# Supplemental Experimental Procedures

### Glacial meltwater samples

### Field Collections

Samples were isolated by filtering 15-30L on 0.22 µm filters at the field site. All samples had high loads of glacial sediments (~0.3-1 µm). One quarter of each filter was set aside for DNA extraction and the remaining filter portions were used to isolate bacteria from the glacial sediments for proteomics.

### DNA Samples: preparation and analysis

PowerSoil extraction kits from QIAGEN were used and the provided protocol was followed. We isolated >2.5 µg of DNA from each sample to complete the metagenome using Illumina NextSeq sequencing. Samples were submitted to the Northwest Genome Sequencing Center (University of Washington) for library preparation, and sequencing using NextSeq at 150 base pair – paired end reads (PE150). Each sample was barcoded and all 5 were loaded onto 1 lane. Trimming, filtering and assembly were completed following Timmins-Schiffman (1). The metagenome FASTA files for subsurface early melt, subsurface late melt, subsurface midseason melt, surface early melt, and surface midseason melt had 74817, 39151, 86946, 71401, and 36681 entries, respectively.

### Protein Samples: preparation and analysis

In order to isolate the bacterial cells from the glacial sediments on the filters, we modified a protocol published by Kallmeyer et al. (2) to be more amenable to mass spectrometry and the low bacterial cell counts found at these sites. Emulsifier mixture included 100 mM EDTA and 100 mM sodium pyrophosphate (4$^o$C). Additionally, bacterial cell buffer (BCB) was made to stabilize cells during the isolation procedure (4 mL of phosphate buffered saline, 645 µl emulsifier mixture, 516 µl of methanol; 4$^o$C). To kill the samples prior to sample preparation, as it is a very lengthy process, 5 g of the sediment-bacteria mixture was rinsed with 1 ml of 0.1% sodium azide in 1 x phosphate buffered saline (PBS). Samples were vortexed and then iced 5 times (2 minutes per cycle). Bacteria and sediments were pelleted by centrifugation (10,000 x g, 4°C, 2 hours) and supernatant was removed. The sediment pellet was resuspended in 5 ml of 4$^o$C BCB**,** vortexed and then iced 5 times (2 minutes per cycle) until a smooth solution was achieved. Centrifuge tubes were then placed in an ice water bath that had a Bronson sonicating probe inserted into the water (10 min; power 5). The resulting sediment slurry was then divided among 6 different 5mL ultracentrifuge tubes (~1ml of sediment slurry per tube). Using a syringe, 3.5mL of 40% (w/v) Nycodenz solution was injected beneath the slurry to make the sediment "float" on top of the Nycodenz. Samples were ultracentrifuged for 18 hours (50,000 x g, 4°C; Beckman SW50.1 rotor; no deceleration brake). The overlying clear layer contains the bacterial fraction and was collected with needle and syringe, pooled and 60 mL 1x PBS was added to decrease the density of the Nycodenz solution. This allowed us to isolate the bacterial cells using ultracentrifugation. Samples were then centrifuged for 15 hours (10,000 x g, 8°C) and the overlying liquid was removed carefully as to not disturb the cell pellets. Cell pellets were combined for each sample to yield one final cell pellet. These were rinsed and pelleted 3x with 500ul of 50 mM ABC. Cell pellets were digested with 1 µg of trypsin each following Nunn et al. (3).

Samples were analyzed on the Q-Exactive in DDA mode using an inline chromatography system (Waters Nanoacquity) with a 4cm precolumn (100um ID) and 30cm PicoTip analytical column (75um ID) maintained at 50$^{o}$C (Dr. Maisch C18). Peptides were separated with a 120-minute gradient from 2% -35% acetonitrile, 0.1% formic acid, with total runtime of 160 minutes. Total ion current per sample was ~5E9 per injection. MS and MS/MS scans were collected at 70,000 and 35,000 resolution, respectively, using a loop count of 20 per DDA cycle with a dynamic exclusion of 10 sec and exclusion of all +1, >+5 ions.

