# Figures

## **Figure 1**

A diagram of a graph

Description automatically generated with medium confidence

### **Modular sequencing can be used to limit the search space**

**a)** Schematic of data-integration workflow. The approach consists of three stages in which increasingly large sequence segments are sequenced and then used as input for the next stage. Initially, only the framework regions (FRs) are sequenced. Then the FR candidates are converted into extended FR-CDR-FR candidates (*i.e.*, CDRs with adjacent FRs). Finally, the FR-CDR-FRs are recombined into full chain candidates. Each stage follows a flow starting with sequence candidate generation based on input data (“Generate”). These candidates are scored using multiple data streams (“Integrate”), then the best candidates are selected using these scores (“Evaluate”). **b)** Size of the search space throughout the workflow. The approximate number of candidates is shown in the modular approach (pink) *versus* processing the whole sequence at once (teal). By first sequencing smaller segments, the search space can be kept relatively small. The segment candidate pool is expanded at the start of the stage and reduced after scoring. This ensures we never consider more than ~103 segment candidates at the same time, keeping computational cost in balance.

## **Figure 2**



### Increasing complexity leads to an increase in ambiguity in mass determination in middle-down proteomics of antibodies.

LC-trace (top panel), zero-charge deconvoluted mass spectrum (middle panel), and averaged MS2 spectrum (bottom panel). **a)** Pure Trastuzumab with clear signals for the light chain and the Fd chain. **b)** Mixture of Trastuzumab and two other monoclonal antibodies. **c)** Polyclonal sample from plasma. The contributions of the target and paired chains diminish relatively when the background becomes more complex.

## **Figure 3**

****

## Sequencing higher complexity samples lead to a loss of fidelity in sequencing when using **only shotgun proteomics data**.

Sequencing results for the monoclonal (panel a-b) and mixture of 3 mAbs (panel c-d) samples are shown. Each sample was submitted to Stitch twice: as a *template selection* run (light blue, panel a and c) and a *definitive* run (dark blue, panel b and d). Each panel shows residue candidates as letters and depth of coverage as bars. The monoclonal sample had no coverage below the cut-off (pink highlight), very few erroneous residue candidates (grey residues), ambiguity, or errors in the consensus sequence (marked with *“x”*). The mixture of 3 mAbs sample had substantially more stretches of sequence with low coverage (pink bars), resulting in more ambiguity, erroneous candidates, and errors.

## **Figure 4**



### **Sequencing an abundant IgA1 heavy chain in a polyclonal sample converges on a single sequence prediction.**

The sequencing process of the target chain in the polyclonal sample is shown. Residue candidates per position are shown in sequence logos, with rejected residue candidates in grey. Below the sequence logos the number of candidates at the start and end of the stage is shown. **a)** The selected germline template, IGHV3-33\*06, is shown along with the deviations from the final sequence. **b)** In the FR sequencing stage, we reduced all FR candidate pools to 4 candidates or less despite starting with large pools of FR3 and FR4 candidates. **c)** During the CDR sequencing stage, we converged on a single CDR1 and CDR2 candidate, thereby also rejecting the remaining incorrect FR1 and FR3 candidates. The only remaining ambiguity was in between two isobaric CDR3 sequences. **d)** Recombining the remaining FR candidates into chain sequences yielded 975 chain candidates. Two of these contained the previously selected CDRs. These two candidates were isobaric, had highly similar Shotgun-scores and fully overlapping fragment coverage. **e)** middle-down fragment coverage for the final sequence (constant region not shown).

