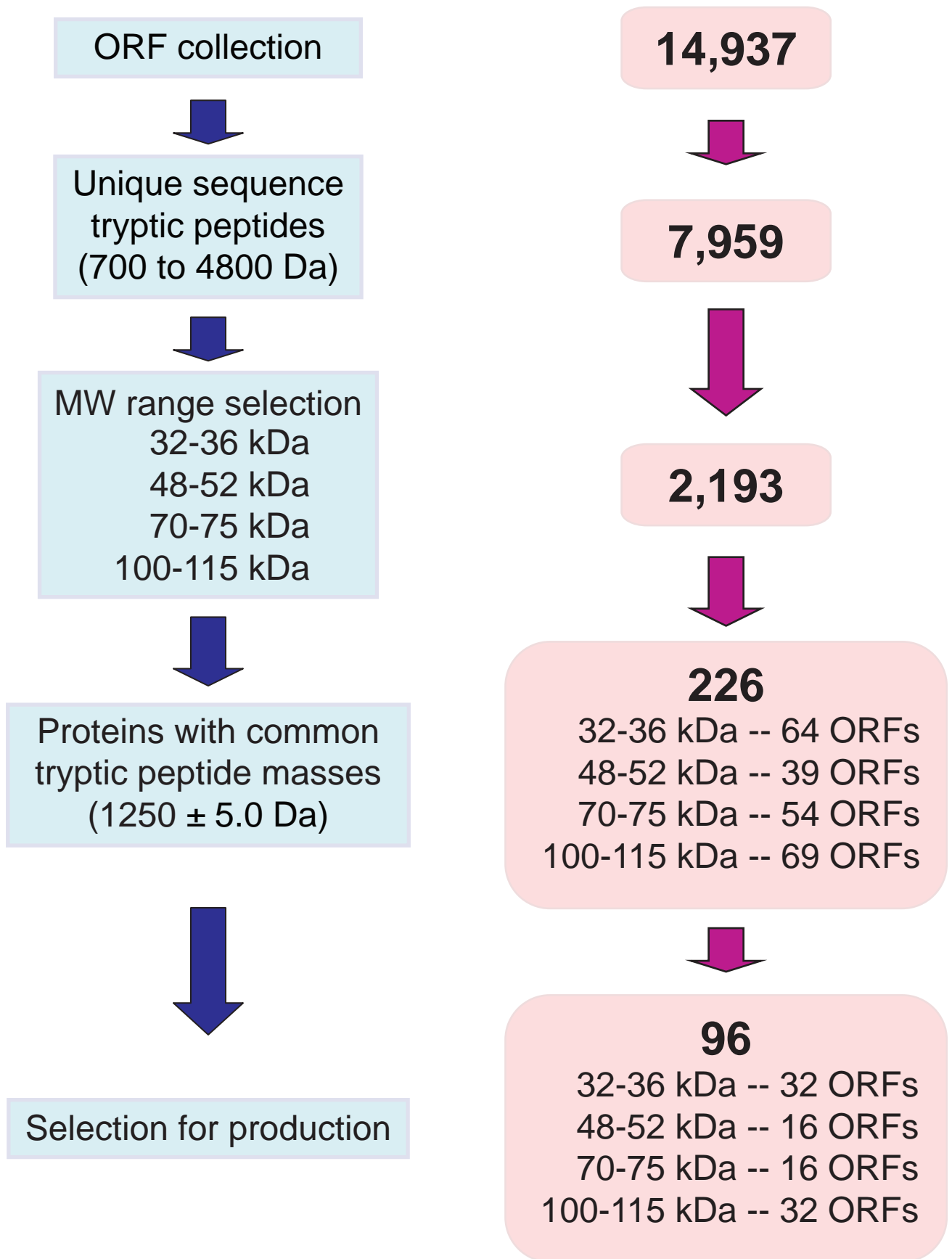


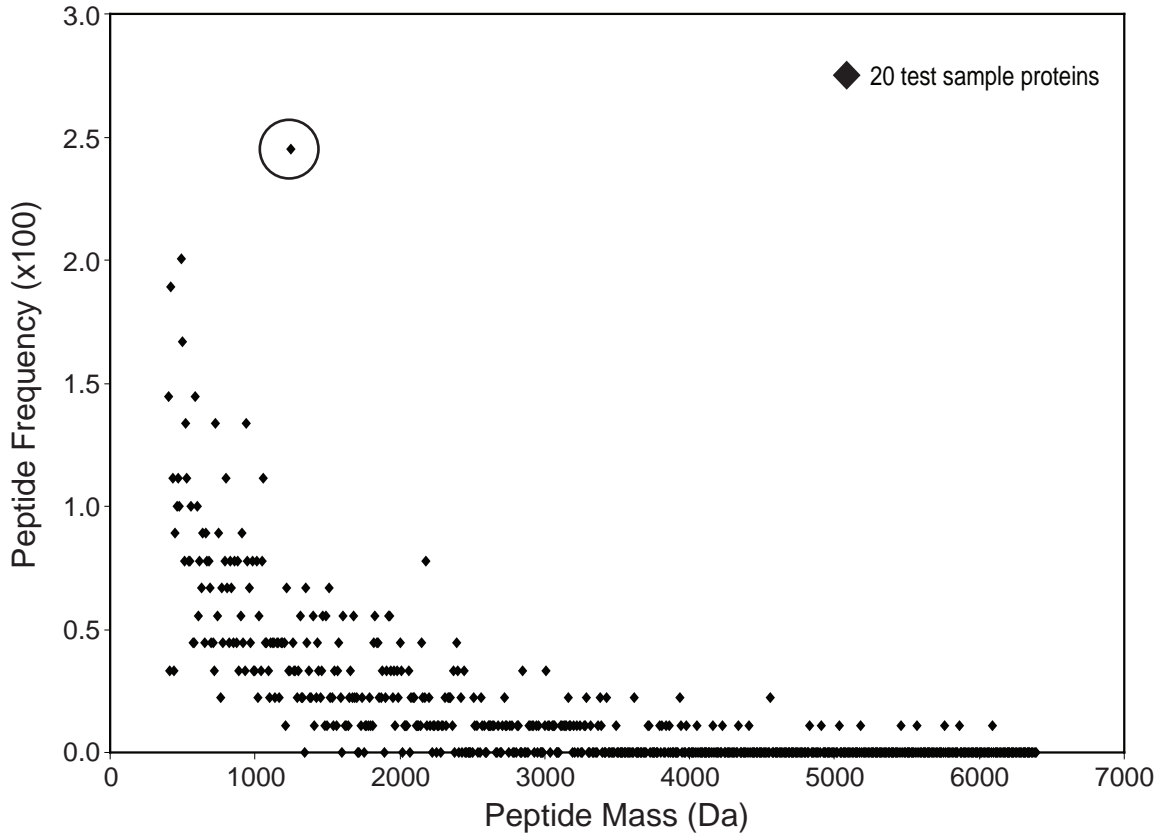
Supplementary Figure 1. Test sample proteins (a) Selection criteria for test sample proteins are depicted on the left, the number of resulting candidates on the right. From an initial set of 14,937 sequences (full-length expression-ready human ORF clones), successive criteria were applied leading to the final selection of 96 for cloning and expression. (b) Tryptic peptide mass distribution of the 20 test sample proteins. (c) Tryptic peptide mass distribution of the 96 proteins (red squares) selected for production compared to the peptide mass distribution of the entire MGC representing the 14,937 ORFs (blue squares). The comparison of the 898 peptides in (b), compared to the 4,482 peptides for the 96 proteins and the 546,213 peptides for the ORFs in (c), reveal a comparable distribution except at a peptide mass of 1250 ± 5 Da. Only the 96 proteins (and the 20 test sample proteins) show a high frequency of the mass range for this tryptic peptide. (d) The 20 purified Test Sample proteins and the Test Sample mixture were resolved by 1D-SDS PAGE on 8% gels. (Left side) Coomassie blue stained. Individual proteins were loaded at 5 pmol (0.16 to 0.56 μ g) each and the test mixture load was 100 pmol total protein (\sim 6.7 μ g). Lanes 1 and 23, Molecular weight standards, lane 2, KHK (\sim 0.16 μ g); lane 3, ATPAF2 (\sim 0.16 μ g); lane 4, SETD3 (\sim 0.17 μ g); lane 5, SPRY2 (\sim 0.17 μ g); lane 6, GLB1L3 (\sim 0.18 μ g); lane 7, FYTTD1 (\sim 0.18 μ g); lane 8, IHPK1 (\sim 0.25 μ g); lane 9, IFRD1 (\sim 0.25 μ g); lane 10, GCNT3 (\sim 0.25 μ g); lane 11, EIF2S3 (\sim 0.26 μ g); lane 12, F2 (\sim 0.35 μ g); lane 13, FARP2 (\sim 0.37 μ g); lane 14, ENOX1 (\sim 0.37 μ g); lane 15, KLHL13 (\sim 0.37 μ g); lane 16, NIBP (\sim 0.50 μ g); lane 17, MARS (\sim 0.51 μ g); lane 18, NUP210 (\sim 0.53 μ g); lane 19, THBS4 (\sim 0.53 μ g); lane 20, KIAA0746 (\sim 0.56 μ g); lane 21, HIRA

(~0.56 µg); lane 22; the test sample containing all twenty proteins (~6.7 µg). Molecular weight values (Da x10⁻³) are shown on the right. (Right side) Silver stain of the 20 recombinant proteins purified as for the Coomassie Blue stained material. Two orders of magnitude differences were loaded. Lane A, 2.2 pmol of each protein; Lane B, 0.22 pmol each protein; Lane C, 0.022 pmol each protein; Lane D, blank; Lane E, molecular weight markers. The major proteins in all cases coincide with the mobilities of the 20 recombinant proteins as seen by Coomassie Blue staining.

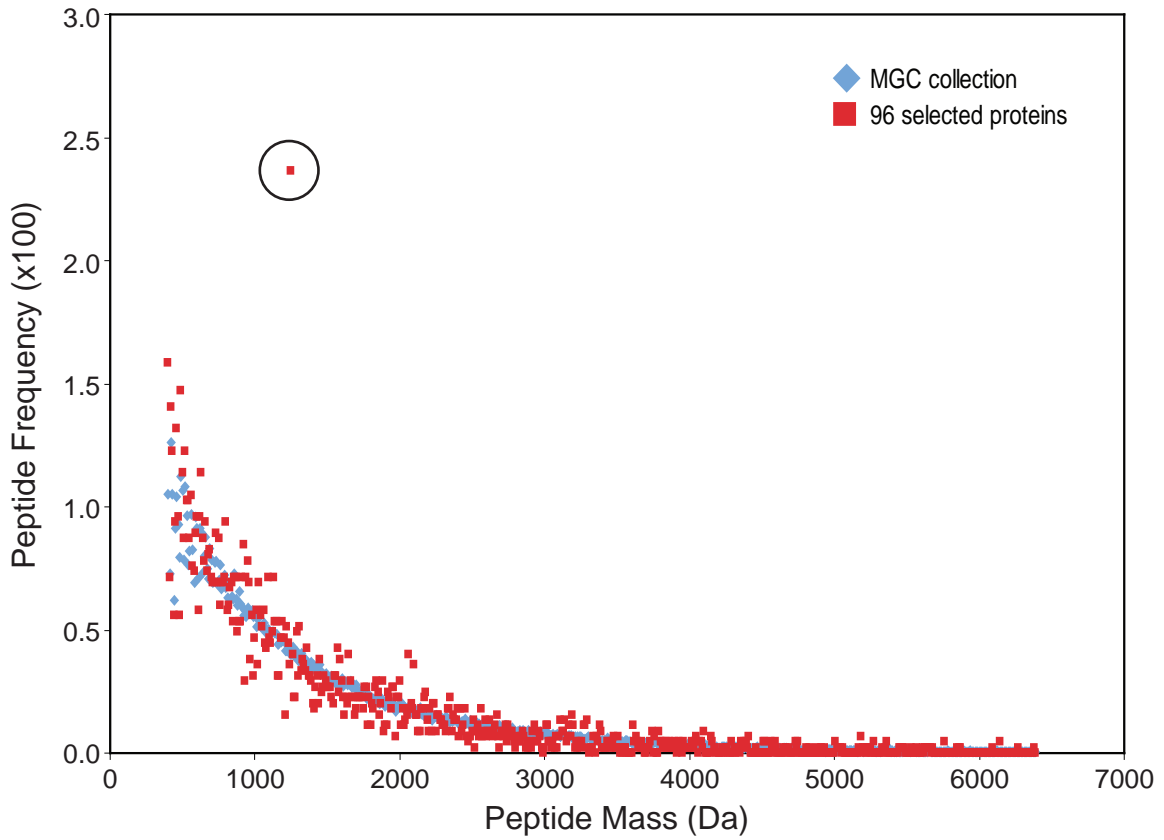
a



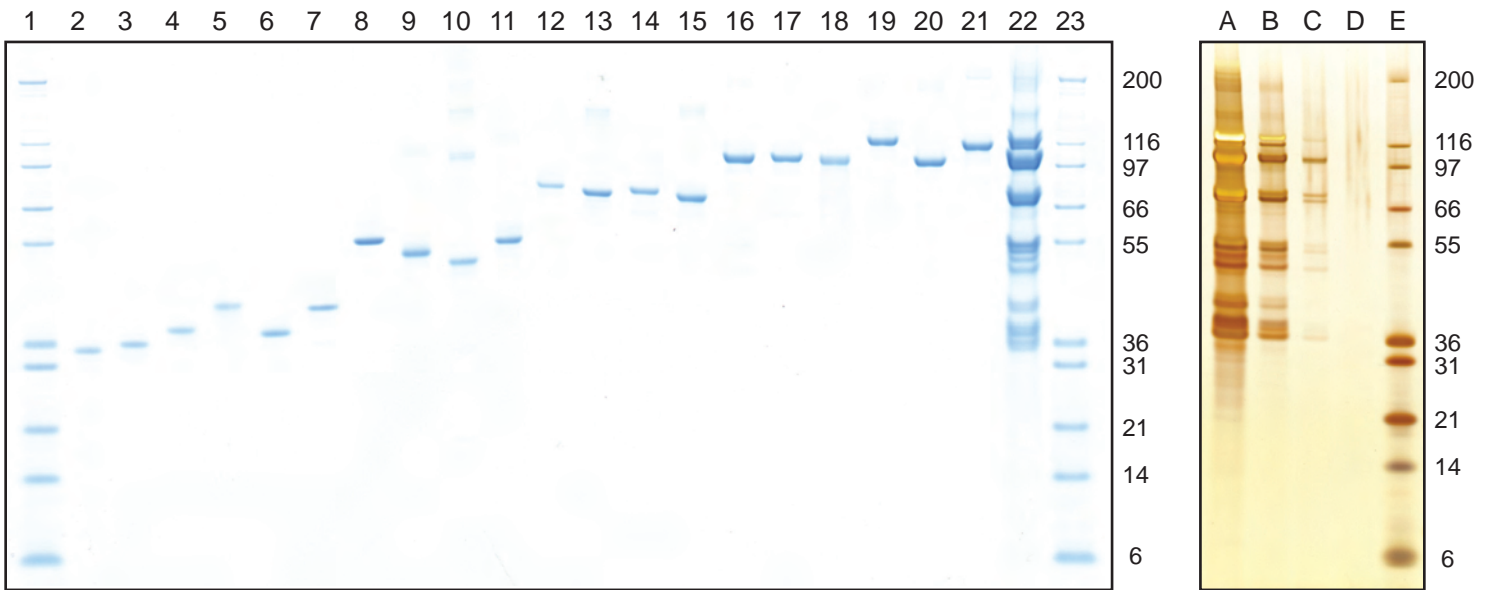
b



c



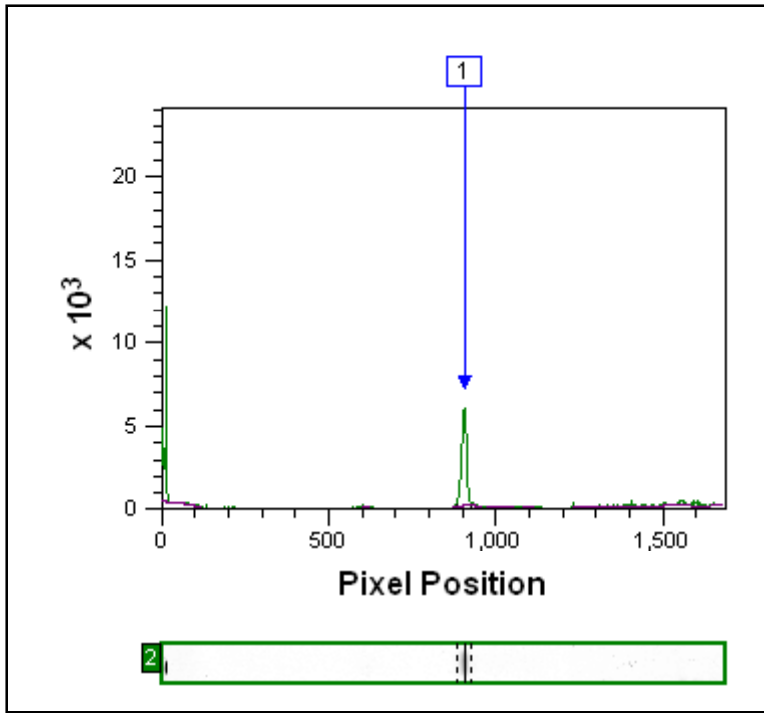
d



Supplementary Figure 2. Image analysis: Test Sample protein purity. Image analyses were performed by using the TotalLab TL100 1D analysis software (Nonlinear Dynamics) as described in the **Online Methods**. Test sample proteins are identified by their gene symbol (**Supplementary Table 1** online) and presented from lowest to highest calculated MW in panels (a) through (t) respectively. Presented are image analysis profiles (upper panel) of the corresponding gel lane (lower panel) for each of the 20 Test Sample proteins as resolved by 1D-SDS PAGE (**Supplementary Fig. 1d** online). Intensity of the Coomassie blue staining (y-axis) and position in the gel (pixels, x-axis) were determined. The image of the gel lane is numbered according to the position in the gel (**Supplementary Fig. 1d** online). Blue arrows (upper panels) denote the center of the detected band and dashed lines (lower panels) denote upper and lower limits of the band in the gel image. Coomassie blue stained bands at higher M_r than the expected proteins were considered oligomeric aggregates (confirmed by LC-MS analysis, data not shown). Lower MW bands were considered contaminants. For the analysis of the FARP2 protein (I), a peak (indicated by *) at the left of the trace was manually excluded in the analysis as it was identified as a gel speck.

