MetaNovo: a probabilistic approach to peptide and polymorphism discovery in complex metaproteomic datasets

Matthys G Potgieter[1,2], Andrew JM Nel[2,], Suereta Fortuin[2], Shaun Garnett[2], Jerome M. Wendoh[3], David L. Tabb[2,4], Nicola J Mulder[1,5], Jonathan M Blackburn[2,5]

1. Computational Biology Division, Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town, South Africa.
2. Division of Chemical and Systems Biology, Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town, South Africa.
3. Division of Immunology, Department of Pathology, University of Cape Town, Cape Town, South Africa.
4. Division of Molecular Biology and Human Genetics, Department of Biomedical Sciences; African Microbiome Institute; South African Tuberculosis Bioinformatics Initiative; Stellenbosch University, Cape Town, South Africa
5. Institute of Infectious Disease & Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa.

Abstract

Metagenome-driven microbiome research is providing important new insights in fields as diverse as the pathogenesis of human disease, the metabolic interactions of complex microbial ecosystems involved in agriculture, and climate change. However, poor correlations typically observed between RNA and protein expression datasets even for single organisms make it hard to infer microbial protein expression with any accuracy from metagenomic data, thus restricting movement beyond microbial catalogues and into functional analysis of microbial effector molecules. By contrast, mass spectrometry analysis of microbiome data at the protein level in theory allows direct measurement of dynamic changes in microbial protein composition, localisation and modification that may mediate host/pathogen interactions in complex microbial ecosystems, but analysis of such metaproteomic datasets remains challenging.

Here we describe a novel data analysis approach, **MetaNovo**, that searches complex datasets against the entire known protein universe, whilst still controlling false discovery rates, thus enabling metaproteomic data analyses without requiring prior expectation of likely sample composition or metagenomic data generation that classically inform construction of focussed, relatively small search libraries. **MetaNovo** directly identifies and quantifies the expressed metaproteomes, and estimates the microbial composition present in complex microbiome samples, using scalable *de novo* sequence tag matching and probabilistic optimization of very large, unbiased sequence databases prior to target-decoy search. We validated MetaNovo against the results obtained from the recently published **MetaPro-IQ** pipeline on 8 human mucosal-luminal interface samples, with comparable numbers of peptide and protein identifications being found when searching relatively small databases. We then showed that using an unbiased search of the entire release of **UniProt** (ca. 90 million protein sequences[1]) MetaNovo was able to identify a similar bacterial taxonomic distribution compared to that found using a small, focused matched metagenome database, but now also simultaneously identified proteins present in the samples that are derived from other organisms missed by 16S or shotgun sequencing and by previous metaproteomic methods**.** Using **MetaNovo** to analyze a set of single-organism human neuroblastoma cell-line samples (*SH-SY5Y*) against

[1] ftp://ftp.uniprot.org/pub/databases/uniprot/previous_releases/release-2017_08/

UniProt we achieved a comparable MS/MS identification rate during target-decoy search to using the **UniProt** human Reference proteome, with 22583 (85.99 %) of the total set of identified peptides shared in common. Taxonomic analysis of 612 peptides not found in the canonical set of human proteins yielded 158 peptides unique to the *Chordata* phylum as potential human variant identifications. Of these, 40 had previously been predicted and 9 identified using whole genome sequencing in a proteogenomic study of the same cell-line.

By estimating taxonomic and peptide level information on microbiome samples directly from tandem mass spectrometry data, **MetaNovo** enables simultaneous identification of human, bacterial, helminth, fungal, viral and other eukaryotic proteins in a sample, thus allowing correlations between changes in microbial protein abundance and change in the host proteome to be drawn based on a single analysis.

Data are available via ProteomeXchange with identifier PXD014214. The **MetaNovo** software is available from **GitHub**[2] and can be run as a standalone **Singularity** or **Docker** container available from the **Docker Hub**[3].

[2] https://github.com/uct-cbio/proteomics-pipelines
[3] https://hub.docker.com/r/thyscbio/metanovo

Introduction

Characterizing complex mass-spectrometry data from clinical or environmental samples allows for potential insights into the complex metabolic pathways involved in processes as diverse as carbon sequestration and climate change to the susceptibility and progression of human diseases. Parallels have been drawn between the role of soil microbiota in the biosphere and the role played by gut microbiota in human health (Ochoa-Hueso, 2017). Soil microbes metabolize organic material derived from plants and store carbon in inert forms, while gut microbiota assist the human host in metabolizing complex dietary constituents, assisting with metabolism and detoxification (Blaser *et al.,* 2016). Microbial ecosystems play a wide-ranging role in the complex determinants of human well-being, with potential disease implications should the delicate balance be disrupted.

Genome and transcriptome sequencing approaches to microbiome analysis allow for functional and taxonomic characterization of the genes and organisms involved in complex microbial communities, but evidence for gene transcription does not in general correlate well with measured protein abundance (Castellana and Bafna, 2010). On the other hand, metaproteomic and clinical proteomic approaches in principle allow researchers to obtain a snapshot of all the proteins present in a complex microbiome sample at a given time, providing a quantitative window into key functional components of complex pathways and organism interactions, including the direct measurement of dynamic changes in microbial protein composition, localisation and modification that may mediate host/pathogen interactions in the context of human health and disease.

Classically, the identification of peptides (and by inference, the parent proteins) from complex mass spectrometry datasets has been dependent on the availability of focussed, representative and relatively small sequence databases or spectral libraries against which to search tandem mass spectra; as a result, the field of proteomics has been primarily directed to the analysis of single organism proteomes to date. However, analysis of complex, multi-species samples – such as human microbiome samples, where the total number of microbial genes may vastly exceed the number of human genes - is far less straightforward

for a variety of reasons, including: the vast majority of specific organisms in any given microbiome are likely to not have been cultured, identified, or characterised in the laboratory, so appropriate reference genomes may not exist to underpin proteomic data analyses; and the fact that massive expansion of proteomic sequence databases used in target-decoy-based peptide-spectrum matching has been shown to lead to dramatically reduced identification rates due to reduced sensitivity and false discovery mis-estimation problems (Heyer *et al.*, 2017; Blackburn & Martens, 2016).

Previous methods for metaproteomic analysis have addressed these issues essentially by relying on prior expectation of the microbial composition of a given microbiome sample, for example using 16S sequencing data to produce a focussed metaproteome library to search. However, the weaknesses in this approach are that: (i) it assumes that the 16S data itself provides a complete description of all bacteria present in the sample, which is probably not true; (ii) it assumes that appropriate reference bacterial proteomes exist and are close enough in sequence to the clinical isolate proteomes for proteomic data analysis to be valid, which may also not be true; & (iii) it automatically ignores all non-bacterial components of the microbiome. Furthermore, identified peptides may belong to multiple homologous proteins, complicating protein inference and quantification. Particular challenges faced by adult gut metaproteomics include inter-subject variability (Li *et al*, 2011), combined with the need to factor in the influence of a very diverse and variable human diet. Thus, whilst matched metagenome bacterial sequencing is the current gold standard to create sequence databases to search metaproteomic mass spectrometry data against, such data is not always available nor necessarily complete. Moreover, such curated databases for a given microbiome typically do not include viral, fungal, host or other eukaryote proteomes that may be present, nor do they account for possible sequence polymorphisms compared to reference strains. The ideal database for mass spectrometry-based peptide identification in metaproteomic datasets would therefore be comprehensive, whilst excluding proteins that are genuinely absent, and would not be reliant on prior expectation of possible sample composition, nor would it require prior generation of DNA/RNA sequencing data to inform construction of a search database.

Various tools have been developed to address the challenges of metaproteomic MS/MS identification. Tang et al. (2016) used *de Bruijn* graphs generated by metagenome assembly to produce protein databases for metaproteomics, showing the value of incorporating metagenome and metatranscriptome data in protein and peptide identification. Zhang *et al.* (2016) showed that an iterative search of the human gut microbial integrated gene catalogue (IGC) produces comparable results to using a matched metagenome database.

De novo sequencing of peptides from mass spectrometry data has long been used for database filtration, allowing for rapid searches of very large search spaces with sequence tags prior to peptide-spectral matching (Frank *et al*, 2007). Particularly considering significant inter-subject variation and the polymorphism of clinical strains, *de novo* sequencing has been suggested as a viable strategy for peptide identification in gut metaproteomics (Blackburn & Martens, 2016). However, scalable and robust *de novo* sequencing-based pipelines need to be able to process the rapidly expanding amount of proteomics and genomics information available.

