










# De novo protein sequencing of antibodies for identification of neutralizing antibodies in human plasma post SARS-CoV-2 vaccination

Received: 13 March 2024

Accepted: 2 October 2024

Published online: 10 October 2024

 Check for updates

Thierry Le Bihan , Teresa Nunez de Villavicencio Diaz , Chelsea Reitzel, Victoria Lange , Minyoung Park , Emma Beadle, Lin Wu , Marko Jovic , Rosalin M. Dubois, Amber L. Couzens , Jin Duan, Xiaobing Han, Qixin Liu & Bin Ma  

The antibody response to vaccination and infection is a key component of the immune response to pathogens. Sequencing of peripheral B cells may not represent the complete B cell receptor repertoire. Here we present a method for sequencing human plasma-derived polyclonal IgG using a combination of mass spectrometry and B-cell sequencing. We investigate the IgG response to the Moderna Spikevax COVID-19 vaccine. From the sequencing data of the natural polyclonal response to vaccination, we generate 12 recombinant antibodies. Six derived recombinant antibodies, including four generated with de novo protein sequencing, exhibit similar or higher binding affinities than the original natural polyclonal antibody. Neutralization tests reveal that the six antibodies possess neutralizing capabilities against the target antigen. This research provides insights into sequencing polyclonal IgG antibodies and the potential of our approach in generating recombinant antibodies with robust binding affinity and neutralization capabilities. Directly examining the circulating IgG pool is crucial due to potential misrepresentations by B-cell analysis alone.

Antibodies are highly selective molecules, making them valuable for biotherapeutics and diagnostic assays. Therapeutic antibodies treat cancer, infections, and autoimmune diseases<sup>1</sup>. Polyclonal antibodies (pAbs) are usually produced from immunized animals, which leads to batch-to-batch variability<sup>2</sup>. Monoclonal antibodies (mAbs) address this issue and are traditionally produced by hybridoma<sup>3</sup>. Recently, mAbs have been developed via phage display<sup>4,5</sup>. There is a growing interest in discovering antibodies by interrogating the natural immune response of animals or humans using ex vivo single B cell sequencing<sup>4,6</sup> alongside integrating B cell sequencing and proteomics<sup>7–10</sup>. This manuscript explores de novo protein sequencing of mAbs directly from a pAb mixture as a complementary approach to existing technological limitations.

Phage display technology uses bacteriophages to incorporate genetic sequences encoding antibodies. It enables the display of

antibody variants on the phage surface, facilitating antibody discovery and affinity maturation. The application of phage display has significantly influenced antibody engineering, particularly following the approval of adalimumab. In vitro display technologies offer broad immune libraries for antibody production, providing flexibility in heavy and light chain pairing that can pose challenges in affinity optimization. This inherent flexibility also allows diverse pairing combinations that could surpass the limited scope of natural pairings. While these pairing possibilities may present challenges, they have also improved affinity and stability in certain cases<sup>11</sup>. Around 18% of FDA-licensed monoclonal antibodies have been developed through phage display<sup>12</sup>.

More therapeutic antibodies were found in the B cell repertoire of transgenic animals and human patients. 19 out of the 28 FDA-approved human monoclonal antibodies from 2002 to 2018 were derived from

“humanized” transgenic mice<sup>1</sup>. However, these mouse-derived human antibodies are not fully non-immunogenic in humans<sup>13</sup>. Antibodies generated through natural human immune responses undergo tolerization, making them safer and more effective than those from other species<sup>4</sup>. The SARS-CoV-2 pandemic has led to the isolation of B cells from convalescent patients, resulting in antibodies like Tixagevimab-Cilgavimab, Bamlanivimab, and Bebtelovimab. Furthermore, Ofatumumab, Daratumumab, and Ustekinumab target CD20, CD38, and IL-12/23, respectively<sup>14,15</sup>. Additionally, VRC01, a neutralizing HIV antibody, was isolated from HIV-positive B cells<sup>16</sup>.

While B cell sequencing is effective for antibody discovery, peripheral B cells may not capture the full diversity of antibodies. Most Long-Lived Plasma B cells (LLPCs) reside in the bone marrow and lymphoid organs, producing high-affinity antibodies for an extended period. Obtaining LLPCs from tissues is challenging and peripheral blood, containing a small fraction of B cells, provides limited information. Georgiou’s studies highlight the importance of alternative sources of B cells for antibody sequencing to understand the humoral response to antigens<sup>10</sup>. Studies estimate that 2%<sup>17</sup> to none<sup>18</sup> match the circulating IgG. Not all circulating B cells produce antibodies<sup>19</sup>. Overlap between circulating IgG and peripheral blood B cells varies based on factors like antigen, immune status, or timing. Relying solely on blood-circulating B cells for development candidates can be challenging. Ultimately, immune protection depends on serum antibodies, not B cell receptors.

IgGs are highly abundant in blood circulation. Affinity purification against a specific antigen can be performed on a solid substrate, resulting in high-affinity polyclonal antibodies (pAbs). This reduces the pAbs’ complexity before more effective mass spectrometry (MS) analysis. Matching the MS data with the sequence database produced by B cell repertoire sequencing helps discover antigen-specific antibodies. This proteogenomics approach generates monoclonal antibodies from animals and humans<sup>7–9</sup> and improves on B cell methods by identifying serum antibodies. However, it cannot detect circulating antibodies absent from the B cell sequencing database, which may be significant. This manuscript demonstrates the existence of such antibodies.

One solution to the problem is performing de novo sequencing of antibodies (i.e. no reference database). Various software tools have been developed for de novo peptide sequencing from a peptide’s tandem mass (MS/MS) spectrum<sup>20–24</sup>. A full-length protein can be sequenced by digesting the protein into overlapping peptides, de novo sequencing each peptide with a preferred tool, and assembling the sequences to get the protein sequence. This approach for monoclonal antibodies has been well-researched<sup>25–27</sup> and is now available as a commercial service. However, de novo protein sequencing may have limitations in coverage depth, due to the need for high-intensity peptide fragments in MS/MS mode. Peptide detection accuracy can also vary based on factors like peptide length and hydrophobicity. Additionally, if the sample contains multiple similar sequences, the overlaps between peptides may be insufficient for the sequence assembly.

De novo protein sequencing studies of polyclonal antibodies have been limited by unresolved challenges. Previous efforts have failed to discover functional antibodies with the same affinity as the original pAb. Guthals et al.<sup>18</sup> attempted pAb de novo protein sequencing using a procedure similar to mAb sequencing, but the resulting antibodies had a lower affinity, possibly indicating incorrect sequences. Bondt et al.<sup>28</sup> used a similar approach, including middle-down MS, but did not show data on the affinity or neutralization effectiveness of the resulting antibodies. These studies highlighted the difficulty of pAb de novo sequencing. Our experience shows that accurately pairing CDRs and heavy/light chains is a major challenge. The numerous mAb clones in a pAb create many combinations that earlier methods cannot handle.

This study sequences a COVID-19-vaccinated patient’s polyclonal antibodies using multiple methods. We use de novo protein sequencing by combining mass spectrometry data from various experiments to address challenges, including bottom-up proteomics for local sequences, peptide chemistry for increasing sequence coverage and distinguishing Ile/Leu, and middle-down proteomics for longer sequences. We also use protein separation and label-free quantitation under non-reductive conditions to assemble full-length chains confidently and propose a heavy-light chain pairing strategy. Additionally, B cell repertoire analysis evaluates overlap with the proteomics dataset. Finally, we express and confirm the antibody sequences’ binding affinity and their neutralization capability using Fluc-GFP pseudovirus.

## Results

### Platform overview

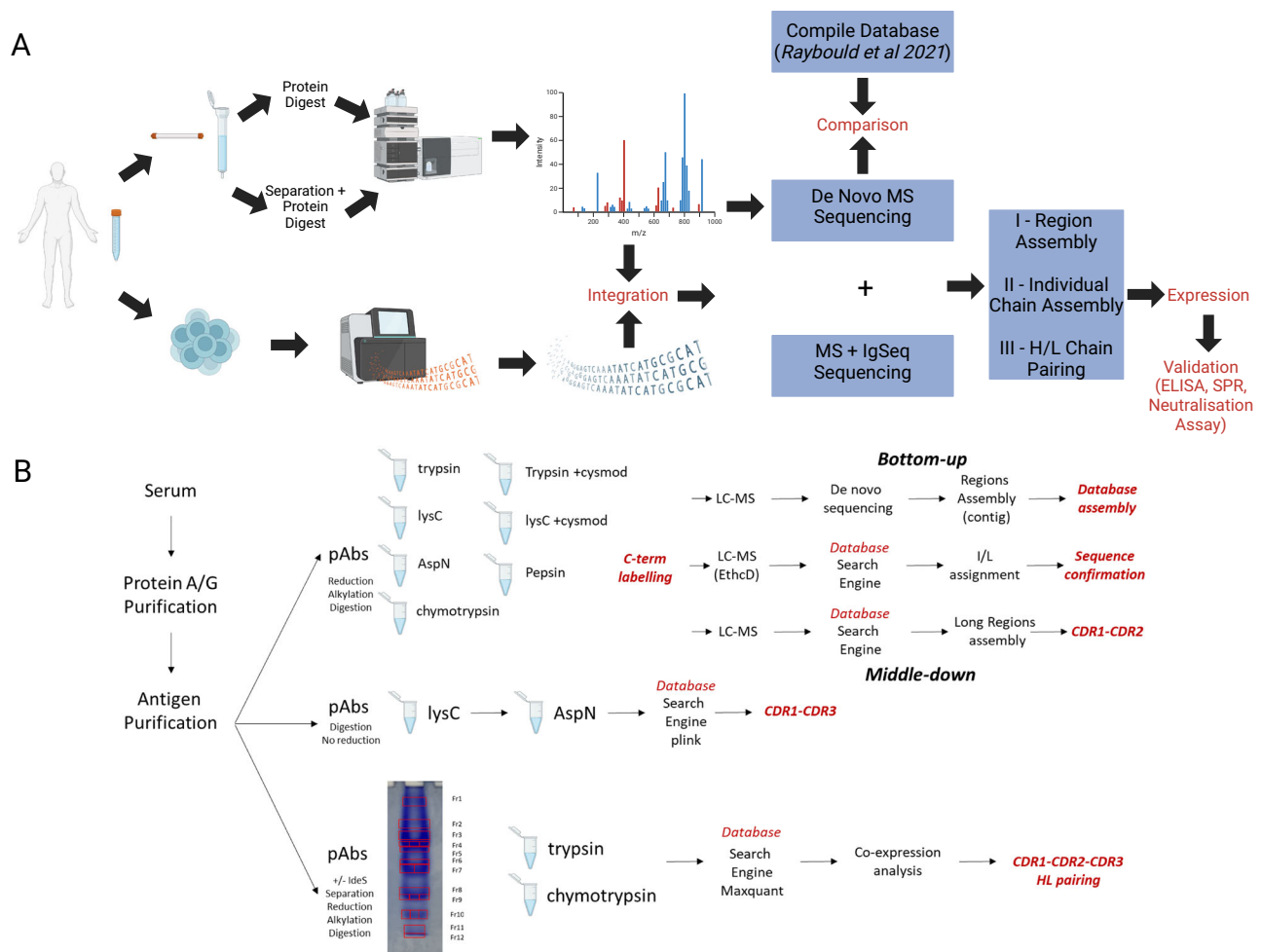
Different processing methods were used on proteomics data (Fig. 1A). The data were initially analyzed conventionally and compared to the IgSeq/B cell database for the same individual. In addition, a de novo protein sequencing analysis was performed on the proteomics dataset. Some recombinant antibodies were generated from this analysis and subsequently tested for antigen binding and neutralization using ELISA, SPR, and pseudovirus neutralization assays. The sequencing protocol for de novo protein sequencing analysis is outlined in Fig. 1B. Multiple proteases were used to generate overlapping peptides from antibody fractions enriched with antigens. Short reads were then used to pair the heavy and light chains’ CDR1, CDR2, and CDR3 regions. We extended the short-read assemblies based on high peptide overlap confidence. The accuracy of longer sequences was verified using a customized FASTA database. Furthermore, this database confirmed the assembly of multiple CDRs in chains through middle-down analysis. Non-reducing experiments were performed to identify sequence pairs for CDR1 and CDR3. Transitioning from short-read to longer-read assembly was facilitated by native gel electrophoresis of the polyclonal antibody, aiding in pairing sequence-specific peptides for the heavy and light chains. In addition, we provide a benchmark study performed on a mix of 5 well-known antibodies to validate several aspects of our proposed platform (See Supplementary Information).

