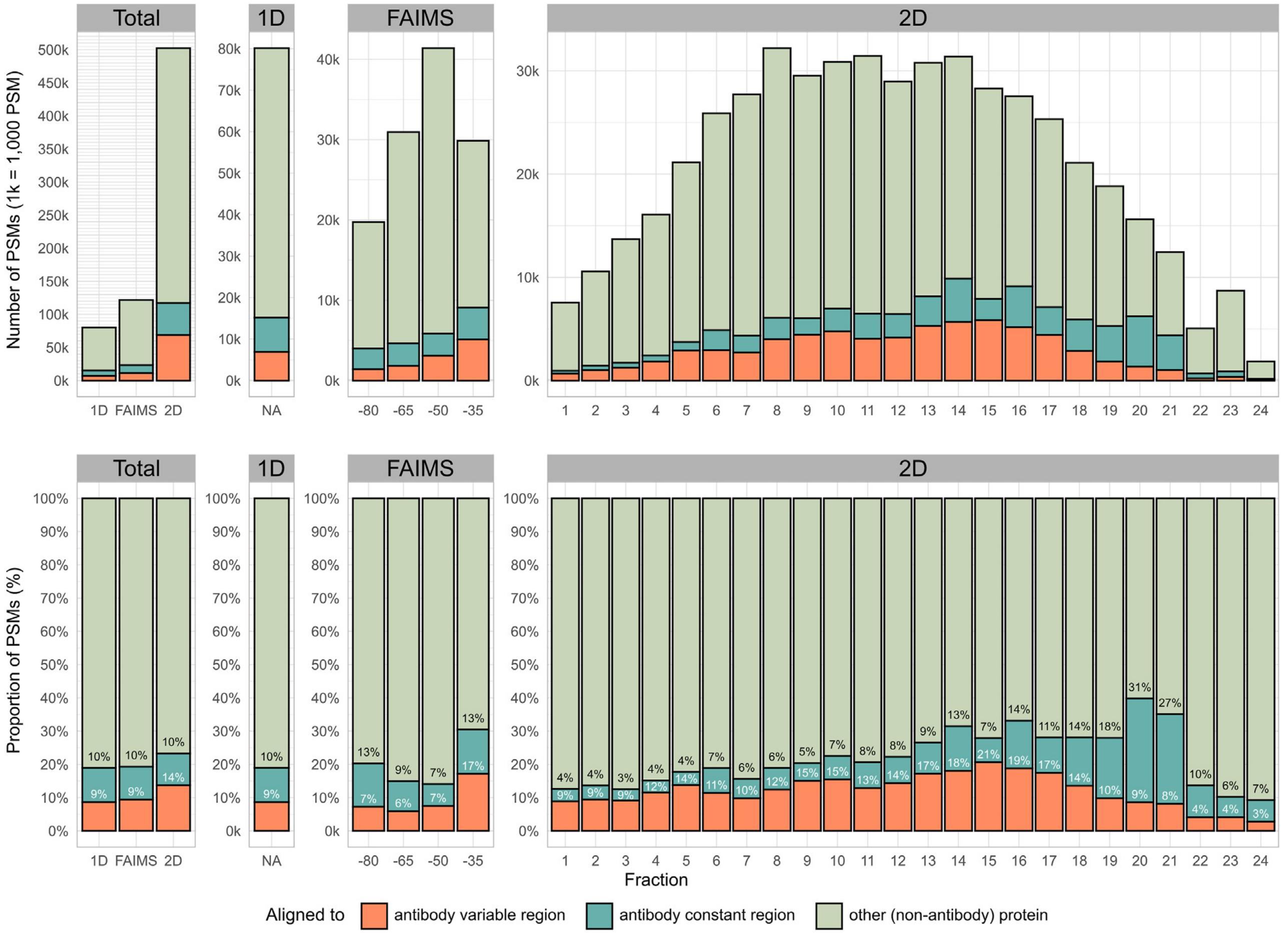


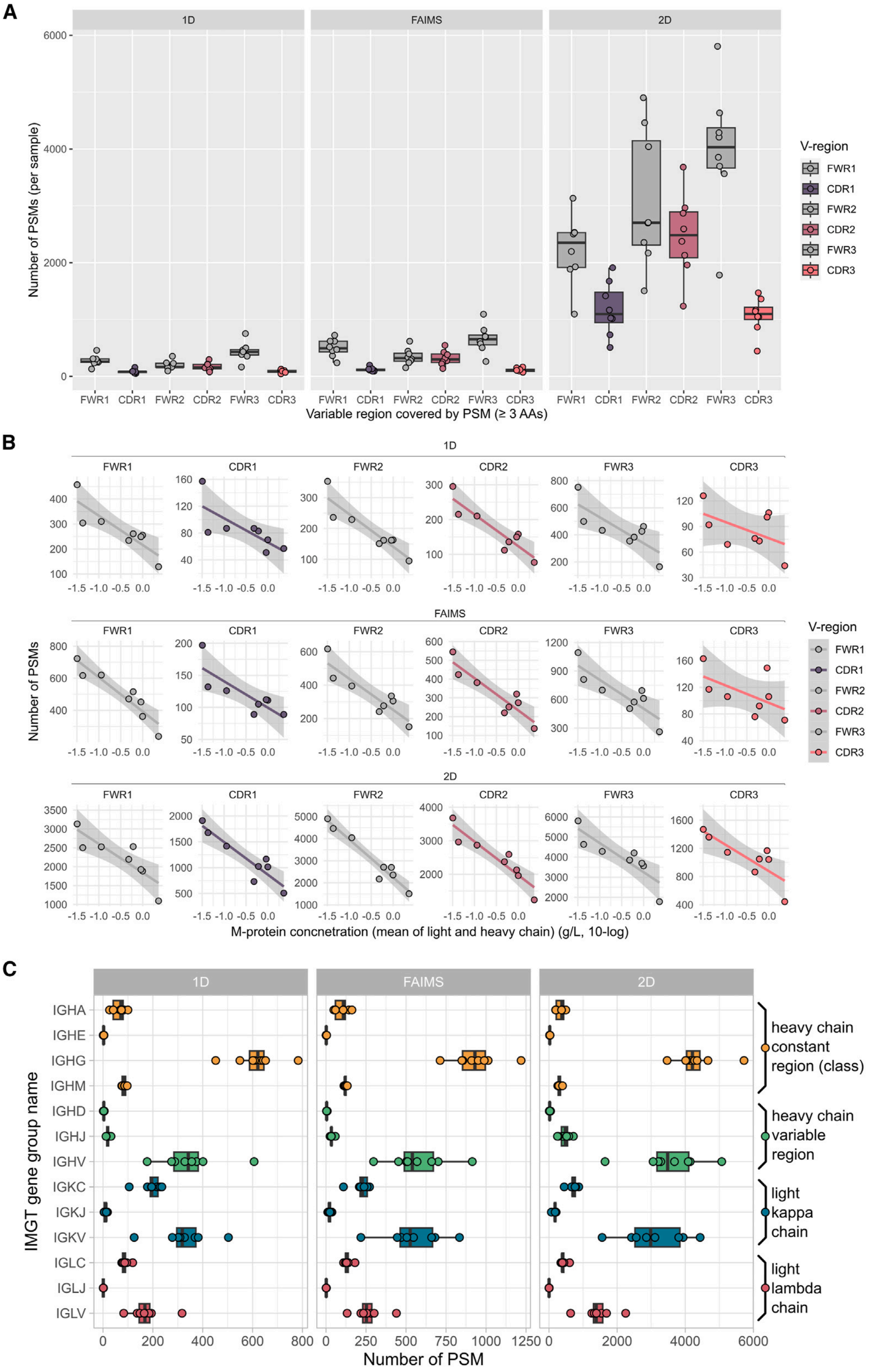
Figure 3. Results of the *de novo* search differentiated to fractions

Distribution of absolute counts (top) and relative proportions (bottom) of PSMs assigned to antibody VRs, antibody constant regions or non-antibody proteins, grouped by LC-MS method (1D, FAIMS, and 2D) and, if applicable, the corresponding fractions (FAIMS: four CVs of −80, −65, −50, and −35 V). In the total of all fractions, there were 6,928 spectra assigned to the VR by 1D, 11,473 by FAIMS, and 68,792 spectra by 2D. The total numbers per dataset and their corresponding proportions are also detailed in Table S5.

count of variable PSM described above, the number of PSMs related to the constant antibody region was also correlated with the residual M-protein concentration (Figure S5). Interestingly, the correlations in IgG and IgA concentrations were clearly stronger than that observed for IgM. Furthermore, these correlations were more accurately captured through the 2D approach compared to the other two methods. In contrast to these correlations, the number of PSMs associated to serum albumin—employed as a reference protein—remained unaffected by the concentrations of M-protein (Figure S5).