# Supplementary figures

## Table S1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | Monoclonal | Mix | Polyclonal |
| Middle down | MS1 | RT window start (min) | 12.9 | 13 | 10.3 |
| RT window end (min) | 13.4 | 13.5 | 11.6 |
| Selected mass (Da) | 24052.28 | 24053.18 | 24811.17 |
| Target mass delta (Da) | -0.6 (26.5 PPM) | 0.3 (10.8 PPM) | NA |
| MS2 | RT window start (min) | 11.8 | 12.5 | 10.3 |
| RT window end (min) | 13.4 | 13.4 | 11.6 |
| No. peaks | 919 | 265 | 469 |
| No. averaged scans | 71 | 22 | 24 |
| Bottom up | | No. raw files | 4 | 8 | 8 |
| Protease 1 | Trypsin | Trypsin | Trypsin |
| Protease 2 | Chymotrypsin | Chymotrypsin | Chymotrypsin |
| Protease 3 | Thermolysin | Thermolysin | Thermolysin |
| Protease 4 | Elastase | Elastase | Pepsin |
| No. Peptides | 14000 | 27421 | 35003 |

### Overview of input data

Middle-down: The MS1 section shows the retention time window over which MS1 scans were averaged before deconvolution to obtain the target precursor mass, the selected mass, and the deviation of that selected mass from the known target mass. The MS2 section shows the retention time window over which MS2 scans were averaged before deconvolution to obtain the fragment masses, how many scans were averaged to achieve, and how many fragment masses were obtained. Bottom-up: The bottom-up section of the table shows the number of raw files that were used as input, which proteases were used for digestion and the number of peptides that resulted from de novo peptide sequencing using PEAKS.

## **Figure S1**

A screenshot of a computer program

Description automatically generated with low confidence

### **Framework region candidate generation**

Candidate FR sequences for each framework region are generated from residue frequency tables from three sources: The definitive Stitch run, the template selection Stitch run and the IMGT (from highest to lowest priority respectively). Residue candidates are selected based on their relative frequency. For each position, residue candidates from the next frequency table are only taken if the depth of coverage in the current table is lower than the depth of coverage at the highly conserved Cysteine 104. After residue candidates have been selected for all positions, all permutations of these candidates are taken for each FR to yield sequence candidates for these FRs.

## **Figure S2**

Timeline

Description automatically generated

### Schematic of the **sliding window fragment matching algorithm**

The sliding window fragment matching algorithm finds the optimum mass offset for an imperfect subsequence for a given fragmentation spectrum (FR2 in the figure). Theoretical fragments are generated at an approximate offset and shifted by a predefined increment (default: 0.01 Da) throughout a predefined range (default: starting position plus and minus 190 Da)). This enables error-tolerant scoring of subsequences and determination of the prefix- and suffix- (N- and C-terminal) masses.

## Table S2

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Monoclonal | | Mix | | | Polyclonal | | |
| Start | **End** | Start |  | **End** | Start |  | **End** |
| FR | **FR1** | 8 | **1** | 240 | 40 | **40** | 8 | 4 | **2** |
| **FR2** | 1 | **1** | 756 | 90 | **7** | 2 | 1 | **1** |
| **FR3** | 1 | **1** | 5 | 1 | **1** | 384 | 14 | **3** |
| **FR4** | 1 | **1** | 4 | 2 | **2** | 64 | 12 | **4** |
| FR-CDR-FR | FR1-**CDR1**-FR2 | 10 | **1** | 1106 |  | **1** | 20 |  | **1** |
| FR2-**CDR2**-FR3 | 10 | **1** | 49 |  | **1** | 30 |  | **1** |
| FR3-**CDR3**-FR4 | 10 | **1** | 20 |  | **3** | 120 |  | **2** |
| Chain | **Chain** | 930 | **1** | 616 |  | **3** | 975 |  | **2** |

### Number of segment candidates throughout the workflow

The table shows the number of segment candidates at the start and end of each stage. For the mix and polyclonal FR sequencing stage, a middle column is included which displays the number of candidates after a “first pass” filtering, for example excluding candidates that do not have highly conserved residues (Cys23 and Cys104 specifically), or candidates with highly unlikely terminal mass offsets.

## Figure S3



### BU- and MD- based scoring and germline-based selection criteria enable effective segment candidate selection throughout the workflow.