### RNA sequencing

This study screened three consenting SARS-CoV-2 vaccines (patient information in Supplementary Table 1). Screening targeted RBD protein-specific IgG antibodies. Each patient provided 10 ml of plasma for analysis and 8 ml of blood for B cell repertoire and IgSeq. A secondary goat anti-human IgG Fc ELISA was used to detect human IgG antibodies. Subject 2 showed the strongest response, followed by “Subj. 1” and “Subj. 3”. Our study focused on Subj. 2 (Fig. 2A). The TakaraBio SMARTer Human BCR IgG IgM H/K/L Profiling Kit was used to create NGS libraries. RNA from Subject 2’s PBMC was used for library construction. We obtained 3,184,854 and 3,250,539 reads for the replicate heavy chains. Kappa and Lambda chains yielded 3,305,559 and 5,174,719 reads, respectively. Quality control parameters and a summary of the IgSeq data collection are shown in Supplementary Data 1. Duplicate acquisitions were made for the heavy chain to ensure reproducibility; an overlap of 64% was observed between the 2 replicates after filtering for 2+ reads per sequence (Supplementary Data 1).

### Mass spectrometry and database searching

Proteomics data processing involved enriching 3 ml plasma with protein G by gravity flow, isolating 22.2 mg of IgG. SARS-CoV-2 spike protein receptor-binding domain (RBD) binding experiments were performed using 20 mg of this IgG and streptavidin beads coupled with photocleavable biotin. Approximately 190 µg per 3 ml plasma was enriched through acid elution, and an additional 10% antibody yield was obtained through UV cleavage.



**Fig. 1 | Schematic overview of the sequencing pipeline. A** Overview analysis of human antibody repertoires from antigen-specific IgG and peripheral blood. Two different types of samples were taken and analyzed; the bottom branch is the procedure of generating a reference database of immunoglobulin V-region by next-generation sequencing (NGS) of the immunized individual's B cell repertoire. The branch above describes the analysis of soluble serum IgG. This comparison and functional characterization of the two antibody repertoires provide a different perspective on studying humoral response. This process involves the purification and the proteomic analysis of affinity-purified serum antibody (MS-based de novo, top) alongside VH: VL pairing and NGS of peripheral B cell V gene repertoires (BCR-

seq, bottom of A). The proteomics dataset was analyzed using the B cell repertoire. A MS-based de novo sequencing analysis of the proteomics dataset was performed in parallel. The two datasets were compared to the Compile database from Raybould et al. 2021 (CoV-AbDab)<sup>46</sup>, comprising a comprehensive list of antibody sequences developed in the COVID context. The process of assembling antibodies can be divided into three steps: (1) Different CDR regions assembled, (2) Intact individual chain assembly performed, and (3) heavy and light chain pairing. The generated sequences were expressed recombinantly, and their performance was evaluated. **B** A breakdown of the different proteomics experiments performed under the MS-based de novo sequencing. Created in BioRender.

Parallel protease digestions allow the production of overlapping peptides. Five proteases—Trypsin, LysC, AspN, Chymotrypsin, and Pepsin were used to digest the sample. The sample was split into two sub-samples. The first sample was reduced and alkylated with iodoacetamide. In the second sample, the cysteine residues were converted into a lysine analog using 2-bromoethylamine hydrobromide, rendering the cysteine residues susceptible to cleavage by trypsin and LysC enzymes<sup>29</sup>. The digested samples were then analyzed on an Orbitrap Exploris 240 instrument.

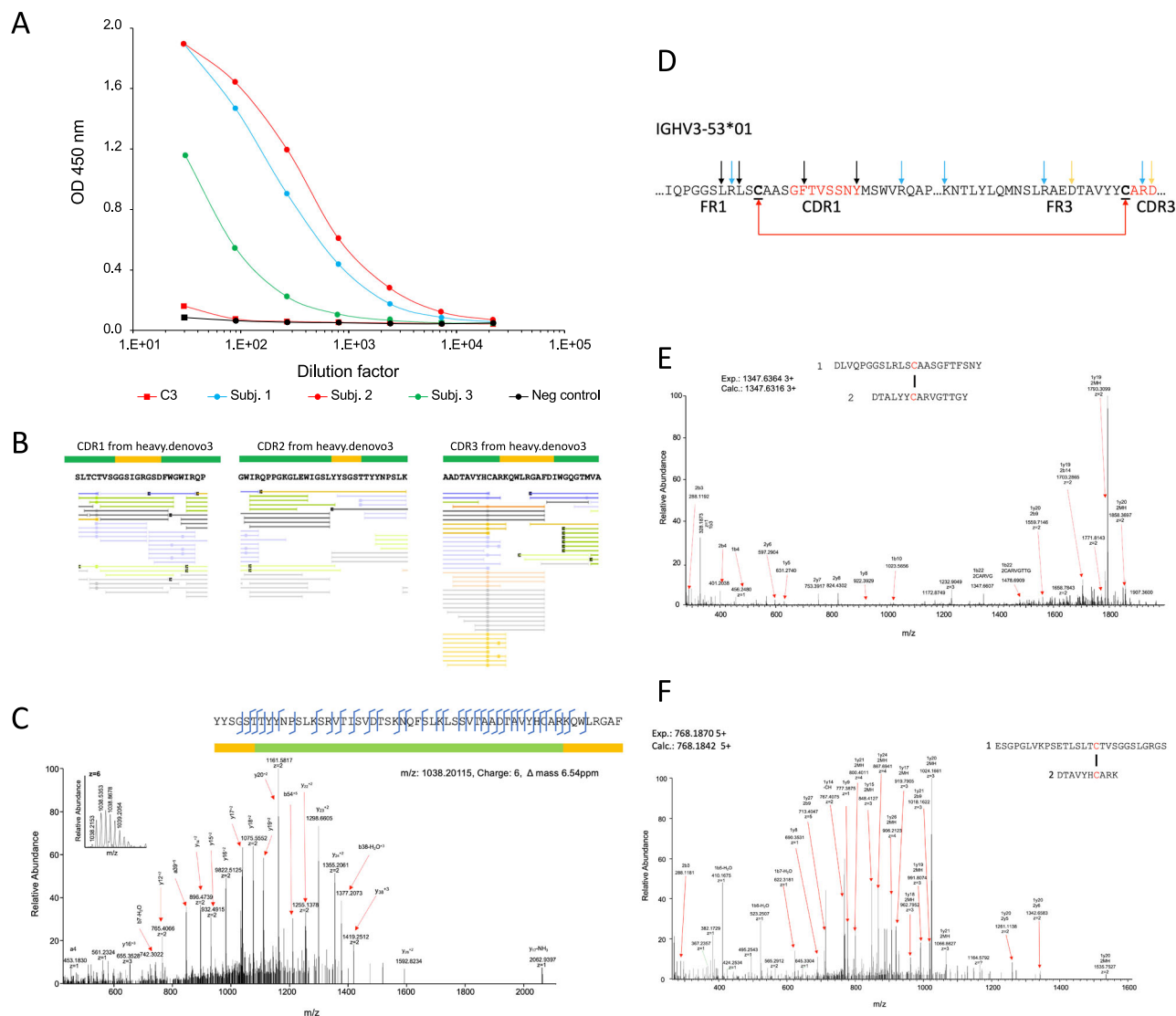
The mass spectrometry data were searched against the IgSeq dataset. Sequences containing two or more confident peptide-spectrum matches were selected and grouped according to their CDR3 similarities. Sequences of which the CDR3 regions differ by at most one amino acid are grouped together. This results in 183 heavy clusters, 203 kappa clusters, and 290 lambda clusters, respectively (Supplementary Data 2). Most of these sequences were matched in the framework regions, and only a handful containing full MS coverage of CDR3 were found, including 6 heavy clusters, 25 kappa clusters, and 13

lambda clusters. These matching CDR3 sequences are shown in Supplementary Data 1, with only one of the matches having a single NGS read while the median number of reads per sequence is 405. After further filtration based on sequence coverage and heavy/light chain pairing, only 2 heavy clusters, 2 kappa clusters, and 2 lambda clusters remained in the final list of recombinant antibodies (Supplementary Data 4).

### De novo protein sequencing

De novo protein sequencing allowed the extraction of additional sequences not found in the IgSeq chain clusters. This resulted in 4 additional heavy chain clusters and 4 additional lambda chain clusters with different CDR3s compared to the IgSeq clusters and partial sequences covering different CDR regions. The selected IgSeq sequences and the full and partial de novo protein sequences are available in Supplementary Data 3 and 4.

Figure 2B exemplifies short contigs with distinct peptide overlaps in three CDR regions. The left sequence in Fig. 2B represents a de novo



**Fig. 2 | Sample assessment and the first steps of MS-based de novo sequencing.**

**A** Plasma titration for antibodies against the receptor-binding domain, RBD of Sars-Cov2. Three plasma samples (Subj. 1, 2 and 3) were benchmarked against an internal in-house standard (C3) and a negative control. **B** Construct of the MS-based de novo assembly of “contig” around different CDRs (CDR1 left, CDR2 middle, and CDR3 right). The green regions are the framework region (FR) while the orange ones are the CDRs for the antigen-enriched polyclonal antibody. The proposed sequence is shown below the highlighted regions, and the bottom parts are short peptides corresponding to elements of the sequences. Green lines are peptides generated from the digestion with Chymotrypsin; light orange is LysC, blue is

protein sequence CDR1 region with possible neighboring sequences (“CTVSGGSLGRGSDFWGW” in Supplementary Data 3). The center of Fig. 2B presents a CDR2 sequence (“WLGSLYSGSTTYNPSLK”), while the right of Fig. 2B displays a CDR3 sequence (“YHCARKQWLRFAGFDLWGQG”). Overlapping peptides provide additional confidence in these proposed sequences. Figure 2C shows a peptide-spectrum match (PSM) between a middle-down MS/MS spectrum and an extended assembly covering significant portions of the CDR2 and CDR3 regions shown above. This spectrum confirms that this pair of CDR2 and CDR3 regions should belong to the same antibody.