Here we present a novel metaproteomic analysis pipeline - **MetaNovo** - based on mapping *de novo* sequence tags to very large protein sequence databases using a high performance and parallelized computing pipeline, generating compact databases for target-decoy analysis and False Discovery Rate (FDR) controlled protein identification, suitable for integration with existing MS/MS analysis pipelines. The pipeline has three main components based on custom and existing open-source tools - generating *de novo* sequence tags, mapping the tags to a protein sequence database, and probabilistic protein ranking and filtering based on estimated species and protein abundance. We benchmark **MetaNovo** against **MetaPro-IQ** (Zhang *et al*, 2016) using publicly-available mass spectrometry datasets and also demonstrate the ability to identify coding polymorphisms and mutations in cultured human cells that differ from the reference human genome by ~3.9 million single nucleotide variants and ~750,000 indels. Finally, we demonstrate the utility of **MetaNovo** in analysing temporal changes in an adult gut microbiome, including the unexpected identification of helminth-derived peptides, as well as correlated changes in the human gut proteome.

Materials and Methods

MetaNovo database generation workflow

UniProt database statistics. The August 2017 release of **UniProt**[4] was downloaded, and a combined FASTA database created from **UniProt SwissProt**, **TREMBL**, and **SwissProt** alternative splice variant sequences (ca. 90 million sequences). See *Table 1. Taxonomic distribution by Kingdom of the **UniProt** database.*

		SwissProt	TREMBL	Total	%
Archaea		19463	1701862	1721325	1.91
Bacteria		333178	58831274	59164452	65.77
Viruses		16734	3138225	3154959	3.51
Other		0	1124964	1124964	1.25
Eukaryota	**Human**	20219	140090	160309	0.18
	Other Mammalia	46508	1360304	1406812	1.56
	Other Vertebrata	18141	2575565	2593706	2.88
	Viridiplantae	39030	4783321	4822351	5.36
	Fungi	32833	7080775	7113608	7.91
	Insecta	9092	2763937	2773029	3.08
	Nematoda	4832	1377572	1382404	1.54
	Other	15396	4518427	4533823	5.04
Total		555426	89396316	89951742	100.00

Table 1. Taxonomic distribution by Kingdom of the UniProt database. The August 2017 release of **UniProt** consists of ca. 90 million protein sequences, the majority of which are Bacteria.

[4] ftp://ftp.uniprot.org/pub/databases/uniprot/previous_releases/release-2017_08/knowledgebase/

Generation of *de novo* sequence tags. Raw MS/MS files were converted to *Mascot Generic Feature* (MGF) format prior to analysis. *De novo* sequence tags were generated using **DeNovoGUI** version 1.15.11 (Muth *et al.*, 2014), with **DirecTag** (Tabb *et al.*, 2008) selected as the sequencing engine. General **MetaNovo** parameters included fragment and precursor ion mass tolerance of 0.02 Da, with fixed modification '*Carbamidomethylation of C'* and variable modifications '*Oxidation of M'* and *'Acetylation of protein N-term'*. For **DirecTag**, a tag sequence length of 4 amino acids was required, and the top 5 sequence tags per spectrum were selected (taking into account alternative possible charge states).[5] The output of **DirecTag** was parsed with a custom **Python** script and all sequence tags for each MS/MS across replicates were stored in a SQLite3 database. The distinct set of sequence tags (N-terminus mass gap, amino acid sequence, and C-terminus mass gap) was obtained using an SQL query, combining identical tags across multiple MS/MS into a single non-redundant list.

Mapping sequence tags to a FASTA database. PeptideMapper (Kopczynski et al., 2017) included in **compomics-utilities** version 4.11.19 (Barsnes *et al*, 2011) was used to search the sequence tag set against the protein sequence database. The same mass error tolerance and post-translational modification settings were used as for **DeNovoGUI** above**.** The open-source **GNU parallel** tool (Tange, 2011) was used to search the set of sequence tags against the fasta database in configurable chunk sizes, using a configurable number of cores per node in parallel.

Enzymatic cleavage rule evaluation. A custom script was used to evaluate the cleavage rule of the peptide sequences of identified sequence tags**.** Only peptides passing the tryptic cleavage rule were selected for downstream analysis (cleavage after *Lys* or *Arg* except when following *Pro*). The corresponding sequence tags for the validated peptide sequences were queried against the **SQLite3** database to obtain a mapping of MS/MS ids to protein ids, allowing for the estimation of protein abundance based on mapped MS/MS in a similar manner to spectral counting in a target-decoy search.

---

[5] Preliminary runs with a sequence tag length of 3 proved impractical with very large databases, requiring more than 5 days with 24 cores on a high-performance cluster to search 8 MGF files against **UniProt**. Similarly, increasing sequence tag options per spectrum from 5 to 30 considerably increased processing time without a corresponding gain in sensitivity, possibly due to the inclusion of low-scoring matches.

Normalized spectral abundance factor calculation. Protein abundance was estimated using the *Normalized Spectral Abundance Factor* (NSAF) approach developed for shotgun proteomics (Zybailov *et al*, 2007). The set of MS/MS IDs in each sample per protein was obtained, and the number of mapped MS/MS ids divided by protein length to obtain a *Spectral Abundance Factor* (SAF). The SAF values for each sample were divided by the sum of the SAF values in each sample, to obtain a *Normalised Spectral Abundance Factor* (NSAF) for each protein in each sample/replicate.

Probabilistic ranking of database proteins. Protein NSAF values were summed across replicates, and each summed protein NSAF value was divided by the sum of the column to obtain a protein posterior probability based on estimated relative protein abundance across replicates, with a maximum value of 1 for each protein. Database proteins were ranked by protein posterior probability, and a minimal set of ranked proteins obtained such that each protein in the list contained at least one unique MS/MS scan ID relative to the set of proteins above its position in the list. The organism name was obtained from the **UniProt** FASTA headers, and summed NSAF values aggregated by organism in the minimal set and divided by the total to obtain an organism level posterior probability. Following this, the proteins in the unfiltered set were annotated with their respective calculated organism posterior probability. The unfiltered set of proteins was re-ranked based on the joint probability of the organism and protein probability values, which allows low abundance proteins with a high estimated relative abundance at the organism level to be favoured compared to proteins where both the protein and organism abundance was estimated to be low. Following the second ranking step, the database was re-filtered to obtain a minimal subset of proteins that could explain all mapped *de novo* sequence tags, based on estimated relative protein and organism abundance.

The pipeline can be obtained from GitHub or downloaded as a standalone Singularity-compatible Docker container from the **Docker Hub**.[6] The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (Perez-Riverol *et al.*, 2019) partner repository with the dataset identifier PXD014214.

[6] https://hub.docker.com/r/thyscbio/metanovo

Mucosal-luminal interface samples

Mucosal-luminal interface (MLI) **sample metagenomic and proteomics data.** Metagenome and proteomics data of 8 MLI samples from adolescent volunteers obtained during colonoscopy were downloaded from PRIDE[7] and through author correspondence. The sample processing, mass spectrometry and metagenomics database creation methods have already been described (Zhang *et. al,* 2016).

***MetaNovo* database generation.** The August 2017 release of **UniProt**[8], containing ca. 90 million protein entries, was used to create a database containing 67246 entries. The default **MetaNovo** settings described above were used.

Database search using MaxQuant. MaxQuant version 1.5.2.8 was used to search the **MetaNovo** database, with the same settings as for the **MetaPro-IQ** publication. *Acetyl (Protein N-term)* and *Oxidation (M)* were selected as variable modifications, and *Carbamidomethyl (C)* was selected as a fixed modification. Specific enzyme mode with *Trypsin/P* was selected with up to 2 missed cleavages allowed, and a PSM and protein FDR of 0.01 was required.