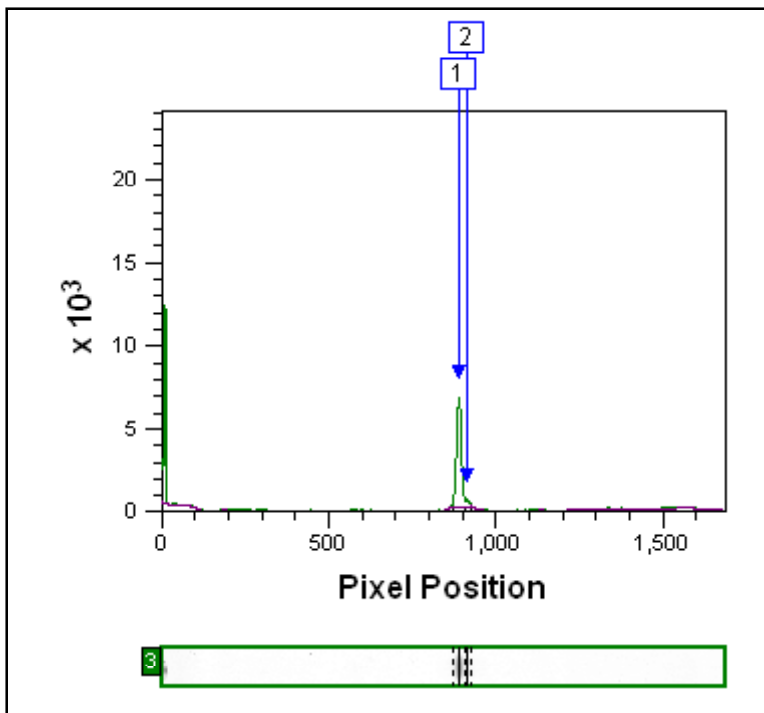
Pixel Intensity $\times 10^3$

a



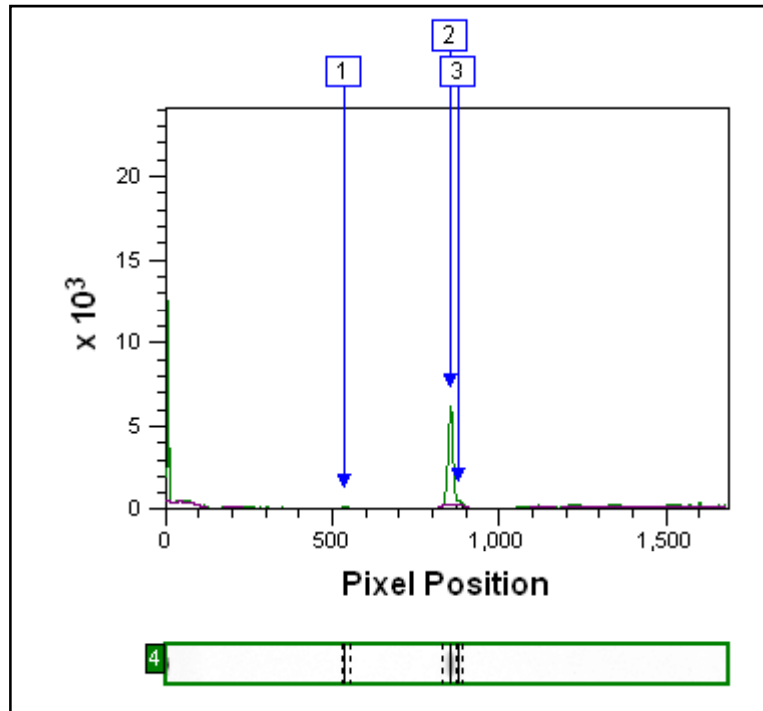
KHK

b



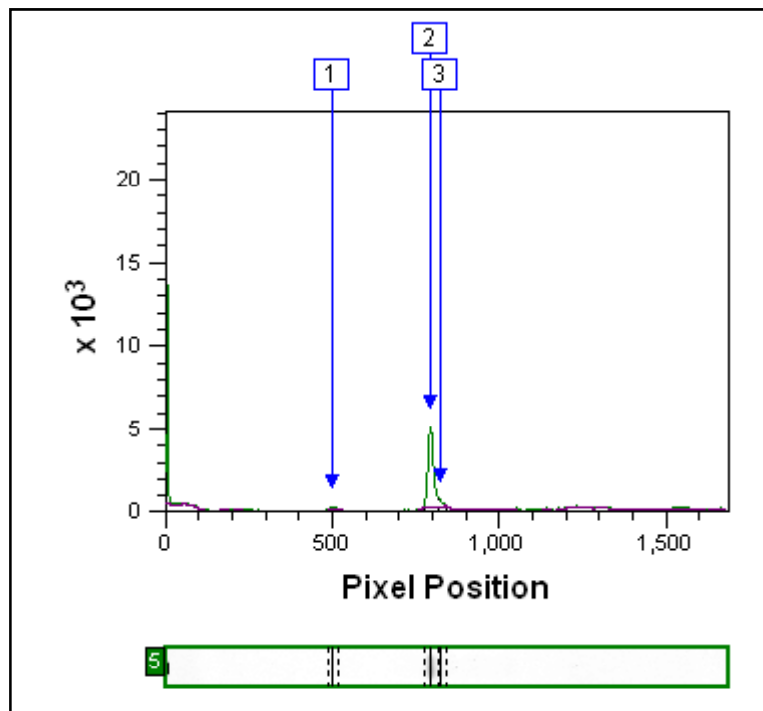
ATPAF2

c



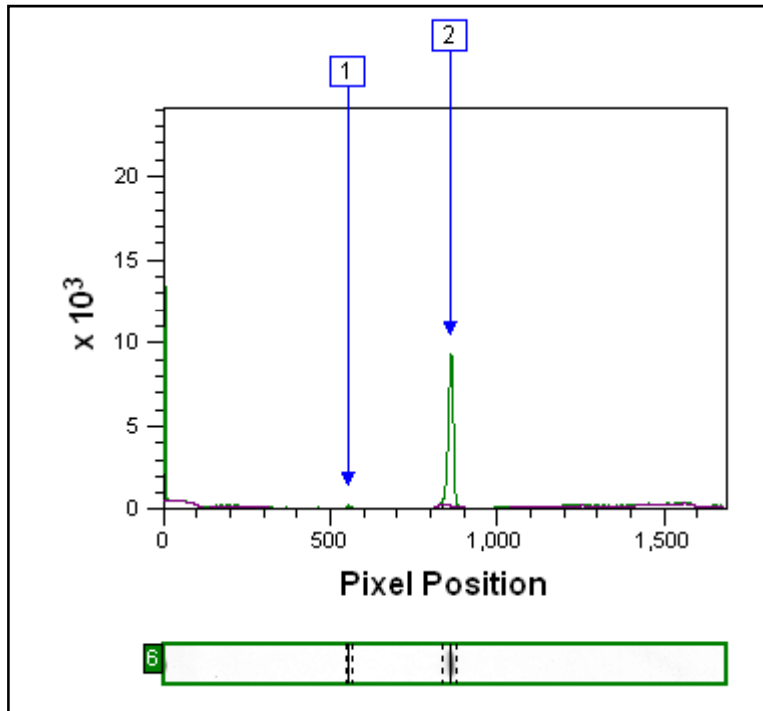
SETD3

d



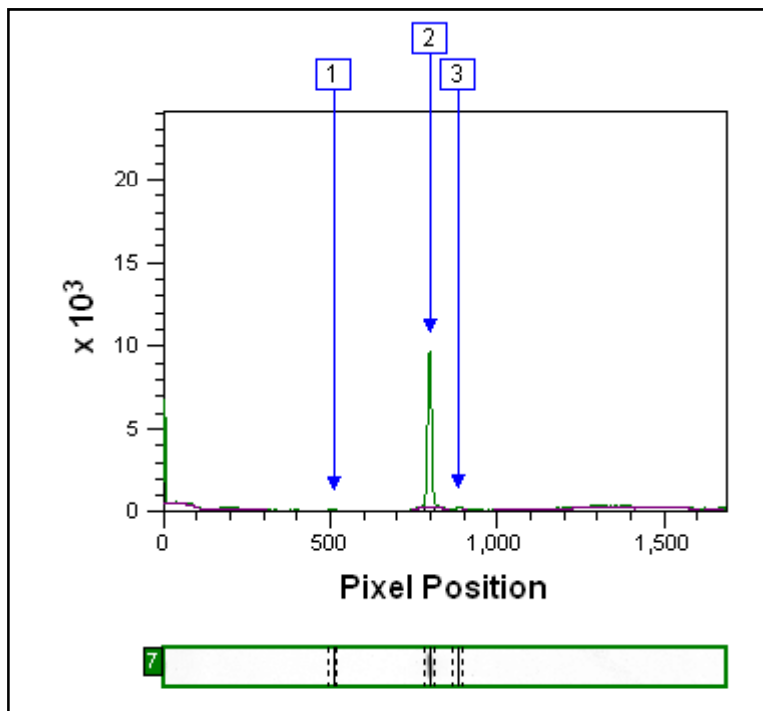
SPRY2

e



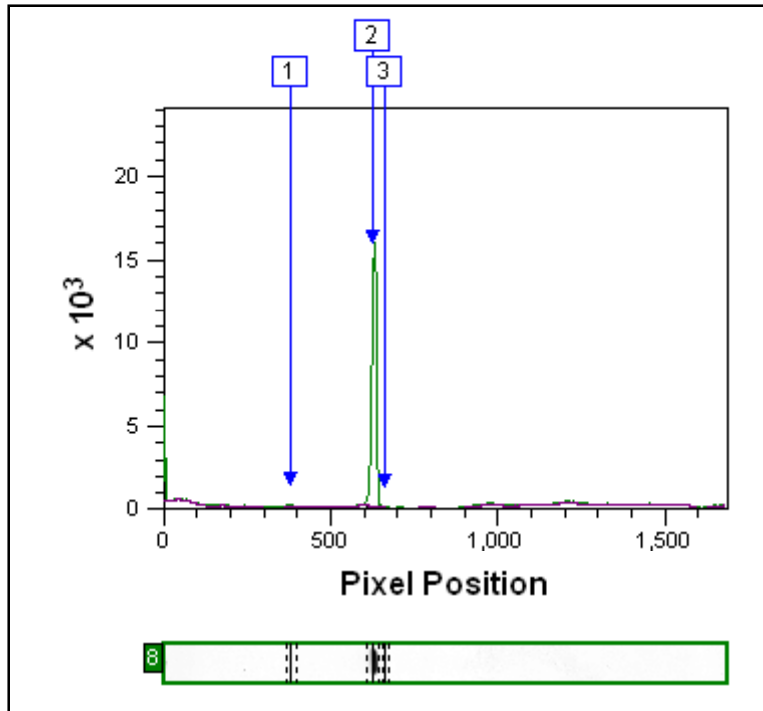
GLB1L3

f



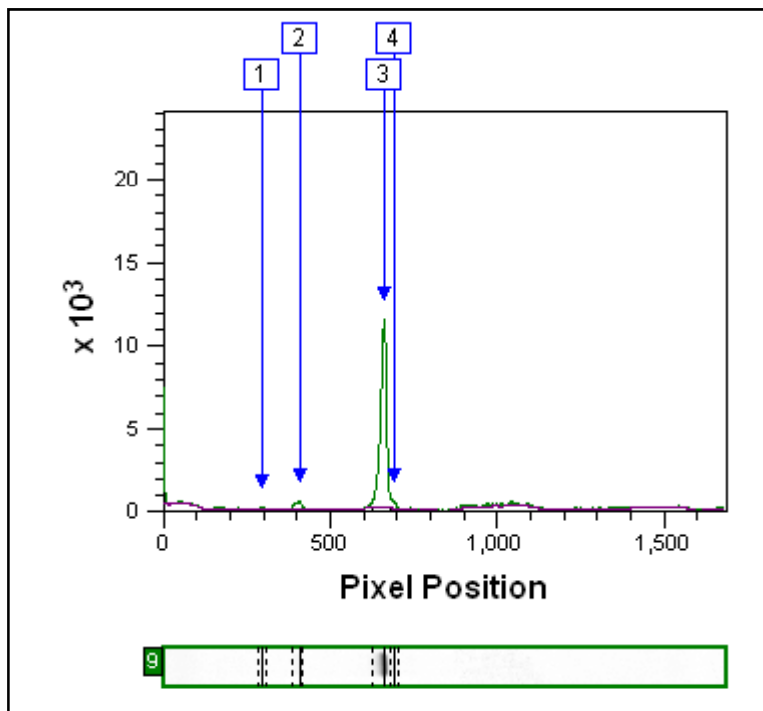
FYTTD1

g



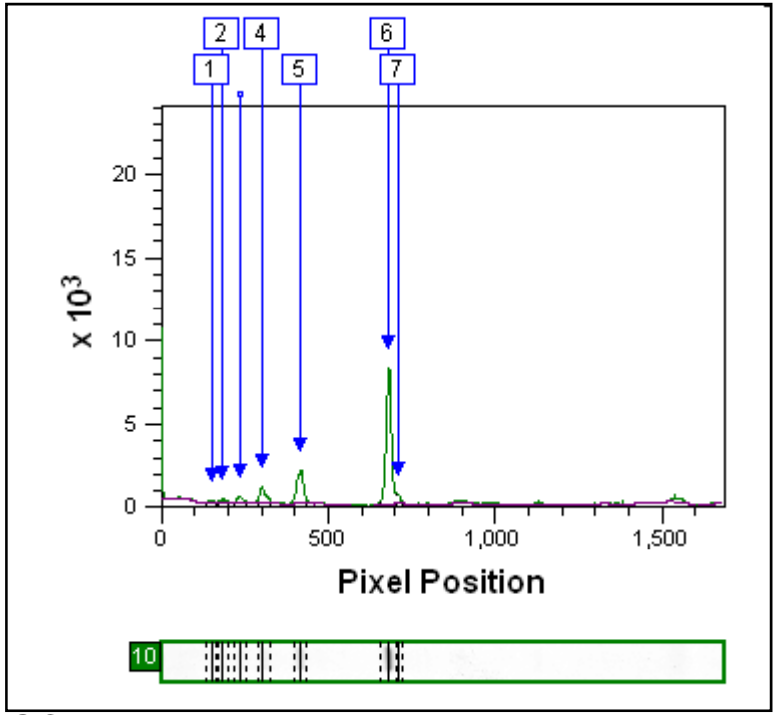
IHPK1

h



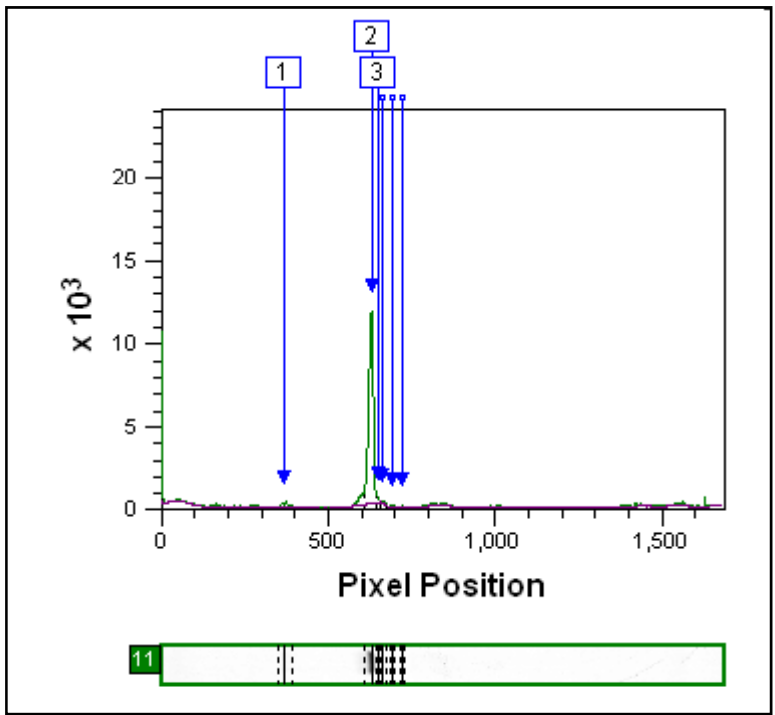
IFRD1

i



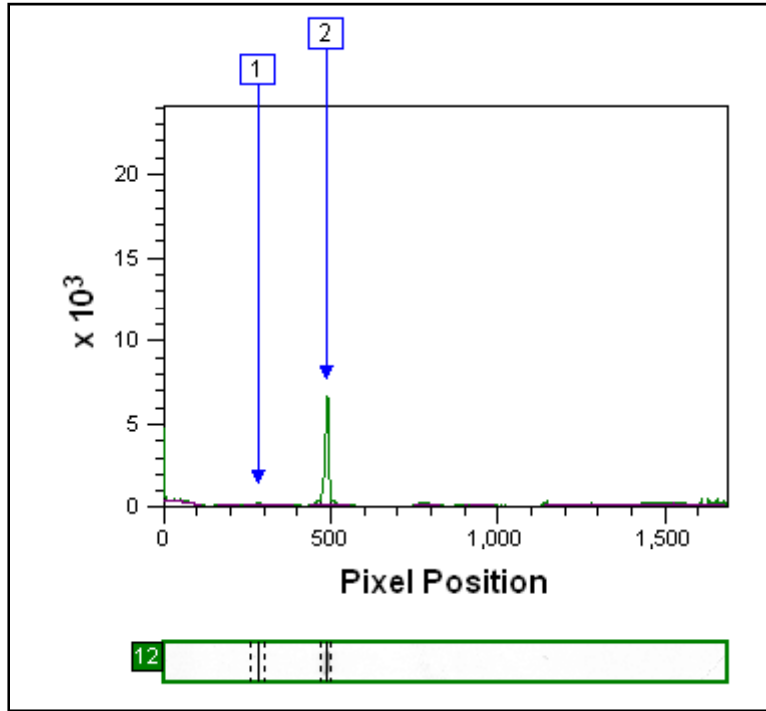
GCNT3

j



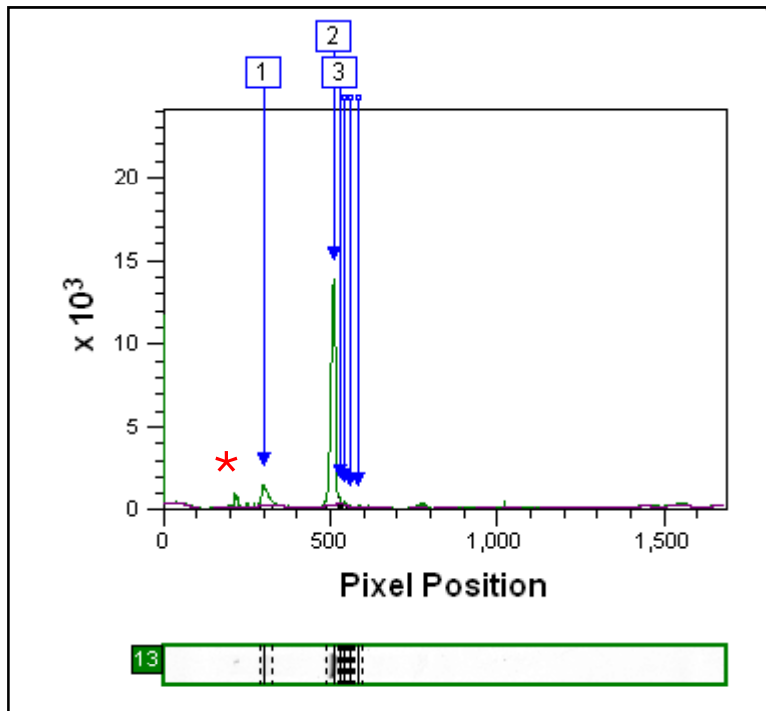
EIF2S3

k



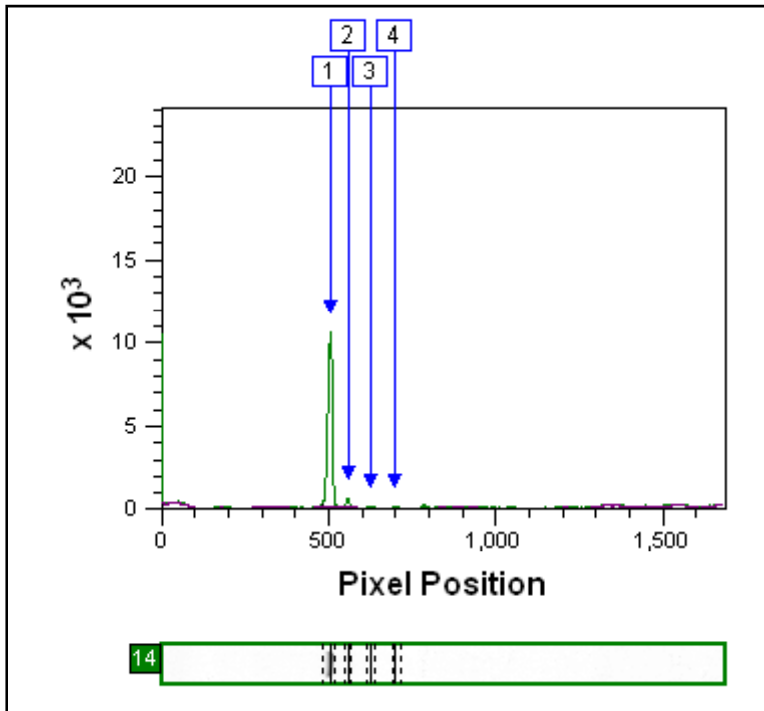
F2

l



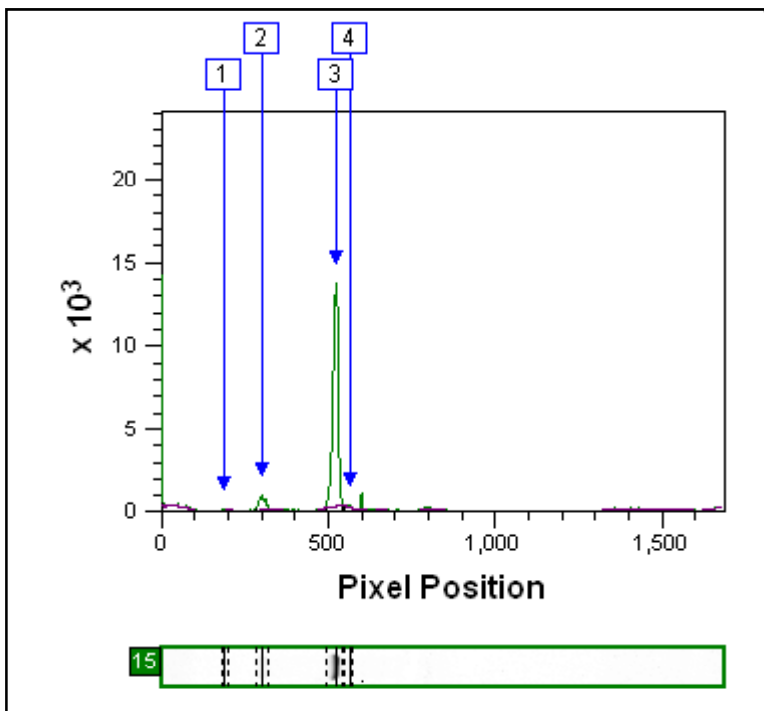
FARP2

m



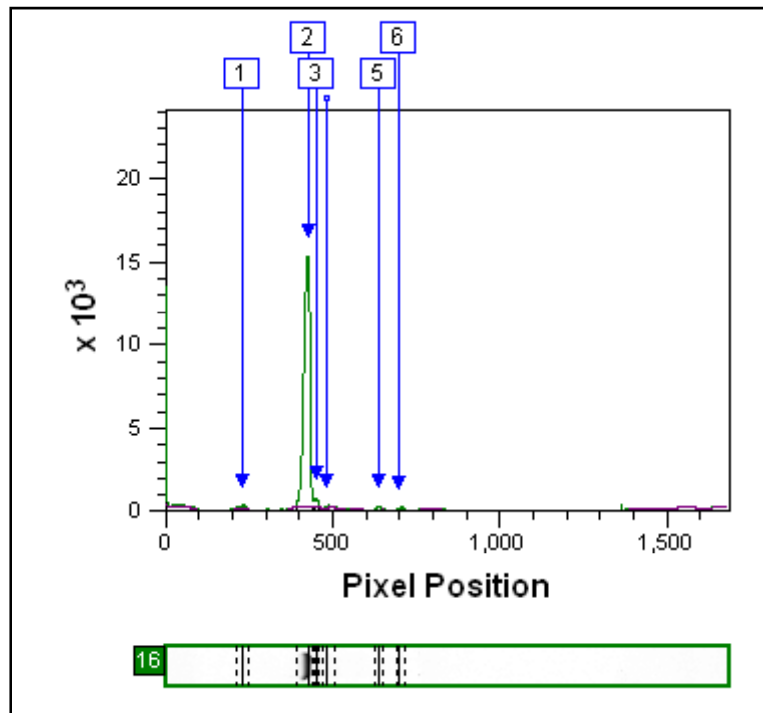
ENOX1

n



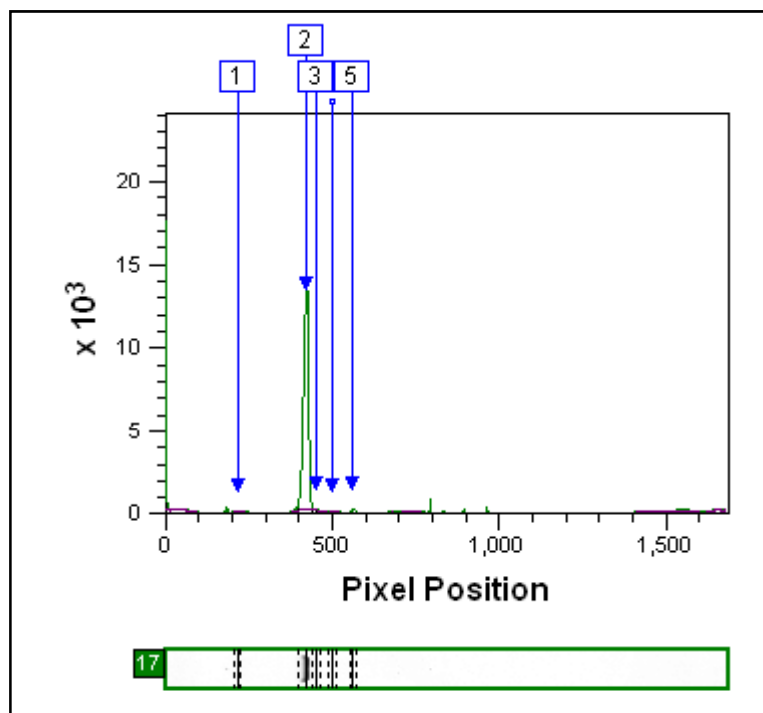
KLHL13

o



NIBP

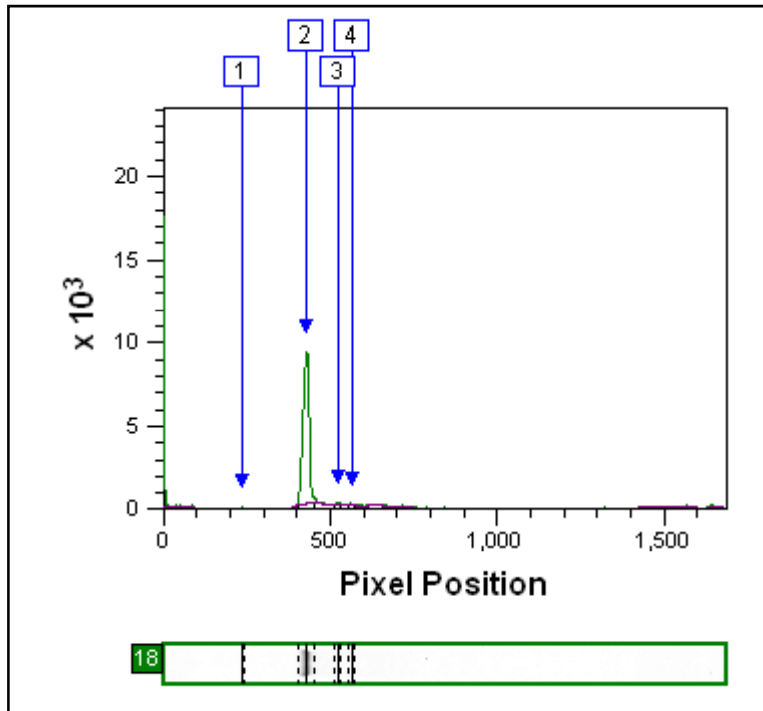
p



MARS

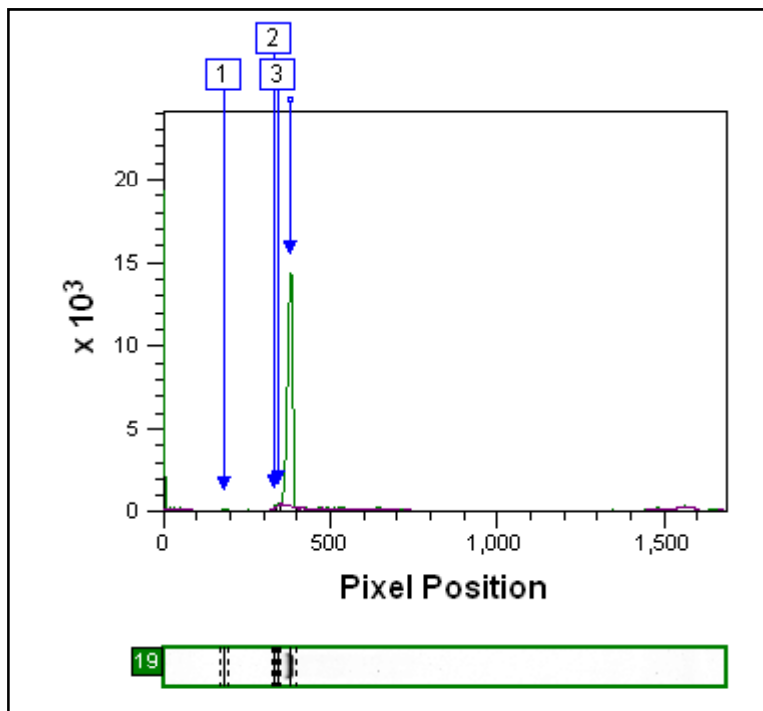
Bell et al. Supplementary Figure 2

q



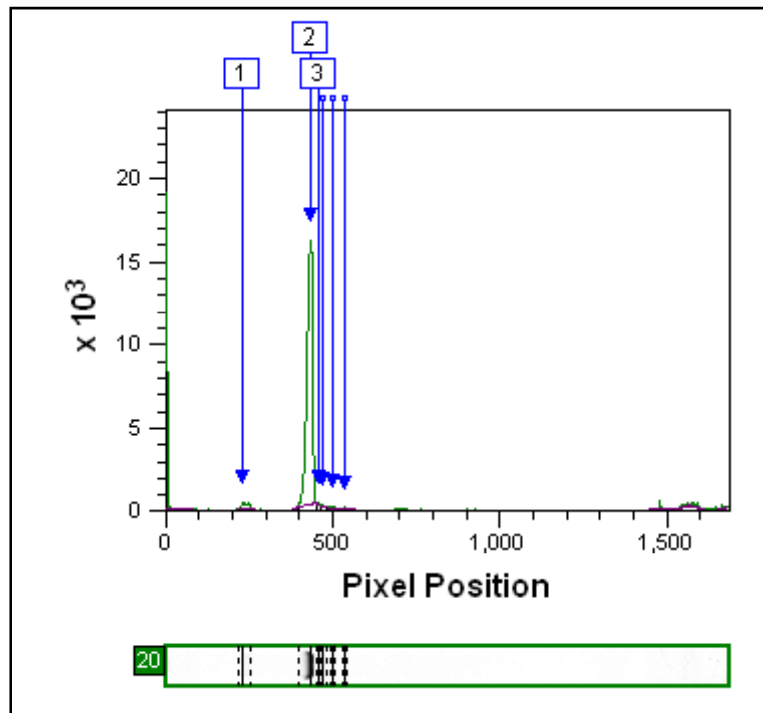
NUP210

r



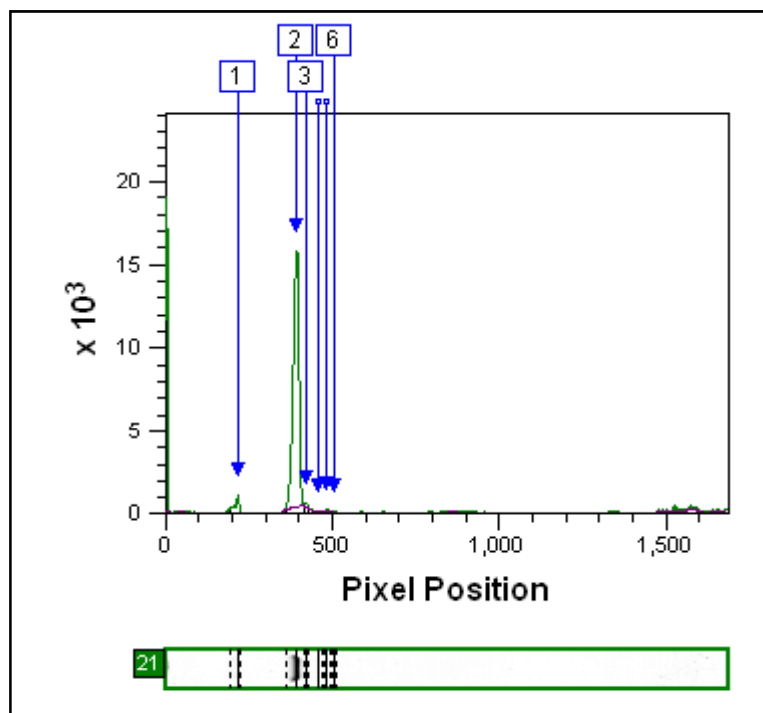
THBS4

s



KIAA0746

t

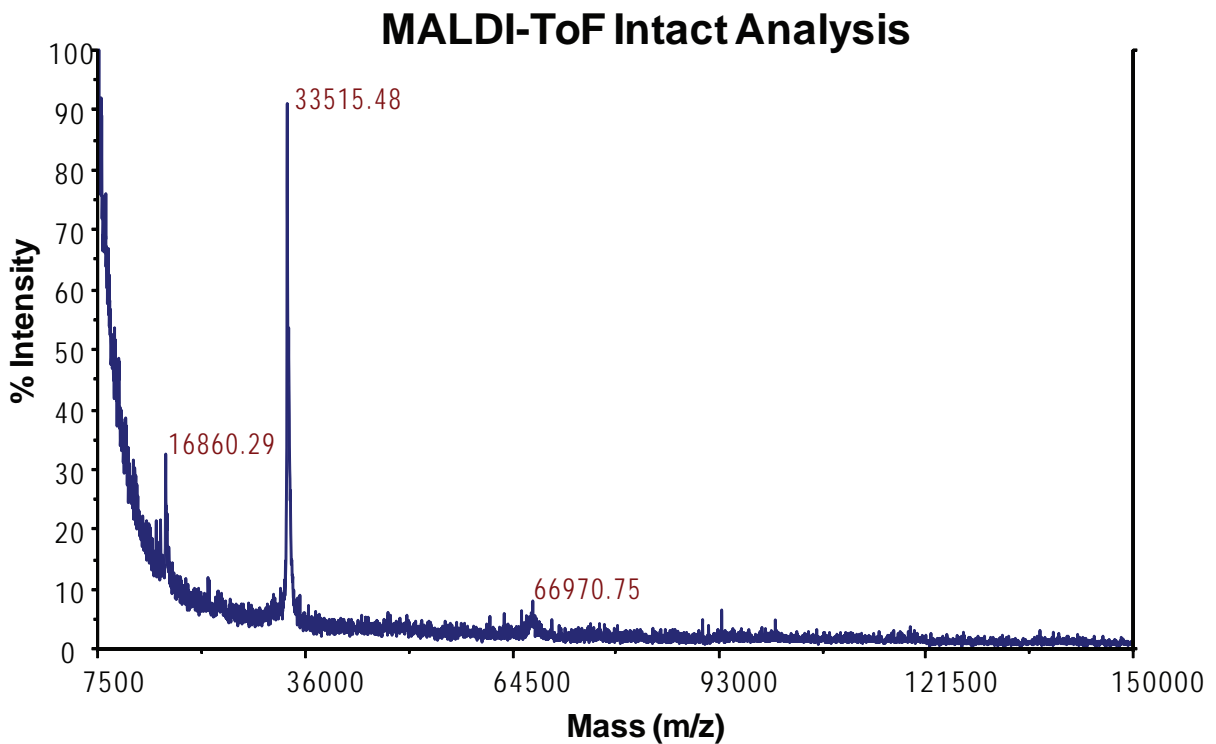


HIRA

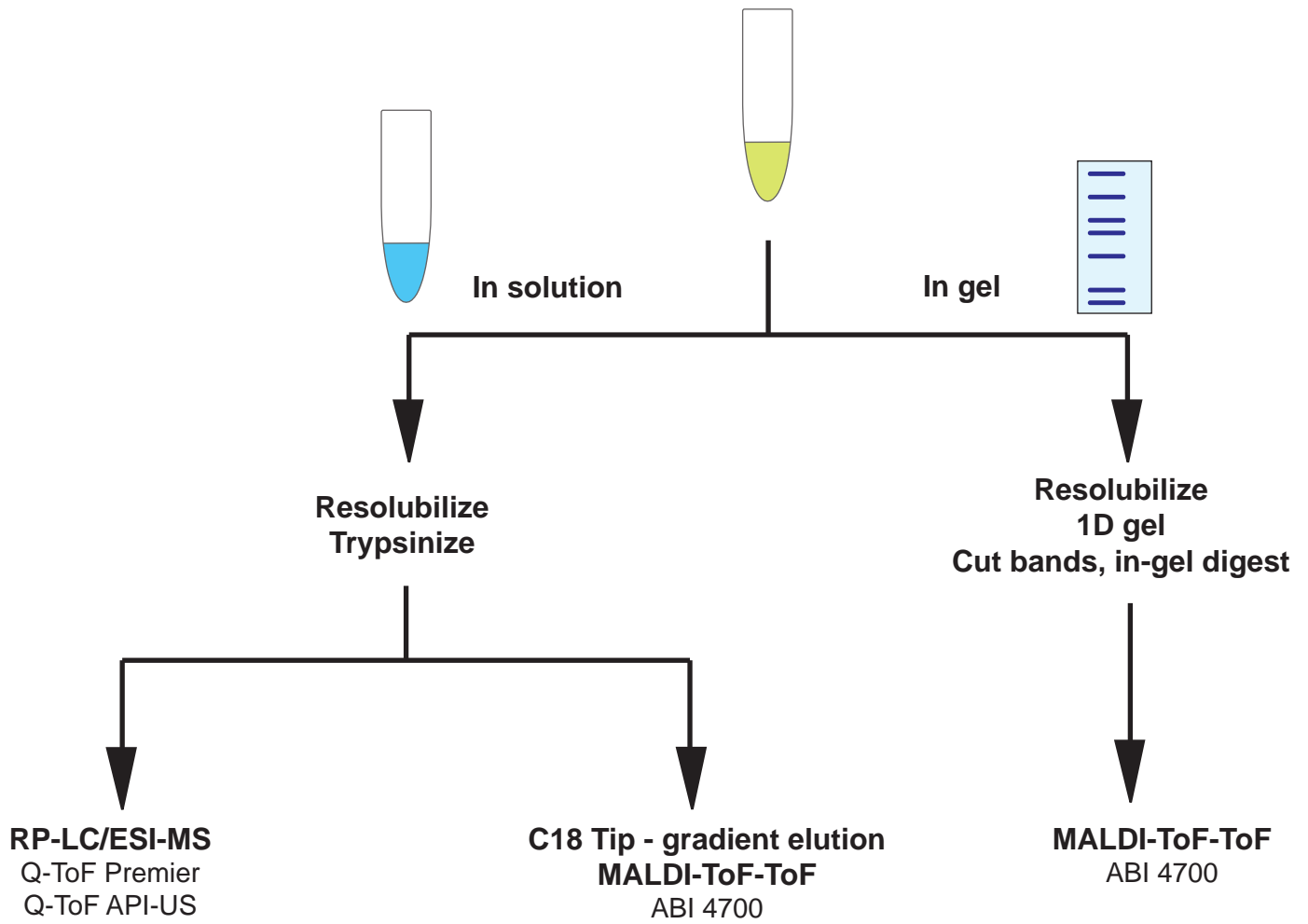
Supplementary Figure 3. Ketoheokinase (KHK) characterization. The predicted amino acid sequence and calculated MW (32,749 Da, 298 amino acids) of KHK is shown above the MALDI-ToF mass spectrum (relative intensity versus m/z) of purified KHK. The mass difference of ~766 Da between the predicted mass of the cloned sequence and experimental mass (33515 Da) is accounted for by an N-terminal extension in the expressed KHK due to the initiation of translation at a vector derived methionine¹. By employing the pENTR221 entry/pET-DEST42 destination clones, a vector derived leader sequence of 7 amino acids; MYKKAGT that corresponds to 780 Da is incorporated with each ORF.

Ketohexokinase: [1-298] mass = 32,749

| | | | | | |
|------------|------------|------------|------------|------------|-----|
| MEEKQILCVG | LVVLDVISLV | DKYPKEDSEI | RCLSQRWQRG | GNASNSCTIL | 50 |
| SLLGAPCAFM | GSMAPGHVAD | FVLDDLRRYS | VDLRYTVFQT | TGSVPIATVI | 100 |
| INEASGSRTI | LYYDRSLPDV | SATDFEKVDL | TQFKWIHIEG | RNASEQVKML | 150 |
| QRIDAHNTRQ | PPEQKIRVSV | EVEKPREELF | QLFGYGDVVF | VSKDVAKHLG | 200 |
| FQSAEEALRG | LYGRVRKGAV | LVCAWAEEGA | DALGPDGKLL | HSDAFPPPRV | 250 |
| VDTLGAGDTF | NASVIFSLSQ | GRSVQEALRF | GCQVAGKKCG | LQGF DGIV | 298 |

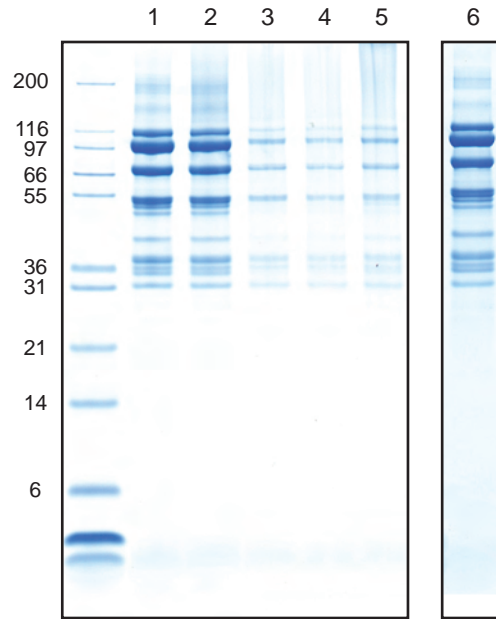


Supplementary Figure 4. MS workflows for protein quality assurance prior to distribution. Shown are the proteomics analysis strategies for quality assurance analysis of the test sample proteins and the Test Sample mixture. Trypsin proteolysis was performed on solubilized proteins and individual bands following SDS PAGE separation and Coomassie blue staining. Analyses of tryptic peptides were by nano-LC-ESI Q-ToF and MALDI-ToFToF MS. For MALDI-ToFToF analysis, a step gradient pre-fractionation of tryptic peptides captured on a C18 Zip Tip microcolumn was employed.



Supplementary Figure 5. Test sample stability assessed by 1D-SDS PAGE.

