

Complete De Novo Assembly of Monoclonal Antibody Sequences

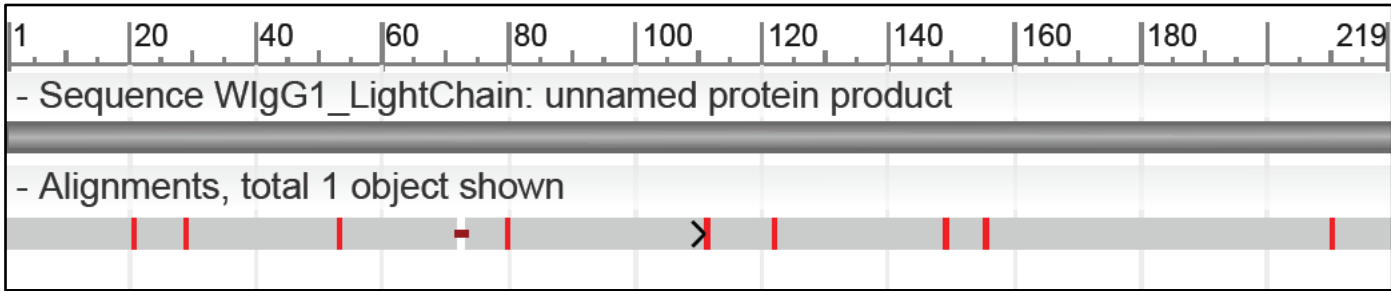
Ngoc Hieu Tran, M. Ziaur Rahman, Lin He, Lei Xin, Baozhen Shan, Ming Li

Supplementary Data

The supplementary data was submitted in a folder named “ALPS_Data” together with the manuscript. The folder “ALPS_Data” includes four subfolders for the four samples in our study. Each subfolder contains three csv files “peptides_0.csv”, “peptides_db.csv”, and “peptides_spider.csv”, corresponding to three peptides lists PSM-DN, PSM-DD, and PSM-DDS, respectively, in the manuscript. In addition, the subfolders also include the assembled contigs from three peptides lists and their positional confidence scores. The target light chains and heavy chains are also provided for evaluation. Finally, a java executable file is included in the folder “ALPS_Data” with detailed instructions how to perform the assembly on the peptides lists. The database that we used for the database and homology search is also provided.

Supplementary Figures

(A)

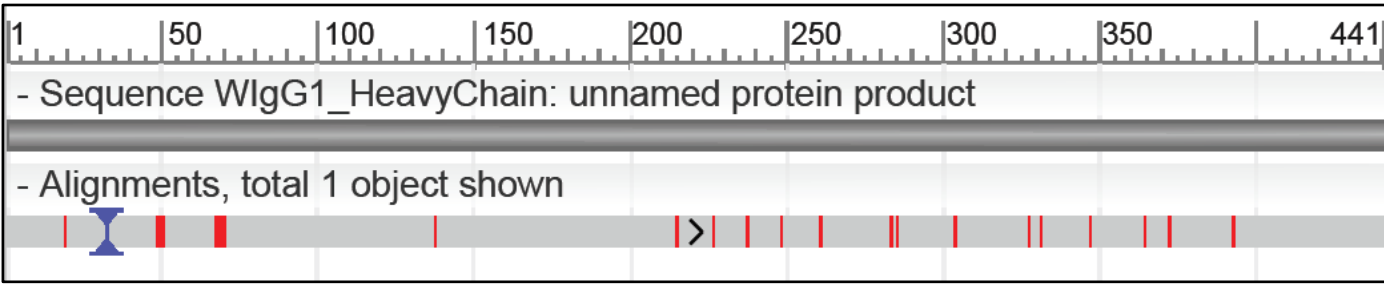


(B)