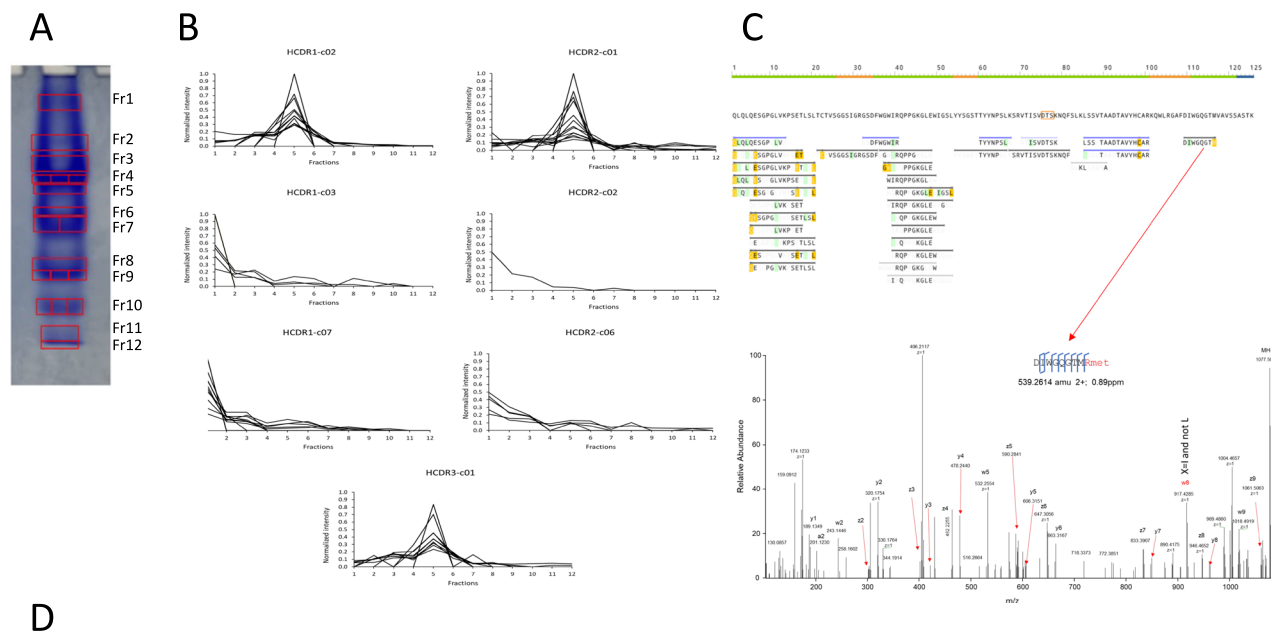
The disulfide bond between the two Cysteine residues near CDR1 and CDR3 was used to identify certain pairs of CDR1 and CDR3 (as shown in Fig. 2D). Figure 2E, F displays two examples identified by pLink software using our MS/MS data under non-reducing conditions.

Trypsin, black is Pepsin, and dark orange is AspN. **C** Typical “middle-down” peptide allowing in this particular case, the assembly of a CDR2 with a CDR3 within the same chain. **D** Resulting peptides from the action of a given protease on a specific germline (IGHV3-53\*01) under non-reduced conditions. The black arrows represent possible pepsin sites, the blue ones are peptides resulting from the digestion with trypsin, and the orange ones are peptides from the protease AspN. **E, F** Some disulfide-linked peptide analysis spectra. Both the experimental mass (Exp.) and the theoretical mass (Calc.) are indicated in both spectra. Those two peptides allow pairing each a specific CDR1 sequence with a specific CDR3 sequence. Source data are provided as a Source Data file.

Supplementary Table 2 lists non-reduced peptides found in this study. Some peptides in Supplementary Table 2 confirm the initial entry assembly from Supplementary Data 3. However, extended digestion under neutral conditions may cause disulfide bonds to scramble<sup>30</sup> and generate false positives, though these are usually associated with low-intensity peptides. Therefore, combining the disulfide bond information with other evidence is necessary to determine the CDR pairing definitively.

The separation experiment, performed under non-reduced conditions, provides additional pairing information. We used IdeS digestion and a Native gel to separate antibodies by charge and mass. This method yielded 12 bands (Fig. 3A), subsequently subjected to Trypsin digestion. Figure 3B shows an example of the normalized quantification vectors of peptides for selected CDR contigs (sequences in





D

Ab ID	Chain ID		Chain Source		Germline Usage		ELISA IC50%	Neutralisation assay <sup>3</sup>		
								ACE2 competition assay		<i>in vitro</i>
	Heavy <sup>1</sup>	Light <sup>1</sup>	Heavy	Light	Heavy	Light		RBD	S1	Cell-based
R1	>heavy-c01.h1RNA	>kappa-c016.k2	RNA	RNA	IGHV3-53	IGKV1-9	~0.6nM	yes	weak	yes
R2	>heavy-c01.h1RNA	>kappa-c016.k3	RNA	RNA	IGHV3-53	IGKV1-9	~0.2nM	yes	yes	yes
R3	>heavy-c01.h22.denovo2	>kappa-c016.k2	<i>de novo</i>	RNA	IGHV3-53	IGKV1-9	no	n/a	n/a	n/a
R4	>heavy-c01.h22.denovo2	>kappa-c016.k3	<i>de novo</i>	RNA	IGHV3-53	IGKV1-9	~6nM	no	no	no
R5	>heavy-c06.denovo1	>lambda-c02.denovo1	<i>de novo</i>	<i>de novo</i>	IGHV3-13	IGLV3-25	~0.3nM	yes	yes	yes
R6	>heavy-c06.denovo1	>lambda-c02.denovo2	<i>de novo</i>	<i>de novo</i>	IGHV3-13	IGLV3-25	~0.3nM	yes	yes	yes
R7	>heavy.denovo4	>lambda-c09.1.denovo	<i>de novo</i>	<i>de novo</i>	IGHV3-23	IGLV3-21	no	n/a	n/a	n/a
R8	>heavy.denovo4	>lambda-c09.3.denovo	<i>de novo</i>	<i>de novo</i>	IGHV3-23	IGLV3-21	no	n/a	n/a	n/a
R9	>heavy-c02.h1RNA	>lambda-c04.l1	RNA	RNA	IGHV3-30	IGLV3-21	no	n/a	n/a	n/a
R10	>heavy-c02.h1RNA	>lambda-c04.l2	RNA	RNA	IGHV3-30	IGLV3-21	no	n/a	n/a	n/a
R11	>heavy.denovo3	>lambda-c09.1.denovo	<i>de novo</i>	<i>de novo</i>	IGHV4-39	IGLV3-21	~0.3nM	yes	yes	yes
R12	>heavy.denovo3	>lambda-c09.3.denovo	<i>de novo</i>	<i>de novo</i>	IGHV4-39	IGLV3-21	~0.3nM	yes	yes	yes
pAb <sup>4</sup>							~0.3nM	n/a	n/a	n/a

notes

- Chain naming based on structure info; h:heavy, k:kappa, l:lambda, RNA IgSeq, de novo
- Approximate estimation of the IC50% value
- Qualitative interpretation of the neutralisation assay, "yes" indicate neutralisation was detected
- Natural polyclonal mixture

**Fig. 3 | Final step in sequencing: CDRs assembly, isobaric ambiguity resolution, and recombinant antibody testing.** **A** Separation of the polyclonal antibody on a native gel. Initially, the enriched polyclonal antibody was digested with IdeS, and the resulting sample was fractionated into 12 fractions and further digested with trypsin. The separation was carried out once with IdeS and once without IdeS, respectively. Separation without Ides resulted in fewer fractions (see Source Data file). **B** Quantitative profiling of various peptide “contigs” across those 12 fractions, where contigs were grouped based on their similarity to assemble chains and pair heavy-light (HL) chains (more information on the contig sequence can be found in SI Table S3). **C** top: shows an example of the sequence coverage resulting from

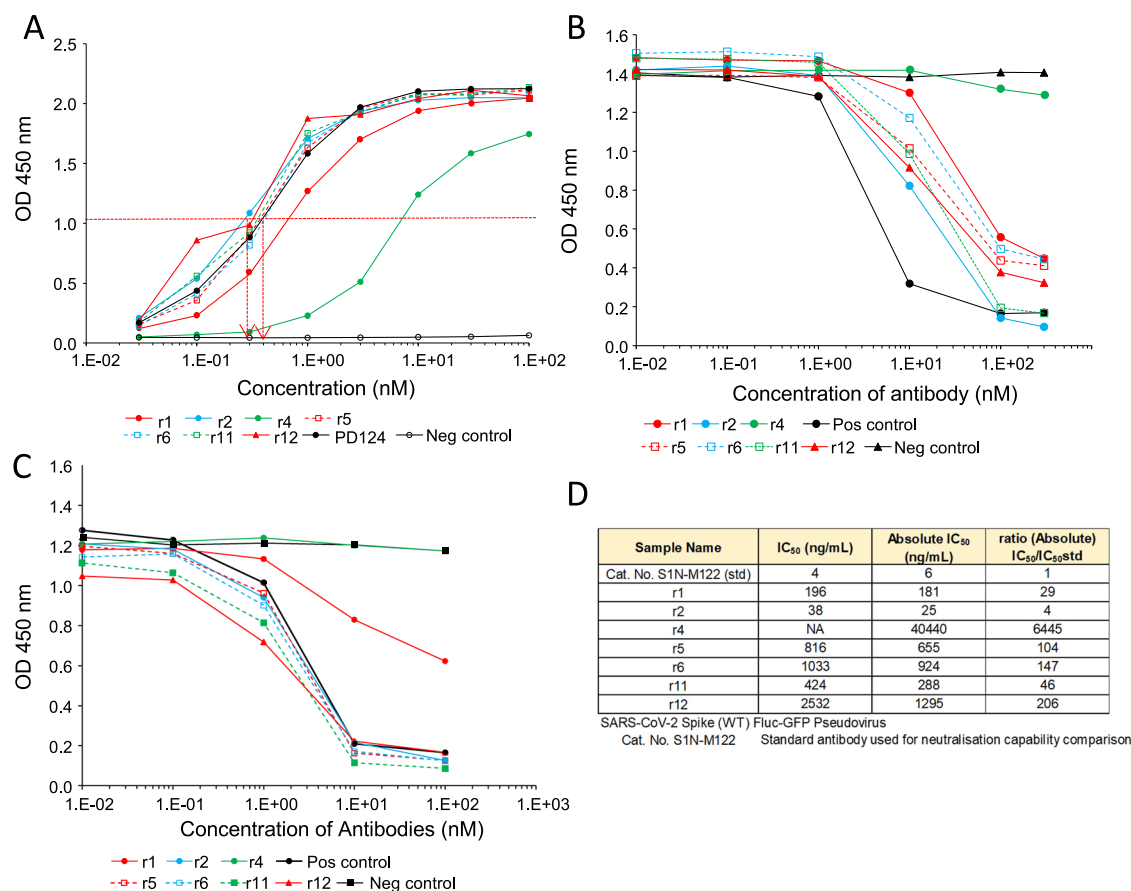
EThcD analysis of trypsin and pepsin digestion. Peptides resulting from pepsin digestion were labeled with Arginine methyl ester. Bottom: displays EThcD spectra acquired to resolve specific Isoleucine/Leucine assignments; in that case, the generation of w-ions under EThcD fragmentation allows the assignment of w8 to an isoleucine. Additionally, in this case, a positive charge (i.e., Methyl Arginine as a y1 ion) was added to the C-terminal end of the non-trypsinic peptide to facilitate the C-terminal fragment to be detectable in MS/MS mode. **D** Expressed candidates and their affinities in ELISA Assay, Angiotensin-converting enzyme 2, ACE2 competition binding assay and in vitro neutralization assay (Using either the Receptor-Binding Domain, RBD or the protein S1). Source data are provided as a Source Data file.