## References

1. Timmins-Schiffman, E., May, D. H., Mikan, M., Riffle, M., Frazar, C., Harvey, H. R., Noble, W. S., and Nunn, B. L. (2017) Critical decisions in metaproteomics: Achieving high confidence protein annotations in a sea of unknowns. *ISME J.* 11, 309–314

2. Kallmeyer, J., Smith, D. C., Spivack, A. J., and D'Hondt, S. (2008) New cell extraction procedure applied to deep subsurface sediments. *Limnol. Oceanogr. Methods* 6, 236–245

3. Nunn, B. L., Slattery, K. V., Cameron, K. A., Timmins-Schiffman, E., and Junge, K. (2015) Proteomics of Colwellia psychrerythraea at subzero temperatures - a life with limited movement, flexible membranes and vital DNA repair. *Environ. Microbiol.* 17, 2319–2335

## Supplemental Tables

Table S1: Novor results for high mass accuracy fragment ion MS/MS spectra using an orbitrap. A tryptic digest of a human cell line was subjected to either beam CID or resonance CID. Database searches with FDR < 0.001 and Comet E-value < 0.001 gave 16, 062 (beam CID) and 14,344 (resonance CID) PSMs. Recall values and the Novor sequence score for a few precision values are shown, along with the number of *de novo* sequences with an equal or greater score. Sequencing rate is the number of correct *de novo* sequences for a given precision divided by the total number of spectra.

| **Precision** | **Recall** | **Novor score** | **# *de novo* sequences** | **Sequencing rate (%)** |
|---|---|---|---|---|
| Beam CID<br>16,062 spectra total | | | | |
| 0.99 | 0.64 | 74 | 8,395 | 52.3 |
| 0.98 | 0.77 | 67 | 10,204 | 63.5 |
| 0.95 | 0.91 | 58 | 12,342 | 76.8 |
| 0.90 | 0.98 | 46 | 14,037 | 87.4 |
| 0.85 | 0.99 | 30 | 15,171 | 94.5 |
| Resonance CID<br>14,344 spectra total | | | | |
| 0.99 | 0.74 | 59 | 7,654 | 53.4 |
| 0.98 | 0.79 | 55 | 8,310 | 57.9 |
| 0.95 | 0.86 | 50 | 9,372 | 65.3 |
| 0.90 | 0.94 | 42 | 10,778 | 75.1 |
| 0.85 | 0.98 | 35 | 11,868 | 82.7 |

Table S2: Novor results for low mass accuracy fragment ion MS/MS spectra using a linear ion trap. A tryptic digest of a human cell line was subjected to either beam CID or resonance CID. Database searches with FDR < 0.001 and Comet E-value < 0.001 gave 9,303 (beam CID) and 11,082 (resonance CID) PSMs. Recall values and the Novor sequence score for a few precision values are shown, along with the number of *de novo* sequences with an equal or greater score. Sequencing rate is the number of correct *de novo* sequences for a given precision divided by the total number of spectra.

| **Precision** | **Recall** | **Novor score** | **# *de novo* sequences** | **Sequencing rate (%)** |
|---|---|---|---|---|
| Beam CID<br>9,303 spectra total | | | | |
| 0.99 | 0.15 | 85 | 633 | 6.8 |
| 0.98 | 0.26 | 81 | 1,067 | 11.5 |
| 0.95 | 0.37 | 78 | 1,575 | 16.9 |
| 0.90 | 0.54 | 72 | 2,452 | 26.4 |
| 0.85 | 0.68 | 68 | 3,245 | 34.9 |
| Resonance CID<br>11,082 spectra total | | | | |
| 0.99 | 0.48 | 82 | 3,278 | 29.6 |
| 0.98 | 0.62 | 78 | 4,307 | 38.9 |
| 0.95 | 0.79 | 70 | 5,637 | 50.9 |
| 0.90 | 0.87 | 64 | 6,535 | 59.0 |
| 0.85 | 0.92 | 60 | 7,309 | 66.0 |

Table S3: Effect of bad data on ability of Novor to assign *de novo* sequences.  Starting with a human tryptic digest LCMSMS data file, ion m/z values were randomized (adding or subtracting the integers 1 to 9) to varying extents, and the number of unique *de novo* sequences with sequence scores >= 60 were counted.

| **Fraction of randomized fragments** | **# Unique sequences** |
|---|---|
| 0% | 20,453 |
| 5% | 17,485 |
| 10% | 13,351 |
| 25% | 4,084 |
| 50% | 322 |

Table S4: Novor results summary for glacial melt water proteomics.  Automated *de novo* sequencing by Novor (sequence score cutoff of 60) shows that 20-25% of the MS/MS spectra were high quality peptide spectra.