| | | | | | | | |
|-------|-----|----------------|----------|--------------|---------------------------|-------------------------|-----|
| Sbjct | 1 | DVLMTQTPLSLPVS | LGDAQAS | ISCRSSQYIVHS | NGNTYLEWYLQKPGQSPKLLIYKVS | NR | 60 |
| Query | 1 | DVLMTQTPLSLPVS | LGDAQAS | ISCRSSQYIVHS | NGNTYLEWYLQKPGQSPKLLIYKVS | NR | 60 |
| Sbjct | 61 | SGVPDRFSGSGS | GTDFTLKI | SRVEAEDLG | VYYCFQGS | SHVPLTFGAGTKLEIKRADAAPT | 120 |
| Query | 61 | SGVPDRFSGSG | TDFTLKI | SRVEAEDLG | VYYCFQGS | SHVPLTFGAGTKLEIKRADAAPT | 118 |
| Sbjct | 121 | SIFPPSSEQLT | SGGASVVC | FLNNFY | PKDINVK | WKIDGSERQNGVLNSWTDQDSK | 180 |
| Query | 119 | SIFPPSSEQLT | SGGASVVC | FLNNFY | PKDINVK | WKIDGSERQNGVLNSWTDQDSK | 178 |
| Sbjct | 181 | SSTLT | LT | KDEYERHNS | YTCEATHKTST | SPIVKS | 219 |
| Query | 179 | SSTLT | LT | KDEYERHNS | YTCEATHKTST | SPIVKS | 217 |

Supplementary Figure S1. Assembly results from list PSM-DD with k=6 for the WlgG1 light chain. (A) BLAST alignment of the contig against the target light chain. (B) Details of the alignment in (A).

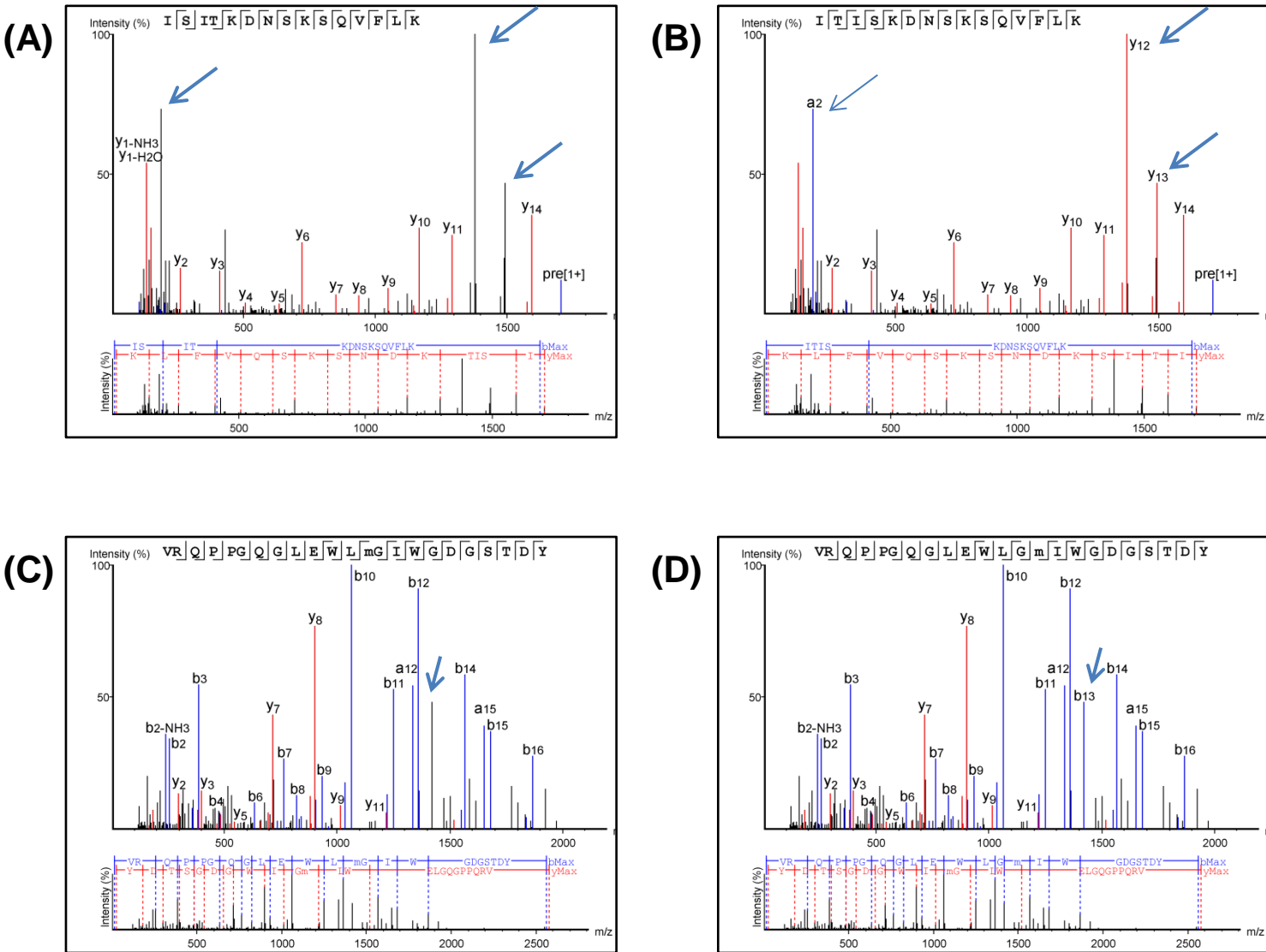
(A)