Supplementary Table 3). The curves for HCDR1-c02, HCDR2-c01, and HCDR3-c01 show similarity, suggesting they belong to the same antibody protein and should be assembled as part of the same chain. Pearson correlation analysis confirms the highest similarity scores between HCDR1-c02, HCDR2-c01, and HCDR3-c01, supporting their assembly as components of the same chain. Co-elution profile analysis also enables the pairing of a heavy chain with its cognate light chain. Orthogonal separation could improve pairing. If heavy and light chain pairing is uncertain, multiple combinations of heavy and light chains can be generated and tested.

In addition, we used semi-denatured protein on a gel to separate the antibodies and the Fab2 fragments (Supplementary Fig. 1A). Seven

heavy chains, nine kappa chains, and six lambda chains were correlated using Native and semi-denaturing gel data. Supplementary Fig. 1B profile correlations showed that five heavy chains might pair with unique light chains. The remaining two heavy chains were associated with weak MS signals and unrelated to the light chains. Additionally, ten light chains had no significant correlation with heavy chains and were excluded from the analysis. Some sequences were ambiguous and required multiple H/L combinations.

Figure 3C shows the EThcD-based isoleucine (I) and leucine (L) residue assignment method we used in this study, as described by Zhokhov et al.<sup>31</sup> This method produces z- and w-ions from the C-terminus. The w-ions help distinguish Isoleucine and Leucine



**Fig. 4 | Performance and properties of the investigated antibodies.** The Materials and Methods section describes how the various ELISA procedures were performed. **A** Conventional ELISA assay comparing recombinant antibodies with the natural polyclonal antibody PD124 affinity enriched in this study as a benchmark measure. The red dashed arrows are the IC<sub>50</sub> for the best recombinant antibody (R2 at approx. 3 nM) versus the natural polyclonal antibody (PD124 at approx. 4 nM). Nonbinding recombinant antibodies (r3, r7, r8, r9, r10) were eliminated from further comparisons. Due to its restricted availability, the natural polyclonal antibody PD124 was excluded from other investigations. Figure 4B illustrates an ACE2 competitive binding assay conducted on the recombinant forms found in (A), utilizing the RBD domain. **C** ACE2 competitive binding assay that targets the S1 protein as the

bait in the test. For **A–C**, although these ELISA assays focus on different aspects, they all utilized the horseradish peroxidase assay with TMB (3,3',5,5'-tetramethylbenzidine, TMB, resulting in the measurement outcomes at 450 nm across the three different assays. **D** In vitro neutralization test that utilizes SARS-CoV-2-Spike/Flu-GFP pseudovirus to neutralize a positive standard S1N-M122. The assay measures the ratios of Absolute IC<sub>50</sub>. The positive control in **B**, **C** is AS35, an Anti-SARS-CoV-2 Spike RBD Neutralizing Antibody, Human IgG1, while in (**D**), it is the S1N-M122, a known anti-Spike RBD Neutralizing Antibody, Chimeric mAb, Human IgG1 (AM122), both sourced from Acro Biosystems. Source data are provided as a Source Data file.

residues. This method works nicely for Tryptic and LysC digests. Other proteases like Pepsin or Chymotrypsin may produce inconsistent results. To overcome this limitation, we have developed a method that introduces a positive charge on the C-terminal amino acid. This modification generates z-ions to w-ions and helps sequence and identify Isoleucine or Leucine residues. Figure 3C, bottom part, shows a spectrum of an isoleucine-specific w8 ion detected using this method. Most I/L ambiguities can be resolved using this labeling technique, EThcD, germline analysis, and IgSeq data.

### Sequencing results and functional assays

The 12 resulting antibodies (R1–R12) are combinations of 6 heavy and 8 light chains. All six heavy chains contain different CDR3. Four of the six heavy chain CDR3 sequences are absent from IgSeq data and constructed by de novo protein sequencing only. The eight light chains belong to four different clusters. Sequences within each cluster share highly similar CDR3 and V-region sequences. Four light chain sequences (two clusters) are from de novo protein sequencing. Together, six of the 12 antibodies have both heavy and light chains constructed from de novo protein sequencing, while eight have at least one chain constructed from de novo protein sequencing. The

sequences of these 12 antibodies and their MS coverage are provided in Supplementary Data 4.

We expressed the 12 recombinant antibodies and tested their functions with the following assays:

- ELISA and SPR analyses for affinity to the RBD domain.
- In vitro ACE2 competition binding assay against RBD and S1 domains, respectively.
- In vitro neutralization assays against Fluc-GFP pseudovirus.

The results are summarized in Fig. 3D.

ELISA assays were used to assess recombinant antibody affinity for the Receptor-Binding Domain (RBD) (Fig. 4A). Seven of the 12 recombinant antibodies (R1, R2, R4, R5, R6, R11, R12) exhibited high affinities towards the RBD. Six recombinant antibodies have affinities similar to or better against the RBD than the natural polyclonal antibody named PD124 (with an IC<sub>50</sub> at 0.3–0.6 nM Fig. 4A red arrows). Note that PD124 was affinity purified with RBD from plasma of Subj. 2, which also serves the starting material to derive the recombinant antibodies. Four of the six strongest binders (R5, R6, R11, R12) were obtained by de novo protein sequencing.

The seven binders were further assessed using Surface Plasmon Resonance (SPR) methods, including complex stability analysis<sup>32</sup> and off-rate screening<sup>33</sup>. Complex stability, which indicates the fidelity between the antibody and antigen, was evaluated by plotting the stability early vs. stability late values (Supplementary Fig. 2A–C). This analysis showed that R11 and R12 bind with the highest complex stability, followed by R2, R6, and R5. The off-rate screening was performed as a complementary analysis to further investigate the top candidates' kinetic properties. Multiparametric evaluation was completed with this method. The top binders (R2, R5, R6, R11, and R12) were categorized with slow off-rates (Supplementary Fig. 2D), indicating their prolonged binding capabilities. The dissociation rates of R4 and R1 were much faster in comparison.

ACE2 competition binding assays were performed on the seven binders to determine if they could interfere with the ACE2/RBD and ACE2/S1 interactions (Fig. 4B, C, respectively). Due to its limited availability, the natural affinity-purified polyclonal antibody (PD124) was not used as a neutralization benchmark. Among the tested antibodies, R1, R2, R5, R6, R11, and R12 exhibited neutralization capability, while R1 demonstrated reduced neutralization ability against S1 compared to the RBD protein. R4 did not show any significant neutralization capability.

An in vitro cell-based neutralization assay was performed to test those seven recombinant antibodies' neutralization activities against Fluc-GFP pseudovirus (Fig. 4D). All tested recombinant antibodies displayed varying degrees of neutralization activities. Relative to the positive control SIN-M122, R2 showed the closest neutralization efficacy, followed by R1 and R11.

### Benchmark study with five known mAbs

The methods were benchmarked with a mixture of five known mAbs: Adalimumab (Ada), Bevacizumab (Bev), Cetuximab (Cet), Rituximab (Rit), and Trastuzumab (Tra). All five mAbs were successfully sequenced from the mixture. The detailed results and data analysis can be found in the Supplementary information. The assembly program constructed all CDR sequences correctly, followed by human inspection and selection (Supplementary Data 7). By combining the correlation matrix (Supplementary Fig. 5) and the disulfide-linked CDR1/CDR3 peptides (Supplementary Data 8), CDRs of each chain were correctly paired. The full sequences of all chains were constructed correctly after combining the contigs containing the correct CDRs of the same chain, followed by minimal manual curation. The heavy and light chains of the five mAbs were also paired correctly by the correlation matrix (Supplementary Fig. 9). In most cases, the isobaric ambiguities between I and L were resolved correctly by the presence of w-ions and by comparing with the germline genes in the remaining cases.

In particular, the correlation matrixes (Supplementary Figs. 5 and 9) could clearly separate the five mAbs into four clusters: Ada, Bev, and Cet are each in a separate cluster, whereas Rit and Tra are in another cluster. The correlation (similarity) scores between the CDRs or the heavy/light chains within the same cluster were significantly higher than those across different clusters. This fact makes the pairings of the CDRs or the heavy/light chains straightforward for Ada, Bev and Cet. The pairings for Rit and Tra were not as obvious. However, a careful comparison of the total scores between different combinations and the combined consideration of disulfide-linked CDR1/CDR3 peptides could still reveal the correct pairing.

The overall benchmark study is presented in detail in the Supplementary Information. An MS metafile describing the run performed for the benchmark study is presented in Supplementary Data 5. A de novo peptide list generated from the benchmark study is included in Supplementary Data 6, and the contig assembly is presented in Supplementary Data 7. Supplementary Data 9 and 10 contain the curated sequence FASTA file from the study and the known sequence, respectively.

## Discussion

The majority (approximately 90%) of human antibody discovery related to SARS-CoV was achieved through B cell sequencing-only approaches, as evidenced by a survey of the CoV-AbDab database (Supplementary Fig. 1C). Although this approach has been successful, there is often limited correspondence between the identified B cells and the circulating IgG population, indicating the potential for missed valuable candidates<sup>17,19</sup>. It should be noted that IgG molecules in circulation, rather than the B cells themselves, serve as the final product and primary effectors responsible for humoral immunity. Most of the circulating IgGs are produced by plasma cells in the bone marrow rather than circulating B cell<sup>34</sup>. While combined proteomics and B cell repertoire analysis can confirm the presence of specific antibodies in circulation, it cannot identify antibodies that may have been missed by B cell sequencing. A de novo protein polyclonal sequencing, although technically challenging, offers the most promising approach to comprehensively understanding the final products of the humoral immune response.