| **Sample** | **Total ms2** | **Total high scoring *de novo* sequences** | **% ms2 spectra with high scoring *de novo* sequences** | **# unique *de novo* sequences** |
|---|---|---|---|---|
| Early season, surface melt | 23,261 | 4,432 | 19.1% | 3,715 |
| Early season, subsurface melt | 22,249 | 4,207 | 18.9% | 3,437 |
| Mid season, surface melt | 21,879 | 5,443 | 24.9% | 4,347 |
| Mid season, subsurface melt | 22,292 | 5,919 | 26.6% | 4,521 |
| Late season, subsurface melt | 18,873 | 4,494 | 23.8% | 3,246 |

Table S5: An example where the correct FASTA-derived sequence is ranked below a *de novo* sequence.  The difference in the order of the two N-terminal amino acids was sufficient to give the *de novo* sequence a slightly higher cross-correlation score.

| **Rank** | **Sequence** | **Xcorr** | **FASTA / de novo** |
|---|---|---|---|
| 1 | **KS**EEAEALYHSK | 4.351 | De novo |
| 2 | **SK**EEAEALYHSK | 4.108 | FASTA |
| 3 | SHYLAEAEEKSK | 0.960 | FASTA |

Table S6: Human LCMSMS data searched against FASTA files of various other species to which high scoring (>=60) *de novo* sequences had been appended. The column "Database unique (%)" is plotted in Fig. 4.

| Species | Common name | Taxonomy (phylum/class/order/family) | Total PSMs | Total unique sequences | Novor PSMs | Novor unique sequences | Database PSMs (%) | Database unique sequences (%) |
|---|---|---|---|---|---|---|---|---|
| H. sapiens | human | Chordata/mammalia/primate/hominidae | 35,488 | 19,693 | 2,937 | 1,804 | 91.7 | 90.8 |
| H.sapiens, shuffled | Randomized human | | 24,800 | 13,793 | 24,732 | 13,737 | 0.3 | 0.4 |
| P. troglodytes | chimp | Chordata/mammalia/primate/hominidae | 34,636 | 19,142 | 3,679 | 2,284 | 89.4 | 88.1 |
| G. gorilla | gorilla | Chordata/mammalia/primate/hominidae | 34,198 | 18,902 | 4,555 | 2,734 | 86.7 | 85.5 |
| P. abelii | orangutan | Chordata/mammalia/primate/hominidae | 33,659 | 18,555 | 5,086 | 3,147 | 84.9 | 83.0 |
| O. garnettii | galago | Chordata/mammalia/primate/galagidae | 30,676 | 16,769 | 9,614 | 5,622 | 68.7 | 66.5 |
| F. catus | cat | Chordata/mammalia/carnivora/felidae | 30,935 | 16,844 | 8,993 | 5,404 | 70.9 | 67.9 |
| S. harrisii | Tasmanian devil | Chordata/mammalia/dasyuromorphia/dasyuridae | 28,709 | 15,639 | 13,597 | 7,942 | 52.6 | 49.2 |
| O. anatinus | platypus | Chordata/mammalia/monotremata/ornithorhynchidae | 28,090 | 15,345 | 15,850 | 9,195 | 43.6 | 40.1 |
| C. brachyrhynchos | crow | Chordata/aves/passeriformes/corvidae | 27,735 | 15,155 | 17,508 | 10,119 | 36.9 | 33.2 |
| X. laevis | frog | Chrodata/amphibian/anura/pipidae | 26,370 | 14,402 | 17,061 | 9,766 | 35.3 | 32.2 |
| G. aculeatus | fish (stickleback) | Chordata/actinopterygii/gasterosteiformes/gasterosteidae | 26,288 | 14,486 | 19,502 | 11,061 | 25.8 | 23.6 |
| B. floridae | lancelet | Chordata/leptocardii/amphioxiformes/branchiostomidae | 25,356 | 13,985 | 22,653 | 12,753 | 10.7 | 8.8 |
| T. Asiatica | tape worm | | 25,859 | 14,342 | 24,445 | 13,764 | 5.6 | 4.0 |

Table S7: C. elegans LCMSMS data searched against FASTA files of various other nematode species to which high scoring (>=60) *de novo* sequences had been appended.  The column “Database unique (%)” is plotted in Fig. S3.