Uniqueness of tryptic antibody peptides and their viability as antibody protein surrogates

Interpreting the results acquired with bottom-up proteomics analysis of antibodies differs from interpreting those of the common proteome. Due to the sheer and almost unlimited number of varying but yet highly homologous sequences, a complete and comprehensive protein database does not exist *a priori*. In bottom-up analysis, theoretically, a single peptide, and, in practice, a small set of peptides, is sufficient to confidently identify a protein. This is under the simplifying condition that no differentiation between variants, such as SNP forms or splicing variants, is required. The likelihood that a peptide serves as a unique surrogate for a protein increases with peptide length, and peptides longer than 6 aa already have a greater than 95% chance of being unique (Figure S6A). Furthermore, the tryptic peptide length is roughly exponentially distributed, which indicates an overall random distribution of the amino acids Lys and Arg, which are of relevance for tryptic cleavage. Inherent to the exponential distribution, just less than 5% of the peptides are longer than 30 aa (or less than 1% longer than 50 aa). Compared to this theoretical distribution, the actual length distribution of acquired peptides shows that short peptides are identified less frequently, and there is a higher degree of similarity between the theoretical and measured distributions for longer peptides. This is why a discriminatory bias or apparent cutoff for longer peptides (>20 aa) was not observed (Figure S6B).



(legend on next page)

The uniqueness of tryptic peptides within the context of the circulating antibody repertoire is a crucial measure to characterize their potential as specific proxies for individual antibodies. In order to estimate this uniqueness measure, we downloaded a large number of publicly available sequences from the Observed Antibody Repository (OAS).[28,29] We then compiled three distinct datasets with varying numbers of individuals and sequences. One dataset comprised sequences from 50 individuals, each contributing up to 50,000 sequences (50 × 50,000). A second dataset included sequences from 20 individuals, each contributing up to 1 million sequences (20 × 1 million). The third dataset encompassed sequences from 200 individuals, with each individual contributing up to 100,000 sequences each (200 × 100,000). Each individual corresponded to one OAS data unit of unpaired sequences, either heavy or light chains, and each dataset contained equal numbers of data units of heavy and light chains (Table S6). Antibody VR sequences were used to generate a database of tryptic peptides, annotated with peptide length, the associated CDR or FWR, and number of mutations. The length distribution of tryptic peptides of antibodies (Figure 5) differed clearly from that of the general proteome (Figure S6B), and differences between the various CDRs and FWRs were observed as well. As an example, the first tryptic cleavage sites of the germline sequence occur at the N-terminal positions 19 (21% of all V-regions), 13 (19%), 12 (17%), 18 (13%), and 5 (5%). Peptides with lengths around 5, 12, and 18 aa were particularly prevalent in the *in silico* digest. Furthermore, the distribution of tryptic peptides of CDR3 exhibited a shift toward longer peptides, with the maximum occurring around 25 aa. For peptides longer than 10 aa, the distribution of peptide lengths of acquired data corresponded to the *in silico*-generated data. However, acquired CDR3 peptide sequences longer than 20 aa are underrepresented in comparison to tryptic *in silico* peptides (Figure 5A). Because this bias was not observed for peptides originating from the common plasma proteome, which were identified through database search (Figure S6), it is more probable that the absence of long CDR3 peptides is a result of data analysis rather than being related to data acquisition.

As a next step, we assessed the uniqueness of peptide sequences. With the exception of CDR3, the largest group consisted of non-unique peptide sequences shared among half or more of the individuals. For peptides originating from these regions, the likelihood of uniqueness varies based on the number of individuals within the dataset, ranging from 15% (CDR1 in 50 individuals with sequence depth of 50,000) to 3% (FWR1 in 200 individuals with a sequence depth of 100,000). CDR1-covering peptides were more (11%–15%) often unique compared to peptides covering the shorter CDR2 (8%–10%) as well as peptides covering the FWR regions. In contrast, tryptic peptides encompassing CDR3 exhibited a considerably higher degree of uniqueness, with the number of unique peptides diminishing as the dataset increased (2.5 Mio sequence in 50 × 50,000, 45% unique; 20 Mio sequences in 20 × 1 million, 38%; and 20 Mio sequences in 200 × 100,000, 29%). Based on these results, we estimate that any given tryptic peptide covering a CDR3 region has at best a one-third chance of being unique and, in a strict sense, being suitable as an antibody surrogate peptide. For other CDRs, the estimated chance is roughly 10%; for FWR it is below 5% (Figure 5B). Moreover, it must be noted that the assignment to the region is an essential factor for the description of the uniqueness of tryptic peptides differentiated by regions. For the purpose of this analysis, we have chosen that a sequence overlap of at least 3 aa defines a CDR sequence. This enabled us to include all regions, including the light-chain CDR2, which is predominantly 3 aa long, and further allows that all peptides with a minimum length of 6 aa can be aligned to at least one CDR or FWR. An increase in the minimum overlap can therefore be expected to result in a reduction in the number of sequences and, at the same time, in a higher degree of uniqueness. In a further step, we examined the relationship between the uniqueness of peptides and the degree of mutations, defined as the number of amino acids deviating from the expressed germline sequence. Overall, around two-thirds of all peptide sequences of the VR corresponded exactly to the germline sequence and, as expected, these sequences were not unique. The number of mutations per tryptic peptides sequence ranged in 95% of cases from zero to five mutations, and just 1% of all sequences had more than eight mutations. Within the range of zero to five mutations, there was a steep increase in the proportion of unique sequences. Less than 10% of unique peptides had just one mutation, and around 50% had five mutations. Interestingly, as the number of mutations further increased (between 5 and 15 mutations per peptide sequences), the increase in the proportion of unique sequences was much lower than that in sequences with low mutation counts, and did not reach 100% uniqueness. These peptides, with high calculated mutation counts, typically have longer insertions in CDR3 compared to the germline sequence. Because each amino acid of this insertion was counted as one mutation, these peptides are represented as highly mutated non-unique peptide sequences (Figure S7).

In a further analysis, we calculated the frequency for antibody sequences to have at least one unique tryptic peptide that could serve as a suitable antibody surrogate. On average, each antibody VR was cleaved by trypsin into five peptides (with lengths between 5 and 34 aa). Thereby, a wide variation was determined, ranging from one to 14 peptides per antibody. The occurrence of tryptic peptides that are unique within the antibody repertoire of an individual depended on the sampling depth (or number of antibody sequences per individual). The

Figure 4. *De novo* search results of antibody PSMs

(A) The number of *de novo* PSMs corresponding to VRs (CDR1-3 and FWR1-3) determined through 1D, FAIMS, or 2D LC-MS. The lower and upper hinges correspond to the first and third quartiles, and the whisker extends from the hinge to the lowest and largest value, respectively, no further than 1.5 times the interquartile range.

(B) Correlation between the number of *de novo* PSMs and the residual M-protein concentration, grouped by dataset and VR.