Accelerated aging (see **Supplementary Methods** online) of the test sample mixture was employed to assess long term protein stability. Test Samples were incubated at -20, 22, 37, 42 and 70 °C for 2.7 days followed by 1D-SDS PAGE and Coomassie blue staining. Left most lane contains MW marker proteins; lanes 1-5 contain an aliquot of the material incubated at -20, 22, 37, 42 and 70 °C, respectively; and lane 6 corresponds to the test sample as analysed at time of production (0 days). Equivalent days at -20 °C were calculated using the formula: $2^{[(\text{high temp} - \text{low temp})/10]} \times (\text{days at high temp}) + (\text{days already spent at } -20 \text{ } ^\circ\text{C})$ ^{2,3} (lower panel). No evidence of lower MW degradation products were found in the aged samples but resolubilization was compromised at the elevated temperatures. MW markers ($\times 10^{-3}$) indicated.



| Lane | Temp (°C) | Days at -20°C | Days at elevated Temp | Equivalent storage at -20°C (days) |
|----------|-----------|---------------|-----------------------|------------------------------------|
| 1 | -20 | 53 | 0 | 53 |
| 2 | 22 | 50 | 2.7 | 100 |
| 3 | 37 | 50 | 2.7 | 190 |
| 4 | 42 | 50 | 2.7 | 248 |
| 5 | 70 | 50 | 2.7 | 1432 |
| 6 | -20 | 0 | 0 | 0 |

Supplementary Figure 6. Database Matching by blastp. The 20 test sample protein sequences were matched to the NCBI, UniProt and IPI databases by employing blastp (www.ncbi.nlm.nih.gov/) and restricting taxonomy to human. Only the NCBI nr database (nr_human_20061127) contained a complete set of identical sequences (**Supplementary Table 5** online). Shown are 3 examples of best matches to sequences in the IPI database that reveal <100% identity. **(a)** Alignment 1: Alignment of IOH29199 (the KIAA0746 protein, 979 amino acids) to IPI00470809.3 (1097 amino acids) reveals 89.2% sequence identity. The predicted IPI00470809.3 protein contains an N-terminal extension of 118 amino acids relative to the KIAA0746 protein and with a further difference of 1 amino acid (Y790I). **(b)** Alignment 2: Alignment of IOH6288 (KHK, 298 amino acids) to IPI00029488.1 (298 amino acids) reveals 99.7% sequence identity, with a single amino acid difference of V49I. **(c)** Alignment 3: Alignment of IOH46182 (NIBP, 912 amino acids) to IPI00783996.1 (1246 amino acids) reveals 73.1% sequence identity. The IPI00783996.1 protein contains an N-terminal extension of 325 amino acids plus a 9 amino acid insertion (position 387), with a single amino acid difference (V647I) relative to the KHK protein. Extensions and insertions in the database sequence relative to the test sample protein are recognized as dashed lines in the test sample sequence. Mismatched amino acid residues are identified by the absence of vertical lines between the alignments. Amino acid residue numbering indicated.

a

Alignment 1

```

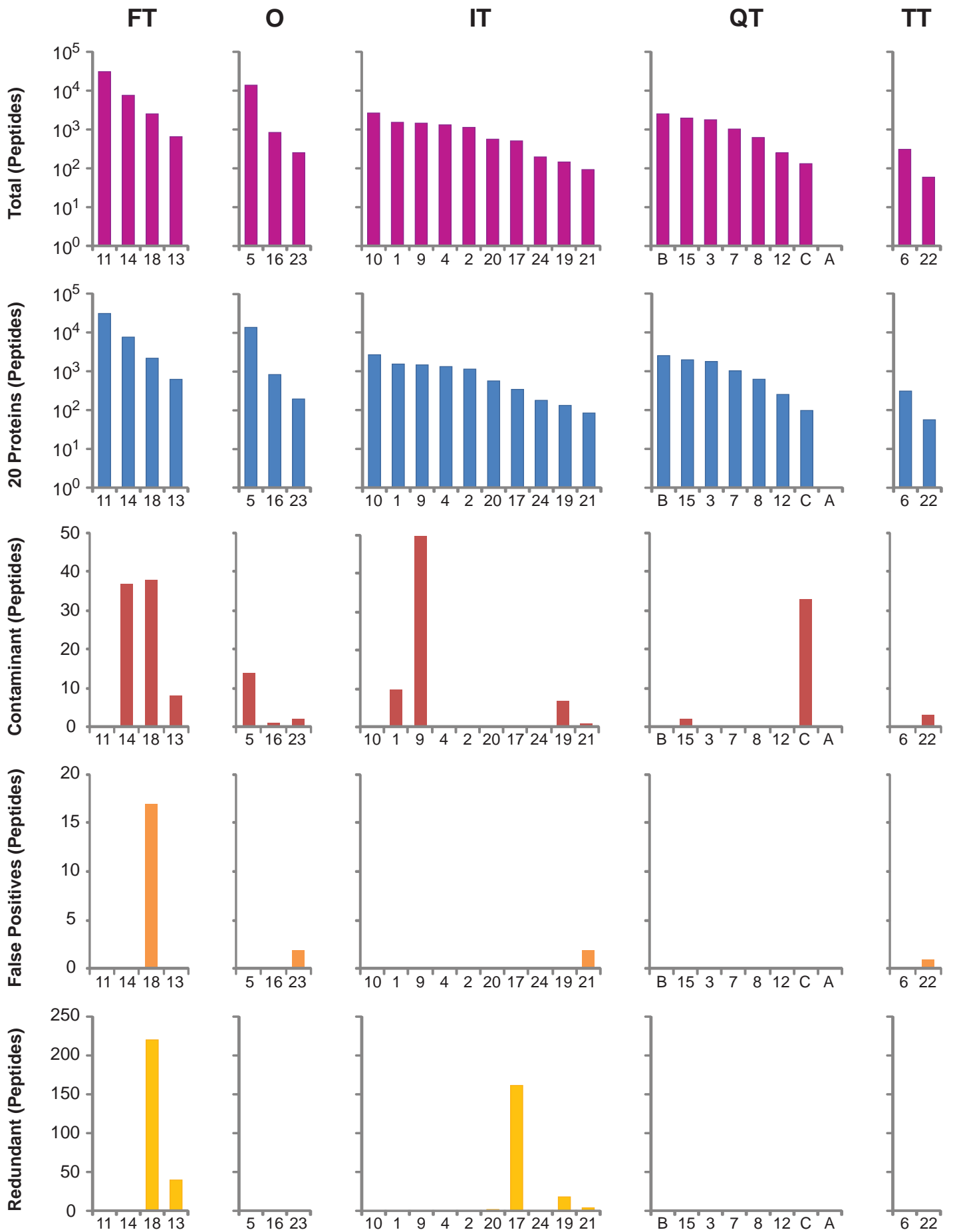
=====
# Aligned_sequences: 2
# 1: /IPI00470809.3/readseq.in.8979
# 2: IOH29199,
# Matrix: EBLOSUM62
# Gap_penalty: 12
# Extend_penalty: 2
#
# Length: 1097
# Identity:      978/1097 (89.2%)
# Similarity:   978/1097 (89.2%)
# Gaps:         118/1097 (10.8%)
# Score: 4984
=====