| Species | Common name | Taxonomy (phylum/class/ order/family) | Total PSMs | Total unique sequences | Novor PSMs | Novor unique sequences | Database PSMs (%) | Database unique sequences (%) |
|---|---|---|---|---|---|---|---|---|
| C. elegans | C. elegans | Nematoda/Chromadorea/Rhabditida/Rhabditoidea | 25,273 | 13,888 | 2825 | 1,568 | 88.8 | 88.7 |
| C. elegans, shuffled | C. elegans | | 18,933 | 10,420 | 18,906 | 10,395 | 0.1 | 0.2 |
| C. briggsae | Most related to C. elegans | Nematoda/Chromadorea/Rhabditida/Rhabditoidea | 21,324 | 11,610 | 10,456 | 5,855 | 51.0 | 49.6 |
| N. americanus | Human hookworm | nematoda/chromadorea/rhabditida/Ancylostomatidae | 20,048 | 10,892 | 16,335 | 9,009 | 18.5 | 17.3 |
| D. viviparus | Lung worm | Nematoda/chromadorea/Rhabditida/Dictyocaulidae | 20,109 | 10,956 | 16,713 | 9,238 | 16.9 | 15.7 |
| O. dentatum | Nodular worm | Nematoda/chromadorea/rhabditida/cloacinidae | 19,843 | 10,827 | 16,634 | 9,188 | 16.2 | 15.1 |
| H. bacteriophora | Entomopathogenic nematode | nematoda/chromadorea/rhabditida/heterorhabditidae | 19,903 | 10,896 | 18,090 | 9,976 | 9.1 | 8.4 |

**Supplemental Figures**

Figure S1: Modeling molecular weight normalized cross-correlation scores using first and second ranked decoy matches. Correct database hits can sometimes have slightly lower xcorr scores than nearly correct *de novo* sequences due to incidental matching of an extra fragment ion or two (e.g., Table S5). Likewise, the top decoy sequence matches an extra fragment ion or two compared to the second best decoy sequence. Here we show the molecular weight normalized xcorr score differences between slightly incorrect *de novo* sequences and correct but lower scoring database sequences (orange), compared to score differences between first and second ranked decoy hits (blue). The left panel shows results from a human tryptic digest, and the right panel shows results from a *C. elegans* tryptic digest. These results suggest that the decoy xcorr values provide a useful model for determining score cutoffs for deciding when a second ranked database sequence has a sufficiently close xcorr value to be considered correct. In these data a molecular weight normalized xcorr score difference of 0.0002 would capture about 95% of the cases where a correct database sequence is ranked second after a *de novo* sequence.