(C) The number of *de novo* PSMs grouped by Ig class and chain.(Hinges and whiskers of the box plot have been defined in A)

Table 1. Improvements in the detection of VR peptides using FAIMS and 2D LC-MS

Region	$M_{(1D)}$	$M_{(FAIMS)}$	$gain_{(FAIMS)}$	$p_{(FAIMS)}$	$M_{(2D)}$	$Gain_{(2D)}$	$p_{(2D)}$
FWR1	258.0	493.5	1.91	0.005	2,350.0	9.11	<0.001
CDR1	82.0	111.5	1.36	0.05	1,095.0	13.35	<0.001
FWR2	162.5	320.0	1.97	0.023	2,702.0	16.63	<0.001
CDR2	154.0	297.0	1.93	0.015	2,483.5	16.13	<0.001
FWR3	429.5	652.5	1.52	0.052	4,031.0	9.39	<0.001
CDR3	84.0	106.0	1.26	0.123	1,096.5	13.05	<0.001
Of which CDR3j	27.5	54.5	1.98	0.002	644.5	23.44	<0.001
Alignments	1,176.5	1,937.5	1.65	0.019	13,449.5	11.43	<0.001
V region	828	1,380	1.67	0.026	8,303	10.0	<0.001

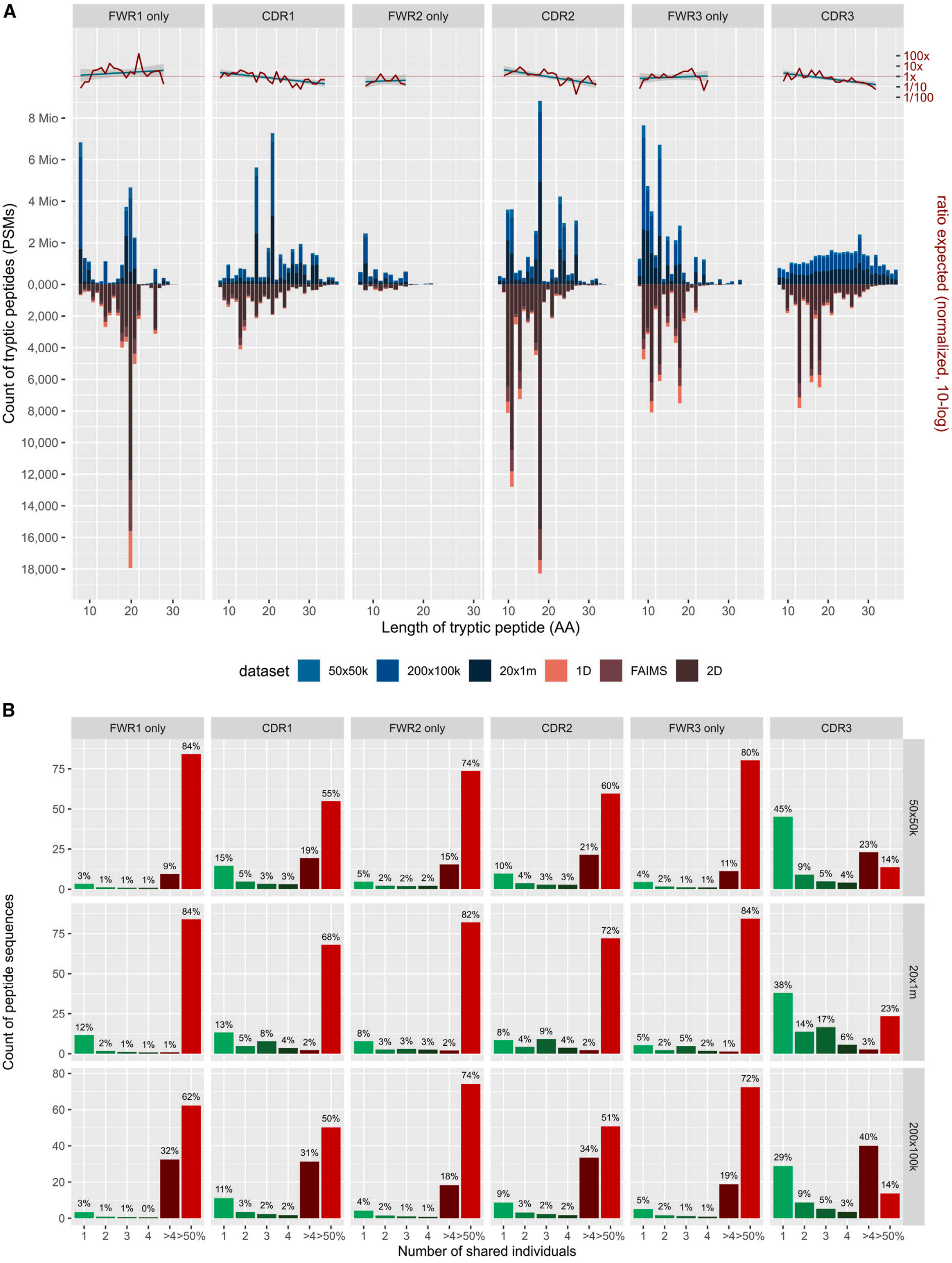
The gains are calculated as the median (M) ratios of the number of PSMs in FAIMS or 2D relative to 1D. The corresponding p values (t test) are provided in columns $p_{(FAIMS)}$ and $p_{(2D)}$. CDR3j counts the CDR3 alignments specifically to the J segment; Alignments counts the sum of all alignments (including multiple alignments to one PSM), and V region counts PSMs aligned to the VR.

occurrence decreased from 66% for datasets of 50,000 sequences to 56% for datasets of 100,000 sequences and 44% for datasets of 1 million sequences. When further assessing the more restrictive dataset-wide uniqueness, the frequency of finding unique peptides dropped to 49% for datasets of 50 individuals with 50,000 sequences and to respectively 34% and 35% for datasets of 200 individuals with 100,000 and 20 individuals with 1 million sequences each. The analysis also showed that a considerable proportion (6%–14%) of antibodies have more than one unique peptide and that, in rare cases (one in 20 million), antibodies with up to 10 unique peptides exist (Table 2).

DISCUSSION

The starting point of this work was the question of to what extent methodological developments in plasma proteomics in the field of peptide fractionation can have an impact on the detection of antibody-related peptides by bottom-up proteomics. Consequently, a workflow was established and validated that enabled us to link the acquired spectra to antibody VRs. Through an *in silico* analysis, we also addressed the potential and possible limitations of trypsin-based bottom-up proteomes to analyze the repertoire of circulating antibodies. On the basis of these three aspects—data acquisition, data analysis, and *in silico* analysis—and with regard to future analyses of monoclonal, oligoclonal, and eventually polyclonal antibodies in clinical samples, we have drawn the following conclusions.