```

| | | | |
|---------------|------|--|------|
| /IPI00470809. | 1 | MAPRPKKQPKDNPLHGRELNVVPSLGRQTSLTTSVIPKAEQSVAYKDFIY | 50 |
| IOH29199, | 1 | M----- | 1 |
| /IPI00470809. | 51 | FTVFEGRVNRNVSEVSVVEYLCSQPCVVNLEAVVSSEFRSSIPVYKKRWKNE | 100 |
| IOH29199, | 2 | ----- | 1 |
| /IPI00470809. | 101 | KHLHTRSRTQIVHVKFPSIMVYRDDYFIRHSISVSAVIVRAWITHKYSGRD | 150 |
| IOH29199, | 2 | -----VYRDDYFIRHSISVSAVIVRAWITHKYSGRD | 32 |
| /IPI00470809. | 151 | WNVKWEENLLHAVAKNYTLLQTIPPFERPFKDHQVCLEWNGYIWNLRAN | 200 |
| IOH29199, | 33 | WNVKWEENLLHAVAKNYTLLQTIPPFERPFKDHQVCLEWNGYIWNLRAN | 82 |
| /IPI00470809. | 201 | RIPQCPLENDVVALLGFPYASSGENTGIVKKFPRFRNRELEATRRQRMDY | 250 |
| IOH29199, | 83 | RIPQCPLENDVVALLGFPYASSGENTGIVKKFPRFRNRELEATRRQRMDY | 132 |
| /IPI00470809. | 251 | PVFTVSLWLWYLLHYCKANLCGILYFVDSNEMYGTPSVFLTEEGYLHIQMH | 300 |
| IOH29199, | 133 | PVFTVSLWLWYLLHYCKANLCGILYFVDSNEMYGTPSVFLTEEGYLHIQMH | 182 |
| /IPI00470809. | 301 | LVKGEDLAVKTKFIIPLKEWFRLDISFNGGQIVVTTSIGQDLKSYHNQTI | 350 |
| IOH29199, | 183 | LVKGEDLAVKTKFIIPLKEWFRLDISFNGGQIVVTTSIGQDLKSYHNQTI | 232 |
| /IPI00470809. | 351 | SFREDPHYNDTAGYFIIIGSRYVAGIEGFFGPKYRLRSLHPAQIFNPL | 400 |
| IOH29199, | 233 | SFREDPHYNDTAGYFIIIGSRYVAGIEGFFGPKYRLRSLHPAQIFNPL | 282 |
| /IPI00470809. | 401 | LEKQLAEQIKLYYERCAEVQEIIVSVYASAACHGGERQEACHLHNSYLDLQ | 450 |
| IOH29199, | 283 | LEKQLAEQIKLYYERCAEVQEIIVSVYASAACHGGERQEACHLHNSYLDLQ | 332 |
| /IPI00470809. | 451 | RRYGRPSMCRAFPWEKELKDKHPSLFQALLEMDLLTVPRNQNESVSEIGG | 500 |
| IOH29199, | 333 | RRYGRPSMCRAFPWEKELKDKHPSLFQALLEMDLLTVPRNQNESVSEIGG | 382 |
| /IPI00470809. | 501 | KIFEKAVKRLSSIDGLHQISSIVPFLTDSSCCGYHKASYLAVFYETGLN | 550 |
| IOH29199, | 383 | KIFEKAVKRLSSIDGLHQISSIVPFLTDSSCCGYHKASYLAVFYETGLN | 432 |
| /IPI00470809. | 551 | VPRDQLQGMLYSLVGGQGSERLSSMNLGKHYQGIDNYPLDWELSYAYYS | 600 |
| IOH29199, | 433 | VPRDQLQGMLYSLVGGQGSERLSSMNLGKHYQGIDNYPLDWELSYAYYS | 482 |
| /IPI00470809. | 601 | NIATKTPLDQHTLQGDQAYVETIRLKDDEILKVQTKEDGDVFMWLKHEAT | 650 |
| IOH29199, | 483 | NIATKTPLDQHTLQGDQAYVETIRLKDDEILKVQTKEDGDVFMWLKHEAT | 532 |
| /IPI00470809. | 651 | RGNAAAQORLAQMLFWGQGVAKNPEAAIEWYAKGALETEDPALIYDYAI | 700 |
| IOH29199, | 533 | RGNAAAQORLAQMLFWGQGVAKNPEAAIEWYAKGALETEDPALIYDYAI | 582 |
| /IPI00470809. | 701 | VLFGQGQVKKNRRLALELMKKAASKGLHQAVNGLGWYHFKKKNYAKAAK | 750 |
| IOH29199, | 583 | VLFGQGQVKKNRRLALELMKKAASKGLHQAVNGLGWYHFKKKNYAKAAK | 632 |
| /IPI00470809. | 751 | YWLKAEEMGNPDASYNLGVHLHDGIFPGVPGRNQTLAGEYFHKAAQGGHM | 800 |
| IOH29199, | 633 | YWLKAEEMGNPDASYNLGVHLHDGIFPGVPGRNQTLAGEYFHKAAQGGHM | 682 |
| /IPI00470809. | 801 | EGTLWCSLYYITGNLETFPRDPEKAVVWAKHVAEKNGYLGHVIRKGLNAY | 850 |
| IOH29199, | 683 | EGTLWCSLYYITGNLETFPRDPEKAVVWAKHVAEKNGYLGHVIRKGLNAY | 732 |
| /IPI00470809. | 851 | LEGSWHEALLYVLAETGIEVSQTNLAHICEERPDLARRYLGVNCVWRY | 900 |
| IOH29199, | 733 | LEGSWHEALLYVLAETGIEVSQTNLAHICEERPDLARRYLGVNCVWRY | 782 |
| /IPI00470809. | 901 | YNFSVFQIDAPSFAYLKMGDLYYGHQNSQDLELSVQMYAQAALDGDSQ | 950 |
| IOH29199, | 783 | YNFSVFQIDAPSFAYLKMGDLYYGHQNSQDLELSVQMYAQAALDGDSQ | 832 |
| /IPI00470809. | 951 | GFFNLALLIEGTIIPHHILDPLEIDSTLHNSNISILQELYERCWSHSNE | 1000 |
| IOH29199, | 833 | GFFNLALLIEGTIIPHHILDPLEIDSTLHNSNISILQELYERCWSHSNE | 882 |
| /IPI00470809. | 1001 | ESFSPCSLAWLYLHLRLLWGAILHSALIYFLGTFLLSILIAWTVQYFQSV | 1050 |
| IOH29199, | 883 | ESFSPCSLAWLYLHLRLLWGAILHSALIYFLGTFLLSILIAWTVQYFQSV | 932 |
| /IPI00470809. | 1051 | SASDPPRPSQASPDSTATSTASPAVTPAADASDQDQPTVTNNPEPRG | 1097 |
| IOH29199, | 933 | SASDPPRPSQASPDSTATSTASPAVTPAADASDQDQPTVTNNPEPRG | 979 |

Supplementary Figure 7. Number of tandem mass spectra assigned (redundant peptides) to tryptic peptides.

Data are from **Supplementary Table 6** online, based on the initial reports submitted by each lab. Types of mass spectrometers: FTICR (FT), Orbitrap (O), Iontrap (IT), QToF (QT), ToFToF (TT). The technology of lab A does not provide tandem mass spectra (see **Online Methods**).



Bell et al. Supplementary Figure 7

Supplementary Figure 8. Naming errors, redundant identifications and false positive identifications. Potential sources of error are presented as examples for erroneous results reported by Laboratory 18 (**Supplementary Table 7b** online). **(a)** Incorrect naming. The Mascot search result offers LOC55068 as the top hit whereas the correct result is either PIG38 or proliferation inducing protein 38. **(b)** False positive. Peptides have been assigned to the cytosolic ovarian carcinoma antigen 1 which is incorrect. In fact all tandem MS (queries 28, 50, 75, 91, 183, 195 and 201) have been assigned to peptides assigned to PIG38 (panel **a**). **(c)** Redundant Identifications. The Mascot search result reports two proteins that are identified by the identical set of queries but with a single peptide difference (asterisk). The higher score is for the correct peptide (above). However the correct protein is the NIBP protein (arrow). **(c)** Shown are selected segments of Mascot peptide reports (www.matrixscience.com) as recorded by one of the laboratories.

a

4. [gi18922227](#) Mass: 73755 Score: 1136 Queries matched: 28
hypothetical protein LOC55068 [Homo sapiens]
 Check to include this hit in error tolerant search or archive report

| Query | Observed | Mr(expt) | Mr(calc) | Delta | Miss | Score | Expect | Rank | Peptide |
|---|----------|----------|----------|-------|------|-------|---------|------|---------------------------------------|
| <input checked="" type="checkbox"/> 9 | 385.65 | 769.29 | 769.37 | -0.08 | 0 | 31 | 0.14 | 1 | K.HSEELR.N |
| <input checked="" type="checkbox"/> 13 | 395.17 | 788.32 | 788.40 | -0.08 | 1 | 45 | 0.01 | 1 | R.KLEEDR.L |
| <input checked="" type="checkbox"/> 27 | 417.67 | 833.33 | 833.41 | -0.08 | 0 | 15 | 9.6 | 1 | R.TEENLT.KD |
| <input checked="" type="checkbox"/> 28 | 422.69 | 843.36 | 843.47 | -0.11 | 0 | 32 | 0.19 | 1 | R.NEVELLK.Q |
| <input checked="" type="checkbox"/> 42 | 445.66 | 889.31 | 889.41 | -0.10 | 0 | 5 | 80 | 4 | K.AWDHFSKA |
| <input checked="" type="checkbox"/> 50 | 471.71 | 941.40 | 941.50 | -0.10 | 0 | 37 | 0.048 | 1 | K.AIYLSGYR.M |
| <input checked="" type="checkbox"/> 75 | 528.72 | 1055.42 | 1055.55 | -0.13 | 0 | 40 | 0.022 | 1 | R.LHVDFAQAR.D |
| <input checked="" type="checkbox"/> 89 | 558.19 | 1114.36 | 1114.50 | -0.14 | 0 | 50 | 0.0023 | 1 | R.FAEFMDK.A |
| <input checked="" type="checkbox"/> 91 | 566.20 | 1130.38 | 1130.50 | -0.12 | 0 | (39) | 0.026 | 1 | R.FAEFMDK.A + Oxidation (M) |
| <input checked="" type="checkbox"/> 107 | 623.73 | 1245.45 | 1245.61 | -0.16 | 0 | 76 | 6.2e-06 | 1 | R.NAQSEQLMGIR.R |
| <input checked="" type="checkbox"/> 112 | 631.74 | 1261.46 | 1261.61 | -0.15 | 0 | (48) | 0.0035 | 1 | R.NAQSEQLMGIR.R + Oxidation (M) |
| <input checked="" type="checkbox"/> 116 | 640.25 | 1278.49 | 1278.65 | -0.16 | 0 | 66 | 4.6e-05 | 1 | K.QEFTGVGATLEK.R |
| <input checked="" type="checkbox"/> 119 | 642.27 | 1282.52 | 1282.65 | -0.13 | 0 | 62 | 0.00011 | 1 | K.EEQSHTQALLK.V |
| <input checked="" type="checkbox"/> 121 | 651.72 | 1301.42 | 1301.56 | -0.14 | 0 | 57 | 0.00037 | 1 | K.ATHEQEEMEEAK.E |
| <input checked="" type="checkbox"/> 142 | 710.29 | 1418.56 | 1418.73 | -0.17 | 0 | 88 | 2.8e-07 | 1 | K.ISANEIEMLLMR.L |
| <input checked="" type="checkbox"/> 145 | 718.29 | 1434.56 | 1434.72 | -0.16 | 0 | (68) | 3e-05 | 1 | K.ISANEIEMLLMR.L + Oxidation (M) |
| <input checked="" type="checkbox"/> 157 | 725.29 | 1448.57 | 1448.75 | -0.18 | 0 | 81 | 1.6e-06 | 1 | R.VDESALAAQAYALKE |
| <input checked="" type="checkbox"/> 158 | 726.28 | 1450.54 | 1450.72 | -0.18 | 0 | (64) | 6.3e-05 | 1 | K.ISANEIEMLLMR.L + 2 Oxidation (M) |
| <input checked="" type="checkbox"/> 183 | 592.92 | 1775.74 | 1775.92 | -0.18 | 1 | 29 | 0.19 | 1 | R.WQLDAYRNEVELLK.Q |
| <input checked="" type="checkbox"/> 185 | 607.52 | 1819.55 | 1819.80 | -0.26 | 1 | (44) | 0.0074 | 1 | K.ATHEQEEMEEAKENFK.N |
| <input checked="" type="checkbox"/> 188 | 612.87 | 1835.59 | 1835.80 | -0.21 | 1 | 46 | 0.0038 | 1 | K.ATHEQEEMEEAKENFK.N + Oxidation (M) |
| <input checked="" type="checkbox"/> 195 | 642.57 | 1924.68 | 1924.88 | -0.20 | 0 | (46) | 0.0034 | 1 | R.SANQFYSMVQSANSVHR.R |
| <input checked="" type="checkbox"/> 201 | 647.89 | 1940.66 | 1940.88 | -0.22 | 0 | 49 | 0.0015 | 1 | R.SANQFYSMVQSANSVHR.R + Oxidation (M) |
| <input checked="" type="checkbox"/> 215 | 690.26 | 2067.76 | 2068.04 | -0.28 | 1 | 59 | 0.00021 | 1 | K.SELEQAKEQSHTQALLK.V |
| <input checked="" type="checkbox"/> 240 | 764.95 | 2291.82 | 2292.12 | -0.30 | 1 | 89 | 1.3e-07 | 1 | R.VDESALAAQAYALKEENDSLR.W |
| <input checked="" type="checkbox"/> 263 | 640.02 | 2556.03 | 2556.32 | -0.28 | 0 | 63 | 5.2e-05 | 1 | R.LRPPSPPAIMHYSEEAALLAEK.L |
| <input checked="" type="checkbox"/> 289 | 1044.73 | 3131.16 | 3131.51 | -0.35 | 0 | 115 | 2.6e-10 | 1 | K.ELVETNGHSHEDSNEINVLTVLVNQDR.E |
| <input checked="" type="checkbox"/> 290 | 783.80 | 3131.17 | 3131.51 | -0.34 | 0 | (43) | 0.0039 | 1 | K.ELVETNGHSHEDSNEINVLTVLVNQDR.E |

Proteins matching the same set of peptides:
[gi18848204](#) Mass: 73815 Score: 1136 Queries matched: 28
PIG38 protein [Homo sapiens]
[gi146981973](#) Mass: 73729 Score: 1136 Queries matched: 28
proliferation-inducing protein 38 [Homo sapiens]
[gi155666034](#) Mass: 73144 Score: 1136 Queries matched: 28
OTIUMP00000018332 [Homo sapiens]

b

10. [gi|32528291](#) Mass: 70835 Score: 202 Queries matched: 7
cytosolic ovarian carcinoma antigen 1 isoform b [Homo sapiens]
 Check to include this hit in error tolerant search or archive report

| Query | Observed | Mr(expt) | Mr(calc) | Delta | Miss | Score | Expect | Rank | Peptide |
|---------------------|----------|----------|----------|-------|------|-------|--------|------|--------------------------------------|
| 28 | 422.69 | 843.36 | 843.47 | -0.11 | 0 | 32 | 0.19 | 1 | R.NEVLLK.Q |
| 50 | 471.71 | 941.40 | 941.50 | -0.10 | 0 | 37 | 0.048 | 1 | K.ALYLSGYR.I |
| 75 | 528.72 | 1055.42 | 1055.55 | -0.13 | 0 | 40 | 0.022 | 1 | R.LHVDFAR.D |
| 91 | 566.20 | 1130.38 | 1130.50 | -0.12 | 0 | 26 | 0.5 | 2 | R.FAEEYMVDK.A |
| 183 | 592.92 | 1775.74 | 1775.92 | -0.18 | 1 | 29 | 0.19 | 1 | R.WQLDAYRNEVLLK.Q |
| 195 | 642.57 | 1924.68 | 1924.88 | -0.20 | 0 | (33) | 0.072 | 2 | R.SANFYSMIQSANSVHR.R |
| 201 | 647.89 | 1940.66 | 1940.88 | -0.22 | 0 | 38 | 0.019 | 2 | R.SANFYSMIQSANSVHR.R + Oxidation (M) |

Proteins matching the same set of peptides:

- [gi|32528293](#) Mass: 67377 Score: 202 Queries matched: 7
cytosolic ovarian carcinoma antigen 1 isoform a [Homo sapiens]
[gi|62913976](#) Mass: 53914 Score: 202 Queries matched: 7
COVA1 protein [Homo sapiens]
[gi|80478560](#) Mass: 54296 Score: 202 Queries matched: 7
Unknown (protein for IMAGE:6148379) [Homo sapiens]

C

8. [gi|15620823](#) Mass: 132671 Score: 295 Queries matched: 5
KIAA1882 protein [Homo sapiens]
 Check to include this hit in error tolerant search or archive report

| Query | Observed | Mr(expt) | Mr(calc) | Delta | Miss | Score | Expect | Rank | Peptide |
|---|----------|----------|----------|-------|------|-------|---------|------|---------------------|
| <input checked="" type="checkbox"/> 65 | 516.74 | 1031.46 | 1031.56 | -0.10 | 0 | 59 | 0.00035 | 1 | R.VSTLPATSTR.Q |
| <input checked="" type="checkbox"/> 138 | 677.77 | 1353.52 | 1353.64 | -0.12 | 0 | 75 | 8.4e-06 | 1 | K.DVAQSLENYTSK.C |
| <input checked="" type="checkbox"/> 139 | 681.28 | 1360.55 | 1360.67 | -0.12 | 0 | 55 | 0.00069 | 1 | R.FQGSTLPAAANR.H |
| <input checked="" type="checkbox"/> 160 | 505.56 | 1513.67 | 1513.80 | -0.14 | 0 | 63 | 7.6e-05 | 1 | K.SPFIYSPIAHNR.G |
| <input checked="" type="checkbox"/> 202 | 879.92 | 1757.82 | 1757.98 | -0.16 | 0 | 45 | 0.0057 | 1 | K.LLLETLPGYSLSDPK.D |

Proteins matching the same set of peptides:

- [gi|28394692](#) Mass: 105447 Score: 295 Queries matched: 5
T1 [Homo sapiens]
[gi|34782807](#) Mass: 101835 Score: 295 Queries matched: 5
NIBP protein [Homo sapiens]
[gi|40255276](#) Mass: 140761 Score: 295 Queries matched: 5
NIK and IKK(beta) binding protein [Homo sapiens]



9. [gi|52545861](#) Mass: 106232 Score: 268 Queries matched: 5
hypothetical protein [Homo sapiens]
 Check to include this hit in error tolerant search or archive report

| Query | Observed | Mr(expt) | Mr(calc) | Delta | Miss | Score | Expect | Rank | Peptide |
|---------------------|----------|----------|----------|-------|------|-------|---------|------|---------------------|
| 65 | 516.74 | 1031.46 | 1031.56 | -0.10 | 0 | 59 | 0.00035 | 1 | R.VSTLPATSTR.Q |
| 138 | 677.77 | 1353.52 | 1353.68 | -0.16 | 1 | 48 | 0.0039 | 2 | K.KDVASLENYTSK.C |
| 139 | 681.28 | 1360.55 | 1360.67 | -0.12 | 0 | 55 | 0.00069 | 1 | R.FQGSTLPAAANR.H |
| 160 | 505.56 | 1513.67 | 1513.80 | -0.14 | 0 | 63 | 7.6e-05 | 1 | K.SPFIYSPIAHNR.G |
| 202 | 879.92 | 1757.82 | 1757.98 | -0.16 | 0 | 45 | 0.0057 | 1 | K.LLLETLPGYSLSDPK.D |



Supplementary Figure 9. Database matching and acrylamide alkylation. For the potential source of error for laboratory 19, a Mascot search result revealed a single tryptic peptide identified by query 185 that identifies the coagulation factor II protein. The assigned peptide was the propionamide derivative that results from acrylamide alkylation of cysteine residues. Acrylamide alkylation resulted from the preparative electrophoresis purification procedure employed to purify several of the test proteins. For this analysis the coagulation factor II protein would not be identified in this sample if the search did not allow for acrylamide alkylation.