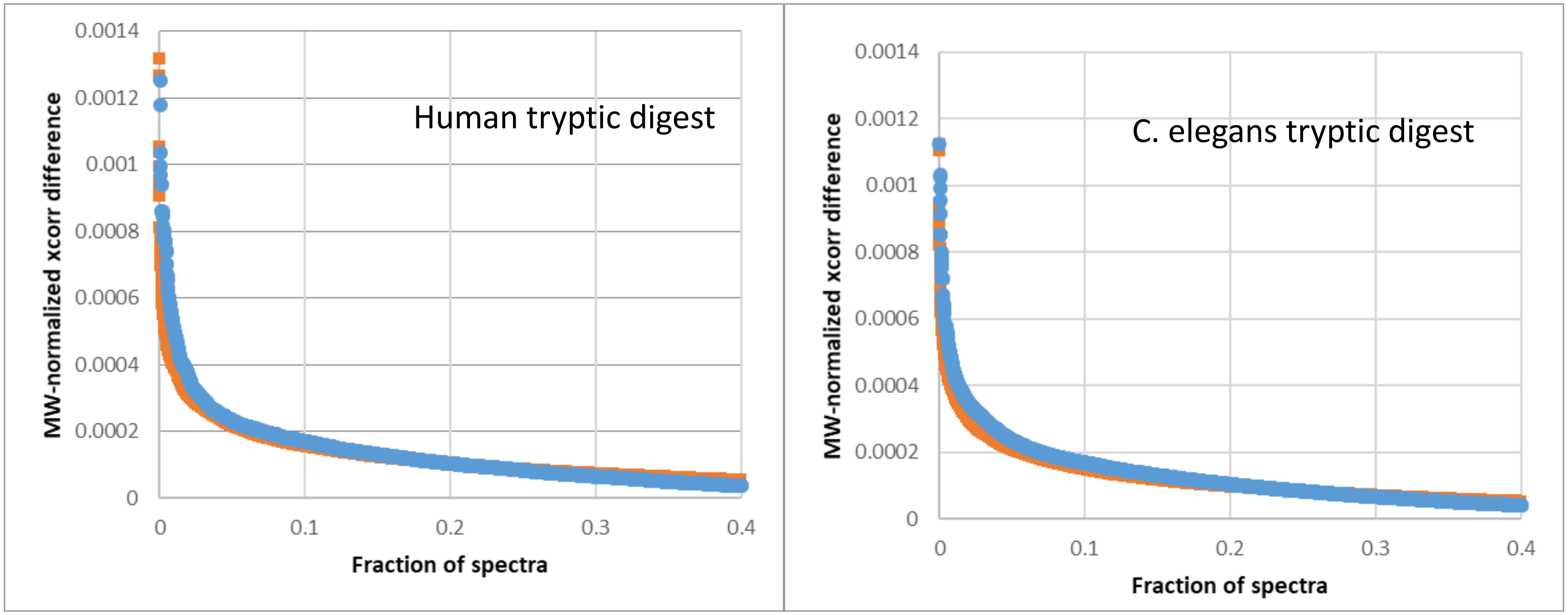
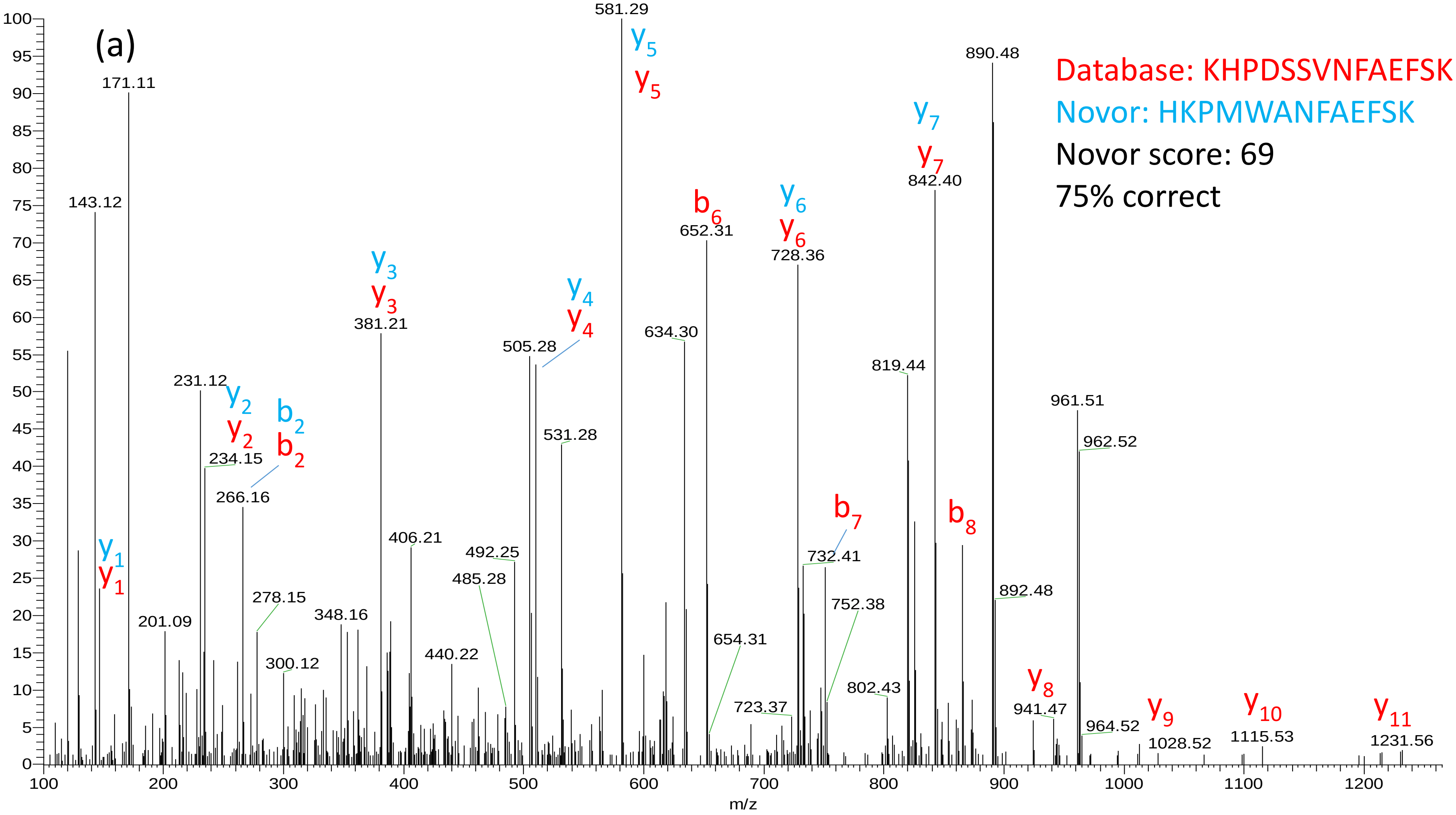