First, through applying FAIMS gas-phase fractionation or 2D-LC high-pH peptide fractionation, we observed significant improvements in two aspects: database-search-based identifications of common plasma proteins and *de novo* search-based identifications of antibody VR peptides. The improvements achieved with 2D-LC were particularly prominent when compared to FAIMS. We attribute the lower improvements seen with FAIMS to the fact that it is concurrently applied on the fly with the regular LC-MS acquisition. While this approach does not extend the measurement time, it also restricts the amount of sample that can be loaded. Hence, the advantages of FAIMS gas-phase fractionation primarily stem from its ability to reduce complexity, and it offers a moderate improvement in extending the sensitivity range. Due to the large dynamic range of the serum proteome, the benefits of FAIMS are more limited in serum compared to the gains typically observed in the analysis of other sample types, such as cell cultures.[26] In contrast, with 2D-LC-MS, the initial sample is split into multiple fractions, each of which must be measured in an individual run. Consequently, this approach necessitates a longer overall instrument acquisition time (in this study, 12 times longer). However, it also enables an increase in the total sample volume introduced to the LC-MS measurement (approximately 30 times larger in this study) and reduces the sample complexity. Altogether, these advancements yielded substantial improvements in the number of identifications, particularly notable for proteins characterized by significantly lower concentrations. The 2D methodology is widely used in biomedicine and plasma proteomics, with the aim to augment both the number of protein identifications and the sensitivity. Our findings regarding the results achieved through 2D techniques closely align with those reported in previously published studies.[36–38] In addition, we have demonstrated that, analogous to the improved detection of plasma proteins, the application of 2D techniques similarly yields a considerable improvement in the number of VR peptides identified. We have also demonstrated that fluctuations in the counts of VR peptides were inversely correlated with the serum concentrations of the remaining M-protein. The association between a lower number of different VR peptides and higher M-protein concentrations and vice versa probably indicated that the remaining M-protein concentration suppresses the serum levels of other Igs, resulting in immunoparesis.[39] Overall, we have demonstrated that peptide fractionation, especially 2D-LC, improves the mapping of antibody VR. From a qualitative point of view, significantly more spectra can be acquired that lead to an alignment of a VR. Based on the serum concentration ranges of the proteins identified by 2D (median 40 μg/L) and calculations about the lowest (polyclonal) antibody concentration required for antigen elimination (~10 μg/L) suggested by Lavinder et al.,[3] we further conclude that targeted 2D LC-MS assays on antibodies are generally capable of covering a significant proportion of the concentration range of circulating antibodies. Due to the higher sensitivity in detecting VR peptides, the proteogenetic sequencing of circulating antibodies can potentially be enhanced and extended to



(legend on next page)

less abundant clones of polyclonal antibodies in neuromyelitis optica[40] and infection.[11]

Second, lacking a tool that combines the processes of sequencing tryptic peptides from MS/MS spectra and then categorizes these according to the VR, we developed a processing method for this purpose on the basis of PEAKS *de novo* search and BLAST alignment.[41–43] Both processing steps were evaluated individually, whereby the *de novo* search was assessed on the basis of a set of confident MS/MS spectra and BLAST alignment on the basis of a set of *in silico*-generated peptides derived from public antibody sequence repository data. In summary, the *de novo* search showed good accuracy in recovering database sequences for shorter peptides. However, for longer peptides, the increasing probability of missing or unrecognized fragments also increased the number of sequencing errors, resulting in a reduction in the number of peptide sequences with an exact (100%) sequence match. The approach of relying on exact sequence matches and the resulting dependency of the correctness to the peptide length closely align with the work of Muth and Renard,[40] where the resulting consequences were discussed in detail. By adopting a less stringent criterion of 90% sequence similarity, we could increase the number of longer *de novo* sequences. We considered a relative tolerant error margin of 10% to be appropriate as the aim of the analysis was to count the number of PSMs that could ultimately be classified according to their antibody VR. As a result, our findings also indicate that 100% sequence accuracy was observed in only a minority of cases. Consequently, the exact determination of antibody sequence regions based solely on tryptic peptides is generally not feasible. So far, this goal has already been achieved in the *de novo* sequencing of monoclonal antibodies. This was accomplished by using multiple enzyme digests to reconstruct longer, error-free sequence regions through the assembly of overlapping peptide sequence.[15,44,45] Subsequent BLAST searches against IMGT (international immunogenetics information system) germline sequences revealed that the recovery of shorter peptides was generally low. However, recovery rates in BLAST searches increased with longer peptide lengths and exceeded 97.5% for peptides longer than 10 aa. Fortunately, it is expected that the majority (three-quarters) of tryptic antibody VR peptides will be longer than 10 aa. Moreover, the distributions of Ig classes and chains, derived from hits on constant regions, conform to the expected pattern.[46] This further supports the validity of the data processing method used.

Third, driven by the question about the applicability and limitation of solely trypsin-based bottom-up proteomics on antibodies, we conducted an *in silico* analysis on datasets of about 20 million sequences from public repository.[28,29] Because of the high sequence homology among antibodies on one hand and the enormous sequence diversity of the CDRs on the other, the enzymatic generation of surrogate peptides in bottom-up experiments poses a challenge when it comes to unambiguously assigning them to the original antibody sequence. In the analysis of the common proteome, this is formally solved by reporting protein groups,[47,48] but this approach is not possible for the vast diversity in antibodies. The suitability of a tryptic peptide to serve as surrogate for an antibody primarily depends on the uniqueness of the peptide, meaning that its sequence is not shared with another antibody (or other protein). Whether a peptide is unique to an antibody or shared by multiple antibodies does not affect sensitivity per se, but it does determine the specificity of detection. The uniqueness of a peptide is particularly important in quantitative assays of antibodies, and different approaches have been described in literature. McDonald et al. defined uniqueness by the number of mutations and classified each peptide with at least one non-isobaric mutation as unique.[16] Noori and co-workers defined peptides as unique if targeted LC-MS measurements on a VR peptide in control serum samples of other individuals are negative.[14] This experimental approach is closely linked to the detection limit and sensitivity of the method and showed good results in monitoring monoclonal M-proteins in minimal residual disease.[14] In the present study, we developed another way of estimating the uniqueness of tryptic peptides, based on publicly available data from comprehensive antibody sequence repositories (OAS). This calculation provides an estimate of the expected presence of specific amino acid sequences but does not provide an assessment of the quantitative impact of ambiguities, if detected. Calculations based on 20 million sequences showed that about one-third of all antibody sequences have at least one unique tryptic peptide and that highly mutated antibody peptides can be shared between several individuals. However, we also found peptides carrying just one mutation that are already unique for one individual among a total of 100 individuals. These findings also make it clear that the question of uniqueness cannot be definitively answered in absolute terms. Instead, it must be considered in relation to the tested population and the aim of the analysis. Furthermore, we demonstrated that an estimated one-third of all antibody proteins contain at least one unique tryptic peptide and that a considerable proportion of antibodies that are unique within an individual's antibody repertoire are shared among individuals and, therefore, they are not unique in datasets representing populations. Next, we determined that the likelihood of a CDR3 peptide being unique (within a dataset of 2–20 million sequences) is three times higher compared to CDR1 or CDR2 peptides. However, it remains limited to approximately 30%. These calculations also confirm the initial assumption that the probability of serving as a suitable antibody

Figure 5. Results of *in silico* analysis on repository antibody sequences

(A) Distribution of peptide length in OAS repository datasets and acquired proteomics LC-MS datasets. The distribution of peptide length for tryptic peptides derived from OAS repository data is represented by blue bars (upper half of plot), while the distribution of peptide lengths from acquired LC-MS proteomics data is depicted by red bars (lower half of plot). Stacked bar plots show OAS subsets and LC-MS sets obtained using three different methods. The experimental acquired peptide counts are displayed upside down and scaled to account for overall differences (scaling factor: 808, derived from 175 million peptide sequences from the OAS repository and 216,000 experimentally determined peptide sequences.) The line chart at the top of the figure indicates the relative enrichment (above zero) or depletion (below zero) of tryptic peptides depending on their length, relative to the theoretical distribution computed based on OAS repository data.

(B) Frequency distribution of peptides based on sequence uniqueness. Peptides are grouped by corresponding VR and originating dataset (OPIG/OAS).