16. [gi|1335344](#) Mass: 69241 Score: 73 Queries matched: 1
unnamed protein product [Homo sapiens]

Check to include this hit in error tolerant search or archive report

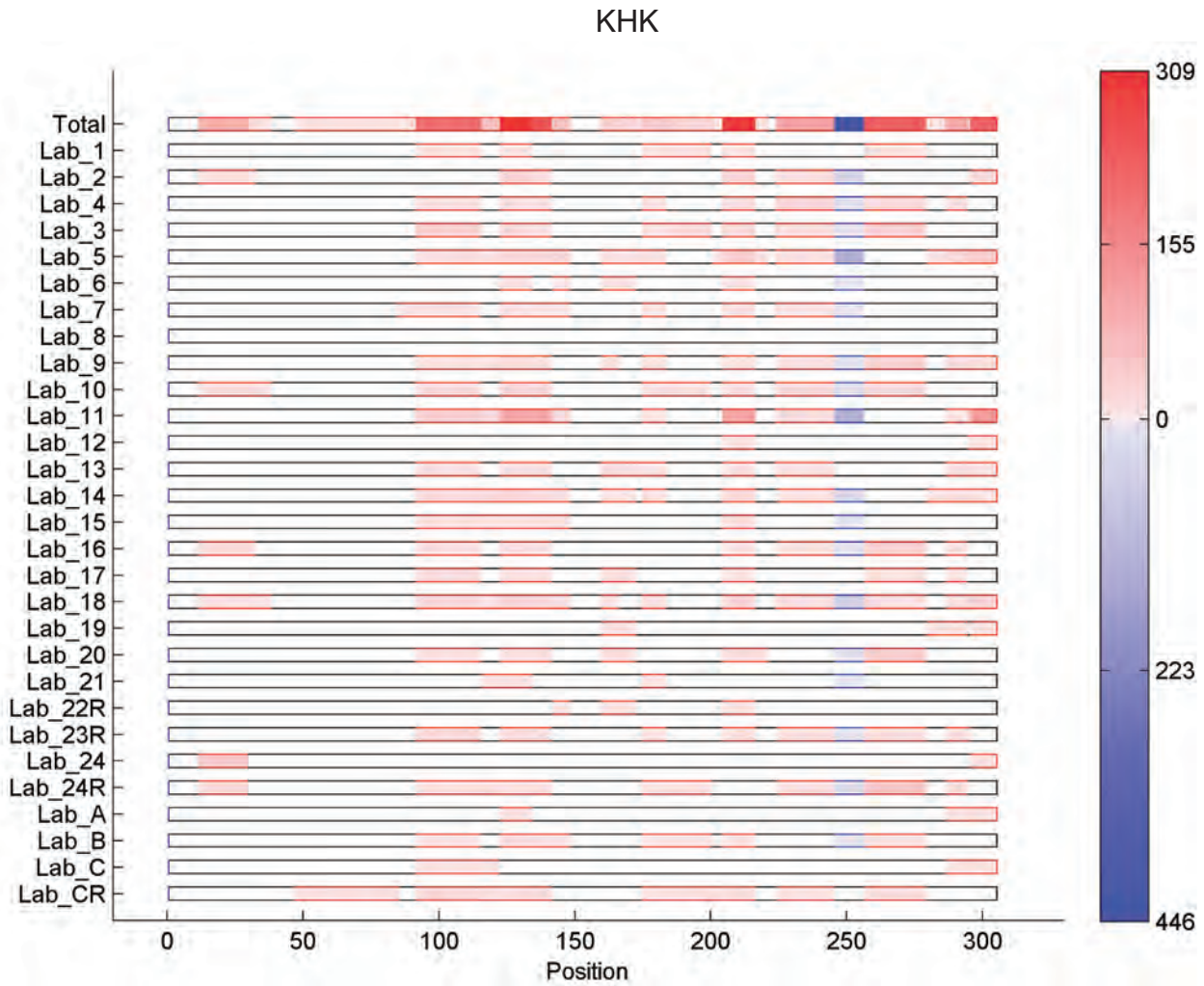
| Query | Observed | Mr(expt) | Mr(calc) | Delta | Miss | Score | Expect | Rank | Peptide |
|---|----------|----------|----------|-------|------|-------|---------|------|---|
| <input checked="" type="checkbox"/> 185 | 749.36 | 2245.06 | 2245.26 | -0.20 | 0 | 73 | 1.4e-05 | 1 | K.GQPSVLQVVNLPIVERPVCK.D + Propionamide (C) |

Proteins matching the same set of peptides:

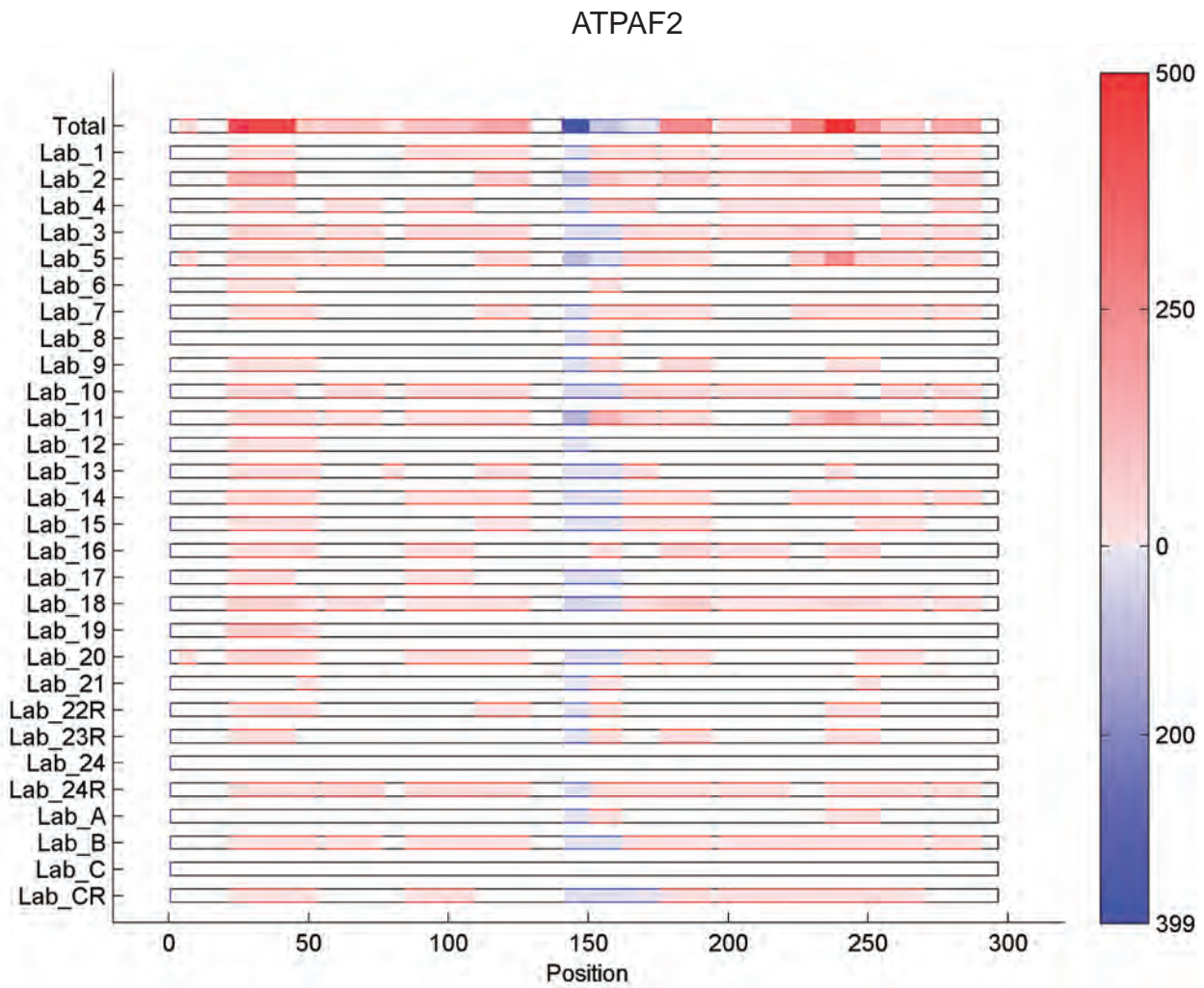
| | | | |
|--|-------------|-----------|--------------------|
| gi 4503635 | Mass: 69992 | Score: 73 | Queries matched: 1 |
| coagulation factor II precursor [Homo sapiens] | | | |
| gi 30802115 | Mass: 69964 | Score: 73 | Queries matched: 1 |
| Coagulation factor II (thrombin) [Homo sapiens] | | | |
| gi 38018090 | Mass: 33616 | Score: 73 | Queries matched: 1 |
| prothrombin [Homo sapiens] | | | |
| gi 38018092 | Mass: 29546 | Score: 73 | Queries matched: 1 |
| prothrombin B-chain [Homo sapiens] | | | |
| gi 62897109 | Mass: 70091 | Score: 73 | Queries matched: 1 |
| coagulation factor II precursor variant [Homo sapiens] | | | |
| gi 62897113 | Mass: 70019 | Score: 73 | Queries matched: 1 |
| coagulation factor II precursor variant [Homo sapiens] | | | |
| gi 67624831 | Mass: 69920 | Score: 73 | Queries matched: 1 |
| prothrombin [Homo sapiens] | | | |

Supplementary Figure 10. Heat Maps for 20 Test Sample Proteins. Heat maps for each protein as deduced from the raw data submitted to Tranche from each lab. Heat maps indicate the frequency of tandem mass spectra assigned to tryptic peptides (red) with the peptides of mass 1250 ± 5 Da indicated in blue. The sum of the data for all labs is also shown for each protein (Total). The scale bar is on the right. After centralized analysis, the data shown here for lab 24 was excluded from the data of **Fig. 1** (as described in **Online Methods**).