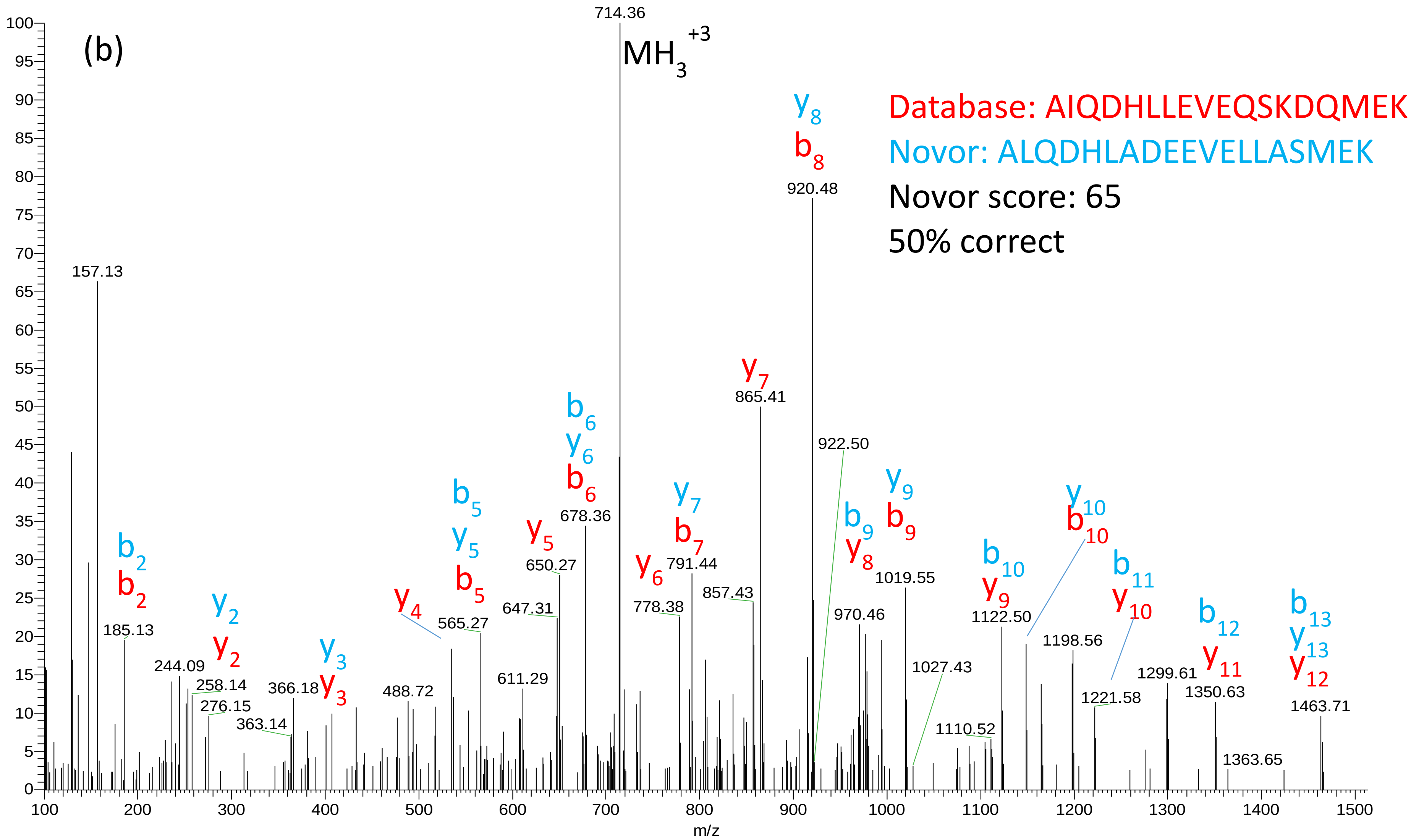