[7] http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD003528
[8] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/

Human neuroblastoma (*SH-SY5Y*) cell line samples

Protein extraction and LC MS/MS analysis. Cells from the *SH-SY5Y* cell line were donated by the Wilkinson Lab at the University of Cape Town (UCT). The cells were cultured in neuroblastoma medium (DMEM), 10% FCS, and 1x PenStrep on a T-150 flask with a canted neck and surface area of 150 cm2 (Sigma CLS3291). When the cells were 80% confluent they were lifted with trypsin. Protein was extracted by FASP in 50 ml 30kDa MWCO filters. Peptides were desalted in a 1ml C18 Supelco SPE cartridge. Samples were run on a Dionex with a 2cm 100µm trap column and a 40cm 75µm analytical column. Sample loading was normalised to an average total ion chromatogram (TIC) peak height of 1×1010. Peptides were separated by a 120 minute gradient from 5%B to 25%B, and Nano ESI set to 2.1kV for analysis on a Q-Exactive Orbitrap mass spectrometer (QE). The QE was operated in full ms1 to data dependent ms2 mode, and peptides were fragmented with a normalised collision energy of 25 and scanned with a resolution of 17 500.

Validation database generation. Sixteen separate human neuroblastoma (*SH-SY5Y)* cell-line MS/MS datasets generated by our lab were analyzed using the **MetaNovo** pipeline with the default settings described above, to obtain a database of 34880 entries from **UniProt**.

Database search with MaxQuant. The human neuroblastoma samples were searched against the human Reference proteome and **MetaNovo** database using **MaxQuant** 1.5.6.5 at 1% protein and PSM level FDR. *Carbamidomethylation of C* was selected as a fixed modification and *Oxidation of methionine* and *Acetylation of protein N-terminal* were selected as variable modifications. Specific digestion with *Trypsin/P* was selected, allowing for a maximum of 2 missed cleavages.

Temporal patterns of diet and microbial diversity in an adult gut microbiome

Ethics statement. The stool samples were provided with written informed consent by a single adult volunteer and, according to the Human Research Ethics Committee of the University of Cape Town, constitute a single case report.

Protein extraction and LC MS/MS analysis. Four stool samples from an adult volunteer were obtained over two one-week windows (two samples per window) separated by one month, representing known differences in diet and clinical symptoms. During the first time window, the volunteer reported vague clinical symptoms, including weight-loss, fatigue and myalgia, while the aetiology remained unknown. The patient reported a predominantly plant-based diet low in animal products at the time, as well as a recent tropical travel history. The second window samples were taken during a self-reported normal diet, with resolved clinical symptoms and a broader variety in diet reported. Each biological replicate was processed using two different sample processing protocols, in solution (insol) and filter-assisted sample processing (FASP) methods. The insol and FASP samples were treated as technical replicates for each sample. Detailed sample processing methodology, as well as LC-MS/MS data acquisition parameters, are described in *Addendum A: Adult Stool Samples*.

Adult gut metaproteome database generation. The **MetaNovo** pipeline was run with default parameters, mapping the *de novo* sequence tags of 16 adult stool samples against **UniProt** to obtain a sequence database of 49047 proteins.

Database search with MaxQuant. The adult stool **MetaNovo** database were searched using **MaxQuant** version 1.5.6.5. *Carbamidomethyl (C)* was selected as a fixed modification and *Oxidation (M)* and *Acetyl (Protein N-term)* were selected as variable modifications. Specific digestion with *Trypsin/P* was selected, allowing for a maximum of 2 missed cleavages, and a PSM and protein FDR of 0.01 was required.

Bioinformatic analysis

Posterior error probability (PEP) score analysis. The PEP scores of peptides identified using **MaxQuant** were obtained from the *peptides.txt* output files. Significant differences between groups of interest were tested for using the *Kruskal-Wallis* non-parametric analysis of variance test followed by *Dunn's* post-hoc test using a custom python script. PEP score distributions were visualized using the python *matplotlib* library.

Taxonomic analysis. Peptide sequences were assigned to a lowest common ancestor using the **UniPept *pept2lca*** tool. Peptide and spectral counts were aggregated at selected phylum levels reported in the ***pept2lca*** output, with custom visualizations done using the Python *matplotlib* module. For the differential abundance analysis of the adult stool samples, intensity-based quantification was performed using normalized ion intensities, aggregated at selected phylum levels in each sample.

Pathway and gene set analysis of adult stool data. The list of protein accessions identified from the ***MetaNovo*** database was used to create a targeted database for *InterProScan* analysis (version 5.29-68.0). The *InterProScan* output was processed using custom *Python* and *R* scripts into gene sets compatible with the *R gage* package for gene set analysis. *InterProScan* annotations used include *IPR*, *GO*, *Reactome*, *MetaCyc*, *EC* and *KEGG* terms.

Differential abundance analysis of adult stool data. A moderated *t-test* was performed on the adult stool data using the R *limma* tool, on quantified taxonomic levels as well as protein and peptide intensities between timepoints. *Benjamini-Hochberg* multiple testing correction was performed using the R *qvalue* package.

Results and discussion

MetaNovo identifies non-bacterial phyla and species, whilst providing comparable bacterial composition data to a matched metagenome approach in human Mucosal-luminal interface samples

Protein and peptide identifications. The **MetaNovo** pipeline was run using the entire, unfiltered **UniProt** database (Aug 2017 release; ca. 90 million entries), resulting in the identification of 69314 target peptides and 15204 protein groups, with 36.29 % of total MS/MS identified. These results are comparable to those of **MetaPro-IQ** on the same samples, with 34 and 33% MS/MS identification rates reported by the authors using a matched metagenome and the **MetaPro-IQ** approach, respectively. PEP score analysis of shared and exclusive peptide identifications in each group revealed significantly different distributions between peptides only identified using **MetaNovo/UniProt** and reverse hits (p-value 1.085e-27), indicating true positive identifications not found using the other approaches. *See Figure 1 C, Supplementary Table 1.*[9]

Taxonomic comparisons. Spectral counts by lowest common ancestor were compared between different analysis runs. The distribution of spectral counts by phylum and species were similar across all runs, with higher numbers of *Chordata* at the phylum level, and higher numbers of *Homo sapiens* and *Bos taurus* identified, using **MetaNovo**. In total, the **MetaNovo** analysis yielded 45 distinct phyla, while the **MetaProt-IQ/IGC** run yielded 14, illustrating the focus on bacterial phyla in the curated IGC database (See *Table 2*). At the species level, **MetaNovo** yielded 405 distinct species, whereas **MetaPro-IQ/IGC** identified 123 distinct species (see *Supplementary Table 2*).

[9]An attempt to run the **MetaPro-IQ** pipeline against the **UniProt** database was not successful due to wall-time restrictions on a high performance cluster causing X!Tandem runs to fail.