Table 2. Results of the VR peptide uniqueness analysis (OAS *in silico* analysis)

Counts and proportions	50 × 50,000	20 × 1 M	200 × 100,000
Number of antibody sequences	2.5 M	19.9 M	19.5 M
Mean number antibody sequences per individual	49,954	994,433	97,550
Number of individuals	50	20	200
Number of tryptic peptides (6–35 aa)	13.1 M	93.9 M	104.9 M
Mean number (range) of tryptic peptides/VR	5.2 (1–13)	4.7 (1–14)	5.4 (1–13)
Mean number of unique tryptic peptides/VR	1.11	0.58	0.89
Uniqueness within individual			
Number of unique tryptic peptides	2.8 M	11.4 M	17.3 M
% antibodies with at least one unique tryptic peptide	66%	44%	56%
Uniqueness within dataset			
% antibodies with at least one unique tryptic peptide	49%	35%	34%
Counts of VR with 0 unique tryptic peptide	1,264,672	12,865,647	12,825,274
Counts of VR with one unique tryptic peptide	892,802	5,735,811	5,222,153
Counts of VR with two unique tryptic peptide	232,919	966,647	1,067,864
Counts of VR with three unique tryptic peptide	75226	236632	293326
Counts of VR with four unique tryptic peptide	24,642	64,519	79,759
Counts of VR with five unique tryptic peptide	6,048	15,821	17,934
Counts of VR with six unique tryptic peptide	1,221	3,052	3,354
Counts of VR with seven unique tryptic peptide	156	481	450
Counts of VR with eight unique tryptic peptide	31	58	51
Counts of VR with nine unique tryptic peptide	8	2	–
Counts of VR with 10 unique tryptic peptide	–	1	1

Three sets of antibody VR sequences (50 × 50,000, 20 × 1 million, and 200 × 100,000) were subjected to *in silico* digested. The resulting peptides were analyzed in terms of their uniqueness and suitability as surrogates for intact antibodies in trypsin-based bottom-up proteomics. M, million.

surrogate is highest for CDR3 peptides, lower for CDR1 and CDR2 peptides, and lowest for peptides covering only the FWR. Based on these results, the question of whether the potential and probable lack of uniqueness of a tryptic peptide justifies the analytical use of such a peptide cannot be answered in an absolute or general way. Instead, such use ultimately requires experimental validation depending on the analytical question.

In summary, while the number detected of antibody VR peptides may appear dwarfed (at best 1,000 CDR peptides measured out of at least 1 million expected) by the expected number of antibodies present, we conclude that the use of 2D LC-MS is a powerful technique for increasing the quantity and sensitivity of detected circulating antibody. When combined with further enrichment and separation procedures (e.g., the purification of specific antibody subgroups), the inclusion of additional proteases, and the application of complementary mass spectrometric methods (e.g., top down or middle down), it has the potential to make a significant contribution to clinical research on antibody-mediated autoimmune disorders, infection, cancer, or vaccine development.

Limitations of the study

We investigated the capability of peptide fractionation to improve the detection of antibody VR peptides using bottom-up proteomics and, to this aim, we measured a set of eight samples with three methods. Each of the three methods utilized a different peptide fractionation technique, and one specific set of parameters was chosen for each method to ensure adequate comparability and applicability of the experiments. This has certainly led to the limitation that the full scope of the many possibilities to define such an experiment—e.g., LC phases, column dimensions, variety of MS analyzer settings, and data-dependent acquisition settings—could not be fully reflected in this study. As a consequence of the method-oriented objective and execution of this work, there are, as already described above, limitations with regard to the conclusions about the repertoire analyzed. The absence of comprehensive antibody sequence information of the samples (e.g., from sequencing data) did not allow a general assessment of the capabilities of mapping the circulating antibody repertoire and the use of only one specific protease in principle did not allow the reconstruction of longer or entire antibody sequence segments, as no overlapping fragments are generated. These major limitations define the main objectives for follow-up research.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - ○ Sample collection
- METHOD DETAILS
  - ○ Sample preparation
  - ○ High pH reversed phase fractionation
  - ○ LC-MS measurements
  - ○ Mass spectrometry data analysis
  - ○ In silico analysis on repository antibody sequences
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2024.100795.

ACKNOWLEDGMENTS

The IFM 2009 trial was funded by the French Ministry of Health Programme Hospitalier de Recherche Clinique and from the French National Research Agency (project ANR-11-PHUC-001-CAPTOR). We thank the OAS authors and contributors for the opportunity to use their antibody sequence data. We thank Ko Hagoort for critical reading of the manuscript.

AUTHOR CONTRIBUTIONS

C.S., conceptualization, methodology, software, investigation, resources, data curation, writing – original draft, writing – review & editing, and visualization; M.M.V., resources and writing – review & editing; T.D., resources and writing – review & editing; P.A.E.S.S., writing – review & editing and supervision; T.M.L., conceptualization, methodology, investigation, resources, writing – review & editing, and supervision.

DECLARATION OF INTERESTS

Sebia has financially supported M.M.V.

Received: November 3, 2023
Revised: March 30, 2024
Accepted: May 20, 2024
Published: June 10, 2024

REFERENCES

1. Rees, A.R. (2020). Understanding the human antibody repertoire. mAbs *12*, 1729683. https://doi.org/10.1080/19420862.2020.1729683.
2. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). Molecular biology of the cell (4th ed.). Biochem. Mol. Biol. Educ. *31*, 212–214. https://doi.org/10.1002/bmb.2003.494031049999.
3. Lavinder, J.J., Horton, A.P., Georgiou, G., and Ippolito, G.C. (2015). Next-generation sequencing and protein mass spectrometry for the comprehensive analysis of human cellular and serum antibody repertoires. Curr. Opin. Chem. Biol. *24*, 112–120. https://doi.org/10.1016/j.cbpa.2014.11.007.
4. Schulte, D., Peng, W., and Snijder, J. (2022). Template-Based Assembly of Proteomic Short Reads For De Novo Antibody Sequencing and Repertoire Profiling. Anal. Chem. *94*, 10391–10399. https://doi.org/10.1021/acs.analchem.2c01300.
5. Georgiou, G., Ippolito, G.C., Beausang, J., Busse, C.E., Wardemann, H., and Quake, S.R. (2014). The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat. Biotechnol. *32*, 158–168. https://doi.org/10.1038/nbt.2782.
6. Liu, H., Pan, W., Tang, C., Tang, Y., Wu, H., Yoshimura, A., Deng, Y., He, N., and Li, S. (2021). The methods and advances of adaptive immune receptors repertoire sequencing. Theranostics *11*, 8945–8963. https://doi.org/10.7150/thno.61390.
7. 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., et al. (2013). High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nat. Biotechnol. *31*, 166–169. https://doi.org/10.1038/nbt.2492.
8. de Graaf, S.C., Hoek, M., Tamara, S., and Heck, A.J.R. (2022). A perspective toward mass spectrometry-based de novo sequencing of endogenous antibodies. mAbs *14*, 2079449. https://doi.org/10.1080/19420862.2022.2079449.
9. 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., et al. (2021). Human plasma IgG1 repertoires are simple, unique, and dynamic. Cell Syst. *12*, 1131–1143.e5. https://doi.org/10.1016/j.cels.2021.08.008.
10. 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., et al. (2012). A proteomics approach for the identification and cloning of monoclonal antibodies from serum. Nat. Biotechnol. *30*, 447–452. https://doi.org/10.1038/nbt.2167.
11. 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., et al. (2019). Sera Antibody Repertoire Analyses Reveal Mechanisms of Broad and Pandemic Strain Neutralizing Responses after Human Norovirus Vaccination. Immunity *50*, 1530–1541.e8. https://doi.org/10.1016/j.immuni.2019.05.007.
12. Barnidge, D.R., Dasari, S., Botz, C.M., Murray, D.H., Snyder, M.R., Katzmann, J.A., Dispenzieri, A., and Murray, D.L. (2014). Using mass spectrometry to monitor monoclonal immunoglobulins in patients with a monoclonal gammopathy. J. Proteome Res. *13*, 1419–1427. https://doi.org/10.1021/pr400985k.
13. Zajec, M., Jacobs, J.F.M., de Kat Angelino, C.M., Dekker, L.J.M., Stingl, C., Luider, T.M., De Rijke, Y.B., and VanDuijn, M.M. (2020). Integrating Serum Protein Electrophoresis with Mass Spectrometry, A New Workflow for M-Protein Detection and Quantification. J. Proteome Res. *19*, 2845–2853. https://doi.org/10.1021/acs.jproteome.9b00705.