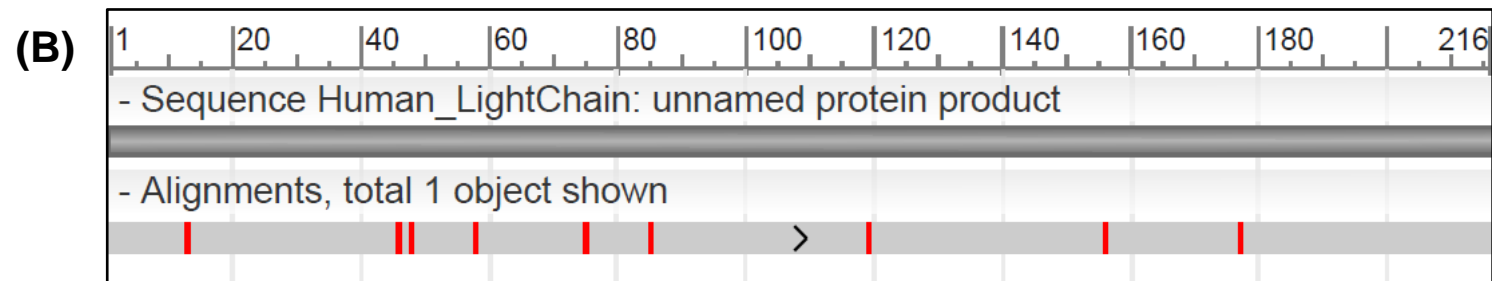
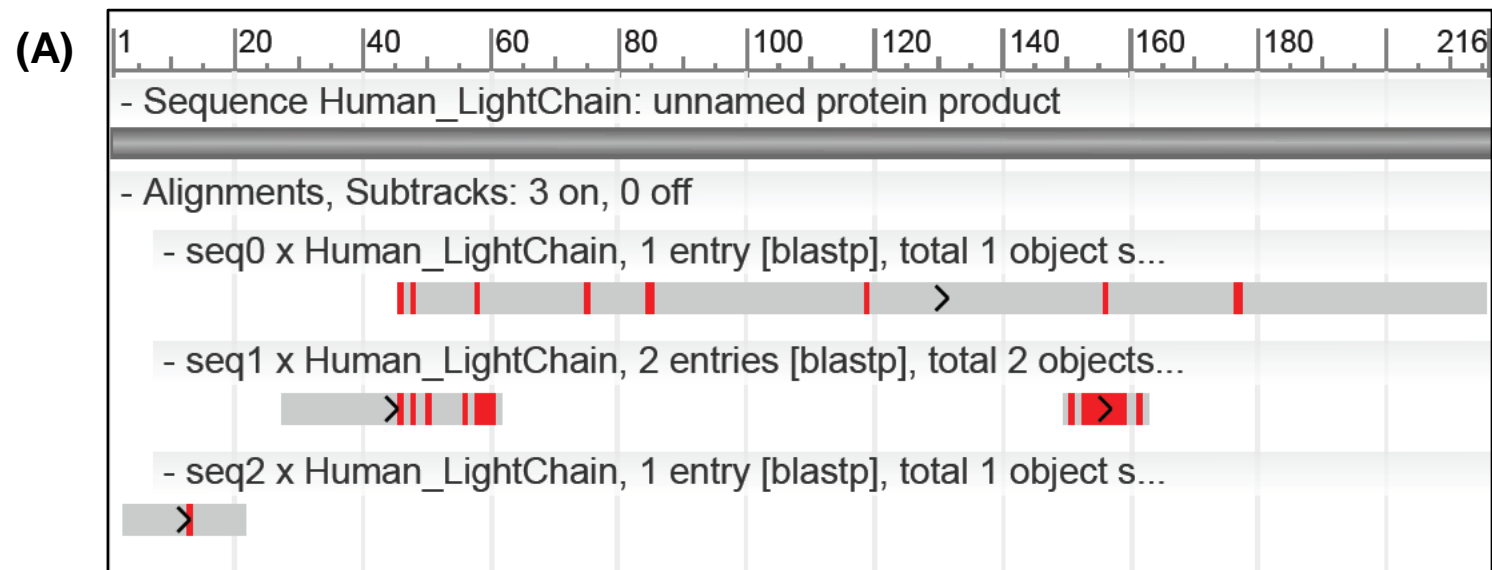
(B)