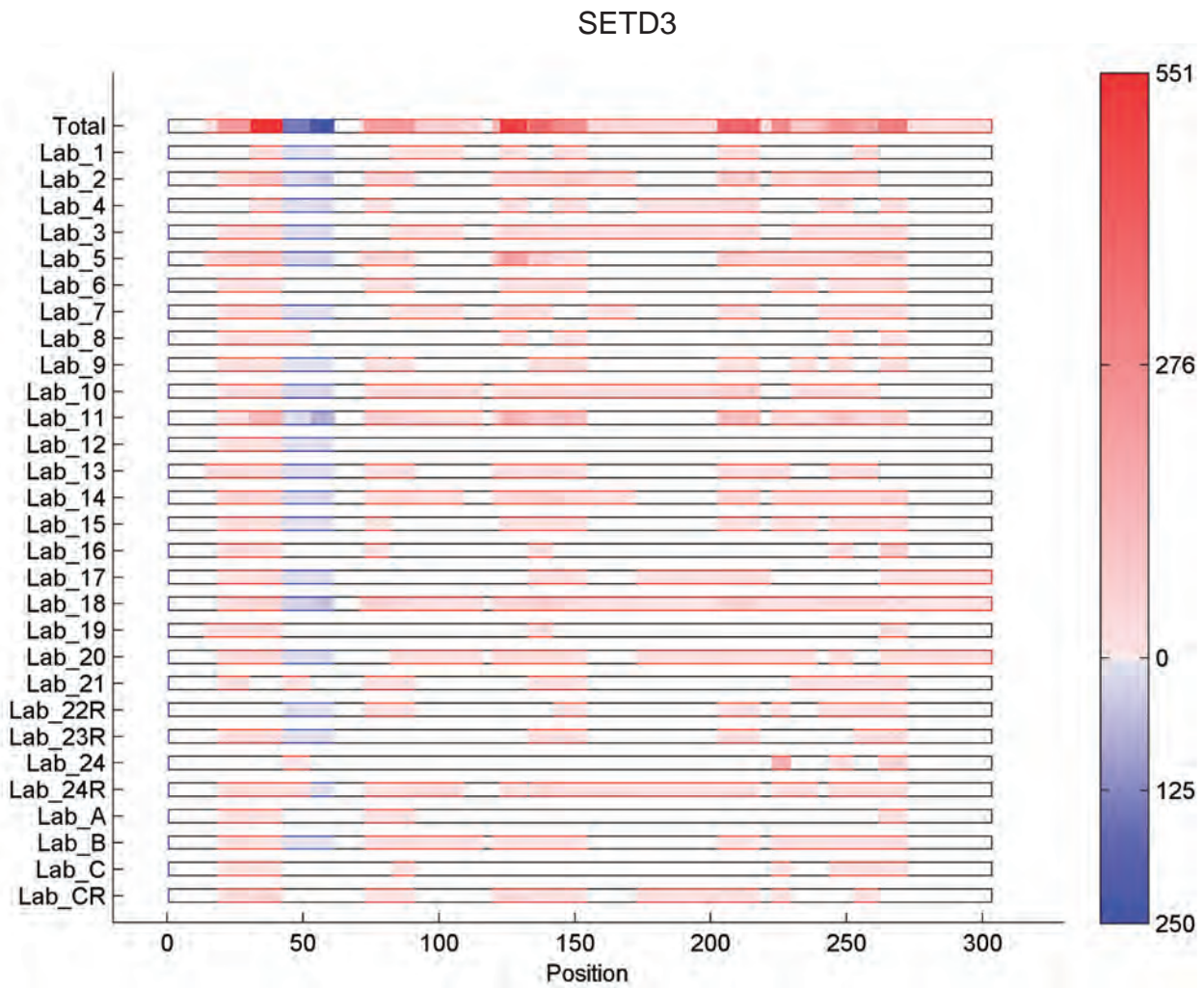
a



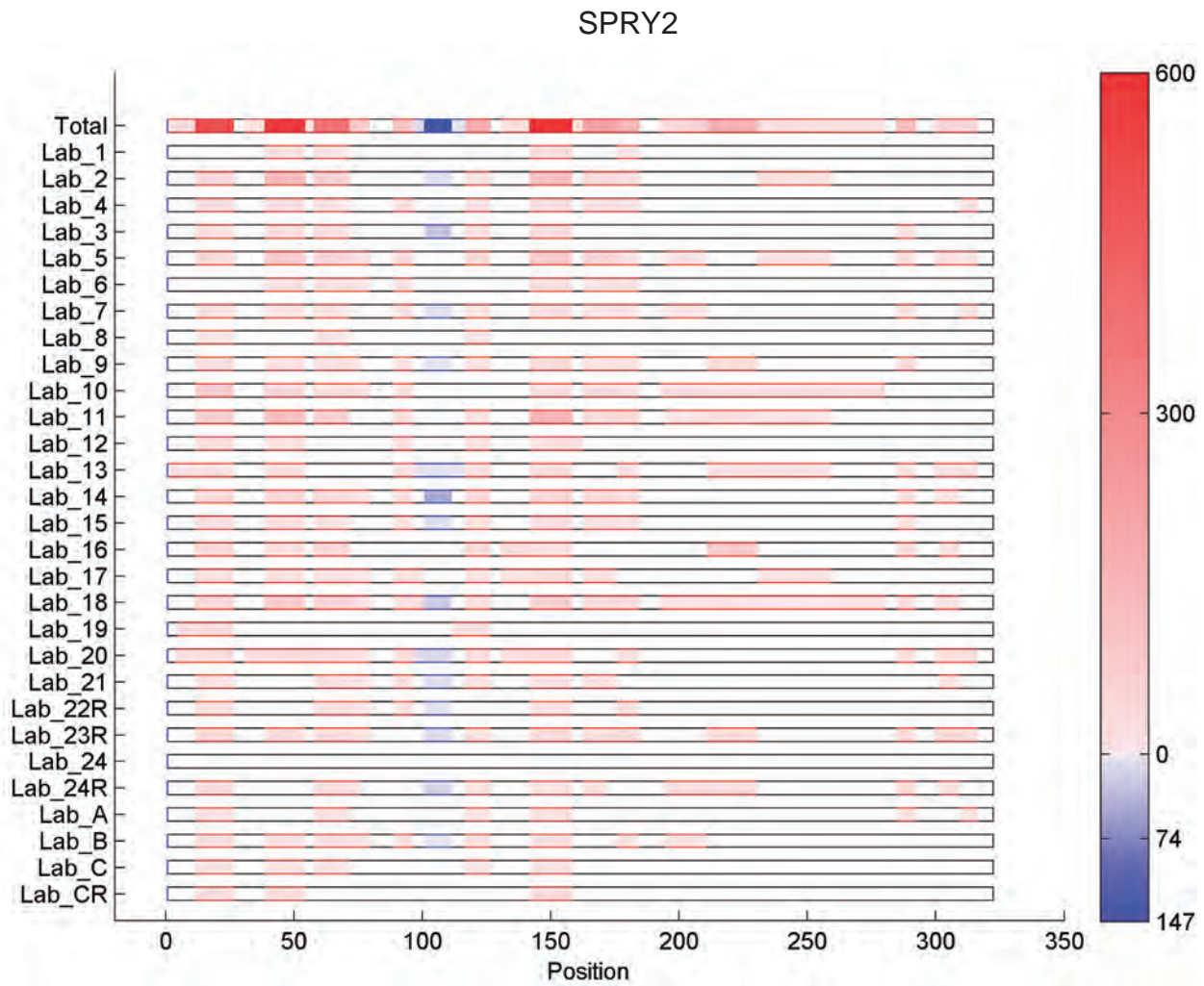
b



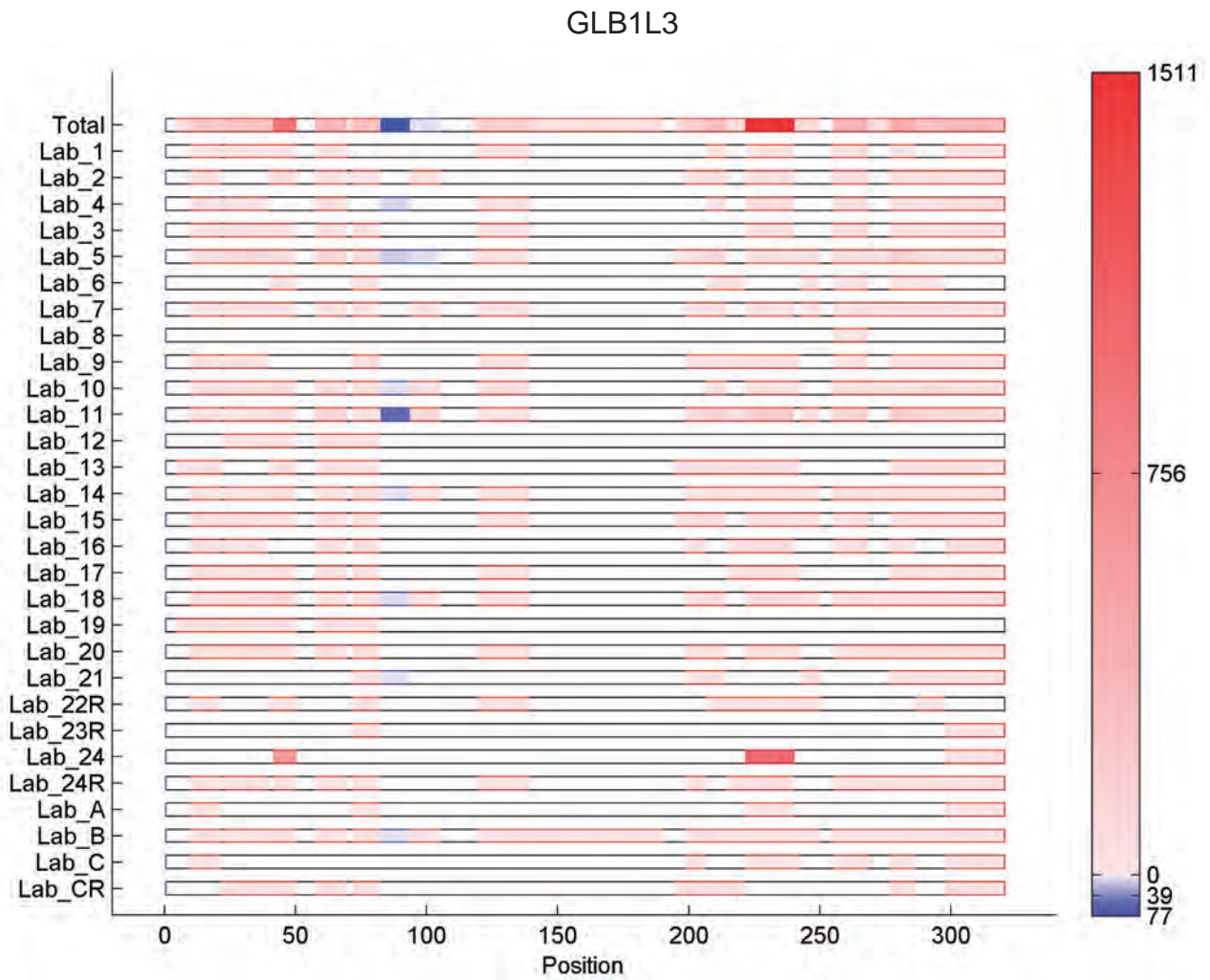
C



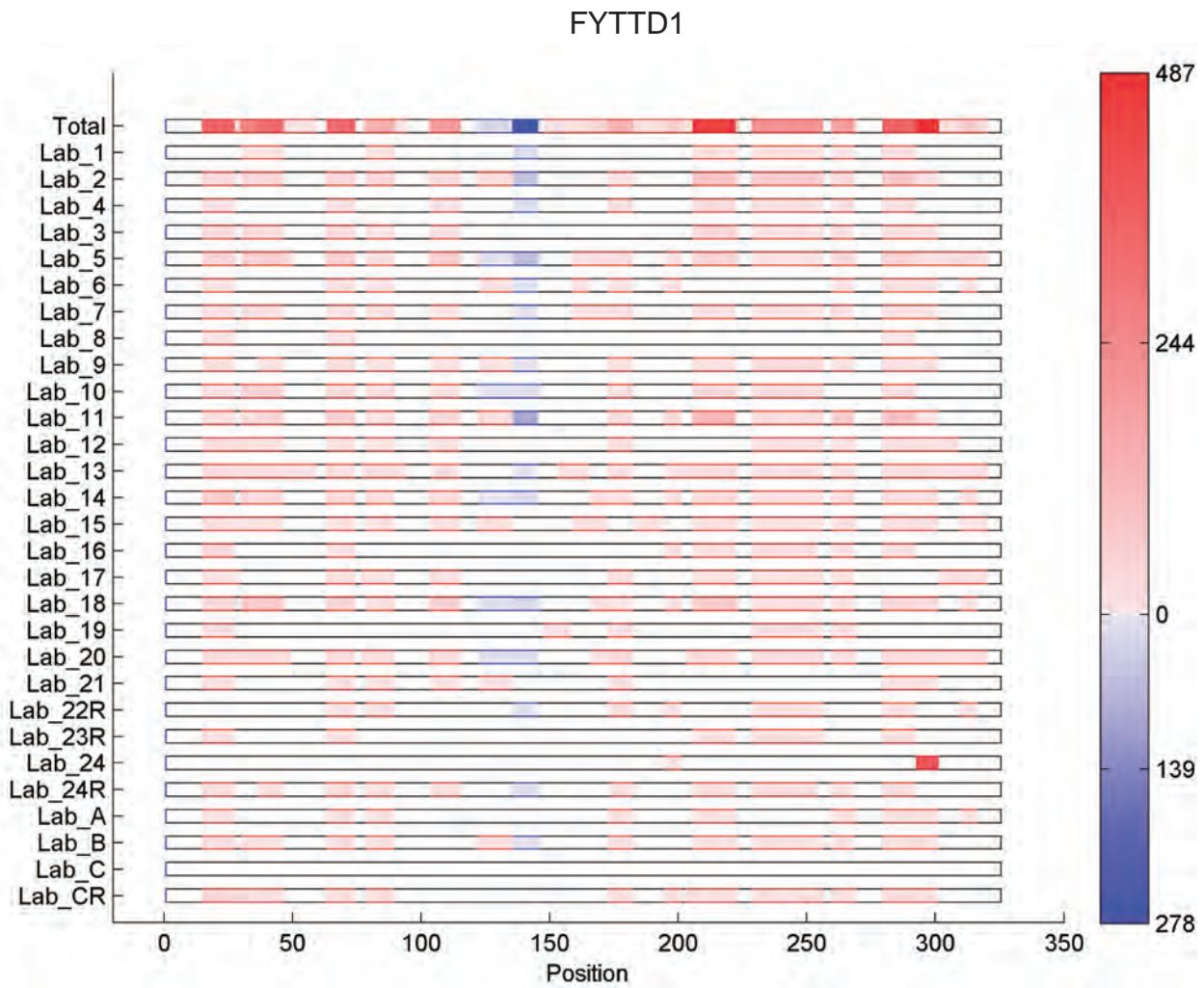
d



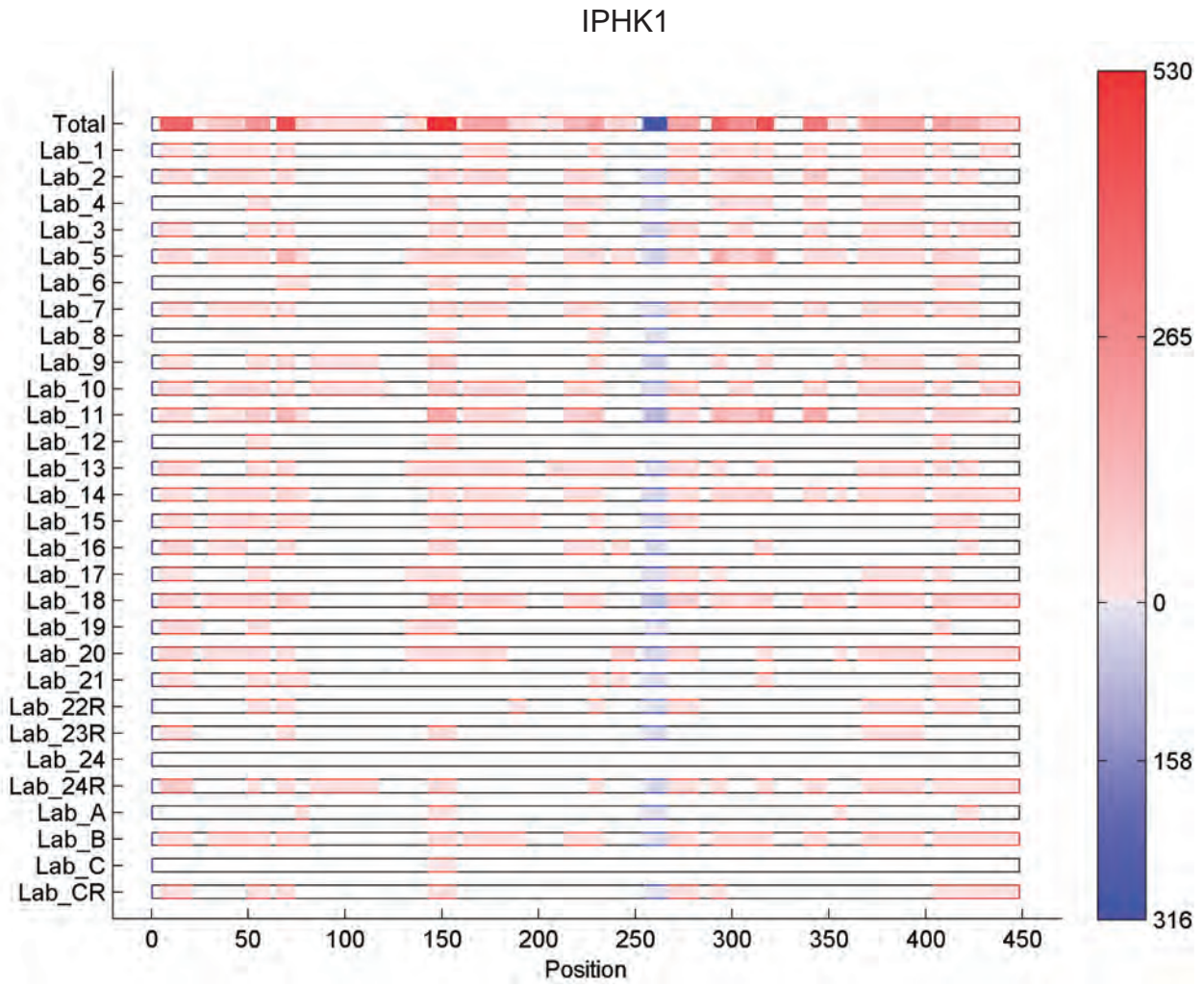
e



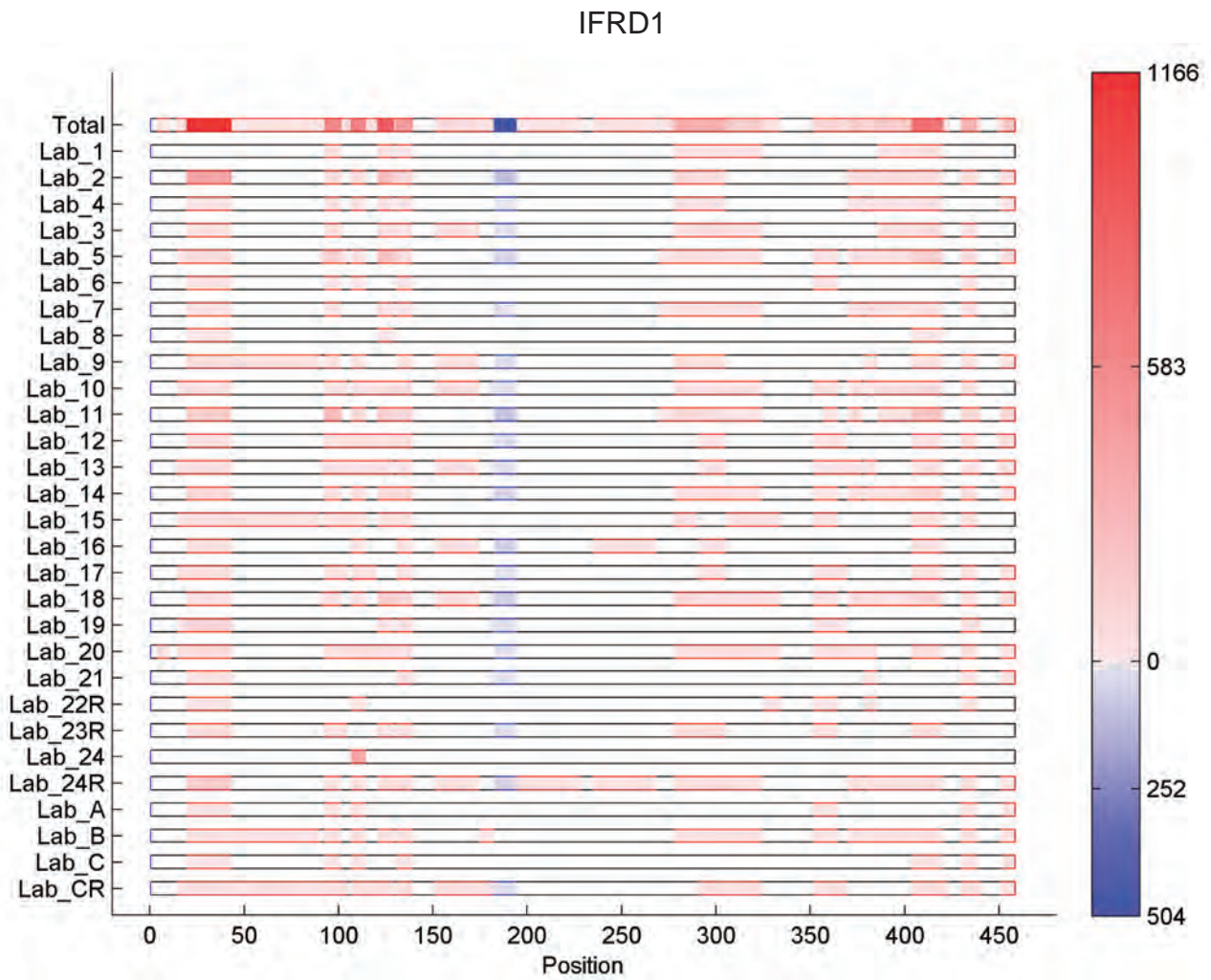
f

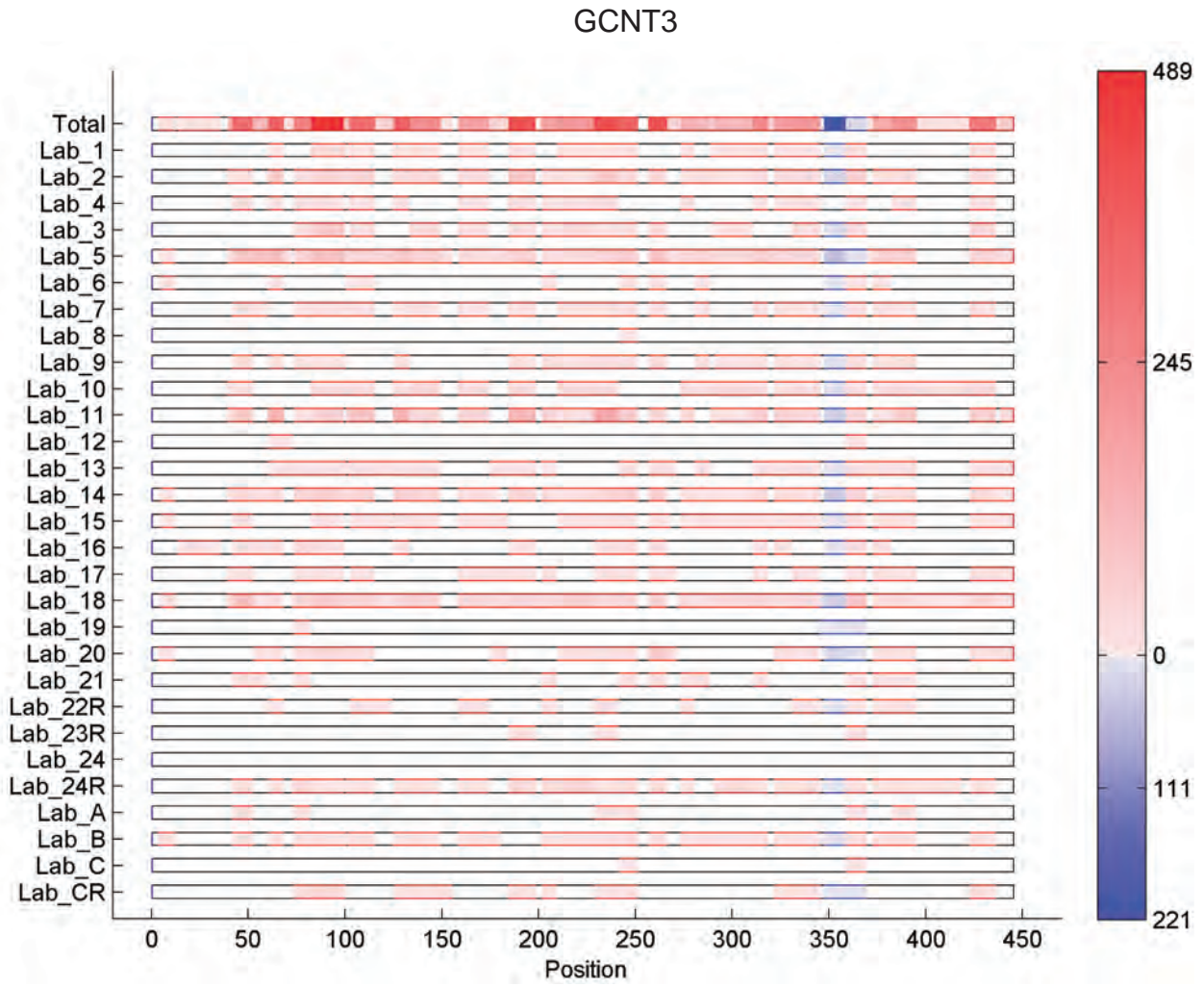


g

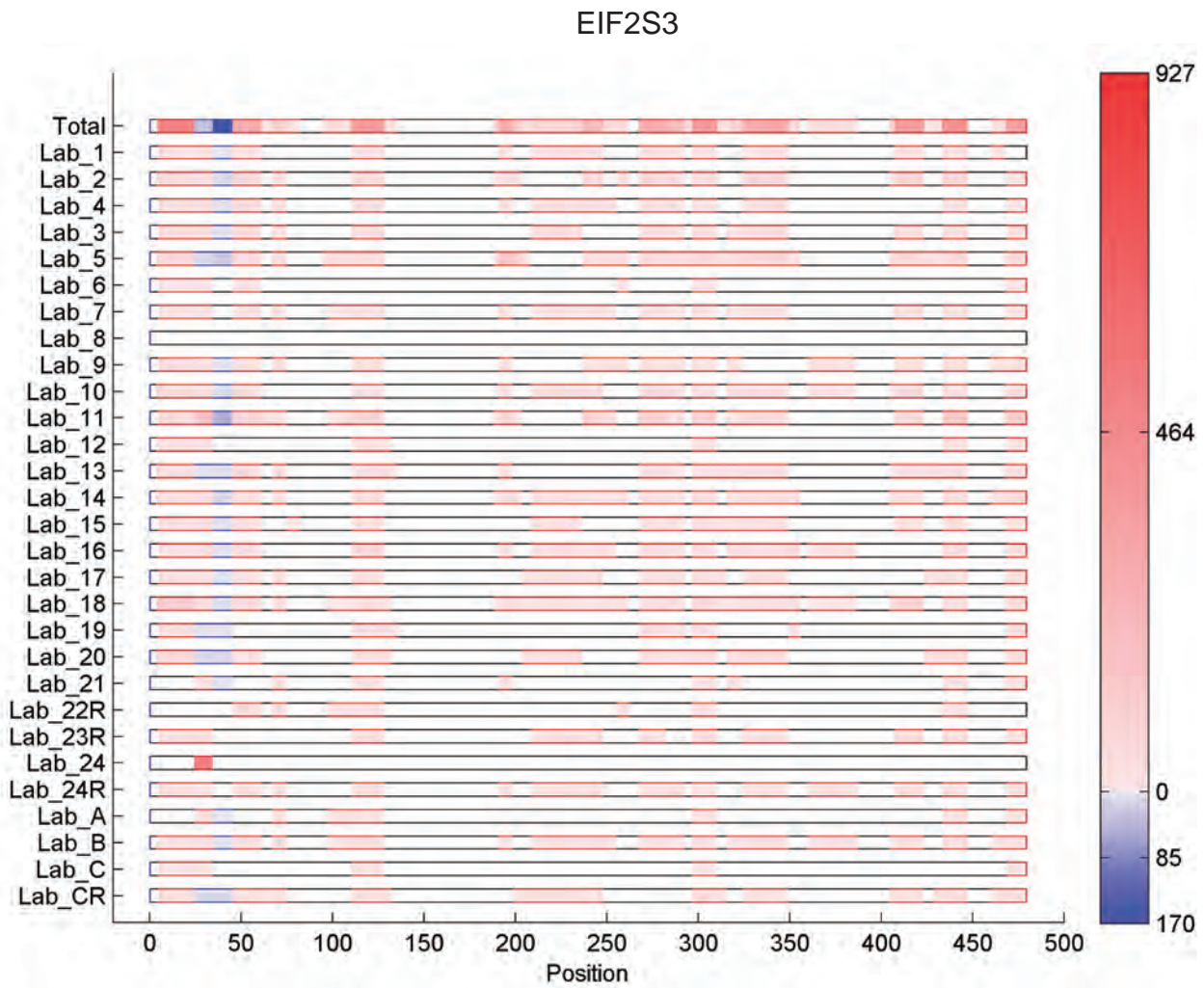


h

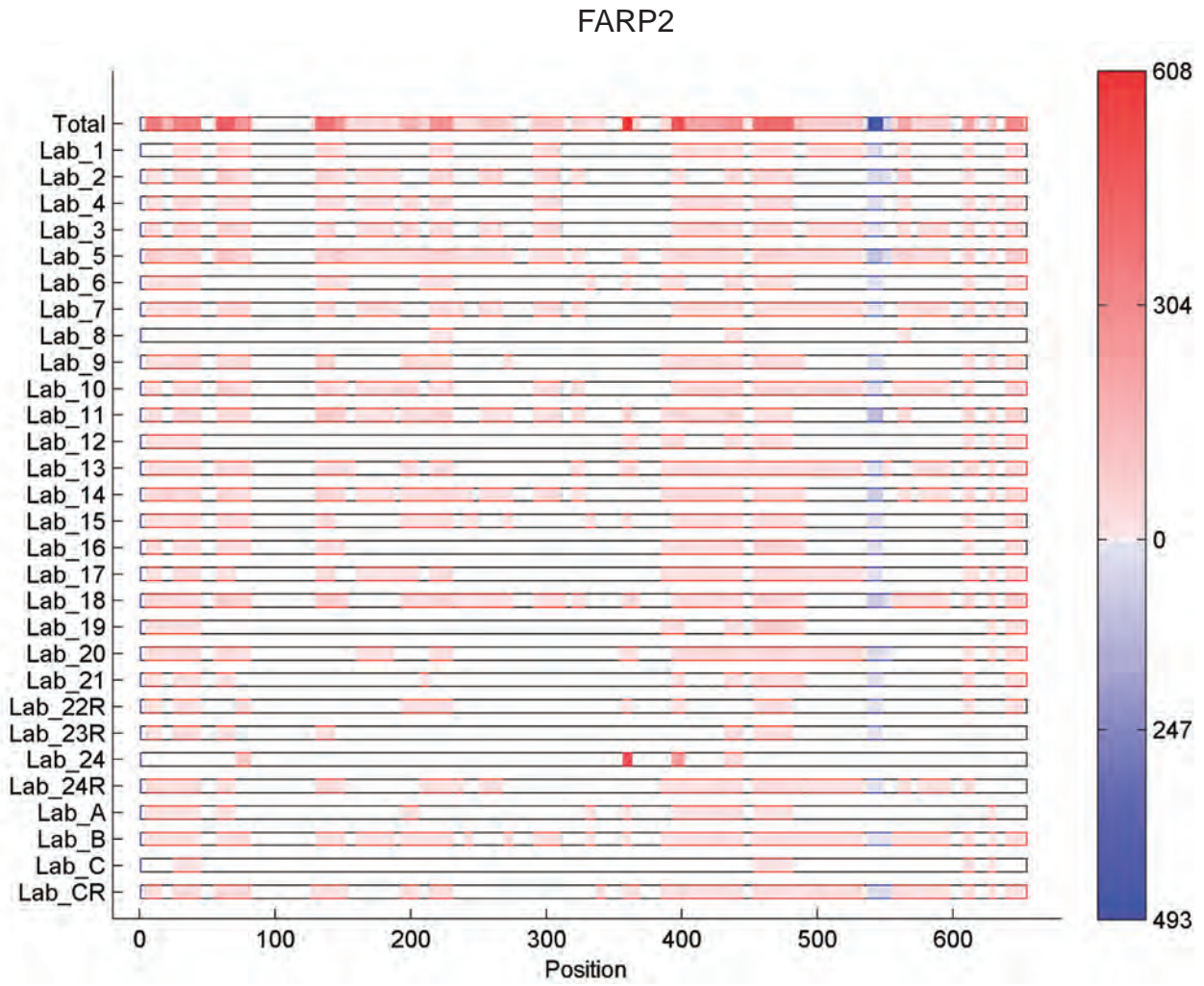




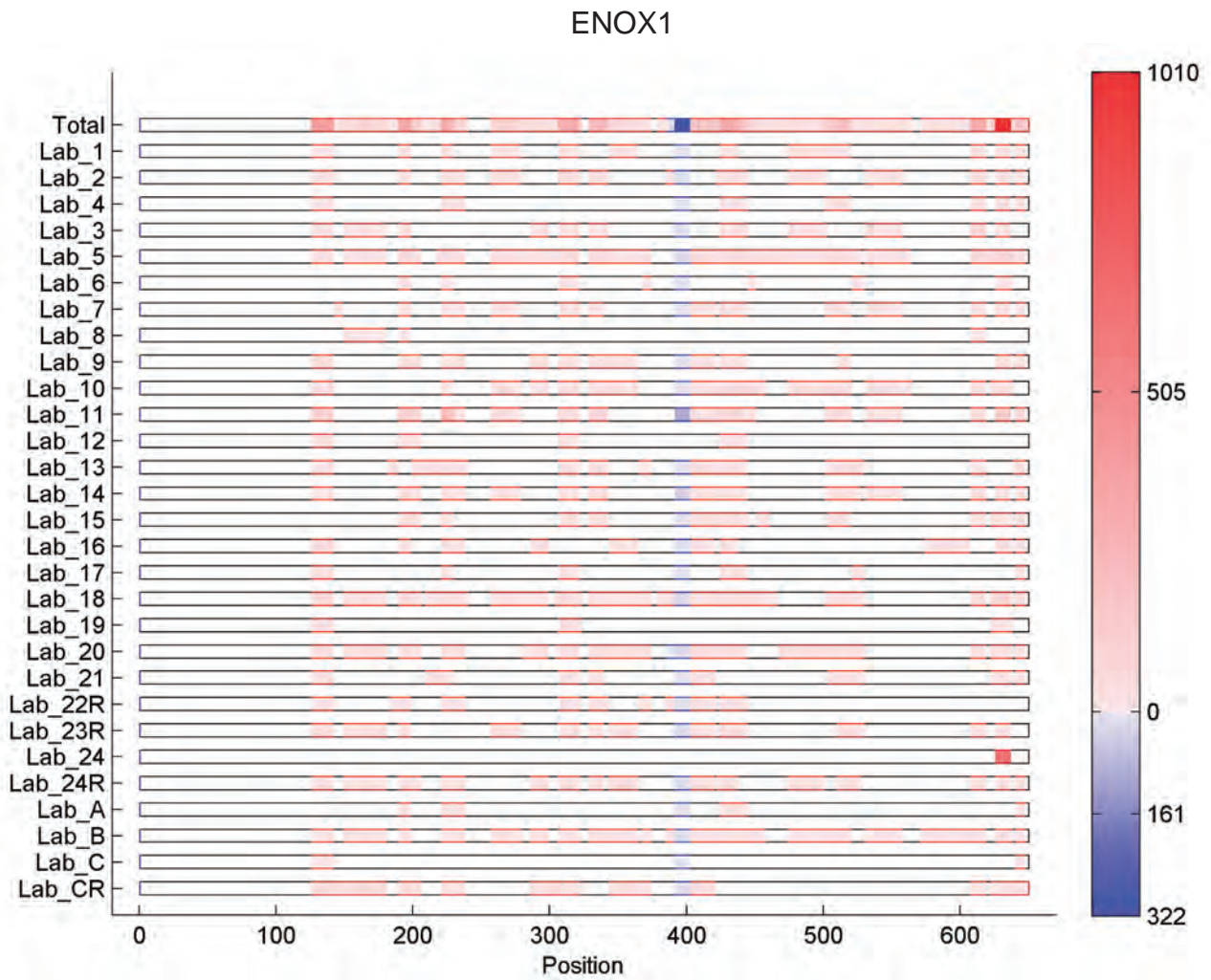
j



I