Figure S2: MS/MS spectra for the alignment examples in Fig. 1(b-d). The red labeling indicates fragment ion assignments for the database derived sequence, and the blue labeling shows the assignments for the *de novo* sequence. Most of the unlabeled fragments are due to co-isolation and co-fragmentation of other peptides. Shown are MS/MS spectra of $MH_3^{+3}$ at m/z 531.6 (a), $MH_3^{+3}$ at m/z 714.4 (b), and $MH_2^{+2}$ at m/z 899.4 (c).



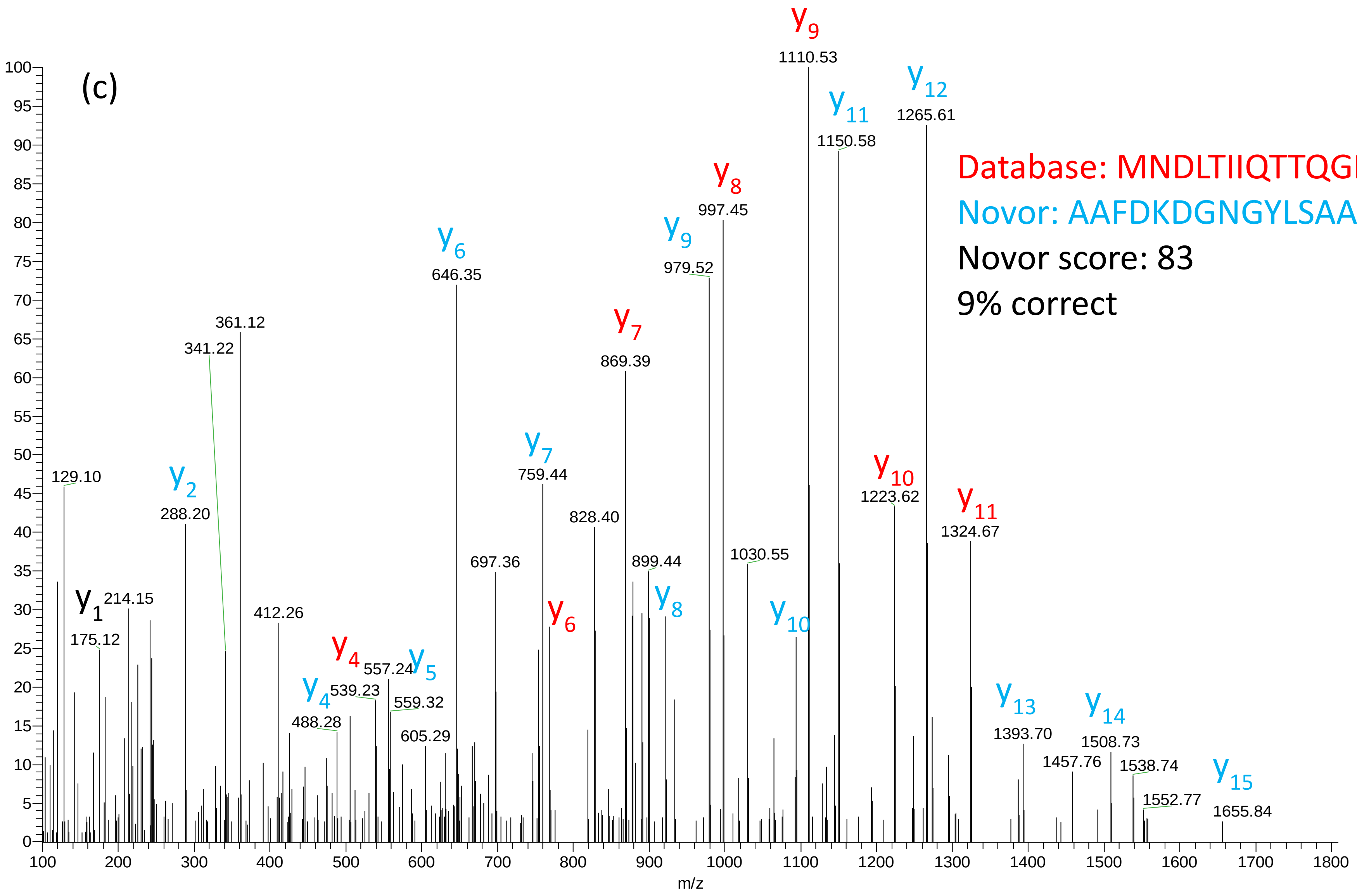


Database: AIQDHLLEVEQSKDQMEK

Novor: ALQDHLADEEVELLASMEK

Novor score: 65

50% correct