| | | | | | | |
|-------|-----|----------------------|-----------|---------------|------------------|-----|
| Sbjct | 1 | QVQLKESGPGLVAPSQSLSI | TCTVSGFSL | LGY----- | GVNWVRQPPGQGLEWL | 48 |
| Query | 1 | QVQLKESGPGLVAPSQSLSI | TCTVSGFSL | LGYNHHQMMCLSL | GVNWVRQPPGQGLEWL | 60 |
| Sbjct | 49 | MGIWGDGSTDYNSALKSRIS | ITKD | NSK | SQVFLKMNSLQ | 108 |
| Query | 61 | GMIWGDGSTDYNSALKSR | ITL | SKD | NSK | 120 |
| Sbjct | 109 | GQGTLVTVSAAKTTPPSVY | PLAPG | SAAQT | DSMVT | 168 |
| Query | 121 | GQGTLVTVSAAKTTPPSVY | PLAPG | SAAQT | NSMVT | 180 |
| Sbjct | 169 | HTFPAVLQSDLYTLSSSV | TVPS | STWP | SETV | 228 |
| Query | 181 | HTFPAVLQSDLYTLSSSV | TVPS | STWP | SETV | 240 |
| Sbjct | 229 | VPEVSSVFI | FPPK | PKD | VL | 288 |
| Query | 241 | VPEVSSVFI | FPPK | PKD | VL | 300 |
| Sbjct | 289 | EQFNSTFRSVSELP | IMH | QD | WLN | 348 |
| Query | 301 | EQFNSTFRSVSELP | IMH | QD | WLN | 360 |
| Sbjct | 349 | PKEQMAKDKVSL | TCMI | T | DFF | 408 |
| Query | 361 | PKEQMAKDKVSL | TCMI | T | DFF | 420 |
| Sbjct | 409 | KSNWEAGNTFTCS | VL | HEGL | HN | 441 |
| Query | 421 | KSNWEAGNTFTCS | VL | HEGL | HN | 453 |

Supplementary Figure S2. Assembly results from list PSM-DD for the WlgG1 heavy chain. (A) BLAST alignment of the contig against the target heavy chain. (B) Details of the alignment in (A).