This present study employed a combination of conventional proteomics and de novo protein sequencing approaches. Our de novo protein sequencing efforts led to the discovery of six additional antibodies, where both the heavy and light chains were absent from IgSeq data. The limited sample size prevents any definitive conclusions from being drawn regarding the superiority of one method over the other, as four out of six antibodies demonstrated equivalent affinity to the natural pAb, while two IgSeq antibodies showed similar affinity levels against the antigen. Additionally, the de novo protein and IgSeq antibodies sequencing exhibited varying degrees of neutralization activity for both in vitro and cell-based neutralization assays. The neutralizing activities against the Fluc-GFP pseudovirus were comparable to those previously reported by He et al.<sup>35</sup>. It is important to note that neutralization depends on binding to the correct epitope, resulting in variations in neutralization abilities even among mAbs with similar binding affinities<sup>36</sup>. The demonstrated binding affinity and neutralizing activities strongly support the accuracy of our sequences. Furthermore, our findings indicate that the circulating B cell repertoire does not encompass all sequences of circulating antibodies, underscoring the necessity of de novo protein sequencing.

In the study by Guthals et al.<sup>18</sup>, 28 antibodies were sequenced and expressed, but only two exhibited binding potencies, albeit significantly lower than the original pAb, indicating potential sequence errors. Notably, due to the absence of a pairing method for heavy and light chains, all combinations of four heavy and seven light chains had to be expressed and tested. More recently, Bondt et al.<sup>28,37</sup> attempted de novo protein sequencing of a polyclonal antibody using a similar approach as in Guthals et al.<sup>18</sup>. However, no binding or neutralization data were provided to validate the accuracy or effectiveness of the generated antibodies. In our experience, numerous possible sequences exist for each CDR, resulting in millions of potential combinations for an antibody's six CDRs. Resolving the assembly ambiguity necessitates more comprehensive information than overlapping peptides, middle-down MS datasets, and intact mass data acquisition. Consequently, we suspect CDRs from different mAbs may have been incorrectly assembled in the works mentioned above, leading to suboptimal binding outcomes. A significant distinction in our approach lies in extensively utilizing various experiments (e.g., middle-down MS, non-reducing digestion, and different antibody separation techniques) to generate orthogonal datasets. We combined these orthogonal datasets to assemble the different CDRs into functional antibodies accurately.

In addition to previous studies<sup>7,8,18,28,38</sup>, we have developed a method for pairing heavy and light chains to produce functional antibodies. Using bulk sequencing data from B cells is challenging as the heavy and light chain sequences are not paired. Previous studies attempted all possible pairings<sup>7,8,18</sup>, but this becomes impractical and

expensive when there are more than 10 heavy and light chains. Single B cell sequencing provides pairing information but is more laborious and costly than bulk sequencing. Additionally, the limited number of cells in a single-cell sequencing experiment reduces the coverage of the B cell repertoire. Utilizing a co-separation MS profile for pairing, our approach offers an alternative solution to overcome these challenges.

Furthermore, we implemented stringent validation measures to confirm the functionality of the assembled antibodies. Extensive binding assays and neutralization tests assessed their functional activities. These thorough experimental validations provide solid evidence for the correctness and effectiveness of our approach.

Our findings are underpinned by a detailed independent benchmark study, which used five well-characterized antibodies to test our MS-based sequencing approach. This study was designed to independently validate the accuracy and reliability of several aspects of the proposed platform. The benchmark results demonstrated that our approach successfully provided precise CDR1/3 linkages and accurate heavy-light chain pairing with no scrambling detected (at a *p*-val of 0.01), reinforcing the robustness of our methodology. Notably, the BEA treatment and Native gel co-fractionation were particularly effective in overcoming challenges related to polyclonal serum analysis. The consistency and reliability of our benchmark results support our confidence in the identified sequences and the subsequent functional assays (see Supplementary Information).

A de novo protein sequencing is vital when B cells cannot be obtained. For instance, immunized animals produce most of the current pAb reagents. However, if the production of the same effective polyclonal antibody ceases in subsequent immunization rounds or is no longer available, de novo protein sequencing of stored protein samples can still recover the original polyclonal antibody. When B cells are available, alternatives such as B cell cloning or the production of phage display libraries from animals producing pAb reagents can be used to recover the original polyclonal antibody sequence.

While previous studies have demonstrated successful de novo protein sequencing of a single protein or monoclonal antibody (mAb)<sup>25,26,38</sup>, it is essential to note that de novo protein sequencing of a polyclonal antibody (pAb) poses more significant challenges. Antibodies within a pAb mixture often share similar framework regions, making it challenging to pair CDRs using peptide overlaps unambiguously. The high sequence diversity within CDRs results in diluted peptide signals, reducing the signal-to-noise ratio and further complicating de novo protein sequencing. Coincidentally, CDRs are the most critical regions of antibodies. Additionally, not all peptides produce equally strong mass spectrometry (MS) signals<sup>39,40</sup>, potentially leading to a failure in sequencing antibodies if even one CDR is missing from the MS data. In our study, we successfully de novo sequenced more CDR sequences than the final list of antibodies, highlighting the challenges involved (see Supplementary Data 3). Additional fractionation of the pAb could improve MS coverage. Still, it would also increase the dataset size and the complexity of handling the more significant number of CDR combinations, requiring more advanced algorithms.

Although the number of reported antibodies in our study may appear small, numerous partial sequences in Supplementary Data 3 suggest that the initial polyclonal antibody (pAb) sample is a highly complex mixture, even after antigen affinity enrichment. This circumstance highlights the difficulty of our pAb de novo protein sequencing approach and the potential for sequencing a larger number of antibodies with further methodological improvements. It is worth noting that intense mass spectrometry (MS) signals are typically generated from peptides resulting from the protease digestion of more abundant proteins, suggesting that the most prevalent antibodies in the pAb mixture are likely to be among the first successfully de novo protein sequenced.

Despite the challenges mentioned above, our study has successfully demonstrated the feasibility of de novo protein sequencing from a human serum polyclonal antibody (pAb), generating multiple monoclonal antibodies (mAbs) that closely resemble the original pAb. This achievement emphasizes the potential of pAb de novo protein sequencing as a valuable strategy for harnessing the natural immune response of animals and humans, enabling the discovery of antibody reagents and therapeutics. We anticipate that with further advancements in mass spectrometry experiments and bioinformatics, the field of pAb de novo protein sequencing will continue to flourish and contribute to antibody research and development.

## Methods

### Sample collection

The blood specimens of three healthy donors were obtained from Discovery Life Sciences. The collection and research use of these specimens were approved by the WCG Clinical IRB. The donors provided informed consent for their participation and the research use of their biological samples. The donors were compensated.

Detailed anonymized information about all donors is available in Supplementary Table 1. Gender information was not used in the study design. Blood and plasma were collected from all 3 donors, but this study focused on donor Subject 2, a 49-year-old female who received the third Moderna shot. Blood for analysis was drawn 60 days later, using BD Vacutainer CPTM mononuclear cell preparation tubes and stored in RNeasy Protect<sup>®</sup> Cell Reagent following the supplier's instructions. For proteomics analysis, blood was centrifuged at  $2,000 \times g$  for 15 min at 4 °C, and samples were aliquoted into 3 ml Matrix cryo vials, totaling 3 ml serum per patient. Additional samples were acquired to benchmark the methods presented in this manuscript. Adalimumab, Bevacizumab, Cetuximab, Rituximab, and Trastuzumab were purchased from Sino Biologicals and combined to create a simple controlled oligoclonal mixture.

### RNA extraction

PBMCs from donor Subject 2 were thawed, centrifuged, and lysed with Buffer RLT (Qiagen) and beta-mercaptoethanol. The lysate was transferred to a QIAshredder spin column and centrifuged for 2 mins at  $19,500 \times g$ . The homogenized lysate was combined with an equal volume of 70% ethanol and proceeded to RNA extraction with the RNeasy Mini Kit (Qiagen) for RNA extraction. The extracted RNA concentration was measured with the RNA BR assay for Qubit fluorometry (Thermo-Fisher), and RNA integrity was assessed by the Agilent TapeStation 4150 (Agilent Technologies), resulting in an RNA Integrity Number (RIN) of 9.6.

### cDNA synthesis, BCR amplification, and sequencing library generation

The SMARTer Human BCR IgG IgM H/K/L Profiling Kit (TakaraBio) was used to create NGS libraries. Following kit instructions, one  $\mu$ g of RNA was used in the first-strand cDNA synthesis. Per the user manual, PCR cycles were increased from 16 to 18 cycles for PCR2. Library concentration was determined with the 1 $\times$  dsDNA HS assay for Qubit fluorometry (Thermo-Fisher), and sizes were assessed using the D1000 ScreenTape on the Agilent TapeStation 4150 (Agilent Technologies), targeting 640 bp for light chains and 690 bp for heavy chains. Peaks at 350 bp indicated adapter dimers and libraries with such peaks underwent size selection with NEBNext Sample Purification Beads (0.65 $\times$  beads to sample volume) per the SPRIselect User Guide.

### Next-generation sequencing

The final products were re-quantified with the 1 $\times$  dsDNA HS assay (Thermo-Fisher), normalized, and pooled to a 4 nM library. Following the Illumina Denature and Dilute Libraries Guide, an 8 pM loading library was prepared. To ensure complete denaturation and efficient



binding to flow cells, the libraries were heat-shocked at 96 °C for 2 min and then cooled in an ice water bath for 5 mins. Sequencing was performed on an Illumina MiSeq platform using the reagent kit v3 for 600 cycle paired-end reads, including a 30% PhiX v3 spike-in.

### IgG enrichment against the antigen

Total IgG was enriched from 3 mL human serum using 3 mL of settled protein G agarose resin (Genscript) in a 20 mL gravity flow column (Biorad). The column was equilibrated with two 15 mL washes of 10 mM phosphate buffered saline (PBS). The serum was centrifuged at  $23,000 \times g$  for 10 min at 4 °C, combined with 9 mL of 10 mM PBS, and passed twice over the protein G resin column by gravity flow. After three 10 mL PBS washes, the IgG fraction was eluted with 12.5 mL of 0.1 M glycine buffer (pH 2.5). Eluted IgG was concentrated and buffer-exchanged into 10 mM PBS using a 30 kDa Amicon filter (Sigma), yielding 22.2 mg of total IgG. Anti-RBD antibody enrichment involved biotinylation of 0.4 mg of SARS-CoV-2 spike protein RBD (Exonbio), followed by coupling to 54  $\mu$ L streptavidin-coated Sepharose beads (Cytiva) for 1 h at 4 °C. After incubation with total IgG, non-specific binders were removed by washing twice with 0.4 mL 10 mM PBS, followed by a wash with 0.4 mL CHAPS 0.5% in 10 mM PBS, and then 6 washes with 0.4 mL 10 mM PBS. Anti-RBD antibodies were eluted twice using 0.4 mL glycine pH 2.5. Beads were washed with 500  $\mu$ L 10 mM PBS to bring pH to 7, then 1 h UV elution was performed in 100  $\mu$ L PBS using UV Stratalinker 2400 (Stratalinker), with 365 nm bulb to elute any remaining anti-RBD antibodies (less than 10% of the total polyclonal were eluted that way). A total of 190  $\mu$ g anti-RBD antibodies were collected.