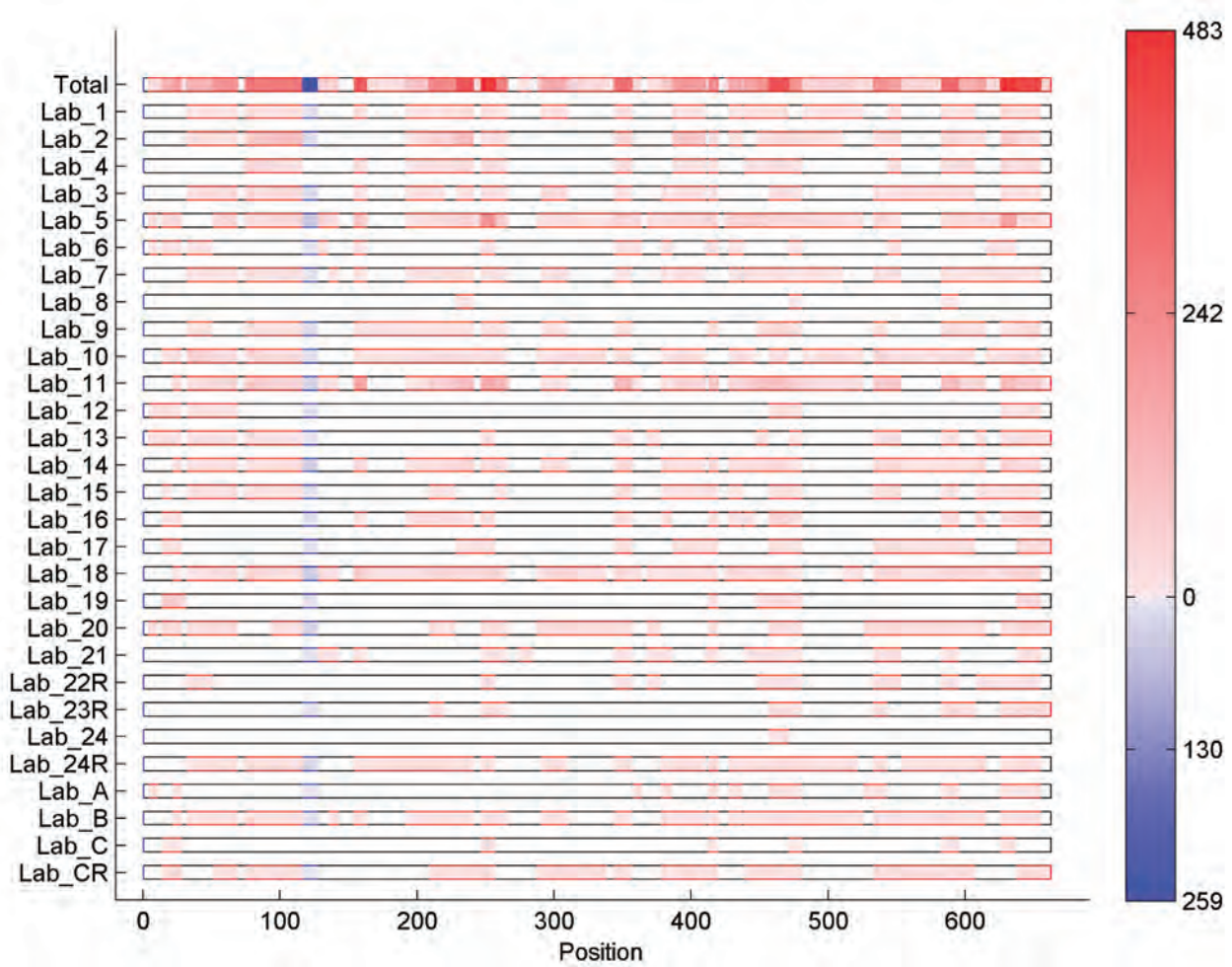


m

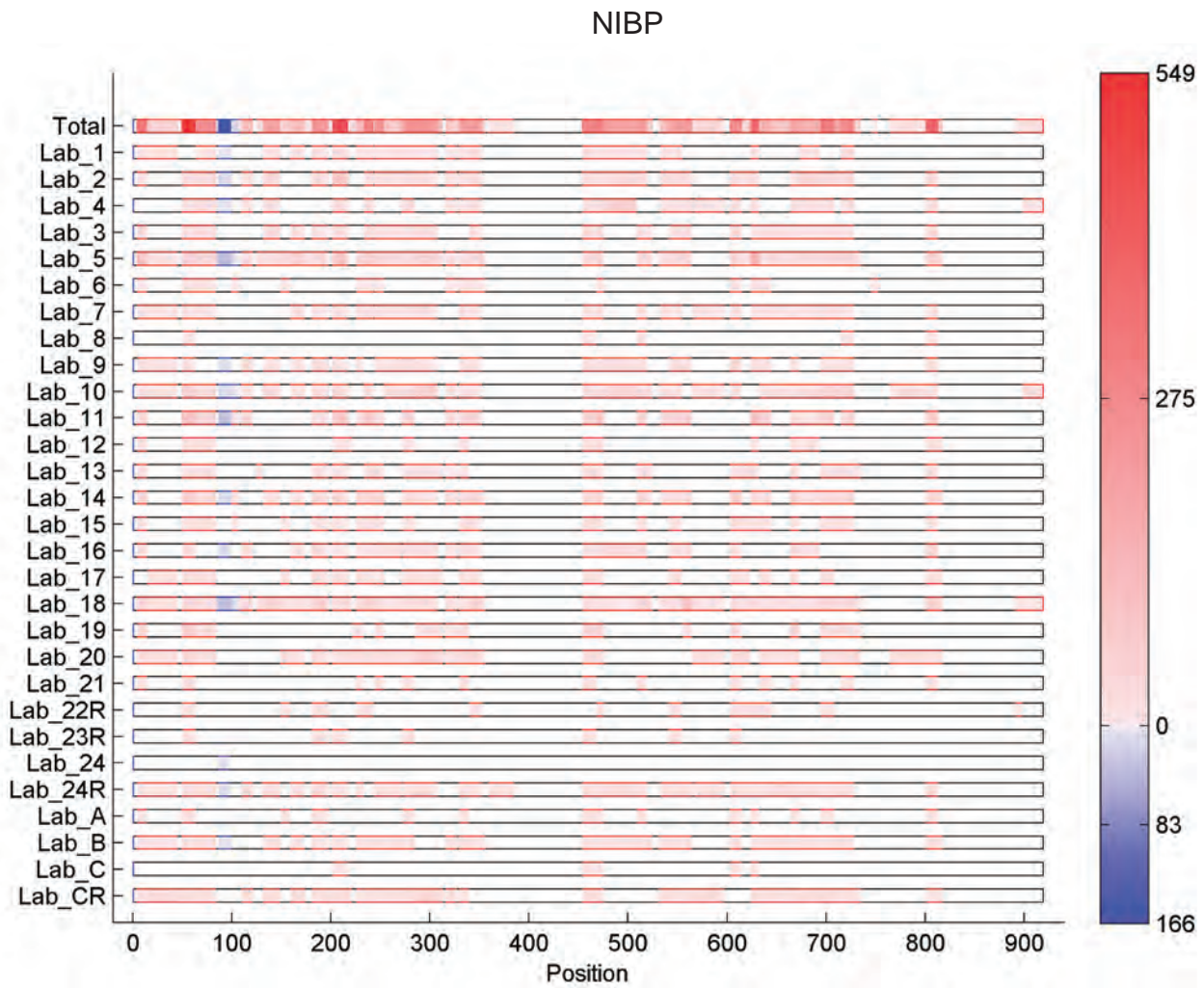


n

KLHL13

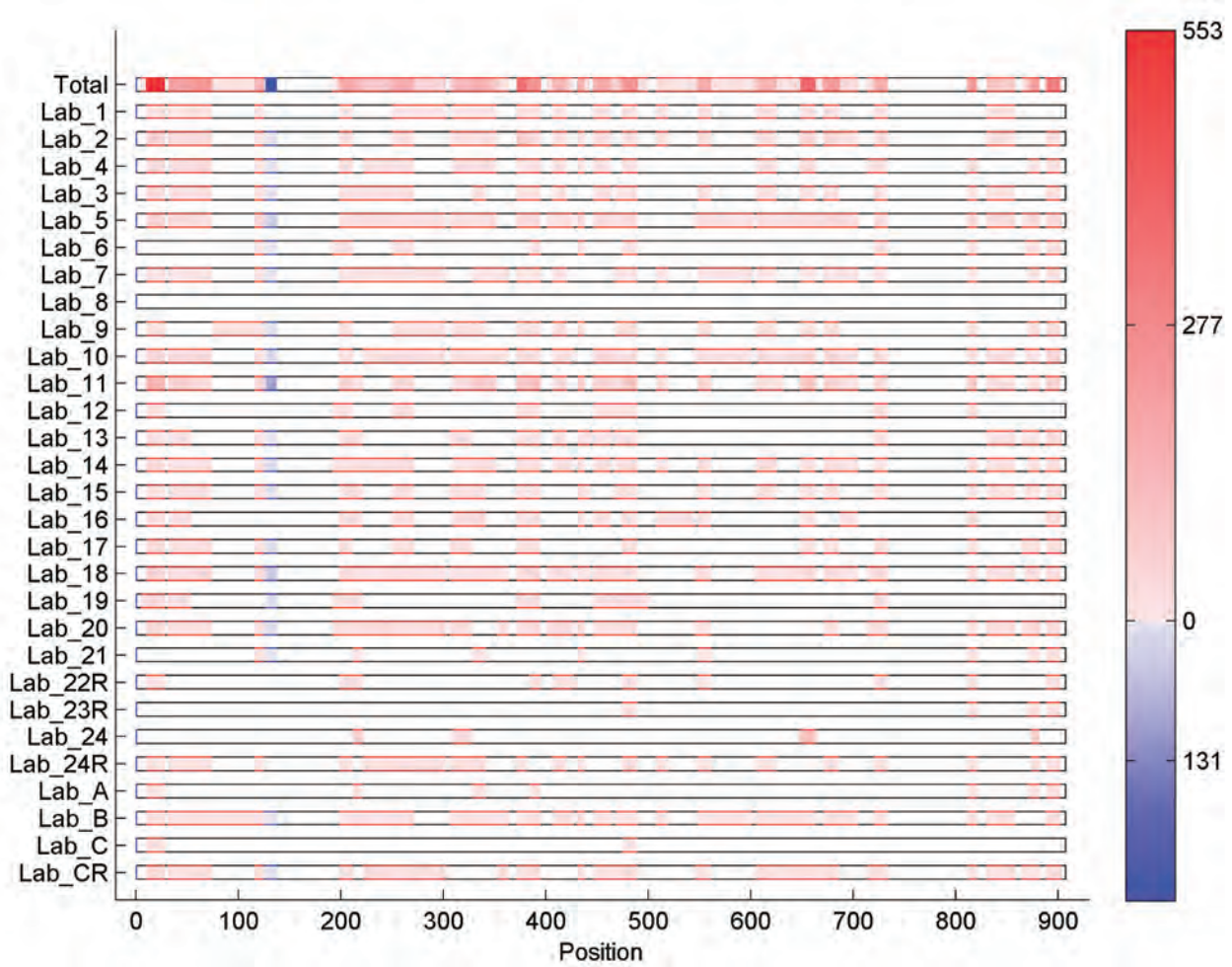


0

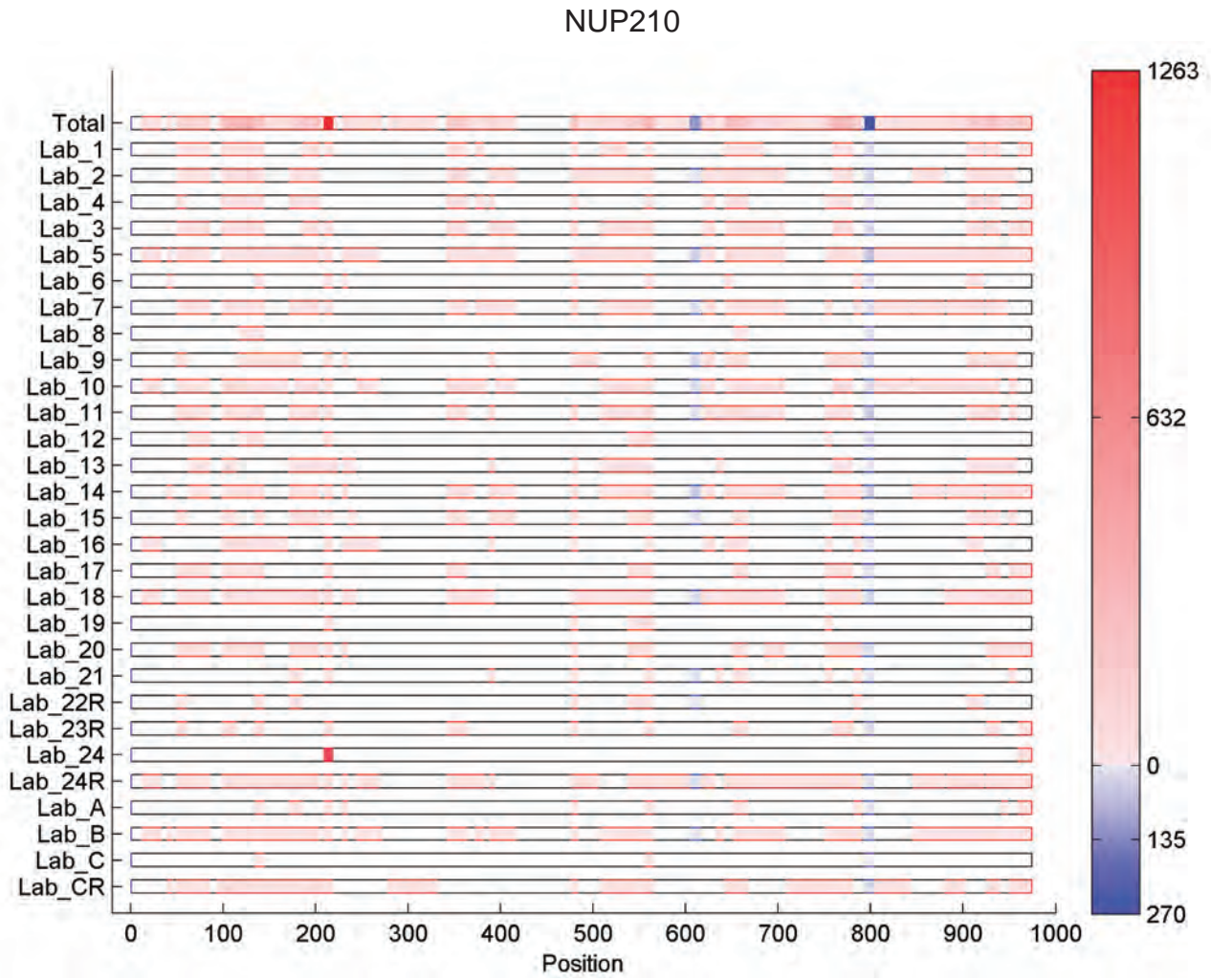


p

MARS

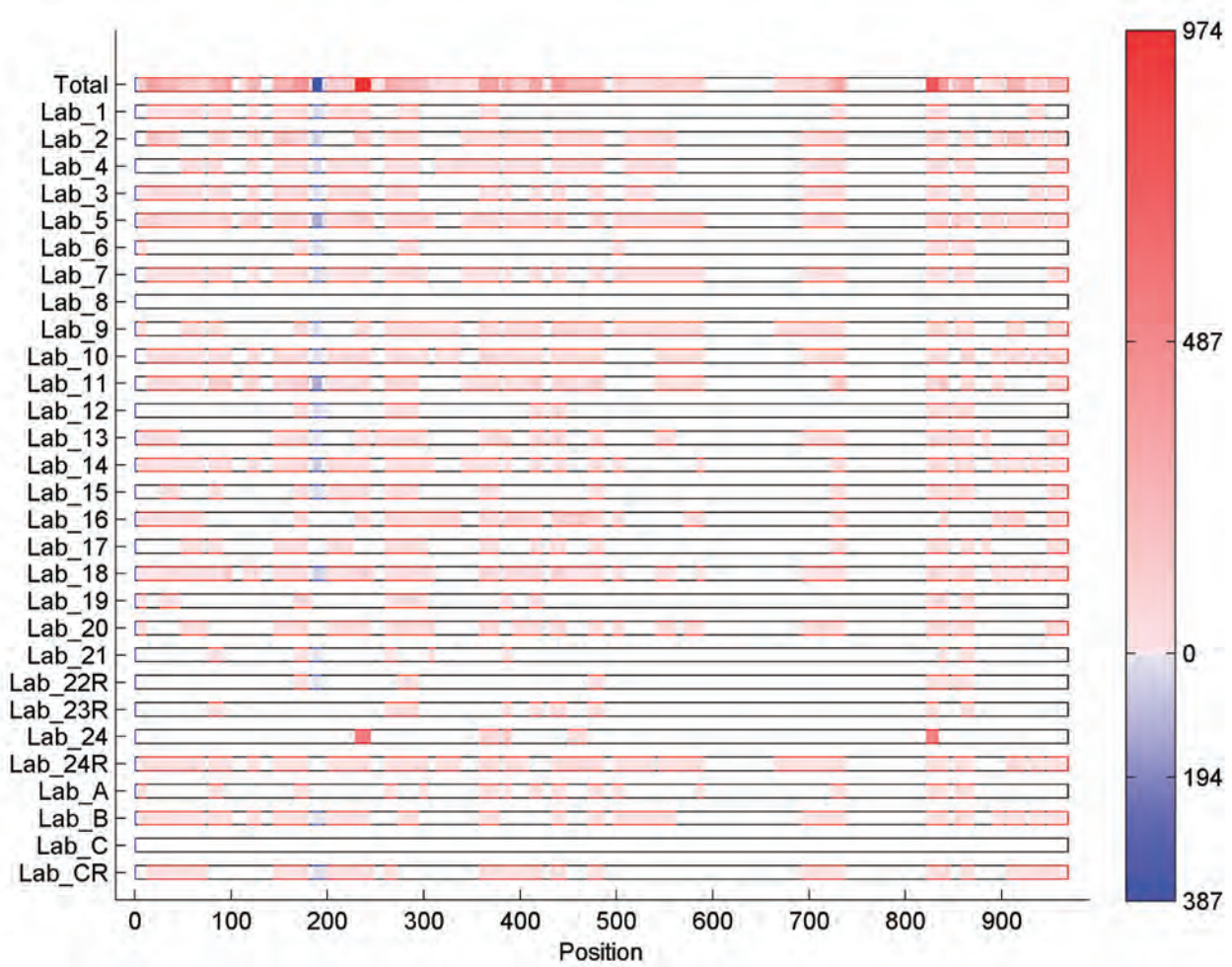


q

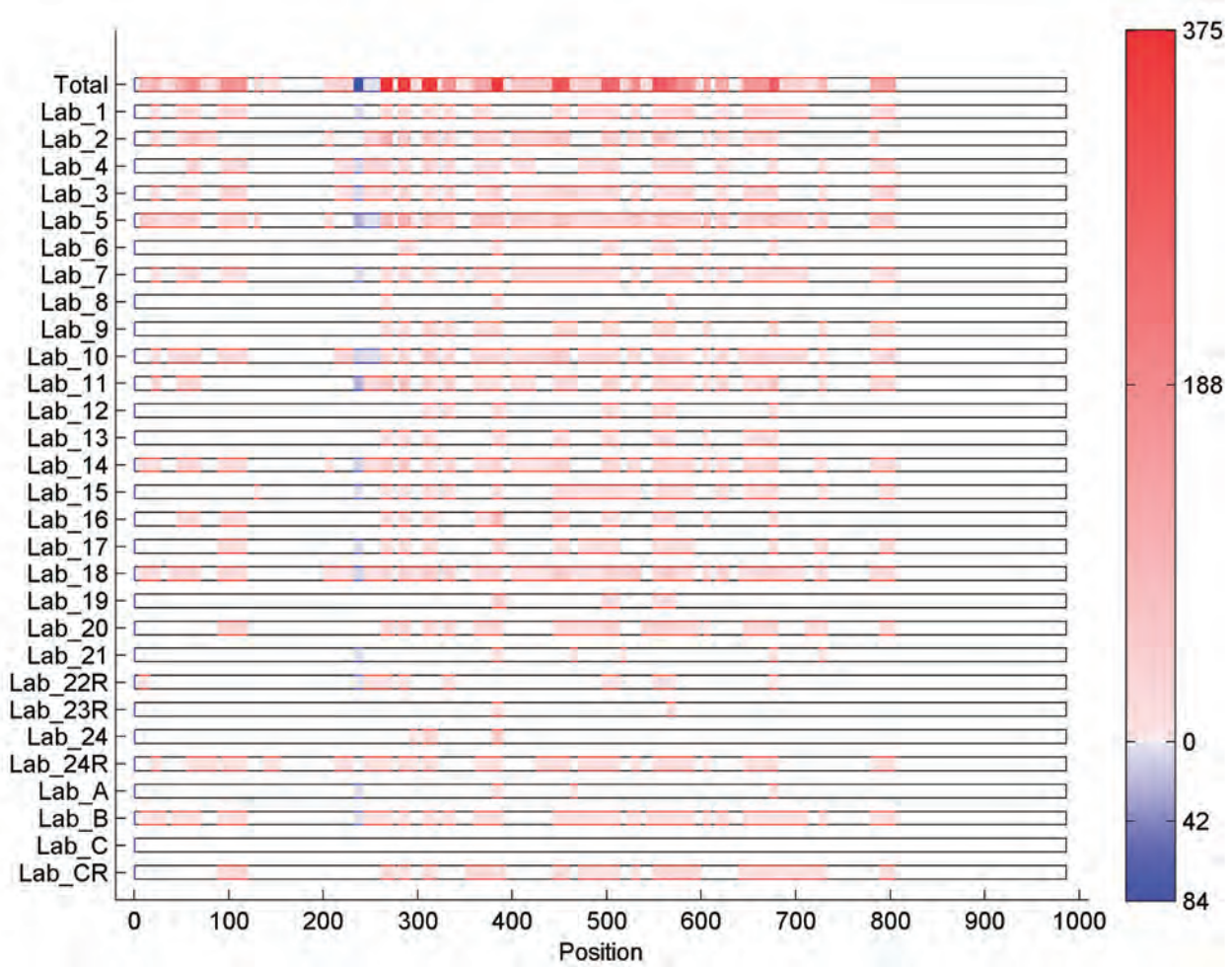


r

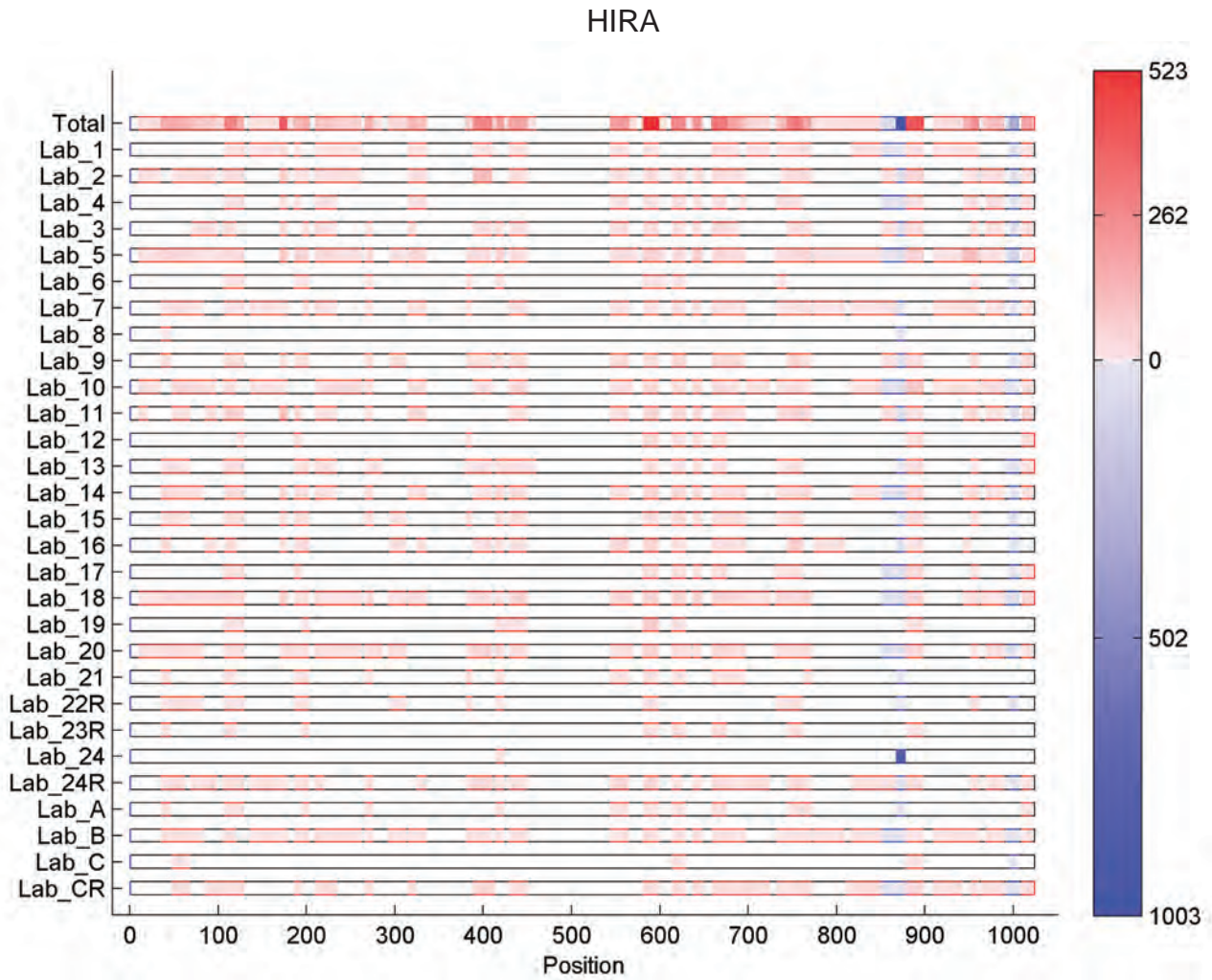
THBS4



KIAA0746



t



Supplementary Figure 11. Comparison of Number of Redundant Peptides Assigned to Keratins and *E. coli* proteins.