14. Noori, S., Wijnands, C., Langerhorst, P., Bonifay, V., Stingl, C., Touzeau, C., Corre, J., Perrot, A., Moreau, P., Caillon, H., et al. (2023). Dynamic monitoring of myeloma minimal residual disease with targeted mass spectrometry. Blood Cancer J. *13*, 30. https://doi.org/10.1038/s41408-023-00803-z.

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

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

17. 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., et al. (2023). Direct Mass Spectrometry-Based Detection and Antibody Sequencing of Monoclonal Gammopathy of Undetermined Significance from Patient Serum: A Case Study. J. Proteome Res. *22*, 3022–3028. https://doi.org/10.1021/acs.jproteome.3c00330.

18. Tu, C., Rudnick, P.A., Martinez, M.Y., Cheek, K.L., Stein, S.E., Slebos, R.J.C., and Liebler, D.C. (2010). Depletion of abundant plasma proteins and limitations of plasma proteomics. J. Proteome Res. *9*, 4982–4991. https://doi.org/10.1021/pr100646w.

19. Lee, J., Paparoditis, P., Horton, A.P., Frühwirth, A., McDaniel, J.R., Jung, J., Boutz, D.R., Hussein, D.A., Tanno, Y., Pappas, L., et al. (2019). Persistent Antibody Clonotypes Dominate the Serum Response to Influenza over Multiple Years and Repeated Vaccinations. Cell Host Microbe *25*, 367–376.e5. https://doi.org/10.1016/j.chom.2019.01.010.

20. Anderson, N.L., Anderson, N.G., Haines, L.R., Hardie, D.B., Olafson, R.W., and Pearson, T.W. (2004). Mass spectrometric quantitation of peptides and proteins using Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA). J. Proteome Res. *3*, 235–244. https://doi.org/10.1021/pr034086h.

21. Hoffman, S.A., Joo, W.-A., Echan, L.A., and Speicher, D.W. (2007). Higher dimensional (Hi-D) separation strategies dramatically improve the potential for cancer biomarker detection in serum and plasma. J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. *849*, 43–52. https://doi.org/10.1016/j.jchromb.2006.10.069.

22. Qian, W.-J., Kaleta, D.T., Petritis, B.O., Jiang, H., Liu, T., Zhang, X., Mottaz, H.M., Varnum, S.M., Camp, D.G., Huang, L., et al. (2008). Enhanced detection of low abundance human plasma proteins using a tandem IgY12-SuperMix immunoaffinity separation strategy. Mol. Cell. Proteomics *7*, 1963–1973. https://doi.org/10.1074/mcp.M800008-MCP200.

23. Zhu, W., Smith, J.W., and Huang, C.-M. (2010). Mass spectrometry-based label-free quantitative proteomics. J. Biomed. Biotechnol. *2010*, 840518. https://doi.org/10.1155/2010/840518.

24. Klaassen, T., Szwandt, S., Kapron, J.T., and Roemer, A. (2009). Validated quantitation method for a peptide in rat serum using liquid chromatography/high-field asymmetric waveform ion mobility spectrometry. Rapid Commun. Mass Spectrom. *23*, 2301–2306. https://doi.org/10.1002/rcm.4147.

25. Cooper, H.J. (2016). To What Extent is FAIMS Beneficial in the Analysis of Proteins? J. Am. Soc. Mass Spectrom. *27*, 566–577. https://doi.org/10.1007/s13361-015-1326-4.

26. Hebert, A.S., Prasad, S., Belford, M.W., Bailey, D.J., McAlister, G.C., Abbatiello, S.E., Huguet, R., Wouters, E.R., Dunyach, J.-J., Brademan, D.R., et al. (2018). Comprehensive Single-Shot Proteomics with FAIMS on a Hybrid Orbitrap Mass Spectrometer. Anal. Chem. *90*, 9529–9537. https://doi.org/10.1021/acs.analchem.8b02233.

27. Sweet, S., Chain, D., Yu, W., Martin, P., Rebelatto, M., Chambers, A., Cecchi, F., and Kim, Y.J. (2022). The addition of FAIMS increases targeted proteomics sensitivity from FFPE tumor biopsies. Sci. Rep. *12*, 13876. https://doi.org/10.1038/s41598-022-16358-1.

28. Kovaltsuk, A., Leem, J., Kelm, S., Snowden, J., Deane, C.M., and Krawczyk, K. (2018). Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. J. Immunol. *201*, 2502–2509. https://doi.org/10.4049/jimmunol.1800708.

29. Olsen, T.H., Boyles, F., and Deane, C.M. (2022). Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. Protein Sci. *31*, 141–146. https://doi.org/10.1002/pro.4205.

30. Muthusamy, B., Hanumanthu, G., Suresh, S., Rekha, B., Srinivas, D., Karthick, L., Vrushabendra, B.M., Sharma, S., Mishra, G., Chatterjee, P., et al. (2005). Plasma Proteome Database as a resource for proteomics research. Proteomics *5*, 3531–3536. https://doi.org/10.1002/pmic.200401335.

31. Nanjappa, V., Thomas, J.K., Marimuthu, A., Muthusamy, B., Radhakrishnan, A., Sharma, R., Ahmad Khan, A., Balakrishnan, L., Sahasrabuddhe, N.A., Kumar, S., et al. (2014). Plasma Proteome Database as a resource for proteomics research: 2014 update. Nucleic Acids Res. *42*, D959–D965. https://doi.org/10.1093/nar/gkt1251.

32. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

33. Schäffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V., and Altschul, S.F. (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. *29*, 2994–3005. https://doi.org/10.1093/nar/29.14.2994.

34. Yu, Y.-K., and Altschul, S.F. (2005). The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. Bioinformatics *21*, 902–911. https://doi.org/10.1093/bioinformatics/bti070.

35. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinf. *10*, 421. https://doi.org/10.1186/1471-2105-10-421.

36. Delmotte, N., Lasaosa, M., Tholey, A., Heinzle, E., and Huber, C.G. (2007). Two-dimensional reversed-phase x ion-pair reversed-phase HPLC: an alternative approach to high-resolution peptide separation for shotgun proteome analysis. J. Proteome Res. *6*, 4363–4373. https://doi.org/10.1021/pr070424t.

37. Cao, Z., Tang, H.-Y., Wang, H., Liu, Q., and Speicher, D.W. (2012). Systematic comparison of fractionation methods for in-depth analysis of plasma proteomes. J. Proteome Res. *11*, 3090–3100. https://doi.org/10.1021/pr201068b.

38. Kulak, N.A., Geyer, P.E., and Mann, M. (2017). Loss-less Nano-fractionator for High Sensitivity, High Coverage Proteomics. Mol. Cell. Proteomics *16*, 694–705. https://doi.org/10.1074/mcp.O116.065136.

39. Wang, L., and Young, D.C. (2001). Suppression of polyclonal immunoglobulin production by M-proteins shows isotype specificity. Ann. Clin. Lab. Sci. *31*, 274–278.

40. Muth, T., and Renard, B.Y. (2018). Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? Brief. Bioinform *19*, 954–970. https://doi.org/10.1093/bib/bbx033.

41. Muth, T., Hartkopf, F., Vaudel, M., and Renard, B.Y. (2018). A Potential Golden Age to Come-Current Tools, Recent Use Cases, and Future Avenues for De Novo Sequencing in Proteomics. Proteomics *18*, e1700150. https://doi.org/10.1002/pmic.201700150.

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

43. McGinnis, S., and Madden, T.L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. *32*, W20–W25. https://doi.org/10.1093/nar/gkh435.

44. Tran, N.H., Rahman, M.Z., He, L., Xin, L., Shan, B., and Li, M. (2016). Complete De Novo Assembly of Monoclonal Antibody Sequences. Sci. Rep. *6*, 31730. https://doi.org/10.1038/srep31730.

45. Guthals, A., Gan, Y., Murray, L., Chen, Y., Stinson, J., Nakamura, G., Lill, J.R., Sandoval, W., and Bandeira, N. (2017). De Novo MS/MS Sequencing of Native Human Antibodies. J. Proteome Res. *16*, 45–54. https://doi.org/10.1021/acs.jproteome.6b00608.