### In-solution digestion

The in-solution digestion process was performed on the antigen-enriched fraction described above (25  $\mu$ g). Initially, the sample was concentrated to 100  $\mu$ L via vacuum centrifugation in a low-pressure centrifuge (i.e., Speedvac). Following this, the sample underwent reduction using dithiothreitol (DTT) at a final concentration of 30 mM for 15 min at 95 °C. Subsequently, the sample was divided into two tubes for further treatment. In the first fraction (5/7th of the sample), alkylation with Iodoacetamide (IAA) was carried out at a final concentration of 50 mM for 30 min in darkness at room temperature. After the IAA treatment, the sample was precipitated using three volumes of acetone, followed by incubation at -20 °C, centrifugation at  $23,000 \times g$  for 10 mins at 4 °C, and drying of the pellet in a low-pressure centrifuge. The pellet was reconstituted with 10  $\mu$ L of 4 M urea and incubated at 37 °C for 10 min. Subsequently, 90  $\mu$ L of water was added, and the sample was divided into five tubes, with four tubes receiving 30  $\mu$ L of 50 mM ammonium bicarbonate for overnight digestion with Trypsin, LysC, AspN, and Chymotrypsin at a 1:20 ratio. In the fifth tube, Pepsin digestion was performed by adding 30  $\mu$ L of HPLC-grade water with 2  $\mu$ L of 1 N HCl and Pepsin at a 1:20 ratio, followed by digestion at 37 °C for 15 min and inactivation at 95 °C for 3 min. Meanwhile, the remaining fraction (2/7th) was treated with 0.5 M 2-bromoethylamine hydrobromide (BEA) in a 100 mM Tris buffer at pH 8 for 4 h at 25 °C. BEA converted cysteine into a lysine-like amino acid, making it a potential protease cleavage site for both Trypsin and LysC<sup>29</sup>. The BEA-labeled samples were precipitated by adding trichloroacetic acid to a final concentration of 20% (w/v). The precipitated sample was washed with acetone, dried, reconstituted, and digested with Trypsin and LysC overnight at 1:20 in 30 mM ammonium bicarbonate. All protease digests were dried under low pressure, reconstituted in 40  $\mu$ L of 0.1% formic acid, and loaded onto Evotips in a method similar to Bache et al.<sup>41</sup> and analyzed using HCD mode on Evosep-Exploris 240 (Thermo-Fisher). The proteases utilized in the polyclonal digestion were sourced from Promega.

For non-reduction in solution digestion, 20  $\mu$ g of polyclonal protein was reconstituted in 20  $\mu$ L of 8 M urea in 100 mM Tris buffer at pH 6.5. N-ethylmaleimide (NEM) was added to a final concentration of

2 mM and incubated at 37 °C for 2 h. Endoprotease LysC was added at a 1:20 protease to protein ratio and incubated overnight at 37 °C, followed by 4-h digestion with AspN at the same ratio and conditions. After protease digestion, the sample was dried under low pressure, reconstituted in 40  $\mu$ L of 0.1% formic acid, and loaded with 2.5  $\mu$ g of the digested samples onto Evotips for analysis using a Thermo Orbitrap Exploris 240 mass spectrometer, as detailed in the Mass spectrometry analysis section.

### Gel-based separation

Gel-based separation was conducted using Native and “Non-reducing Room Temperature” (NRT) gels. Native gel: polyclonal antibodies were separated on a Biorad precast 7.5% polyacrylamide Mini-Protean TGX gel with and without IdeS treatment, respectively. IdeS treatment involved incubating 50  $\mu$ g of the polyclonal antibody with 50 units of IdeS (Promega), and the volume was reduced to 30  $\mu$ L using vacuum centrifugation under low pressure. For the non-IdeS sample, 25  $\mu$ g was loaded onto the gel. NativePAGE 4 $\times$  buffer (Thermo-Fisher) was added in a 4:1 ratio for both IdeS-treated and non-treated samples, and the gel was run using a Biorad power unit set to 130 V for 180 min. The running buffer was diluted from 10 $\times$  stock Tris/Glycine Buffer (Biorad) to 1 $\times$  using Milli-Q water. Non-reducing Room Temperature gel (NRT, as shown in Supplementary Fig. 1A): The polyclonal antibody underwent separation in a modified non-reducing SDS-PAGE procedure without pre-heating, enabling different denaturation and migration patterns. A total of 10  $\mu$ g sample per lane was combined with Laemmli buffer (Biorad), loaded onto a precast 7.5% Mini-PROTEAN TGX gel, and run at 150 V for about 60 min in Tris/Glycine/SDS electrophoresis buffer. All gels and buffers were sourced from Biorad. Coomassie brilliant blue (Biorad) staining for 30 min and overnight destaining with Biorad Destain were performed for all gels.

### In-gel digestion

The procedure used was a modified version of Mann’s method<sup>42</sup>. After gel separation, bands were cut and washed twice with 200  $\mu$ L of HPLC-grade water. The bands were dehydrated with 200  $\mu$ L of 100 mM tetraethylammonium bicarbonate (TEAB) and acetonitrile (ACN) (1:1 ratio). Subsequently, they were reconstituted in 25 mM DTT in 100 mM TEAB; samples were reduced at 56 °C for 30 min. The DTT solution was removed, and 55 mM IAA was added to alkylate for 30 min at room temperature. After that, the bands were washed twice with 0.4 mL of HPLC-grade water and dehydrated with 200  $\mu$ L of 100 mM TEAB/ACN in a 1:1 ratio. For digestion, trypsin was diluted to a concentration of 6 ng/ $\mu$ L in 100 mM TEAB, of which 100  $\mu$ L was added to gel pieces. The digestions were allowed to proceed at 37 °C overnight. The supernatant containing the digested peptides was collected in fresh tubes, and the gel pieces were dehydrated again to extract additional peptides using 100  $\mu$ L of a 60% ACN, 0.1% formic acid (FA) solution. The supernatant from this extraction was combined with the digestion supernatant and dried down under low pressure. The dried samples were then resuspended in 40  $\mu$ L of 0.1% FA, and 100% of the sample was loaded on Evotips according to the manufacturer’s instructions. Finally, the samples were analyzed in High Collision Dissociation (HCD) mode on an Evosep-Exploris 240 mass spectrometer, as detailed in the Mass spectrometry analysis section.

### Isobaric ambiguity resolution by ETHcD mass spectrometry

To resolve Isoleucine/Leucine isobaric ambiguity, three selected samples underwent ETHcD mode analysis as Zhokhov et al.<sup>31</sup> proposed. These samples included in-solution digested trypsin and pepsin, with, in addition, the pepsin sample labeled at the C-terminal with arginine methyl ester dihydrochloride in a procedure similar to the one presented by Krusemark et al.<sup>43</sup>. For the labeling process, the coupling reagent 7-Azabenzotriazol-1-yloxy)tripyrrolidinophosphonium hexafluorophosphate, PyAOP, was utilized along with arginine methyl ester

dihydrochloride. After resuspending the digestions in 10  $\mu\text{L}$  of DMSO, 6  $\mu\text{L}$  of a coupling solution was added (66 mg PyAOP into 132  $\mu\text{L}$  DMSO) and incubated for 5 min. Following this, 14  $\mu\text{L}$  of a Reagent Solution, consisting of 100 mg of Methyl Arginine dissolved in 50  $\mu\text{L}$  of ddH<sub>2</sub>O plus 26  $\mu\text{L}$  of N-Methyl morpholine, was added and incubated for 1 h. The reaction was then stopped by adding 320  $\mu\text{L}$  of 0.1% formic acid and further subjected to liquid-liquid phase separation with 640  $\mu\text{L}$  of chloroform for clean-up. The processed sample (i.e., the water phase) was dried and resuspended in 0.1% FA, and 2.5  $\mu\text{g}$  of the digested samples were subsequently loaded onto Evotips according to the manufacturer's instructions and then analyzed on an Evosep-Eclipse in EThcD mode as detailed in the next section.

### Mass spectrometry analysis

Mass spectrometry analysis was conducted using an Orbitrap Exploris 240 (Thermo-Fisher) for HCD-MS experiments and a Thermo-Fisher Orbitrap Eclipse Tribid Mass spectrometer for EThcD-MS experiments. The HCD-MS experiments involved 30 samples per day on a 15 cm PepSep column from Bruker Daltonics, with data acquired at a resolution of 60,000 for 400–2000  $m/z$  precursors. A standard AGC target and maximum injection time set to Auto were used, along with an intensity threshold of 2.5e4 and charge states 2–8 included. Dynamic exclusion was set to 15 s, and MS/MS fragmentation with fixed 30% HCD at a resolution of 7500. On the other hand, EThcD-MS experiments were carried out on an Orbitrap Eclipse Tribid Mass spectrometer connected to an Evosep with similar settings as the HCD-MS experiments but using ETD for fragmentation with EThcD settings and fixed energy at 55. MS/MS spectra were acquired at 7500 resolution with a maximum injection time of 50 ms.

### Recombinant expression

Fab Sequences were sent to Sino Biologicals for protein expression using human IgG1 Fc as a backbone. The target gene was amplified, inserted into an expression vector, confirmed through sequencing, and passed on to downstream processes. Expression was performed using HEK293 cells for transient transfection in a serum-free medium. After 6 days of culture, proteins were purified using an affinity purification protein G column and analyzed by SDS-PAGE.

### Surface plasmon resonance, SPR

Affinity was carried out with RBD immobilized on the OpenSPR-XT instrument (Nicoya Lifesciences). RBD was immobilized in the active channel using the standard EDC/NHS amine coupling chemistry for amine coupling to high-capacity carboxyl sensors and 1 M Tris-based solution for reaction quenching (Supplementary Fig. 2A). The immobilization setup provided sufficient RBD levels (3150 RU). In contrast, PBS with 0.05% Tween20 was used as the running buffer to minimize non-specific binding (Supplementary Fig. 2B). A regeneration screen identified 10 mM Glycine-HCl pH 1.5 as a universal regeneration solution for subsequent analysis. Screening of a nanomolar concentration range for antibody analytes identified 100 nM as the optimal concentration for evaluating all candidates. Normalized sensorgrams for each concentration set were overlaid for comparative analysis (Supplementary Fig. 2C), and measurement of complex stability was plotted as immediately post-injection (early stability) vs 30 s after (late stability). The same experimental setup performed an off-rate screen with 300 s contact time and 600-s dissociation of 100 nM analytes normalized for their relative binding responses. It ranked based on relative dissociation percentages (Supplementary Fig. 2D).

