# Figures

## **Figure 1**



### Increasing complexity leads to an increase in ambiguity in LC-MS/MS experiments.

LC traces (top panel), MS1 trace for a chain (middle panel), and MS2 trace for a selected and fragmented chain (bottom panel). **(a)** Low complexity sample containing Trastuzumab. **(b)** Mix sample containing XXX. **(c)** Polyclonal sample from plasma.

## **Figure 2**

**Timeline

Description automatically generated**

### **Loss of sequence extraction fidelity for higher complexity samples using only shotgun proteomics data**

The outcome of a two-stage shotgun sequencing approach using STITCH is shown for the monoclonal (panel a-b) and mixed sample (panel c-d). Each panel (a-d) shows the residue candidates (letters, *green: correct, grey: incorrect*) and depth of coverage (bars, *grey: sufficient coverage, pink: low coverage*) per position. The light blue panels (a and c) show results from the initial *template selection* runs, where high scoring reads are aligned to germline templates from the IMGT. The blue panels (b and d) depict results from the *definitive* runs, where all reads are aligned to a selection of consensus sequences from the *template selection* run. For the monoclonal sample (panel a and b), there are no residues that fall below the lower coverage limit, resulting in very few ambiguous positions (i.e., positions with more than one residue candidate), and very few errors in the consensus sequence. All of the latter are limited to the CDRs (marked with “x”). This excellent performance is however not retained in the mixture of 3 mAbs sample (panel c and d). Stretches of sequences fall below the coverage cut-off, which results in erroneous predictions in the initial *template selection* run (panel c) and ambiguity in the *definitive* run (panel d) (again marked with “x”).

## **Figure 3**



### **Modular sequencing limits the search space**

A) Schematic of the generalized workflow that consists of three stages in which increasingly large sequence segments are sequenced and then used as input for the next stage. First the framework regions (FRs) are sequenced. Then the FR candidates are expanded into 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 in which data are expanded into a set of sequence candidates (“Generate”). These candidates are scored using multiple data streams (“Integrate”), then filtered using these scores (“Evaluate”). By using a modular workflow, i.e., first sequencing smaller segments of the sequence, we limit the search space throughout the workflow. B) candidate pool size changes throughout the workflow, both for the entire chain (teal) and the considered segment (pink). After we select a germline template, we have prohibitively large chain candidate pool (~109), as we have many ambiguous residues, especially in the CDRs. On the other hand, the segment candidate pool is expanded at the start of every stage and then reduced to a manageable size after scoring. This ensures that throughout the workflow we never consider more than ~103 segment candidates at the same time, thereby keeping computational cost in balance.

## **Figure 4**



### **Sequencing an abundant IgA1 heavy chain in a polyclonal sample**

The full sequencing process for the most abundant IgA1 clone in the polyclonal sample is shown, which converges on 4 isobaric sequence candidates. For each sequence segment (FR, FR-CDR-FR, chain), residue candidates per position are shown as stacked letters, with rejected residue candidates in grey. Below these plots, the number of segment candidates at the start and end of each stage is given. On the right side of panels a-d, the number of possible combinations between the segment candidates at the start and end of the stage is given. A) selected germline template, IGHV3-33\*06, with the deviations from the final sequence. B) FR sequencing stage. Initially, there are nearly 400,000 permutations of FR candidates, largely due to large ambiguity in FR3 and FR4. Nearly all candidates werecould be rejected, leaving only 2, 1, 2, and 1 sequence candidate(s) for FR1, FR2, FR3 and FR4, respectively. C) CDR sequencing stage. Each FR-CDR-FR segment starts with ~20 candidates and the majority of these can be rejected. During this stage, one of the remaining FR3 candidates is also rejected. D) chain sequencing stage. Recombining the remaining FR candidates into chain sequences yields 1868 candidates, but only 4 of these candidates contain previously selected CDRs. These 4 candidates are isobaric, show highly similar scores and fully overlapping fragment coverage. E) middle-down proteomics fragment coverage of one of the selected chain candidates.

# Supplementary data

## 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 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. 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, determination of the N- and C-terminal suffix masses and offset from previous positions.

## **Figure S3**



### **FR candidate generation**

Candidate FR sequences for each framework region are generated from 3 residue frequency tables: The definitive stitch run, the template selection stitch run and the IMGT (from highest to lowest priority). What candidate residues are taken for each position is 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 Cysteine 104, a highly conserved residue which we have found to be the lowest covered residue consistently. After residue candidates have been selected for all positions, all permutations of these candidates are taken for each framework region, excluding gaps, to yield sequence candidates.

## Table S4

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | 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 sequence candidates throughout the workflow

The table shows the number of sequence candidates at the start and end (in bold) of each stage. For the Mix and Polyclonal FR sequencing stage, a middle column is included which displays the number of contigs 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 X



### **S**egment candidate score distributions throughout the workflow

Green circles highlight selected candidates for to the next stage. FR-CDR-FR points are coloured according to their most diverse flanking FR.Each column is a sample, monoclonal, mix and polyclonal respectively. Each row is a target segment, FR1-4, CDR1-3 and then chain respectively.

## **Table S5**

|  |  |  |
| --- | --- | --- |
| **Score name** | **Description** | **Usage** |
| Multiscore | This score quantifies the number and length of unbroken subsequences exactly matching the prediction. For each position, the score counts the number of subsequences (of length >= 6) in peptides \* the length of the subsequence, thereby placing a higher weight on longer matching subsequences. It does not account for K/Q and I/L. | Scoring relatively long (>20 residues) contigs against bottom-up data |
| Top-down fragmentation score at the best scoring offset termed spectrum score | This score quantifies the overlap between the theoretical spectrum for a contig and the recorded middle down spectrum. The score is a probability based score, where the likelihood of a random match is calculated based on the number of detected peaks vs the number of identified peaks (Olsen & Mann, 2004). | Scoring all contigs against MD fragmentation data |
| PEAKS support score termed peaks 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 bottom up data (before they are rescored as FR-CDR-FR contigs) |
| Template support 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 at the considered position. Positions for the templates are numbered according to the IMGT numbering convention and the prediction is numbered by aligning to the consensus sequence based on the provided reads/templates/contaminants during initial scoring. | Manual inspection |
| Background support score | This score quantifies the support by the provided germline sequences in the IMGT. It is calculated by summing the number of germline sequences that have a residue matching the prediction at the considered position. Positions for the germlines are numbered according to the IMGT numbering convention and the prediction is numbered by aligning to a consensus sequence based on the provided reads/templates/contaminants during initial scoring. | Manual inspection |

### **Different scores used in the algorithm**.

## Figure S6



### **CDR candidate generation**

Candidate CDR sequences are generated by joining a pair of adjacent FR candidates (e.g. an FR1 candidate and an FR2 candidate) using overhanging reads from both FRs (first panel). Reads containing the 3 flanking residues are taken for the left (e.g. FR1) and right FR (e.g. FR2). The N-terminal overhanging reads are then combined with the C-terminal (e.g. FR2) overhanging reads, generating all possible combinations (second panel). The mass of the CDR is calculated using the FRs and the s*liding window score* (Figure S4) and used to filter the CDR candidates, retaining only those matching the target mass within a 5 Da tolerance (third panel).