46. Schroeder, H.W., and Cavacini, L. (2010). Structure and function of immunoglobulins. J. Allergy Clin. Immunol. *125*, S41–S52. https://doi.org/10.1016/j.jaci.2009.09.046.

47. Nesvizhskii, A.I., Keller, A., Kolker, E., and Aebersold, R. (2003). A statistical model for identifying proteins by tandem mass spectrometry. Anal. Chem. *75*, 4646–4658.

48. Plubell, D.L., Käll, L., Webb-Robertson, B.-J., Bramer, L.M., Ives, A., Kelleher, N.L., Smith, L.M., Montine, T.J., Wu, C.C., and MacCoss, M.J. (2022). Putting Humpty Dumpty Back Together Again: What Does Protein Quantification Mean in Bottom-Up Proteomics? J. Proteome Res. *21*, 891–898. https://doi.org/10.1021/acs.jproteome.1c00894.

49. Attal, M., Lauwers-Cances, V., Hulin, C., Leleu, X., Caillot, D., Escoffre, M., Arnulf, B., Macro, M., Belhadj, K., Garderet, L., et al. (2017). Lenalidomide, Bortezomib, and Dexamethasone with Transplantation for Myeloma. N. Engl. J. Med. *376*, 1311–1320. https://doi.org/10.1056/NEJMoa1611750.

50. Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D.J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., et al. (2019). The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. *47*, D442–D450. https://doi.org/10.1093/nar/gky1106.

51. Ye, J., Ma, N., Madden, T.L., and Ostell, J.M. (2013). IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res *41*, W34–W40. https://doi.org/10.1093/nar/gkt382.

52. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the Tidyverse. J. Open Source Softw. *4*, 1686. https://doi.org/10.21105/joss.01686.

53. Langerhorst, P., Noori, S., Zajec, M., De Rijke, Y.B., Gloerich, J., van Gool, A.J., Caillon, H., Joosten, I., Luider, T.M., Corre, J., et al. (2021). Multiple Myeloma Minimal Residual Disease Detection: Targeted Mass Spectrometry in Blood vs Next-Generation Sequencing in Bone Marrow. Clin. Chem. *67*, 1689–1698. https://doi.org/10.1093/clinchem/hvab187.

54. R Core Team (2020). R: A Language and Environment for Statistical Computing.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Serum sample from MM patients (*N* = 8)	Attal et al.[49]	IFM/DFCI2009, https://clin.larvol.com/trial-detail/NCT01191060
Chemicals, peptides, and recombinant proteins		
1,4-dithiothreitol	Merck/Sigma Aldrich	PN D9779-1G
Acetonitrile (ACN)	Biosolve BV	PN 1204102
Ammonia solution	Merck/Sigma Aldrich	PN 5.33003.0050
Ammonium formate	Merck/Sigma Aldrich	PN 78314-100mL-F
Formic acid	Biosolve BV	PN 06914143
S-methyl methanethiosulfonate	Merck/Sigma Aldrich	PN 64306-1ML
Sodium deoxycholate	Merck/Sigma Aldrich	PN 30970-25G
Triethylammonium bicarbonate (TEAB)	Merck/Sigma Aldrich	PN T7408-100ML
Trifluoroacetic acid, TFA	Biosolve BV	PN 20234131
Trypsin	Promega	PN V5280
Water, ULC/MS grade	Biosolve BV	PN 23214102
Deposited data		
Mass spectrometry data	ProteomeXchange/PRIDE[50]	10.6019/PXD046072
Software and algorithms		
PEAKS X	Bioinformatics Solutions Inc., Waterloo, Canada	
NCBI BLAST	McGinnis et al.[43]	https://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/
NCBI IgBlast	Ye et al., 2013[51]	https://ftp.ncbi.nih.gov/blast/executables/igblast/release/LATEST
Perl, version v5.32.1	Perl.org	https://www.perl.org/
R, version 4.2.1	R Software Foundation	https://www.r-project.org
R package tidyverse	Wickham et al.[52]	https://www.tidyverse.org/
Antibody sequence repository	Observed Antibody Space[28,29]	https://opig.stats.ox.ac.uk/webapps/oas/
vreg-anno (mass spectrometry analysis)	this paper	https://github.com/cstingl/vrpep-anno https://doi.org/10.5281/zenodo.11203821
vreg-uniq (*in silico* analysis)	this paper	https://github.com/cstingl/vrpep-uniq https://doi.org/10.5281/zenodo.11203918
Other		
96-wellplates	Axygen, Corning, NY	Axygen X50 500μL V96 PP, PN 12507927
Kinetex EVO C18 column	Phenomenex	PN 00F-4725-AN
PepMap nano LC column	Thermo Fisher Scientific	PN 164941
PepMap trap column	Thermo Fisher Scientific	PN 160454

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Christoph Stingl (c.stingl@erasmusmc.nl).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[50] partner repository with the dataset identifier PXD046072 and 10.6019/PXD046072.
- The code generated during this study to process the mass spectrometry can be accessed on https://github.com/cstingl/vrpep-anno. The code used to conduct the *in silico* data analysis of antibody sequences is available on https://github.com/cstingl/vrpep-anno. Archival DOIs are listed in the key resources table.
- Any additional information necessary to re-analyze the data reported in this paper is available via the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Sample collection

Serum samples were collected in the IFM 2009 study (clinical trial ID NCT01191060) under written informed consent.[49] Information about gender and age was not disclosed for this study. We used eight serum samples from four multiple myeloma (MM) patients at two different time points during treatment and remission. These blood samples were selected as they contained, characteristically for MM, a patient-specific antibody (the M-protein) with a known sequence and concentrations in an otherwise polyclonal antibody background. The M-protein sequences used as reference antibodies were constructed from mRNA sequencing data in a previous study, as described by Langerhorst and co-workers.[53] Information about the samples, including antibody germline genes, sampling days, and serum concentrations, is provided in Table S7.

METHOD DETAILS

Sample preparation

Sample preparation was carried out in 96-wellplates (Axygen, Corning, NY) using a 12-channel multichannel pipet (Mettler-Toledo Rainin, Oakland, CA). During all reaction and incubation steps at an elevated temperature, the wellplate was closed with heat-seals (3 s at 160°C). Two μL of each serum sample were dissolved in 80 μL digest buffer, which consisted of 0.5% sodium deoxycholate, 50 mM triethylammonium bicarbonate, and 5% acetonitrile (ACN). Stable isotope-labelled peptides specific to the patient's M-protein were spiked at concentrations corresponding to 5 pmol/μL serum. The samples were reduced in the presence of 5 mM 1,4-dithiothreitol (20 μL of 25 mM solution) for 30 min at 56°C, alkylated with 10 mM S-methyl methanethiosulfonate (20 μL of 60 mM solution) for 15 min at room temperature, and subsequently digested by adding 4 μg trypsin (20 μL of 200 ng/μL solution) and incubating at 37°C over-night. The next day, the digests were stopped, and detergent was precipitated by adding 0.5% trifluoroacetic acid (TFA, 20 μL of 4% solution). The precipitate was then spun down (10 min, 4,400 rpm), and the supernatant was filtered through a 0.25 μm membrane to remove any remaining SDC precipitate. Unless specified otherwise, all reagents were purchased from Sigma-Aldrich/Merck. Next, the samples were twice diluted with an oxidative solution containing 0.5% TFA and 0.5% hydrogen peroxide to completely oxidize methionine, which is typically only partially oxidized. Half of the volume (~120 μL) was directly employed for subsequent high pH reversed-phase fractionation, while the second half was further diluted 10-fold in 0.1% TFA for 1D and FAIMS LC-runs. Figure 1A shows an overview of the prepared samples, their quantities, and their utilization in the various measurements and methods.