### ELISA

Indirect ELISAs were performed on patient plasma and antibodies (i.e., PD124 polyclonal and mAbs sequenced from PD124), respectively. For the ELISA performed on plasma (Fig. 2A), 0.1  $\mu\text{g}$  of recombinant RBD (ExonBio) was immobilized on Maxisorp 96-well

plates (Thermo-Fischer) overnight at 4 °C (2  $\mu\text{g}/\text{mL}$  in sodium carbonate-bicarbonate buffer pH 9) whereas 0.2  $\mu\text{g}$  of recombinant RBD was immobilized to plate for the recombinant mAb ELISA (Fig. 4A). For the negative controls in ELISA testing of the recombinant antibodies, we used human IgG antibodies purchased from Sigma Aldrich (I4506) (Fig. 4A–C). Human plasma was obtained from Innovative Research Inc (IPLASK2E50ML, single donor human plasma with K2 EDTA) for testing the titer shown in Fig. 2A. For the positive controls, we used the following antibodies: Anti-SARS-CoV-2 Spike RBD Neutralizing Antibody, Human IgG1 AS35 (Acro Biosystems, clone AS35, lot no. S35-21CHF1-ZS) (see Fig. 4B, C), and Anti-Spike RBD Neutralizing Antibody, S1N-M122 (Clone AM122) from Acro Biosystems (see Fig. 4D). Uncoupled antigen was removed by washing wells three times with 200  $\mu\text{L}$  of 0.03% Tween in PBS (PBS-T), followed by blocking with SuperBlock buffer (Thermo-Fischer). Plasma or recombinant mAbs were serially diluted with Stabilzyme diluent, then 100  $\mu\text{L}$  was added to wells in duplicate and incubated for 1 h at room temperature. Uncoupled antibodies were removed by three washes with 200  $\mu\text{L}$  PBS-T. Goat anti-human IgG Fc-horseradish peroxidase-conjugated secondary antibody (Sigma Aldrich, catalog no. A0170, lot no. 0000181901) was diluted 1:5000, then 100  $\mu\text{L}$  was added per well and incubated for 1 h at room temperature. The unbound secondary antibody was removed by washing three times with PBS-T, then 100  $\mu\text{L}$  3,3', 5,5' tetramethylbenzidine (TMB) substrate (Thermo-Fischer) was added and allowed to react for 3 min, followed by quenching with 100  $\mu\text{L}$  0.2 N sulfuric acid (Sigma Aldrich) to stop the reaction. The plates were read in a spectrophotometer (Biotek Synergy LX) at 450 nm, and the average OD450 was plotted for each sample to generate the sigmoid binding curve.

### ACE2 competition binding assay

As described above, 0.2  $\mu\text{g}$  of recombinant RBD (ExonBio) was immobilized on Maxisorp 96-well plates (Thermo-Fischer) overnight at 4 °C. After washing and blocking, wells were incubated with serially diluted recombinant antibodies. Positive and negative controls included Sars-CoV2 Spike RBD neutralizing antibody AS35 (Acro Biosystems, clone AS35, lot no. S35-21CHF1-ZS) and human IgG from serum (Sigma Aldrich), respectively. Biotinylated ACE2 (Sino Biologicals) (70 ng/mL) and streptavidin-HRP conjugate (Millipore) (16.67 ng/mL) were added sequentially. TMB substrate (Thermo-Fisher) was used, followed by quenching with 0.2 N sulfuric acid (Sigma Aldrich). Plates were read at 450 nm on a Biotek Synergy LX spectrophotometer to generate the neutralization curve based on the average OD450 for each sample. The positive control is AS35, an Anti-SARS-CoV-2 Spike RBD Neutralizing Antibody, Human IgG1.

### In vitro, cell-based neutralization assay

For the in vitro, cell-based neutralization assay, HEK293/Human ACE2 Stable Cell Line (Acro Biosystems) was cultured in complete DMEM medium (Shanghai BasalMedia Technologies) supplemented with 10% fetal bovine serum (VivaCell) at 37 °C with 5% CO<sub>2</sub>. Recombinant antibody samples were serially diluted in a complete DMEM medium and added to a 96-well plate. Pseudovirus SARS-CoV-2 Spike (WT) (PSSW-HLGB001, Acro Biosystems) was diluted and added to the plate. For the cell control group, 25  $\mu\text{L}$  of the complete DMEM medium was added instead. After 60 min of incubation, cell density was adjusted, and cells were seeded into the plate for a 48-h incubation. A detection reagent (Britelite plus Reporter Gene Assay System) was added to each well. The positive control is S1N-M122, a known anti-Spike RBD Neutralizing Antibody, Chimeric mAb, Human IgG1 (Clone AM122), sourced from Acro Biosystems.

### Luminescence measurement

The luminescence value of each well in the plate was measured using a microplate reader. The detection time for each well was set to

0.1 s/well.

$$\text{Inhibition rate} = \left(1 - \frac{X - \overline{CC}}{\overline{VC} - \overline{CC}}\right) \times 100\%$$

$X$ , the luminescence value (RLU) of a given well.

$\overline{CC}$ , cell control, only adds cells.

$\overline{VC}$ , virus control, and only cells and pseudovirus are added.

$\overline{CC}$ , the mean value of the cell control group.

$\overline{VC}$ , the mean value of the virus control group.

The data were analyzed using GraphPad Prism 8 software (Non-linear regression). IC50 or relative IC50 is defined as the incubation concentration of a sample required to bring the curve down to the point halfway between the top and bottom plateaus of the curve. Absolute IC50 is defined as the incubation concentration of a sample that provokes 50% inhibition.

## Data analysis

**IgSeq data processing.** The protein database for searching the MS data was generated by translating the merged R1 and R2 reads from the Illumina MiSeq run using the human IMGT germline database as a reference. After translation, identical protein sequences were merged into unique sequences.

**MS-based database searching.** Novor.Cloud (<http://novor.cloud>) was used to match the mass spectrometry data to the antibody sequence database that is translated from the NGS reads. The matching antibody sequences were sorted by their MS coverage in CDR regions and total sequence coverage, and further organized by clustering based on their CDR3 sequences.

**De novo protein sequencing.** De novo antibody protein sequencing is performed at three different scale levels: (1) Contig assembly, (2) within-chain CDR pairing/chain assembly, and (3) heavy-light chain pairing.

**Scale 1, contig assembly.** The contig assembly consists of five steps. Step 1 - de novo peptide sequencing: The MS/MS spectra were de novo sequenced using Novor software to obtain peptide sequences<sup>44</sup>. Additionally, Novor assigns a confidence score for each amino acid within a given sequence. Step 2 - finding the maximum mass block overlap: The algorithm computes the maximum “mass block” overlap between every possible pair of peptides. This involves considering amino acid permutations within a block (e.g., [DF] = [FD]) and isobaric similarities (e.g., [I] = [L], [N] = [GG], [Q] = [GA] = [AG]) and direct similarity [A] = [A]. This concept is best illustrated with an example. Suppose LGAGYDFNH and FDGGHWYQQL are two peptides. The underlined sub-sequences can be divided into blocks [DF][N][H] and [FD][GG][H], respectively. Since mass [DF] = mass [FD], mass [N] = mass [GG], and mass [H] = mass [H], the two peptides overlap by 3 mass blocks. The two overlapping peptides can be merged to form a longer sequence LGAGY $\beta_1\beta_2$ HWYQQL where  $\beta_1$  = DF or FD, and  $\beta_2$  = N or GG are the mass blocks being considered. The one with the highest average Novor amino acid confidence score is used between the two sequence choices of each block. Step 3 - overlap graph construction: To construct the overlap graph, each de novo peptide with a Novor score of 75 or higher corresponds to a vertex/node. An edge is added between each pair of vertices if their peptides overlap by at least 3 mass blocks. Step 4 - contig construction: For each path consisting of  $k$  ( $k \leq 3$ ) vertices in the graph, a contig is constructed by merging the peptides of all vertices on the path. Each contig is then aligned with its best matching germline gene sequence. Contigs that completely cover a CDR region (including two additional amino acids on both sides of the CDR) are retained. The Novor.Cloud engine (available online at <http://novor.cloud>) was used to search the MS/MS spectra in the

contig database. The contigs were grouped by CDR region sequences and sorted according to their database search scores. Step 5 - manual human inspection: The grouped contigs were subjected to human inspection to select a short list of CDR region sequences and their corresponding contigs. Generally, longer contigs, with more PSM coverage and longer peptide overlaps, were given higher priority. Subtle de novo peptide sequencing errors were also corrected during human inspection.

**Scale 2, CDR pairing and chain assembly.** The assembled and selected CDR contigs were subjected to pairing analysis. This process is analogous to the heavy-light chain pairing described below, but with contigs replacing the chains. After computing the similarity scores for all pairs of CDRs, pairs that showed high similarity scores to each other, but relatively low similarity scores to other CDRs, were selected and regarded as being from the same antibody chain. Clusters of high-scoring CDR1, CDR2, and CDR3 pairs were assembled by aligning them to a germline gene. Any gaps in the framework region were filled with either overlapping de novo peptides or germline gene sequences to obtain a full chain sequence. Additionally, long peptides spanning two CDR regions were identified by database searching, and disulfide bridge-containing peptides from non-reduced protease digestion were identified using pLink software<sup>45</sup>. These peptides crossing two CDR regions were utilized to help resolve ambiguities in the quantification-based pairing.

**Scale 3, heavy-light chain pairing.** Novor.Cloud was used to search the MS/MS data for the heavy and light chain sequences to identify PSMs. MaxQuant software (Ver2.1.3) calculated the peak area of each peptide from each fraction in NRT and native gel separation experiments. For a given peptide with peak areas  $y_1, y_2, \dots, y_k$  in the  $k$  fractions of a separation experiment and  $\alpha = \sum_{i=1}^n y_i$ , the normalized quantification vector of the peptide is calculated by  $y = (\frac{y_1}{\alpha}, \frac{y_2}{\alpha}, \dots, \frac{y_k}{\alpha})$ . The co-existence of two different peptides being part of the same antibody was determined using the Pearson correlation coefficient between their normalized quantification vectors. The similarity score between two chains was computed as the average similarity between every pair of unique peptides from the two chains. In cases of multiple separation experiments, the similarity scores from each experiment were summed to obtain the final similarity score. After computing the scores for all pairs of heavy and light chains, pairs with high similarity scores to each other and relatively low similarity scores to other chains were selected and regarded as originating from the same antibody. A Heavy-Light chain pairing matrix can be found in Supplementary Fig. 1B.

**Isobaric ambiguity resolution.** To resolve isobaric ambiguities between leucine and isoleucine using Electron Transfer/Higher-energy Collisional Dissociation (ET<sub>h</sub>CD) mode, z-ions were identified via Novor de novo protein sequencing or database matching. Once isobaric ambiguity was confirmed, specific w-ions indicative of leucine or isoleucine were manually assessed by examining the presence or absence of w-ion peaks unique to either residue. Specific w-ion peaks confirmed the identity of leucine or isoleucine at distinct positions, thereby resolving the isobaric ambiguity in the sequence.