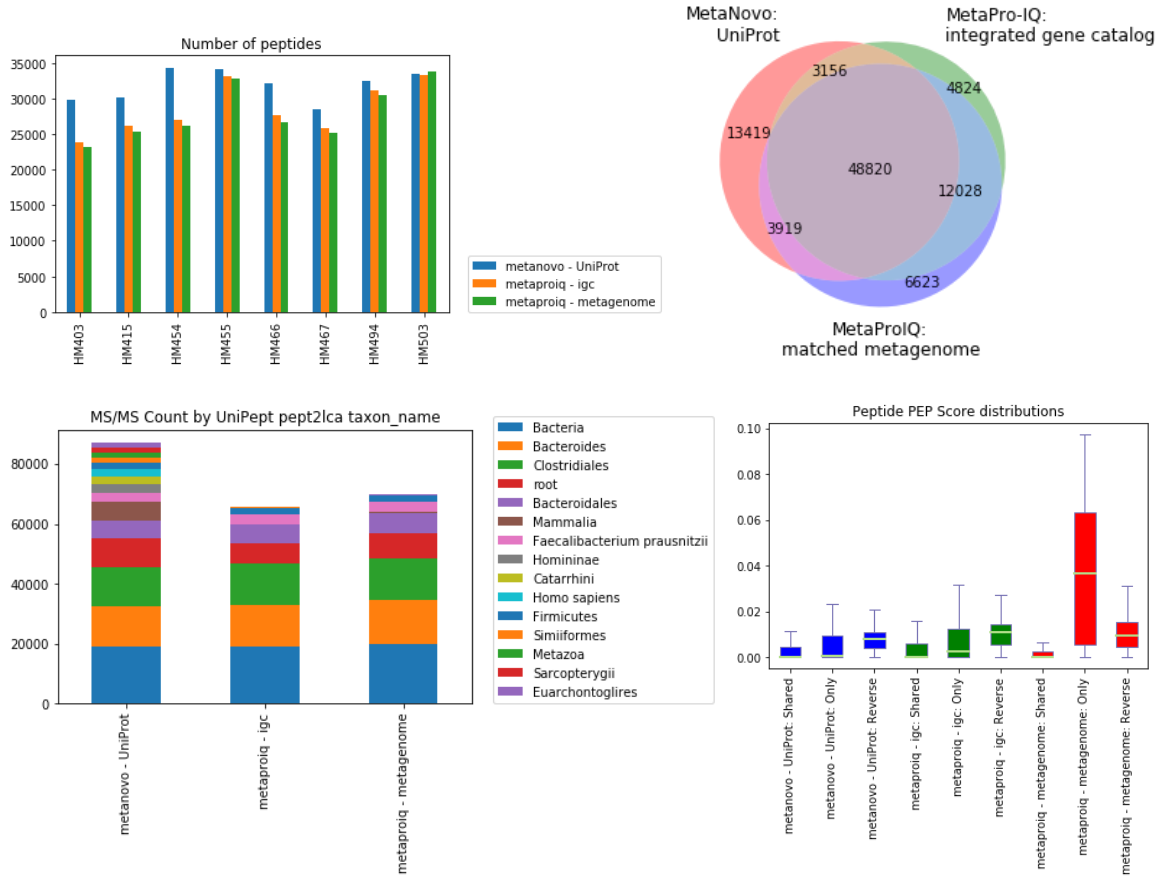

Figure 1. A. Bar chart of peptide identifications. The identification rates of **MetaNovo** are comparable to the previously published results of **MetaPro-IQ** using matched metagenome and integrated gene catalog databases. **B.** Venn diagram showing large overlap in identified sequences using different approaches, with the highest number of sequences identified using **MetaNovo. C.** Spectral counts by *UniPept lowest common ancestor* showed similar taxonomic distributions obtained from different approaches. **D.** Peptides uniquely identified by **MetaNovo** had a significantly different distribution compared to reverse hits (p-value 1.08e-27), indicating true positive identifications not found using the other two approaches. The boxes extend from the lower to the upper quartile, and the whiskers represent 1.5 times the interquartile range (IQR) below and above the first and third quartiles, respectively.

phylum_name	metanovo - UniProt	metaproiq - igc	metaproiq - metagenome
Bacteroidetes	26808	28222	29723
Firmicutes	23476	25166	24671
Chordata	24729	796	858
Proteobacteria	726	658	640
Actinobacteria	273	309	147
Streptophyta	217	21	7
Verrucomicrobia	115	144	143
Arthropoda	122	2	0
Ascomycota	32	4	2
Planctomycetes	31	1	0
Nematoda	23	0	0
Basidiomycota	15	2	1
Spirochaetes	10	1	0
Chytridiomycota	10	0	0
Chlorobi	9	0	0
Mucoromycota	8	0	0
Euryarchaeota	8	5	1
Chlorophyta	6	0	0
Cnidaria	6	0	0
Zoopagomycota	5	0	0
Nitrospirae	5	0	0
Apicomplexa	5	0	1
Cyanobacteria	5	0	0
Brachiopoda	5	0	0
Porifera	5	0	0
Platyhelminthes	5	0	0
Candidatus Uhrbacteria	5	0	0
Tenericutes	4	4	0
Synergistetes	4	0	0
Lentisphaerae	4	0	0
Rotifera	3	0	0
Candidatus Berkelbacteria	3	0	0
Candidatus Doudnabacteria	3	0	0
Chloroflexi	3	0	0
Mollusca	2	0	0
Ignavibacteriae	2	0	1
Candidatus Nealsonbacteria	2	0	0
Candidatus Schekmanbacteria	2	0	0
Candidatus Liptonbacteria	2	0	0
Candidatus Moranbacteria	1	0	0
Microsporidia	1	0	0
Candidatus Andersenbacteria	1	0	0
Candidatus Saccharibacteria	1	0	0
Crenarchaeota	1	0	0
Armatimonadetes	1	0	0

Table 2. Spectral counts by UniPept lowest common ancestor, aggregated by phylum. In total, the **MetaNovo** run yielded 45 distinct phyla, while the **MetaProt-IQ/IGC** run yielded 14 and the **MetaProt-IQ/Metagenome** run yielded 12.

MetaNovo identifies novel polymorphisms in a human neuroblastoma cell-line, whilst providing similar coverage to use of a single reference proteome

Peptide and protein identifications. Using the human Reference proteome obtained from **UniProt**, 25285 target peptides, and 3451 target proteins were identified with a typical **MaxQuant** workflow, with 37.61 % of MS/MS identified in the cell-line. With **MetaNovo**, 23613 target peptides and 3694 target proteins were identified, with 37.2 % MS/MS identification rate. 85.99 % of the total set of identified target peptides between the two runs were shared in common - indicating a sensitive and specific identification rate using the **MetaNovo** pipeline approaching a Reference proteome approach to single organism MS/MS dataset. *See Figure 2 A, B.*

PEP score analysis indicates high-confidence identifications only obtained using MetaNovo. 1030 peptides were only identified using **MetaNovo**, while 2649 were only identified using the human Reference proteome - with 22583 peptides identified in common between the two runs. A statistically significant difference between the PEP score distributions of peptides only identified using **MetaNovo** and reverse hits was found (p-value 0.0488), while the peptides only identified using the Reference proteome did not have a significantly different distribution from reverse hits (p-value 0.447); *see Supplementary Table 3.* This implies that the **MetaNovo**-only identifications contain a significant proportion of true-positive identifications, whereas the Reference proteome-only hits likely contain a significant proportion of false-positive identifications.

Novel polymorphism identifications from orthologous sequences. Peptides identified using the **MetaNovo** approach were compared to the human Reference proteome FASTA database. Sequences mapping exactly to a region in a known Reference protein, regardless of the tryptic nature at the matched position, were considered annotated. After excluding peptides belonging to known contaminants, the remaining 612 peptides were considered potential novel human annotations or sample contaminants. A statistically significant

difference was found between the PEP score distribution of these 612 peptides, and reverse hits (p-value 0.000849).

Taxonomic analysis was used to narrow down the list to 158 peptides annotated to a lowest common ancestor in the *Chordata* phylum, as peptides most likely to be of human origin in this set. *Kruskal-Wallis* analysis of variance followed by *Dunn's test* post-hoc did not reveal a statistically significant difference between the PEP scores of human and other *Chordata* peptides (p-value 8.78e-01) - indicating that the *Chordata* peptide set is likely to contain true positive non-canonical human peptides. Of the 158 non-human *Chordata* peptides identified, 40 have previously been predicted from whole-genome sequencing efforts of the *SH-SY5Y* cell-line, by comparing sequences to the published 6-frame genome translation database (Krishna *et al.*, 2014), with 9 of these peptides identified at the proteomics level by the same authors. Thirteen of the polymorphisms identified by **MetaNovo** corresponded to **UniProt** human natural variant annotations. *See Supplementary Table 4.*

Peptide *VDVETPDINIEGSEGK* is a potentially novel peptide identification not previously reported, assigned to the *Bovidae* taxon by **UniPept**. This peptide shares 100.0 % sequence identity to AHNAK nucleoprotein in *Ovis aries* and *Bos taurus*, but only has 87.5% sequence identity to *Neuroblast differentiation-associated protein AHNAK* in *Homo sapiens*, with 2 adjacent amino acid differences compared to the canonical human protein. The identification of this peptide by the **MetaNovo/UniProt** pipeline was thus made possible by the existence of the orthologous sequences in the **UniProt** database, which made a sequence tag match and consequent database inclusion possible. It is plausible that whilst the variant positions do not correspond to annotated sequence variations in *Homo sapiens*, this peptide identification may instead be the result of accumulated mutations in the heavily passaged, transformed cell line. Another peptide, *AQEALLQLSQALSLMETVK*, was assigned by **UniPept** to *Mammalia* lowest common ancestor group, sharing 100.0 % sequence identity to *Perilipin (Macaca mulatta)*, but 94.7% identity to the same protein in *Homo sapiens* (a single amino acid difference). However, the sequence had previously been predicted from the whole genome sequence of *SH-SY5Y* and also been identified at the MS/MS level (Krishna *et al.*, 2014) , and reported in dbSNP (dbSNP:rs9973235). Thus, a high level of confidence can be placed in this identification as a true human variant.