High pH reversed phase fractionation

Preparative chromatography was conducted on an Ultimate 3000 LC system (Thermo Fisher Scientific) equipped with C18 reversed phase column (Kinetex EVO, 2.1 mm × 150 mm, PN 00F-4725-AN, Phenomenex) operated at an oven temperature of 40°C. Peptides were separated using a binary gradient, increasing from 4% to 50% solvent B over 8 min at a flow rate of 450 μL/min. Solvent A consisted of a 10 mM ammonium formate buffer at pH 10, while solvent B was composed of 80% ACN and 10 mM ammonium formate buffer at pH 10. Twenty-four fractions, each comprising 200 μL and collected over a period of 26 s, were collected in a 96 wellplate (PN P-96-450V-C, Axygen/Thermo Fisher Scientific). These fractions were then dried using a speedvac concentrator (Thermo Savant), resuspended in a solution of 2% ACN/0.1% TFA, split into two aliquots, and transferred to a heat-sealed 384-wellplatefor storage at 4°C until subsequent LC-MS analysis.

LC-MS measurements

LC-MS measurements were conducted on a nano-LC system (Ultimate 3000 RSLC, Thermo Fisher Scientific, Germering, Germany) coupled to an Orbitrap Lumos Tribrid mass spectrometer. For FAIMS measurements, the instrument was equipped with a High Field Asymmetric Waveform Ion Mobility Spectrometry (FAIMS) interface (Thermo Fisher Scientific, San Jose, CA). Ten μL of the sample (for 1D and FAIMS) or fraction (for 2D) were injected and transferred on a trap column (C18 PepMap, 300 μm ID x 5 mm; Thermo Fisher Scientific) using 0.1% TFA at a flow rate of 20 μL/min and further eluted and separated an 25 cm analytical nano-LC column (PepMap C18, 75 μm ID x 250 mm, 2 μm, 100 Å; Thermo Fisher Scientific) using a binary gradient. For 1D and FAIMS measurements, the gradient ranged over 90 min from 4% to 34% solvent B, and for 2D runs over 30 min (for 2D) from 3% to 30% solvent B. Solvent A consisted of 0.1% formic acid, and solvent B contained ACN with 0.08% formic acid. The flow rate was set at 300 nL/min, and the column was operated at a temperature of 40°C. For electrospray ionization we used coated silica nano electro-spray emitters

(New Objective, Woburn, MA) at a spray voltage of 1.8 kV (without FAIMS) or 2.2 kV (FAIMS). For FAIMS measurements, ion mobility fractions were collected at compensation voltages (CV) of −35, −50, −65 and −80 V. A data dependent acquisition MS method was used, with an Orbitrap survey scan (range 375–1550 m/z, resolution of 60,000, AGC target 400,000). The peptide precursors detected in the survey scan were then subsequently steps isolated (using a window of 1.6 amu), fragmented (HCD with 30% normalized collision energy), and detected (Orbitrap, with a resolution of 30,000, maximal injection time of 54 ms and an AGC target of 100,000). This process continued until a cycle time of 3 s (for experiments without FAIMS) or 3.2 s (with FAIMS) was reached. Precursors with single charge and precursors masses that had been selected once for MS/MS were excluded from subsequent fragmentation for 60 s. For 1D and FAIMS experiments, a total of 8 runs each (90-min gradient and about 120 min runtime) were conducted. In the case of 2D experiments, a total of 196 runs (30-min gradients and 60 min runtime) were carried out.

Mass spectrometry data analysis

The acquired RAW data were processed using the software package PEAKS X (Bioinformatics Solutions Inc., Waterloo, Canada). The following parameters were employed for the extraction of MS/MS spectra, *de novo* search, and database searching. For the *de novo* search: fixed modifications included beta-methylthiolation (+45.99 u) of cysteine and oxidation (+15.99 u) of methionine. Variable modifications encompassed deamidation (+0.984 u) of asparagine and glutamine. A parent mass tolerance of 10 ppm and a fragment mass tolerance of 0.02 amu were applied. Semi-tryptic cleavage was utilized. For subsequent PEAKS database searches: the human subset of the Uniprot/Swissprot database (downloaded October 5th, 2021) was applied, optionally extended by the M-protein sequences for a second search. Results from the *de novo* searches were exported as 'peptide.csv' files with a peptide-spectra-match false discovery rate of 1%. The peptide sequences were then subjected to a BLAST search against the IMGT germline database (downloaded June 14th, 2021). No distinction was made between the isobaric amino acid leucine and isoleucine, with the germline databank and *de novo* sequences adjusted accordingly. Furthermore, for all *de novo* sequences containing deamidated asparagine and glutamine, an additional sequence with the corresponding isobaric amino acids aspartic acid and glutamic acid, respectively, was included in the query sequences (input) of the BLAST search. Each *de novo* sequence that aligned with an IMGT germline VR gene was annotated using the corresponding germline amino acid positions in the CDRs and FWRs frameworks, provided there was an overlap of at least 3 amino acids. Data processing steps were conducted using an in-house written Perl script (version 5.32.1), and the generated data files have been included in the supplementary data. For further statistical analysis and plotting we used the statistical programming language R (version 4.2.1) and *Tidyverse* libraries.[52,54]

In silico analysis on repository antibody sequences

Antibody sequences for *in silico* computations were retrieved from the public Observed Antibody Space (OAS).[28,29] The selected OAS data units for download were defined as unsorted PMBC (B cell source and type) from non-diseased human individuals (species) derived from bulk sequencing (isotype). Each OAS data unit corresponded to an individual and contained un-paired sequences from either exclusively light or heavy chains. Based on the download of 200 OAS data units, we created three datasets with varying sequencing depths (numbers of sequences per individual) and varying numbers of individuals. In each dataset, half of the individuals (data units) exclusively contributed heavy chains, while the other half solely contributed light chains. These three datasets, labelled as 50 × 50k, 20 × 1M, and 200 × 100k, were compiled using 50, 20, and 200 OAS data units (individuals), respectively. Each dataset was sampled to include up to 50,000 (50k), 1 million (1M), and 100,000 (100k) sequences, respectively. Details about the datasets, including the OAS dataset identifier, the total number of sequences, and the actual number of sequences used, can be found in Table S6. The entire VR, FWRs and CDRs were extracted (excluding entries with incomplete CDR1-3 and FWR1-3 amino acid sequences). The VR sequences were computationally cleaved into peptides using the specificity of trypsin (cleaving C-terminal to Arg and Lys, except at the N-terminal position of Pro), without allowing for miscleavages. Next, for each peptide, annotations were made regarding its overlap with CDRs and FWRs, and the number of mutations was calculated by comparing the amino acids in the peptide sequence to the corresponding germline sequences. For each peptide sequence longer than 6 and shorter than 35 aa, the number of assigned individuals and the occurrences within the dataset were computed and used to query and calculate the length distributions (Figure 5A) and uniqueness (Figure 5B) of peptides based on their originating CDRs and FWRs. Additionally, the data was used to establish the correlation between the number of mutations and the uniqueness of peptide sequences with the dataset. (Figure S7) For data processing, analysis and plotting we used the programming language Perl (version 5.32.1), relational database server PostgreSQL (version 14), and the statistical programming language R[54] (version 4.2.1), which included the *Tidyverse* libraries.[52]

QUANTIFICATION AND STATISTICAL ANALYSIS

For statistical analyses, calculations and plotting of data we used the statistical programming language R[54] (version 4.2.1) with the *Tidyverse* libraries.[52] All analysis were carried out on basis of a sample set of eight samples (one measurement each) for each of the three methods. A two-sided t-test (R core function *Welch Two Sample t-test*) was used to calculate the significance of quantitative differences in terms of PSM between various fractionation methods (Table 1; Figures 2 and 4). A one-sample t-test was applied to determine fractionation method-dependent gains in PSMs in Table 1 and sequence coverage in Table 2. For all tests, a *p*p-value equal to or below 0.05 indicated a significant difference.