**Non-reduced samples digestion analysis.** The identification of the S-S bridge-containing peptides from non-reduced protease digestion was analyzed using pLink<sup>45</sup>. The software identification parameters were set as follows: Flow Type: Disulfide bond (HCD-SS); Enzyme: LysC-AspN; Peptide mass: 300-9000 Da; Peptide length: 3-90; Fixed modification: Gln->pyro-Glu; and Variable modifications: Nethylmaleimide[C], Oxidation[M], Deamidated[N], Deamidated[Q], and Acetyl [ProteinN-term]. The protein database (Supplementary Table 2) contained the sequences of interest deduced from the polyclonal antibody analysis.



## Statistics & reproducibility

No statistical methods were employed to predetermine sample size, and no data were excluded from the analyses. The experiments were not randomized, and investigators were not blinded to allocation during the experiments or outcome assessment. Due to limited availability of material, experiments using the serum pAb from Subject 2 were carried out only once. A benchmark experiment with a mixture of five mAbs was used to demonstrate the reproducibility of similar results.

## Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The MS raw files related to the human polyclonal sequencing effort data generated in this study have been deposited in the MASSIVE database under FTP://MSV000093480@massive.ucsd.edu; [<https://doi.org/10.2534/CSFT8DWIZ>]. The MS raw files related to the Benchmark sequencing effort data generated in this study have been deposited in the MASSIVE database under ftp://MSV000094953@massive.ucsd.edu [<https://doi.org/10.2534/C58K7576M>]. The IgSeq sequence reads are available at NCBI Biosample accession [SAMN38474421](https://www.ncbi.nlm.nih.gov/biosample/SAMN38474421) Source data are provided with this paper.

## Code availability

The source code for the contig assembly program has been made available at <https://github.com/rnipb/ContigAssembly>.

## References

- Lu, R. M. et al. Development of therapeutic antibodies for the treatment of diseases. *J. Biomed. Sci.* **27**. <https://doi.org/10.1186/s12929-019-0592-z> (2020).
- Hjelm, B., Forsström, B., Löfblom, J., Rockberg, J. & Uhlén, M. Parallel immunizations of rabbits using the same antigen yield antibodies with similar, but not identical, epitopes. *PLoS ONE* **7**, e45817 (2012).
- Köhler, G. & Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* **256**, 495–497 (1975).
- Wilson, P. C. & Andrews, S. F. Tools to therapeutically harness the human antibody response. *Nat. Rev. Immunol.* **12**, 709–719 (2012).
- Frenzel, A., Schirrmann, T. & Hust, M. Phage display-derived human antibodies in clinical development and therapy. *mAbs* **8**, 1177–1194 (2016).
- Pedrioli, A. & Oxenius, A. Single B cell technologies for monoclonal antibody discovery. *Trends in Immunol.* **42**, 1143–1158 (2021).
- Cheung, W. C. et al. A proteomics approach for the identification and cloning of monoclonal antibodies from serum. *Nat. Biotechnol.* **30**, 447–452 (2012).
- Sato, S. et al. Proteomics-directed cloning of circulating antiviral human monoclonal antibodies. *Nat. Biotechnol.* **30**, 1039–1043 (2012).
- Lavinder, J., Wine, Y., Boutz, D., Marcotte, E. & Georgiou, G. Proteomic identification of antibodies. Patent No.: US 10, 175, B2 (2019).
- Georgiou, G. et al. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* **32**, 158–168 (2014).
- Tiller, T. et al. A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties. *MAbs* **5**, 445–470 (2013).
- França, R. K. A. et al. Progress on phage display technology: tailoring antibodies for cancer immunotherapy. *Viruses* **15**, 1903 (2023).
- Doevendans, E. & Schellekens, H. Immunogenicity of innovative and biosimilar monoclonal antibodies. *Antibodies* **8**:21 (2019).
- Zost, S. J. et al. Potently neutralizing and protective human antibodies against SARS-CoV-2. *Nature* **584**, 443–449 (2020).
- Westendorf, K. et al. LY-CoV1404 (bebtelovimab) potently neutralizes SARS-CoV-2 variants. *Cell Rep.* **39**, 110812 (2022).
- Georgiev, I. S. et al. Delineating antibody recognition in polyclonal sera from patterns of HIV-1 isolate neutralization. *Science* **340**, 751–756 (2013).
- Chaudhary, N. & Wesemann, D. R. Analyzing immunoglobulin repertoires. *Front. Immunol.* **9**. <https://doi.org/10.3389/fimmu.2018.00462> (2018).
- Guthals, A. et al. De novo MS/MS sequencing of native human antibodies. *J. Proteome Res.* **16**, 45–54 (2017).
- Chen, J. et al. Proteomic analysis of pemphigus autoantibodies indicates a larger, more diverse, and more dynamic repertoire than determined by B cell genetics. *Cell Rep.* **18**, 237–247 (2017).
- Frank, A. & Pevzner, P. PepNovo: De novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **77**, 964–973 (2005).
- Ma, B. et al. PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **17**, 2337–2342 (2003).
- Liu, K., Ye, Y., Li, S. & Tang, H. Accurate de novo peptide sequencing using fully convolutional neural networks. *Nat. Commun.* **14**, 7974 (2023).
- Eloff, K. et al. De novo peptide sequencing with InstaNovo: accurate, database-free peptide identification for large scale proteomics experiments. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.08.30.555055>.
- Yilmaz, M., Fondrie, W. E., Bittremieux, W., Oh, S. & Noble, W. S. De novo mass spectrometry peptide sequencing with a transformer model. *39th Int. Confer. Mach. Learn.* **162**, 25514–25522 (2022).
- Bandeira, N., Pham, V., Pevzner, P., Arnott, D. & Lill, J. R. Automated de novo protein sequencing of monoclonal antibodies. *Nat. Biotechnol.* **26**, 1336–1338 (2008).
- Liu, X. et al. De Novo protein sequencing by combining top-down and bottom-up tandem mass spectra. *J. Proteome Res.* **13**, 3241–3248 (2014).
- Peng, W., Pronker, M. F. & Snijder, J. Mass spectrometry-based de novo sequencing of monoclonal antibodies using multiple proteases and a dual fragmentation scheme. *J. Proteome Res.* **20**, 3559–3566 (2021).
- Bondt, A. et al. Into the dark serum proteome: personalized features of IgG1 and IgA1 repertoires in severe COVID-19 patients. *Mol. Cell Proteomics* **23**. <https://doi.org/10.1016/j.mcpro.2023.100690> (2023).
- Thevis, M., Ogorzalek Loo, R. R. & Loo, J. A. In-gel derivatization of proteins for cysteine-specific cleavages and their analysis by mass spectrometry. *J. Proteome Res.* **2**, 163–172 (2003).
- Sung, W. C. et al. Evaluation of disulfide scrambling during the enzymatic digestion of bevacizumab at various pH values using mass spectrometry. *Biochim Biophys. Acta Proteins Proteom.* **1864**, 1188–1194 (2016).
- Zhokhov, S. S., Kovalyov, S. V., Samgina, T. Y. & Lebedev, A. T. An EThcD-based method for discrimination of leucine and isoleucine residues in tryptic peptides. *J. Am. Soc. Mass Spectrom.* **28**, 1600–1611 (2017).
- Säfsen, P., Klakamp, S. L., Drake, A. W., Karlsson, R. & Myszkowski, D. G. Screening antibody-antigen interactions in parallel using Biacore A100. *Anal. Biochem.* **353**, 181–190 (2006).
- Murray, J. B., Roughley, S. D., Matassova, N. & Brough, P. A. Off-rate screening (ORS) by surface plasmon resonance. An efficient method to kinetically sample hit to lead chemical space from unpurified reaction products. *J. Med. Chem.* **57**, 2845–2850 (2014).
- Longmire, R. L. et al. In vitro splenic igg synthesis in hodgkin's disease. *N Engl J Med.* **289**, 763–767 (1973).



35. He, B. et al. Rapid isolation and immune profiling of SARS-CoV-2 specific memory B cell in convalescent COVID-19 patients via LIBRA-seq. *Signal Transduct Target Ther.* **6**, 195 (2021).
36. Rouet, R. et al. Broadly neutralizing SARS-CoV-2 antibodies through epitope-based selection from convalescent patients. *Nat. Commun.* **14**. <https://doi.org/10.1038/s41467-023-36295-5> (2023).
37. Bondt, A. et al. Human plasma IgG1 repertoires are simple, unique, and dynamic. *Cell Syst.* **12**, 1131–1143.e5 (2021).
38. Liu, X., Han, Y., Yuen, D. & Ma, B. Automated protein (re)sequencing with MS/MS and a homologous database yields almost full coverage and accuracy. *Bioinformatics* **25**, 2174–2180 (2009).
39. Le Bihan, T., Robinson, M. D., Stewart, I. I. & Figgeys, D. Definition and characterization of a 'trypsinosome' from specific peptide characteristics by Nano-HPLC-MS/MS and in silico analysis of complex protein mixtures. *J. Proteome Res.* **3**, 1138–1148 (2004).
40. Li, Y. F., Arnold, R. J., Tang, H. & Radivojac, P. The importance of peptide detectability for protein identification, quantification, and experiment design in MS/MS proteomics. *J. Proteome Res.* **9**, 6288–6297 (2010).
41. Bache, N. et al. A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics. *Mol. Cell Proteomics.* **17**, 2284–2296 (2018).
42. Mann, M., Hendrickson, R. C. & Pandey, A. Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* **70**, <https://doi.org/10.1146/annurev.biochem.70.1.437> (2001).
43. Krusemark, C. J., Frey, B. L., Smith, L. M. & Belshaw, P. J. Complete chemical modification of amine and acid functional groups of peptides and small proteins. *Methods Mol. Biol.* **753**, 77–91 (2011).
44. Ma, B. Novor: real-time peptide de novo sequencing software. *J. Am. Soc. Mass Spectrom.* **26**, 1885–1894 (2015).
45. Lu, S. et al. Mapping native disulfide bonds at a proteome scale. *Nat. Methods* **12**, 329–331 (2015).
46. Raybould, M. I. J., Kovaltsuk, A., Marks, C. & Deane, C. M. CoV-AbDab: the coronavirus antibody database. *Bioinformatics* **37**, 734–735 (2021).

## Acknowledgements

The authors are grateful to Jenna Kerry for her assistance with figure conversion, to Cindy Xu for running some gels, and to Kyle Suttill for ensuring smooth operation within the lab. All authors' research is entirely supported by the internal research funding from Rapid Novor Inc.

## Author contributions

E.B., V.L., M.J., R.D., A.L.C., J.D., C.R., M.P., and X.H., performed research, reviewed and edited the paper; B.M., designed research, analyze data,

wrote the paper; T.L.B. designed research, analyzed data, wrote the paper; T.N.D.V.D., Q.L., and L.W. analyzed data, reviewed and edited the paper.

## Competing interests

The authors disclose the following competing interests: B.M. and Q.L. have equity interests in Rapid Novor, a company that may potentially benefit from the research outcomes. The remaining authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-53105-8>.

**Correspondence** and requests for materials should be addressed to Bin Ma.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024