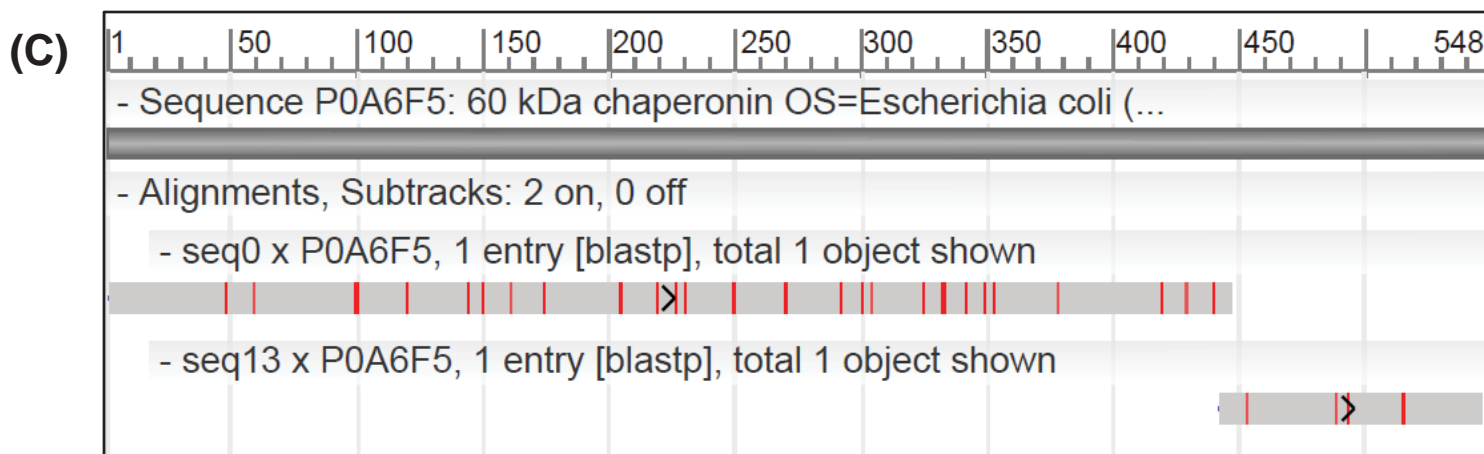
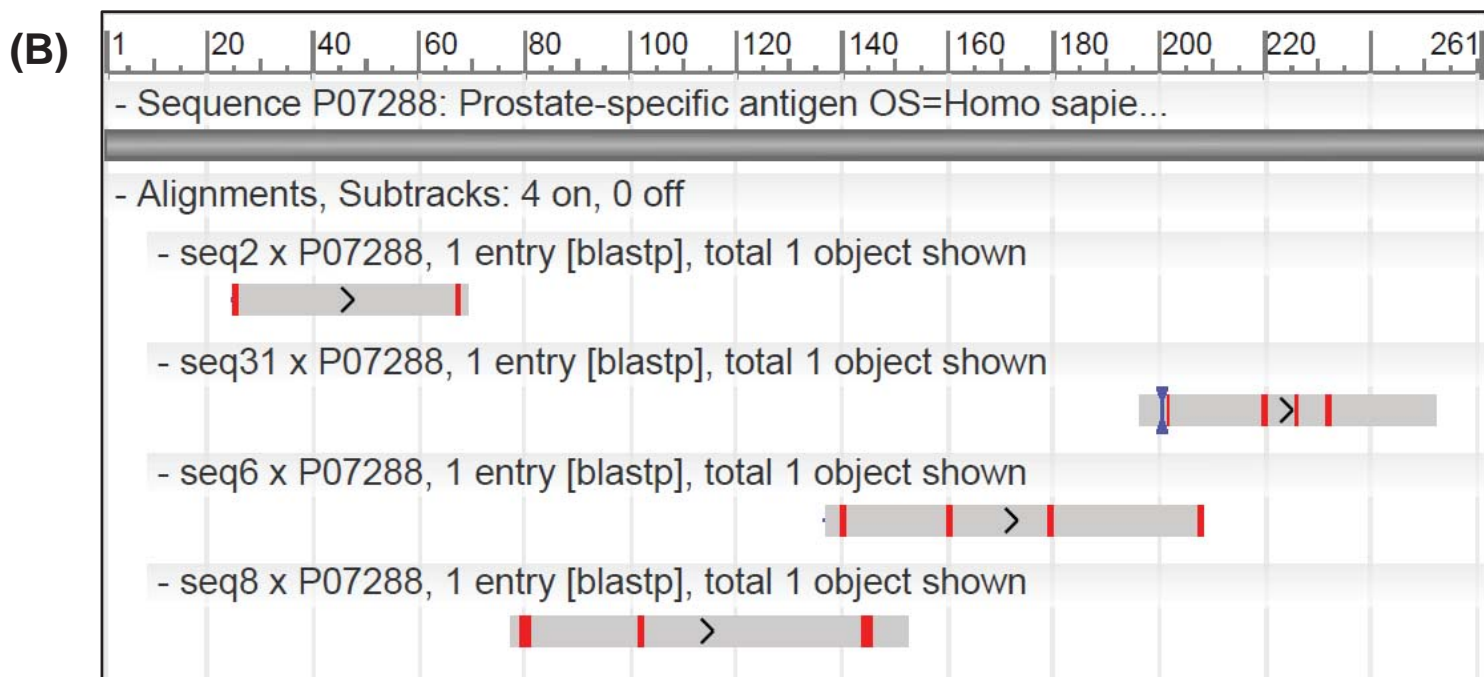
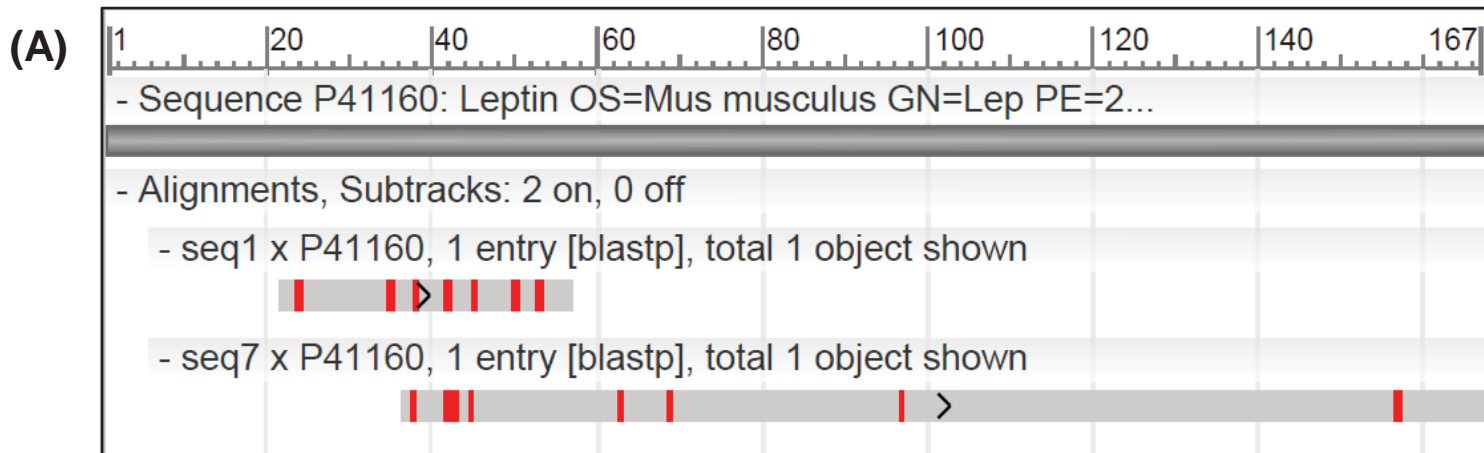
Supplementary Figure S3. High-confidence spectra to support our corrections of the WlgG1 heavy chain. (A) Assignment of residues “ISIT” at positions 67-70 in the WlgG1 heavy chain does not match three peaks in the spectrum. (B) The spectrum is better interpreted with the assignment “ITIS” as in our assembly result. (C) Assignment of residues “MG” at positions 49-50 in the WlgG1 heavy chain does not match one peak in the spectrum. (D) The spectrum is better interpreted with the assignment “GM” as in our assembly result.