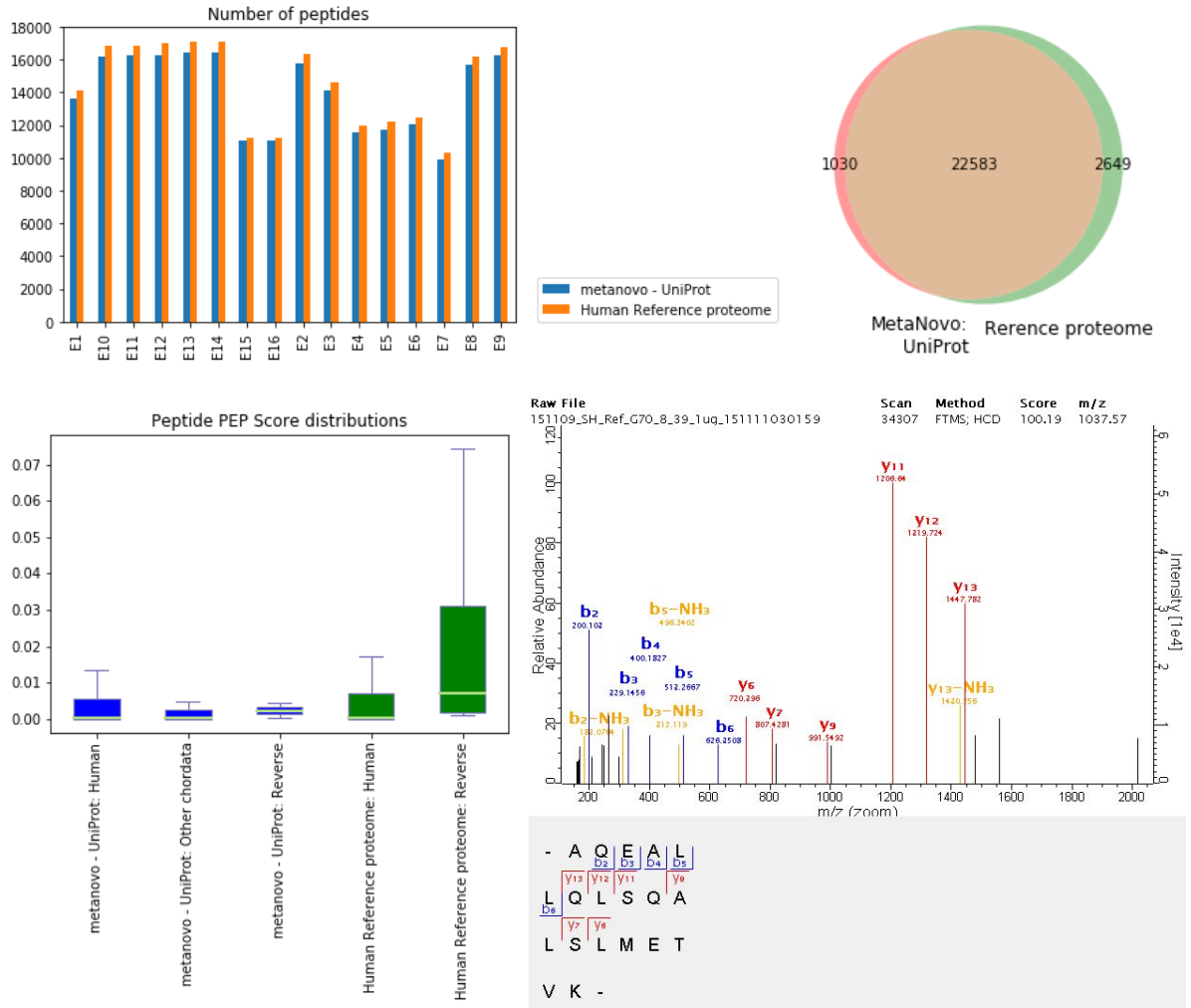

Figure 2. **A.** Bar chart comparing the peptide identification results of **MetaNovo** with the human Reference proteome on human neuroblastoma cell-line samples. Comparable identification rates are evident. **B.** Venn diagram illustrating peptide sequence identification overlap between **MetaNovo** and the human Reference proteome - 85.99 % of identified sequences were shared in common. **C.** PEP score distributions of peptides from the human Reference proteome compared to other *Chordata* were not significantly different, indicating novel human sequences matching orthologous proteins in other species. The boxes extend from the lower to the upper quartile, and the whiskers represent 1.5 times the interquartile range (IQR) below and above the first and third quartiles, respectively. **D.** An identified non-canonical peptide sequence from **MetaNovo** that had previously been identified in a proteogenomics study on the same cell-line. The peptide corresponds to a non-synonymous polymorphism in protein *Perilipin-3* (dbSNP:rs9973235). The **UniPept** lowest common ancestor assignment for this peptide was *Mammalia*, and the peptide shares 100.0% sequence identity with an orthologous protein G7NMI2_MACMU in *Macaca mulatta* (Rhesus macaque) from which it was identified in the **MetaNovo** database.

MetaNovo identifies temporal patterns of diet, disease and microbial diversity in an adult gut microbiome.

Peptide and protein identifications. MetaNovo/UniProt was used to analyze 8 adult stool samples obtained from a single individual at 2 different timepoints, one month apart and corresponding to bouts of health and ill-health. We identified 29724 and 6360 target peptides and proteins, respectively, with 18.92 % of MS/MS identified. 255 of the identified peptides corresponded with a known contaminant in the **MaxQuant** database, and 2544 peptides could be mapped exactly to human Reference proteins. Hierarchical clustering of the normalized and imputed peptide intensities across samples showed clear differences between the two sample collection windows that were 1 month apart and which corresponded to un-diagnosed ill-health and subsequent remission. *See Supplementary Figure 1.*

Alpha diversity analysis. Spectral counts by **UniPept** lowest common ancestor (LCA) assignments for each sample were used to obtain alpha diversity measures using *Shannon's* entropy. The first time point had a lower median entropy measure (2.51) than the second time-point (2.62) indicating a lower diversity of identified taxa. *Kruskal-Wallis* analysis of variance showed a significant difference between the diversity measures of the two groups (p-value 0.021). *See Supplementary Table 5.*

Relative proportions of identified phyla identified. Spectral counts were aggregated to phylum level to obtain relative proportions across all samples. The most abundant phylum identified was Firmicutes (38.66 %), followed by Chordata (36.16 %), Bacteroidetes (17.21 %), Streptophyta (3.45 %), Actinobacteria (1.98 %), Proteobacteria (1.55 %), Ascomycota (0.32 %), Basidiomycota (0.24 %) and Gemmatimonadetes (0.11 %). *See Supplementary Table 6.*

Presence-absence analysis indicates pathogenic organisms uniquely identified in the first time point. Spectra belonging to 227 organisms by UniProt fasta header annotation were uniquely identified in the first (diseased) time point. The most abundant, by summed MS/MS count, were *Aspergillus niger* (170 scans), the pathogenic helminth *Trichinella papuae* (45 scans), *Anacardium occidentale* (31 scans), *Clostridium sp. CAG:306* (24 scans) and *Daucus carota* (22 scans). Of the 149 organisms only identified in the second time point, the most abundant were *Parabacteroides distasonis str. 3776 Po2 i* (61 scans), *Parabacteroides distasonis str. 3999B T(B) 4* (50 scans), *Eggerthella sp. CAG:298* (38 scans), *Glycine max* (29 scans), and *Romboutsia ilealis* (18 scans). *See Supplementary Table 7.*

Differential abundance analysis using UniPept *pept2lca*. Using **UniPept** *pept2lca*, 374 distinct lowest common ancestor IDs were identified. Using summed precursor ion intensities with LIMMA moderated t-test, 52 were significantly different (q-value < 0.05) between the timepoints, with the top 5 taxa all less abundant in the second (well) time point. The top 5 differential lowest common ancestor groups between the first and second time points were *Hyphomonas* (log2FC -6.067, adj. p-value 0.0004), *Trichinella papuae* (log2FC -7.329, adj. p-value 0.0004), *Hypoxylon* (Iog2FC -8.859, adj. p-value 0.0005), *Aspergillus* (log2FC -6.033, adj. p-value 0.0007), and *Blautia obeum* (log2FC -5.279, adj. p-value 0.001). Other significantly different taxa include animal components *Bos*, *Bovinae, Ovis aries,* and *Pecora* that were more abundant in the second time point, and plant components *Malvaceae, Musa acuminata,* and *Cocosae* that were more abundant in the first. *Firmicutes* such as *Roseburia intestinalis* were also more abundant in the first time point, likely a response to an increased proportion of complex carbohydrates in the gut. *See Figure 3, Supplementary Table 8.*