Comparison of the number of redundant peptides assigned to keratin proteins. (a) and *E. coli* proteins (b) in raw data submitted to Tranche from each lab. Gel based analyses are indicated by red symbols. Not all labs identified *E. coli* proteins. The high proportion of *E. coli* proteins for certain labs may be due to carry over from prior experiments. The frequency of tandem mass spectra assigned to tryptic peptides (redundant peptide counts) for keratins or *E. coli* proteins are plotted as a function of redundant peptides assigned to the 20 human recombinant proteins for each lab (**Supplementary Tables 15a** and **15c** online).

Supplementary Note








Instructions for transferring results to Tranche.

Instructions for PREPARATION OF DATA for submission to Tranche:

1. Select a name for your data-package (6 to 20 characters – letters or numbers only please). This name will be known only by you and Alex Bell. Do not use a name that might identify you or your laboratory.
2. Create the following directory tree on your local computer.



3. Rename “New Folder” to the name you choose for your data package. This is the folder you should upload to Tranche.
4. Please prepare a “**methods_file**” which should contain a complete accounting for methodologies for each of the following categories:
 - MS data collection (please ensure that the following are ALSO included: i) time in minutes of all LC separations; ii) for 2D-LC identify first column type (SCX, Wax, other); iii) for MALDI ToFToF the number of spots and the number of ions fragmented).
 - Tandem MS data processing (generation of peaklists; also indicate if temperature fluctuations were monitored and if so was a correction made when processing the data).
 - Database search analysis (software, parameters, etc)
 - Criteria for peptide identification (Mascot ID/homology score, Sequest parameters, other).
 - Criteria for protein identification (e.g. 2 peptides above ID score.)
5. Complete the attached Word document (see additional information below) (HUPO_Test_Sample_reporting_Additional_Info_LIMS). This data is required for the LIMS module of CellMapBase.
6. Place both files described in steps 4 and 5 in the top directory as indicated below:

| | | | |
|---|------|-----------------------|-------------------|
|  methods_file.doc | 0 KB | Microsoft Word Doc... | 7/24/2007 2:40 PM |
|  Info_for_S-CellMapBase-HUP... | 0 KB | Microsoft Word Doc... | 7/24/2007 2:40 PM |
|  search_resultfile | | File Folder | 7/24/2007 2:43 PM |
|  proteinlist | | File Folder | 7/24/2007 2:44 PM |
|  peaklist | | File Folder | 7/24/2007 2:38 PM |
|  mass_spec | | File Folder | 7/24/2007 2:38 PM |
|  excel_proteinidentification_su... | | File Folder | 7/24/2007 2:45 PM |

7. Copy the mass spec files into folder “mass_spec”.
8. Copy the peaklist files into folder “peaklist”.
9. Copy the search engine output files (for the Mascot search engine this is the xxx.dat file) into the “search_resultfile” directory (peptide identifications).

Reminder: Make sure that there is no information in these files that would identify you, or your lab, or the institution.

10. Copy your file containing your protein identifications into directory “proteinlist”. This list was generated by the search engine or your summarizing software from the search results.
11. Copy the Excel file containing your FINAL protein summary list into directory “excel_proteinidentification_su...”. This final summary list (with peptide statistics) was generated after interaction with Alex Bell.

SUBMISSION TO TRANCHE

The data is to be uploaded to Tranche at:

<http://www.proteomecommons.org/dev/dfs/examples/hupo-2007/Tranche-HUPO.jsp>.

At this site, you will need to enter an email address (required to retrieve your unique identifier). After clicking on “upload data”, you will be prompted to identify/select the folder (created above – steps 1-11) that is to be uploaded.

This web-interface will direct your submission to the HUPO-2007 project at Tranche. Subsequent to your submission, you will be issued 1 Tranche hash that uniquely identifies your data. Please communicate the Tranche hash to Alex Bell (alex.bell@mcgill.ca).

For more details about Tranche please see the information and FAQ page:

<http://www.proteomecommons.org/dev/dfs/examples/hupo-2007/index.html>.

Additional information: Tranche transfer

The HUPO_Test_Sample_reporting_Additional_Info_LIMS word document records 1) the fraction of the sample processed for gel-based and gel-free processing and 2) the fraction of the digest analyzed by ESI- or MALDI-MS.

In order to assure confidentiality we suggest that any trace within the files that leads back to the lab that generated this data be removed. It may be necessary to do a search and replace.

Supplementary Methods

Test Sample Preparation, Characterization and Quality Control

The criteria of purity of >95%, lack of tags, unique tryptic peptide sequences, and that every protein contains at least one tryptic peptide of 1250 ± 5 Da to represent complexity, were met as documented below:

Protein selection:

Test sample proteins were designed according to the following criteria (see **Supplementary Figure 1a** online for details): First, the sample should contain 20 human proteins in the 32-115 kDa molecular weight range, since this represents the protein distribution range of the “Ultimate ORF Collection”. Second, none of the proteins should be produced with a purification tag in order to prevent a 20 fold concentration of the tagged tryptic peptide(s). Third, all proteins should be highly stable in order to withstand shipping. Fourth, each protein should be at least 95% pure and with the final mixture containing less than 1% contamination represented by any single non-human protein (i.e. bacterial protein). Fifth, all selected proteins should contain unique tryptic peptide sequences within the 700 to 4800 Da range where tryptic peptides are usually found by LC-MS. For this selection, all available predicted protein sequences provided through the “Ultimate ORF Collection” (~15 000 ORFs⁴) were digested with trypsin, *in silico*. The corresponding peptide masses were then calculated and the sequences were compared. The fifth criteria resulted in the initial selection of 7959 proteins. The sixth criterion was that the proteins cluster in four MW ranges of 32-36 kDa, 48-52 kDa, 70-75 kDa and 100-115 kDa. For this selection, the MW was calculated based on the

predicted sequence and 2193 proteins were selected. In order to attempt some complexity at the peptide level, the seventh criterion was based on tryptic peptide mass. The design here was to ensure that every protein would contribute at least 1 tryptic peptide within a specified mass range.

Subsequently, proteins were grouped into sets greater than or equal to 96 in number, such that each group of proteins had common tryptic peptide masses within 10 Da. This grouping led to the selection of 226 proteins in the peptide mass range 1250 ± 5 Da. A request to characterize all such peptides in this mass range assessed whether different laboratories could address the sampling bottleneck for the stochastic data-dependent acquisition as well as report all 1250 peptides as detected in their raw data. From these 226 proteins, 96 were selected for production.

Emphasis was put on selecting proteins such that they populate the molecular weight range of 32 to 115 kDa as shown in **Supplementary Figure 1a** online and that they should yield 20 peptides (at least 1 from each protein) within 10 Da of each other. The distribution of tryptic peptide masses for the Mammalian Gene Collection (MGC)⁵, is representative of the ORF collection. Both the 96 proteins selected for production as well as the 20 test sample proteins were also comparable to the MGC tryptic peptide mass distribution in their distribution of tryptic peptide masses (**Supplementary Fig. 1b** and **c** online) with the exception that tryptic peptides in the mass range 1250 ± 5 Da revealed a higher frequency exactly as was designed.

Protein Expression and Purification:

The final selection of 20 proteins was based on those that could be expressed and purified from the set of 96 (**Supplementary Fig. 1** online). cDNAs for 85 of the 96 proteins selected for production were successfully cloned (see Cloning below), with 8 failing at the subcloning step and a further 3 failing at the restriction enzyme Quality Control (QC) analysis (not shown). Of these, 43 revealed an expected molecular weight following verification on 1D-SDS PAGE (not shown) and 27 were selected for purification of which 23 were successfully purified (not shown). The final 20 proteins selected are shown in **Supplementary Fig. 1** online following their purification from inclusion bodies. In this way, a final set of 20 proteins was completed as indicated in **Supplementary Table 1** online, with their gel mobility and purity revealed by SDS PAGE analysis of 5 pmol of each (**Supplementary Fig. 1d** online) either with Coomassie Blue staining or silver staining (**Supplementary Fig. 1d**, right side online). Evaluation of the predicted isoelectric points of the 20 Test Sample proteins revealed a representative distribution over the pH range 4.5 – 12.3 (<http://www.invitrogen.com/etc/medialib/en/filelibrary/pdf.Par.72904.File.dat/HumanProteinStandardsforMassSpectrometry.pdf>).

Quality control:

Based on the densitometry of the gel shown in **Figure 1d**, left side, higher molecular weight bands were found as potential contaminants for several proteins (**Supplementary Fig. 2** online). Mass spectrometry analysis (MALDI-ToFToF) revealed in separate experiments that these potential contaminants were in all cases oligomers

of the parent protein (data not shown). Based on the densitometric analysis, the calculated purity of each test sample protein ranged from 95-100% pure (**Supplementary Table 2** online). As part of the Quality Control process, the characterization of ketohexokinase (fructokinase) is shown (**Supplementary Fig. 3** online). Determination of the mass of the intact protein (**Supplementary Fig. 3** online) revealed that the protein product was 766 Da larger than that predicted by the gene sequence. This apparent discrepancy was accounted for by an N-terminal extension of 7 amino acids of the expressed ketohexokinase due to initiation of translation at a vector derived methionine¹. Vector derived synthesis results in the extension of ketohexokinase by MYKKAGT that increases the calculated mass by 780 Da, consistent with the intact protein mass determination. All test proteins were expressed with this N-terminal extension. Silver staining revealed a similar repertoire of proteins (**Supplementary Fig. 1d**, right side online).

The final 20 protein Test Sample mixture was further analyzed by both gel-based and gel-free LC-MS proteomics protocols as summarized in **Supplementary Fig. 4** and **Supplementary Table 3** online. For gel-based analysis, the mixture was resolved by gel-electrophoresis on a 10% polyacrylamide gel and in-gel trypsin digested. Extracted peptides were analyzed by MALDI-ToFToF MS and nano-LC-ESI MS. For gel-free analysis, the protein mixture was digested in solution with trypsin and the tryptic peptide mixture was pre-fractionated by acetonitrile-0.1% TFA step gradients from a reversed phase Zip-Tip® (Millipore) prior to analysis of the eluted pools by MALDI-ToFToF and nano-LC-ESI MS. Tandem MS spectra were matched to the NCBI nr database (release

dates 15August2004 and 13May2006) by employing the Mascot search engine (version 2.1, Matrix Science). As summarized in **Supplementary Table 3** online, on average 7-14 distinct peptides were assigned to each protein with average sequence coverage ranging from 18-33%, with the exception that the GLB1L3 protein was not found by matching the ToFToF data to the 15Aug2004 release of the NCBI database. Subsequent analysis revealed that the error was related to the MALDI-ToFToF data analysis set-up rather than the MS data as the GLB1L3 protein could be identified from the ToFToF data by employing off-line database searching.

Protein stability:

Test Samples (100 pmol total protein) were tested for stability employing an accelerated stability test as follows: Test Samples ready for distribution were incubated for 2.7 days at -20, 22, 37, 42 and 70°C and then analysed by 1D-SDS PAGE followed by Coomassie blue staining and comparison with a gel image as generated at time of preparation (zero days) of the test sample mixture (**Supplementary Fig. 5** online). The SDS PAGE analysis revealed that the proteins were not degraded to lower MW forms (**Supplementary Fig. 5** online) however; the same amount of protein could not be retrieved from vials stored at the higher temperatures as evidenced by Coomassie blue staining. The LC-MS analysis indicated that similar number of distinct peptides and coverage were obtained regardless of incubation temperatures (**Supplementary Table 4** online). Long term stability was determined as equivalent days at -20°C and the following relationship: $2^{[(\text{high temp} - \text{low temp})/10]} \times (\text{days at high temp}) + (\text{days already spent at } -20^\circ\text{C})^{2,3}$. Stability was evaluated based on degradation products as judged by lower

M_r Coomassie blue staining material. Protein solubilization was found to be compromised due to storage at temperatures of 37°C or higher (**Supplementary Fig. 5** online).

Cloning:

The 96 Ultimate™ human ORF entry clones that passed through all three bioinformatic filters were Gateway®-cloned from pENTR221 entry clones (Invitrogen) into the pET-DEST42 vector¹. This expression vector employs a T7 promoter, ampicillin resistance, and encodes a short N-terminal leader peptide of MYKKAGT when employed with pENTR221 entry clones. Expression in *E. coli* was performed in the BL21-Star (DE3) host¹ with conditions that maximize inclusion body formation⁶. Overnight starter cultures (5mL) were used to inoculate 50mL cultures (1:100 dilution) grown at 37°C, 250rpm (Infors shaker incubator). The larger cultures were induced with 1mM isopropylthiogalactoside (IPTG) when the OD-600nm was 0.5-0.7. Growth continued for 3-3.5 hours before harvesting cells by centrifugation at 5000 x *g* for 5 minutes at 4°C. Excess media was removed and cell pellets were stored at -20°C until processed.

Inclusion body isolation:

Cell pellets were lysed by resuspension in 5mL BugBuster (Novagen) containing 50 U/mL benzonase (Novagen), 1X Complete protease inhibitors (Roche), and 0.2 mg/mL lysozyme (Sigma). The *E. coli* lysate was centrifuged for 20 minutes at 16,000 x *g* at 4°C to pellet the crude insoluble fraction. The insoluble pellet was washed by resuspension in 25mL of 20mM Tris pH 8.0, 10mM EDTA, 1% Triton X-100 using a

sonicator (Misonix, mini-tip, setting 3.5) to completely disperse the pellet. The crude inclusion bodies were pelleted by centrifugation for 20 minutes at 16,000 x *g*, 4°C, and the wash buffer was removed. The previous wash step was repeated once before resuspending inclusion bodies by pipetting up and down in 1mL 20mM Tris pH 8.0. The final wash was subjected to centrifugation at 16,000 x *g*, 4°C for 20 minutes and the supernatant removed to produce the final inclusion bodies. Inclusion bodies were stored at -20°C.

Preparative electrophoresis:

Inclusion bodies were solubilized in 1X lithium dodecylsulfate (LDS) sample buffer (Invitrogen) containing 50mM dithiothreitol (DTT) and heated at 70°C for 10 minutes. Preparative electrophoresis was performed using the Bio-Rad 491 Prep Cell with discontinuous Tris-glycine chemistry and separating gels with acrylamide percentages appropriate for the migration of each protein according to manufacturer's instructions. Fractions of the target protein were evaluated on 8% or 10% NuPAGE[®] Bis-Tris Midigels (Invitrogen), MES-SDS running buffer (Invitrogen), and stained with SimplyBlue[™] SafeStain (Invitrogen) using the sensitive staining protocol according to manufacturer's instructions. Briefly, gels were fixed with three, 5 minute water washes, stained for 1 hour, then destained for 15 minutes in 100mL water before adding 20mL of 20% w/v NaCl and continuing the destain incubation for 45 minutes. Fractions containing a single protein were pooled and concentrated by centrifugal ultrafiltration (Vivaspin-20 (Sartorius), MWCO 10,000 for proteins in the 32-36 kDa group, and MWCO 30,000 for proteins in the 48-52 kDa, 70-75 kDa and 100-115 kDa groups) to a final volume of 0.4

to 0.7 mL prior to acetone precipitation. Four volumes of cold (-20°C) acetone (ACS grade, J.T. Baker) was added to the protein concentrate, vortexed, and incubated at -20°C overnight. Proteins were pelleted by centrifugation at 16,000 x *g*, 4°C, for 10 minutes and the supernatant was removed with a pipet. Pellets were air-dried in a laminar flow hood for 10-15 minutes prior to capping. Precipitated proteins were stored at -20°C.

Protein purity analysis:

Purified proteins were analyzed on 10% NuPAGE[®] Bis-Tris gels, 250 ng load per protein, and stained with SimplyBlue[™] SafeStain using the sensitive staining protocol (see above). A 5% contaminant (12.5ng) at this protein load (250ng) is above the limit of detection for this staining procedure (5ng BSA) (Invitrogen) (not shown). Further to this analysis, image analysis of the 1D-SDS PAGE gel of 5 pmol of each of the 20 Test Samples proteins (**Supplementary Fig. 1d** online) was performed by employing the TotalLab TL100 1D analysis software (Nonlinear Dynamics). Volumes for Coomassie blue stained bands (**Supplementary Table 2** online) were determined with software set for automatic lane detection, rolling disc background subtraction (radius 1000), and automatic band detection (minimum slope 50) with manual editing to remove detection of specks and add detection of low abundance or closely spaced bands. Corrected volume corresponds to band peak volume minus a background volume. Bands appearing at higher M_r in the gel lane than expected were confirmed by LC-MS to be oligomeric aggregates (not shown).

Protein quantification:

Acetone precipitated proteins were resuspended in 20mM Tris pH 8.0, 2% SDS, 50mM DTT and quantified by employing the reducing-agent compatible bicinchonic acid (BCA) assay (Pierce) in microplate format. BSA (Pierce) protein standard curve and all samples were run in triplicate.

LC-MS Proteomics Analyses:

In order to ensure complete solubilization of the proteins in the test sample for LC-MS proteomics then for gel-free and 1D-SDS gel-based strategies, the test sample should be solubilized in 8 M urea and SDS sample buffer, respectively.

For the gel-based analysis, the mixture (100 pmol total protein) was dissolved in lithium dodecylsulfate (LDS) sample buffer plus 50 mM dithiothreitol, heated at 70°C for 10 min and the proteins were resolved by 1D-SDS PAGE on a 10% Bis-Tris NuPAGE® gel (Invitrogen). The gel was stained with Coomassie blue and processed manually for in-gel digestion with trypsin as follows: Coomassie blue stained bands were excised, diced into ~1 mm cubes, and washed in 50% acetonitrile in 25 mM ammonium bicarbonate buffer pH 7.8 until clear, then dehydrated with acetonitrile followed by drying in a speed vacuum centrifuge. The gel pieces were rehydrated for 2 h on wet-ice with a minimal volume of trypsin solution (10 µg/mL in 25 mM ammonium bicarbonate buffer pH 7.8) and then incubated overnight for 16-18 hours at 37°C. Peptides from the digested proteins were extracted from the gel pieces in two steps by incubations at

room temperature for 30 min first with 25 μ L of 5% TFA and then with 25 μ L of 5% TFA /50% acetonitrile. The two extracts were pooled and taken to dryness in a speed vacuum centrifuge. The dried extracted peptides were reconstituted in a 50% acetonitrile/0.1% trifluoroacetic acid (TFA) solution for MALDI-ToFToF (Applied Biosystems 4700 Proteomics Analyzer) MS analysis and 15% acetonitrile/0.1% formic acid (FA) solution for nano-LC/ESI-QToF (nano-Aquity UPLC/ESI-QToF Premier MS and a CapLC QToF API-US MS (Waters)) MS analysis. For MALDI-ToF MS analysis, extracted peptides were spotted with α -cyano-4 hydroxycinnamic acid (α -CHCA) prepared in 50% acetonitrile/0.1% TFA and the 20 most intense peptide ions were selected for tandem MS. For nano-LC-ESI MS, peptide separations were performed on 100 μ m x 100 mm column AtlantisTM dC18 employing 3 μ m packing (Waters) and a flow rate of 200 nL/min. A gradient of 5-45% acetonitrile in 0.1% formic acid over 45 min, followed by 45-95% acetonitrile in 0.1% formic acid over 5 min was used. Peptide ion selection was based on charge state, ions of +2 to +4 charges, and the signal to noise intensity criteria of 10 for switching from MS to tandem-MS mode. The threshold for switching from tandem-MS to MS mode was set to >3500 counts/sec. MS/tandem-MS scans alternated between up to 4 tandem MS scans from 100 to 3000 m/z (1.4 sec) for every MS scan (1 sec) from 400 to 1800 m/z in a data directed analysis (DDA) mode.

For the gel-free analysis, the protein mixture was solubilized with deionized 8 M Urea in 25 mM ammonium bicarbonate pH 8.0, diluted 5 fold with 25 mM ammonium bicarbonate pH 8.0 and subjected to trypsin (0.2 μ g) digestion by incubation at 37°C for 16-18 h. For LC-ESI MS, aliquots (20 μ L) of the digest were analysed as indicated

above. For MALDI-ToFToF MS analysis, the resulting tryptic peptide mixtures were fractionated by employing reversed phase C18 Zip Tips® according to manufacturer's instructions and by employing stepwise elution of peptides with increasing concentrations of acetonitrile (in 0.1% TFA) from 10 to 50% (5 steps) prior to analysis of the eluted pools. For MALDI-ToF and for LC-ESI MS the analysis was as indicated above.

Data processing of raw data files from the Q-TOF and MALDI-ToF instruments were processed with Mascot Distiller (Version 2.1, Matrix Science) and GPS Explorer (version 2.0, Applied Biosystems), respectively, without smoothing and by employing charge states as determined from the MS scans. The resulting peaklists for the centroid files were searched against the NCBI nr database (release dates 15August2004 and 13May2006) by employing the Mascot search algorithm (Version 2.1, Matrix Science). Identifications of peptides were constrained to be tryptic with up to one missed cleavage and with variable oxidation of methionine residues, variable alkylation of cysteine residues by acrylamide. Peptide and fragment ion mass tolerances were set at 50 ppm and 0.2 Da, respectively for QToF data and 150 ppm and 0.5 Da, respectively for MALDI-ToFToF data. Peptide identifications were accepted if scored at greater than 95% confidence.

Protein blending and preparation of the Test Samples:

Air-dried, acetone precipitated purified proteins were resuspended in SDS buffer (20mM Tris pH 8.0, 2% SDS, 50mM DTT) for concentration determination based on the BCA

assay. Individual proteins were then pooled together to produce an equimolar blend of the 20 proteins. The master blend was then aliquotted into 100 pmol total protein (5 pmol each) aliquots and acetone precipitated (see above), with recoveries assessed by 1D-SDS PAGE followed by Coomassie blue staining. After centrifugation and removal of the supernatant, precipitated proteins were air-dried in a laminar flow hood for 10 minutes prior to capping and storage at -20°C.

Supplementary References

1. S. Khan, R. Hsu, A. Jones, I. L. Ross, D. N. Hart, and M. Kato. Identification of the dominant translation start site in the attB1 sequence of the pET-DEST42 Gateway vector. *Protein. Expr. Purif.* **49**, 102-7 (2006).
2. K.J. Hemmerich, in *Medical Plastics and Biomaterials* (1998), Vol. July, pp. 16.
3. P Matejtschuk and P Phillips, in *Medicines from Animal Cell Culture*, edited by G Stacey and J Davis (John Wiley & Sons, Ltd., 2007).
4. F. Liang *et al.* ORFDB: an information resource linking scientific content to a high-quality Open Reading Frame (ORF) collection. *Nucleic Acids Res.* **32**, D595-9 (2004).
5. R. L. Strausberg, E. A. Feingold, R. D. Klausner, and F. S. Collins. The mammalian gene collection. *Science* **286**, 455-7 (1999).
6. B. Fahnert, H. Lilie, and P. Neubauer. Inclusion Bodies: Formation and Utilisation. *Adv. Biochem. Eng. Biotechnol.* **89**, 93-142 (2004).