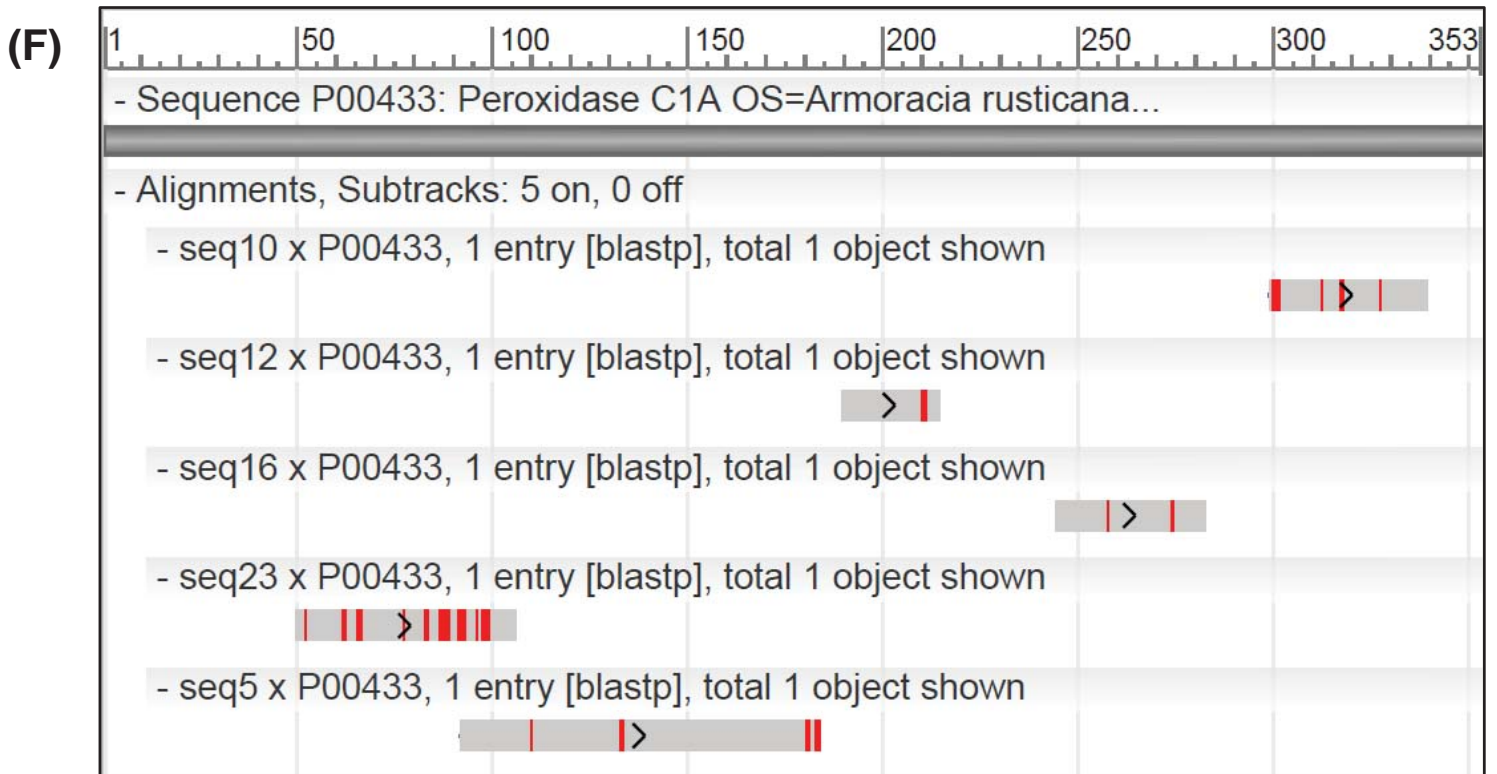
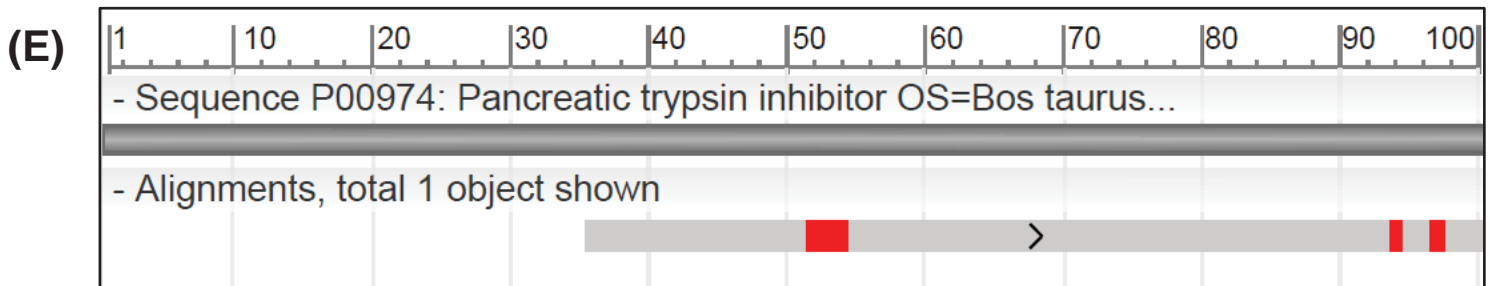
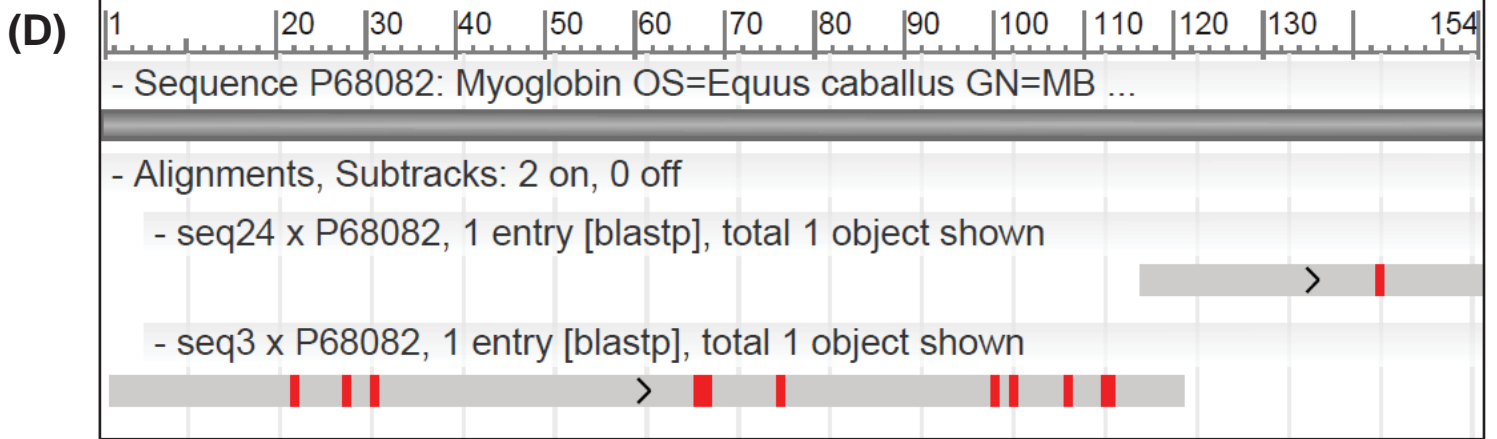
(C)