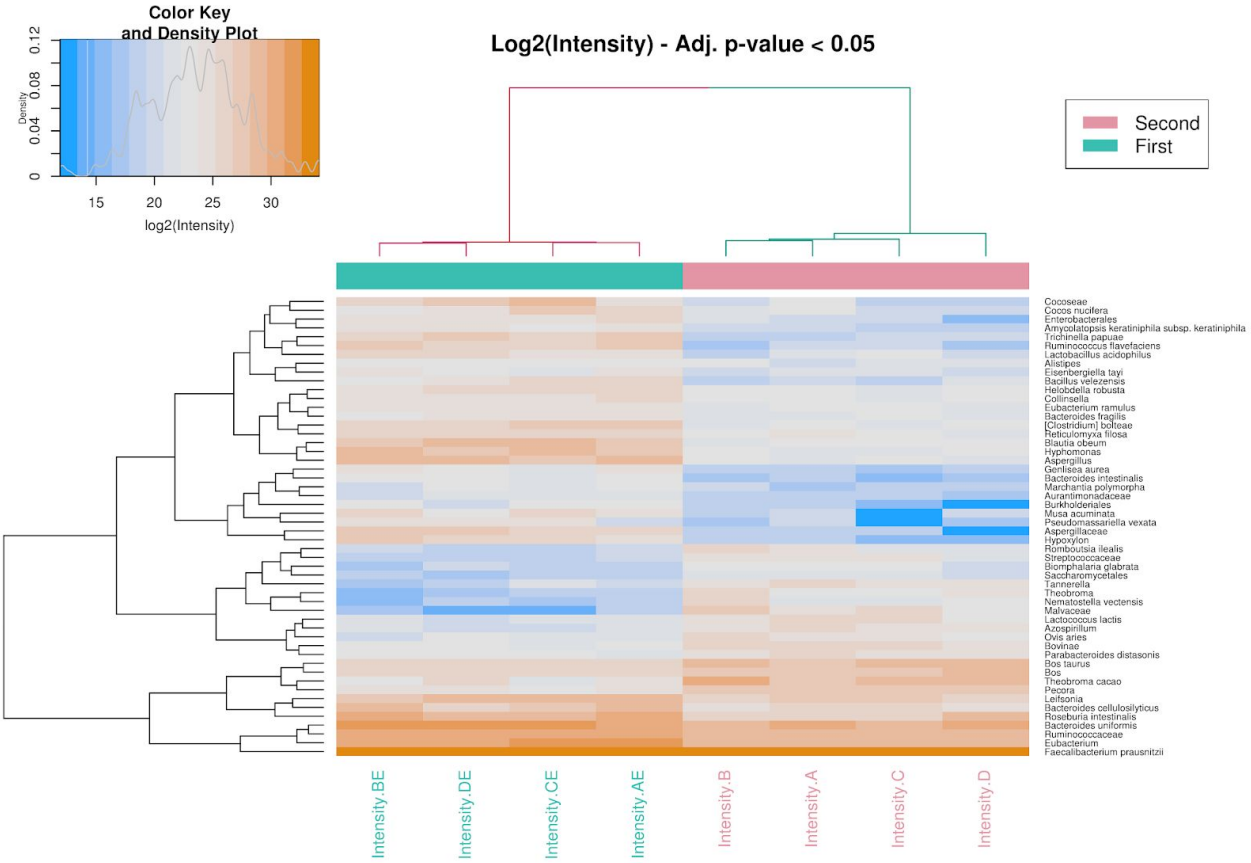

Figure 3: Heatmap of normalized precursor ion intensity of identified peptides from 8 faecal samples. Hierarchical clustering showed a clear grouping by time points, using the summed precursor ion intensity for each lowest common ancestor identified using the *UniPept pept2lca* tool. Normalized peptide intensities were aggregated by taxon, and the log2 transformed summed precursor ion intensity calculated per replicate and taxon. The heat map above shows the lowest common ancestors that are differentially abundant between replicates with a q-value less than 0.05 after performing a *limma* moderated t-test. Statistically significant differences in diet and microbial composition are seen between the two comparison groups, corresponding to known differences in diet and clinical symptoms - fatigue, weight-loss and myalgia accompanied by poor appetite and a predominantly vegetarian diet in time point 1, and clinical resolution and a mixed diet in time point 2. For example, *Bos taurus, Pecora, Ovis aries* assignments are more abundant in time point 2, while plant taxa such as *Cocosae,* and *Firmicutes* such as *Roseburia intestinalis*, are more abundant in time point 1. *Theobroma cacao* (chocolate) is more abundant in time point 2, while the pathogenic helminth *Trichinella papuae* and the fungal genus *Aspergillus* are more abundant in time point 1.

Posterior error probability scores. The PEP scores of human, contaminant, reverse and non-human peptides were compared using *Kruskal-Wallis* analysis of variance test followed by *Dunn's test* post-hoc analysis. A significant difference was found between groups (p-value 5.23e-58). Post-hoc, a significant difference (p-value 4.076e-49) was found between human and non-human peptides with as well as between non-human peptides and reverse hits - an indication that the proportion of false positives in the non-human peptide set is higher than human peptides, but still contains true positive identifications. *See Supplementary Table 9.*

Perturbed KEGG pathways. 6 KEGG terms were differentially perturbed based on *R gage* analysis of ion intensity of the constituent proteins identified using *InterProScan* (q-value < 0.05). *gage* uses log fold changes as "per-gene" statistics, and is based on a parametric gene randomization procedure that allows the identification of gene sets that are perturbed in one or both directions (Luo *et al.*, 2018). *Nitrogen metabolism* (*Metabolism; Energy metabolism*, q-value 0.009, set size 35), *Glyoxylate and dicarboxylate metabolism* (*Metabolism; Carbohydrate metabolism*, q-value 0.013, set size 72), *Arginine biosynthesis* (*Metabolism; Amino acid metabolism*, q-value 0.013, set size 72), *Purine metabolism* (*Metabolism; Nucleotide metabolism*, q-value 0.03, set size 110), *Pantothenate and CoA biosynthesis* (*Metabolism; Metabolism of cofactors and vitamins*, q-value 0.038, set size 49), *Pentose phosphate pathway* (*Metabolism; Carbohydrate metabolism*, q-value 0.04, set size 94) were all upwardly perturbed towards the first (diseased) time point. All perturbed KEGG pathways belonged to the *metabolism* class, with 2 *Carbohydrate metabolism* pathways identified - consistent with the differences in diet reported between time points, possibly due to increased carbohydrate and plant product consumption reported in the first time point. *See Supplementary Data 1.*

Differentially perturbed InterPro terms. 31 **InterPro** domains were enriched in the upwards direction towards the first time point. *IPR014710: RmlC-like jelly roll fold* (q-value 0.0, set size 104), *IPR011051: RmlC-like cupin domain superfamily* (q-value 0.0, set size 103), *IPR022379: 11-S seed storage protein, conserved site* (q-value 0.0, set size 36), *IPR016140: Bifunctional inhibitor/plant lipid transfer protein/seed storage helical domain* (q-value 0.0, set size 34) and *IPR036312: Bifunctional inhibitor/plant lipid transfer protein/seed storage helical domain superfamily* (q-value 0.0, set size 36) describe the top 5

InterPro terms, three of which are related to seed storage - consistent with a self-reported increase in plant products consumed during the first time point. *See Supplementary Data 1.*