Score distributions (MD-score (x axis) and Shotgun-score (y axis)) for all considered segment candidates are shown. Each column depicts segment candidates for a sample.Each graph represents segment candidate scores for a target segment. Each dot in the graphs represents a rejected candidate, whereas stars indicate selected candidates. A yellow outline highlights the correct candidate. The correct candidate was selected in all stages (except for the CDR3 and chain for the polyclonal sample, where it was not present. However, ambiguity could not be fully resolved everywhere. A legend with the consecutive selection criteria is shown at the top of the figure. Grey: Input candidate. Red: First pass criteria met (default criterium: Has conserved Cys23 or Cys104). Blue: Preliminary selection criterium met (default criterium: MD-score was <5 removed from the maximum in the pool). Green: Selected for the next stage (default criterium: Top Shotgun-score). Any additional/alternative criteria are shown in the graph itself (e.g., for the mix FR4 candidate pool, the first pass filtering criterium included the presence of W118). The number of candidates satisfying each criterium is given at the top of each graph. E.g., for the monoclonal, 8 FR1 candidates were considered in total. Of these 8, only 4 had the highly conserved Cys23. Of these 4, only 2 were considered top candidates based on their MD-score. A single candidate was finally selected for the next stage.

## **Table S3**

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Usage** |
| Shotgun-score | This score quantifies the number and length of unbroken subsequences exactly matching the prediction. Each predicted residue is scored by counting the number of subsequences (of length >= 6) in the BU reads multiplied by the length of that subsequence, thereby placing a higher weight on longer matches. The sum of these scores represents the candidate score. Several local metrics such as (weighted) depth of coverage, average supporting read length are calculated for each predicted residue and can be used during manual inspection. | Scoring all segment candidates against BU data |
| MD-score | Probability-based fragmentation score at the best scoring prefix mass. This score quantifies the overlap between the theoretical fragment masses for a given (sub)sequence and the recorded MD spectrum. The score is based on the likelihood of a random match, which is calculated based on the number of detected peaks vs the number of annotated peaks (Olsen & Mann, 2004). | Scoring all segment candidates against MD fragmentation data |
| Local BU-score | This score quantifies the support by PEAKS *de novo* reads. It is calculated by summing the number of reads that have a residue matching the prediction at the considered position. Positions for the prediction are numbered according to the IMGT numbering convention and reads are aligned to the prediction during manual inspection or CDR gap filling, or a consensus sequence based on the provided reads/templates/contaminants during initial scoring. | Internally scoring CDR candidates against BU data (before they are rescored as FR-CDR-FR candidates and manually inspected) |
| Germline score | This score quantifies the support by the provided germline sequences in the IMGT. It is calculated by aggregating the fraction of germline sequences that have a residue matching the prediction residue. | Manual inspection and tiebreaking |
| Template score | This score quantifies the support by the provided template(s).It is calculated by summing the number of templates that have a residue matching the prediction. | Manual inspection and tiebreaking |

### **Different scores calculated throughout the workflow**.

The Template- and Germline- score are not used in the manuscript. The local BU-score is only used for internal ranking before CDR rescoring as the CDR-candidate are too short for the Shotgun-score.

## Figure S4



### **CDR candidate generation**

Candidate CDR sequences are generated by joining a pair of adjacent FR candidates (e.g. a FR1 candidate and a FR2 candidate) using overhanging reads from both FRs. **a)** Reads containing the 3 CDR flanking residues are taken for the left (e.g. FR1) and right FR (e.g. FR2). **b)** The N-terminal overhanging reads (left\_overhangs) are then combined with the C-terminal overhanging reads (right\_overhangs), generating all possible combinations. **c)** The mass of the CDR is calculated using the FR candidates and the s*liding window score* (Figure S2) and used to filter the CDR candidates, retaining only those matching the target mass within a 5 Da tolerance.