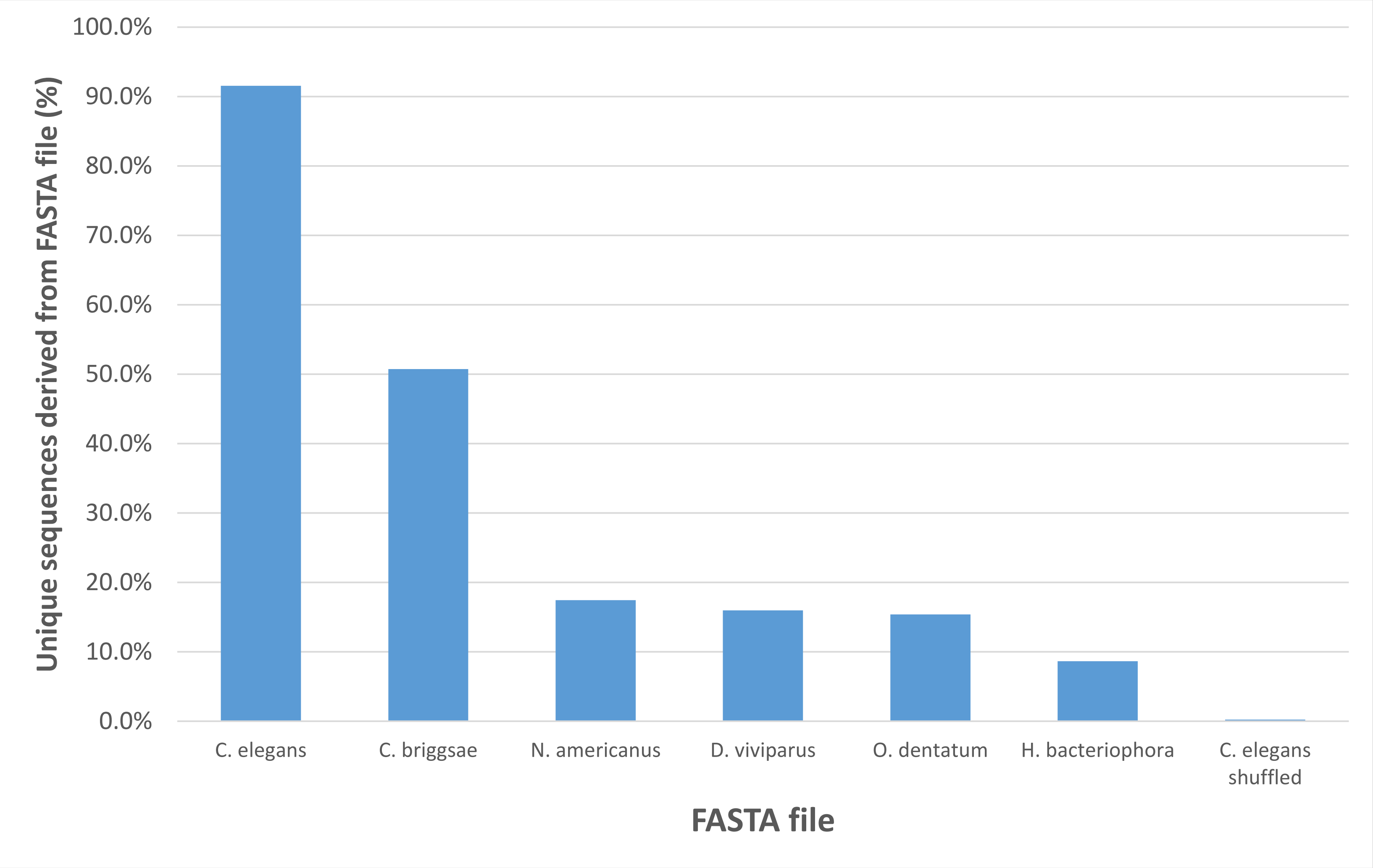
Database: MNDLTIIQTTQGFCR

Novor: AAFDKDGNGYLSAAELR

Novor score: 83

9% correct

Figure S3: Searching LC-MS/MS data from *C. elegans* tryptic peptides against various FASTA files from different members of the Nematoda phylum. Using a PeptideProphet FDR of 0.01 and maximum Comet E-value of 0.01, the fraction of unique peptides that best matched to FASTA file sequences are shown.


Unique sequences derived from FASTA file (%)
C. elegans
C. briggsae
N. americanus
D. viviparus
O. dentatum
H. bacteriophora
C. elegans shuffled
FASTA file