Fifty five **InterPro** terms were dysregulated towards the second time point with q-value < 0.05. *IPR007110: Immunoglobulin-like domain* (q-value 0.0, set size 214), *IPR036179: Immunoglobulin-like domain superfamily* (q-value 0.0, set size 224), *IPR013106: Immunoglobulin V-set domain* (q-value 0.0, set size 203), *IPR003599: Immunoglobulin subtype* (q-value 0.0, set size 174), and *IPR013783: Immunoglobulin-like fold* (q-value 0.0, set size 311) were the top five dysregulated **InterPro** terms, all related to immune function. Changes in gut microbiota composition can cause either a pathological or beneficial outcome mediated by the regulation of CD4+ T cell subtypes induced by the gut microbiota (Wu and Wu, 2012). It has been reported that hormonal and immune responses to carbohydrate ingestion are associated with a decrease in perturbations in blood immune cell counts, and lower inflammatory cytokine responses (Nieman, 1998). It can be speculated that long term differences in carbohydrate composition in the human diet may have an influence on the host immune system. The identification of differentially abundant helminth peptides in the first time point are a clinically significant factor that may explain the differences in immunoglobulin related pathways. Intestinal parasites can lead to a broad range of effects on the host immune system, often leading to a downregulation in the response of peripheral T cells to parasite antigens, as well as a broad suppression in responsiveness to antigens and allergens - affecting both innate and adaptive immune response pathways (McSorley and Maizels, 2012). *See Supplementary Data 1.*

Gut pathogen detection using mass spectrometry. Seven nematoda peptides where identified, of which 5 peptides were assigned to the genus *Trichinella*. Peptide *SELLLGVANYR* (PEP score 4.535300e-07) was significantly more abundant in the first time point (q-value 0.0028). BLAST analysis yielded a match to *Vicilin-like antimicrobial peptides 2-1* of *Trichinella papuae* with 100.0% identity. The presence of this pathogen and associated peptide corresponds with the immune dysregulation identified with gene set enrichment of **InterPro** terms, and the decreased taxonomic diversity identified in the second time point with *Shannon's* entropy diversity metric. *See Figure 4.*

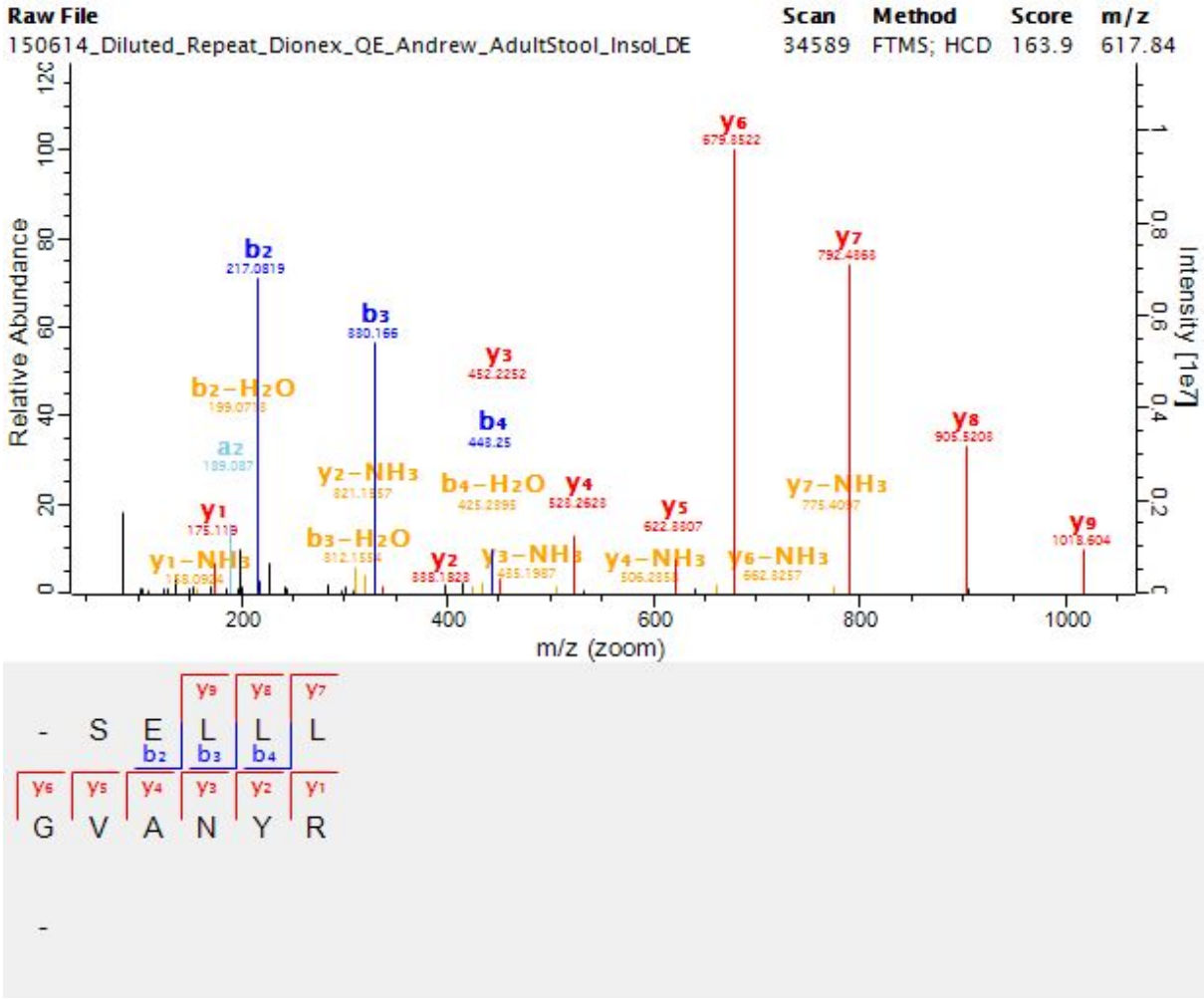

Figure 4. *SELLLGVANYR* (PEP score 4.535300e-07) was significantly more abundant in the second time point (adj.p-value 0.0028). BLAST analysis yielded a match to *Vicilin-like antimicrobial peptides 2-1* of *Trichinella papuae* with 100.0 % identity. The presence of this pathogen and associated peptide corresponds with the immune dysregulation identified with gene set enrichment of **InterPro** terms.

Discussion

We have shown that a *de novo* sequence tag search and probabilistic ranking methodology applied to the entire **UniProt** knowledge base can yield a representative set of proteins for sensitive protein and peptide identification using a target-decoy database search of a given set of samples.

Analyzing samples previously analyzed with **MetaPro-IQ** and a matched metagenome database and integrated gut gene catalog database, we identified comparable numbers of peptides and proteins, with a similar taxonomic distribution. However, we identified a higher component of host (*Homo sapiens*) and dietary components such as *Bos taurus*. Thus, we show that the characterization of complex metaproteomic samples is feasible even when matched metagenome or specialized metaproteomic databases are not available.

Less than 1% difference in MS/MS percentage assignment between the **MetaNovo** database derived from **UniProt** and the human Reference proteome was found in the analysis of the human neuroblastoma cell line data**.** Further, 85.99% of identified peptides across the two runs were shared in common, indicating a highly sensitive and specific peptide identification level using the **MetaNovo** database - a comparable level of sensitivity to the **UniProt** human Reference proteome. The less marked overlap at the protein id level could be due to the inclusion of orthologous proteins from related organisms to human, yet still yielding a representative database at the tryptic peptide sequence level. Thus, it is essential to include the peptide-level taxonomic analysis of **UniPept** to characterize peptides from closely related isoforms. For instance, peptides identified from a *Gorilla gorilla* protein may be shared with a closely related *Homo sapiens* ortholog, and be classified to a taxon that includes both *Homo* and *Gorilla* by the **UniPept** *pept2lca* tool, such as *Primates*. Further, we report the identification of variant peptide sequences in mass spectrometry proteomics without the use of a matched genomics database, such as a variant peptide identification in Human *perilipin 3* which was made possible by the inclusion of an orthologous protein from *Macaca mulatta* (Rhesus macaque) in the **MetaNovo** database.