| | | | |
|-------|-----|--|-----|
| Sbjct | 1 | ELVLTQSPASLSISPGERATLSCRASQSVSSYLAWYQHKPGQAPRILLYDASTRATGIPA | 60 |
| Query | 1 | ELVLTQSPASLSLSPGERATLSCRASQSVSSYLAWYQHKPGQAPRILLYDASTRATGLPA | 60 |
| Sbjct | 61 | RFSGSGSGTDFTLTISSLEPEDFAIYYCQQRSNWPPSFTFGPGTRVDLKRTVAAPSVFIF | 120 |
| Query | 61 | RFSGSGSGTDFTLTISSLEPEDFAIYYCQQRSNWPPSFTFGPGTRVDLKRTVAAPSVFLF | 120 |
| Sbjct | 121 | PPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNAIQSGNSQESVTEQDSKDSTYSISST | 180 |
| Query | 121 | PPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNAIQSGNSQESVTEQDSKDSTYSLSST | 180 |
| Sbjct | 181 | LTLISKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC | 216 |
| Query | 181 | LTLISKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC | 216 |

Supplementary Figure S4. Assembly results for the HUMAN light chain. (A) BLAST alignment of the top assembled contigs from list PSM-DN against the target light chain. (B) BLAST alignment of the full-length contig assembled from list PSM-DD against the target light chain. (C) Details of the alignment in (B).



Supplementary Figure S5. BLAST alignments of assembled contigs for the dataset of 6-protein mixture against the respective target protein sequences. (A) leptin. (B) kallikrein. (C) groEL.



Supplementary Figure S5 (continued). BLAST alignments of assembled contigs for the dataset of 6-protein mixture against the respective target protein sequences. (D) myoglobin. (E) aprotinin. (F) peroxidase.