Furthermore, we have demonstrated that MetaNovo allows quantification of taxa that would not have been identified using 16S rRNA sequencing or by metaproteomic analyses that are dependent on prior knowledge or expectation of likely sample composition. Thus, **MetaNovo** enables simultaneous correlated characterization of the microbiome (including bacteria, virus, fungal and eukaryotes) and the proximal host proteome, whilst also allowing identification of specific microbial proteins that are plausible effectors mediating the interface between microbiome and host, as well as the dietary components in the adult gut. From the adult stool samples analyzed, dietary changes between two time points 1 month apart may be associated with changes in microbial diversity. In particular, an increase in the proportion of plant-based foods in the diet may be associated with an increase in *Firmicute* abundance. Notably, the identification of a *helminth* pathogen in the first time point corresponding with decreased alpha diversity is consistent with the well-known association between pathology and decreased gut microbial diversity. Dysregulation of immune-related **InterPro** terms in the downwards direction in the first time point could be a result of the immune mediating effect of a parasitic helminth infection. Presence-absence analysis indicating that the *Trichinella* peptides were only identified in the first time point indicates that these peptides are not shared with normally occurring gut microbiota. The increased abundance of *Aspergillus* peptides in the first time point is interesting in the light of reports of fungal dysbiosis in inflammatory gut pathology (Liguori *et al.*, 2015).

Assignments by **UniPept** of peptides to organisms that are not likely to be present in either the adult gut metaproteome or in neuroblastoma samples described above, but are in many cases closely related to *Homo sapiens* (such as *Pan troglodytes*, or *Gorilla gorilla*), may arise purely due to false positive identifications at the given FDR threshold. However, another source of error may be inaccuracies at the database export step, such as due to a sequence tag of an MS/MS spectrum arising from non-tryptic degradation in the sample being incorrectly assigned to a fully tryptic sequence in the **UniProt** database, causing a protein sequence that is not actually present in the sample to be included in the **MetaNovo** database - and then identified purely on the basis of non-tryptic fragments present in the sample. However, biologically interesting events such as single amino acid polymorphisms or novel gene isoforms not included in the human Reference proteome may lead to the inclusion of orthologous proteins of closely related organisms in the **MetaNovo** database, allowing for the

identification of novel polymorphisms in clinical or metaproteomic datasets - indicated by similar PEP score distributions between human peptides and other *Chordata* peptides identified in the human *SH-SY5Y* samples, and validation of 40 identified non-canonical peptides by genome sequencing efforts in the same cell line, with 9 variant peptides also previously identified at the MS/MS level.

Conclusion

MetaNovo makes characterizing complex mass spectrometry data with an extremely large, scalable, unbiased search space feasible, allowing for novel peptide and polymorphism identification, even when genomic information is not available, and improving the accuracy and affordability of current clinical proteomic and metaproteomic analysis methods by leveraging the rapidly growing number of available sequences from genome and metagenome sequencing efforts. By tailoring a database for the samples to be analyzed in an unbiased manner, **MetaNovo** allow proteomic identification to be carried out, unlimited by prior expectation. We have shown that using **MetaNovo** a sensitive characterization of the gut metaproteome allows for a simultaneous snapshot of the metabolic processes at play in both the host and the microbiome. Finally, the ability to search an extremely large database such as **UniProt** provides the potential to identify pathogens and biomarkers that may be missed using limited or hand-selected databases, as illustrated by the detection of *helminth* peptides in an adult stool sample, as well as of rare clinically relevant sequence polymorphisms in tumour-derived samples.

By estimating taxonomic and peptide level information on microbiome samples directly from tandem mass spectrometry data, in combination with the entire UniProt database, **MetaNovo** thus enables simultaneous identification of human, bacterial, helminth, fungal, viral and other eukaryotic proteins in a sample, thus potentially now allowing clinically-important correlations to be drawn between changes in microbial protein abundance and change in the host proteome, based on a single analysis. Using the freely-available **MetaNovo** algorithm, researchers will thus be able to focus on downstream analysis to extract biological meaning at the level of expressed proteomes in microbiome samples, whilst minimising reliance on prior knowledge, or external information such as whole genome sequencing.

Author contributions

MGP - design of pipeline; Bioinformatic data analysis; wrote manuscript

AJMN – Bioinformatic data analysis; experimental data generation

SF - Experimental data generation

SG - Experimental data generation

JW - Experimental data generation

DT - DirecTag configuration; Taxonomic weighting strategy

NM - Design of overall study; design of pipeline; wrote manuscript; corresponding author

JB – Design of overall study; design of pipeline; wrote manuscript; corresponding author

Acknowledgments

MGP would like to thank the NRF for an MSc grant (NRF BFG 93665) and Professor Nicola Mulder for a PhD grant. DLT was supported by the South African Tuberculosis Bioinformatics Initiative (SATBBI), a Strategic Health Innovation Partnership grant from the South African Medical Research Council and South African Department of Science and Technology. JMB thanks the NRF for a South African Research Chair grant.

References

1. Ochoa-Hueso, R. (2017). Global Change and the Soil Microbiome: A Human-Health Perspective. *Frontiers in Ecology and Evolution*, *5*, 71. https://doi.org/10.3389/fevo.2017.00071
2. Blaser, M. J., Cardon, Z. G., Cho, M. K., Dangl, J. L., Donohue, T. J., Green, J. L., … Brodie, E. L. (2016). Toward a Predictive Understanding of Earth's Microbiomes to Address 21st Century Challenges. *MBio*, *7*(3), e00714-16. https://doi.org/10.1128/mBio.00714-16
3. Castellana, N., & Bafna, V. (2010). Proteogenomics to discover the full coding content of genomes: a computational perspective. *Journal of Proteomics*, *73*(11), 2124–2135. https://doi.org/10.1016/j.jprot.2010.06.007
4. Heyer, R., Schallert, K., Zoun, R., Becher, B., Saake, G., & Benndorf, D. (2017). Challenges and perspectives of metaproteomic data analysis. Journal of Biotechnology, 261, 24–36. https://doi.org/10.1016/j.jbiotec.2017.06.1201

5. Li, J., Quinque, D., Horz, H.-P., Li, M., Rzhetskaya, M., Raff, J. A., … Stoneking, M. (2014). Comparative analysis of the human saliva microbiome from different climate zones: Alaska, Germany, and Africa. *BMC Microbiology*, *14*(1), 316. https://doi.org/10.1186/s12866-014-0316-1
6. Blackburn, J. M., & Martens, L. (2016). The challenge of metaproteomic analysis in human samples. Expert Review of Proteomics, 13(2), 135–138. https://doi.org/10.1586/14789450.2016.1135058
7. Tang, H., Li, S., & Ye, Y. (2016). A Graph-Centric Approach for Metagenome-Guided Peptide and Protein Identification in Metaproteomics. PLOS Computational Biology, 12(12), e1005224. https://doi.org/10.1371/journal.pcbi.1005224
8. Zhang, X., Ning, Z., Mayne, J., Moore, J. I., Li, J., Butcher, J., … Figeys, D. (2016). MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. Microbiome, 4(1), 31. https://doi.org/10.1186/s40168-016-0176-z
9. Frank, A. M., Savitski, M. M., Nielsen, M. L., Zubarev, R. A., & Pevzner, P. A. (2007). De novo peptide sequencing and identification with precision mass spectrometry. *Journal of Proteome Research*, *6*(1), 114–123. https://doi.org/10.1021/pr060271u
10. Vizcaíno, J. A., Côté, R. G., Csordas, A., Dianes, J. A., Fabregat, A., Foster, J. M., … Hermjakob, H. (2013). The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Research*, *41*(D1), D1063–D1069. https://doi.org/10.1093/nar/gks1262
11. Muth, T., Weilnböck, L., Rapp, E., Huber, C. G., Martens, L., Vaudel, M., & Barsnes, H. (2014). DeNovoGUI: An Open Source Graphical User Interface for de Novo Sequencing of Tandem Mass Spectra. Journal of Proteome Research, 13(2), 1143–1146. https://doi.org/10.1021/pr4008078
12. Tabb, D. L., Ma, Z.-Q., Martin, D. B., Ham, A.-J. L., & Chambers, M. C. (2008). DirecTag: Accurate Sequence Tags from Peptide MS/MS through Statistical Scoring. Journal of Proteome Research, 7(9), 3838–3846. https://doi.org/10.1021/pr800154p
13. Kopczynski, D., Barsnes, H., Njølstad, P. R., Sickmann, A., Vaudel, M., & Ahrends, R. (2017). PeptideMapper: efficient and versatile amino acid sequence and tag

mapping. Bioinformatics, 33(13), 2042–2044. https://doi.org/10.1093/bioinformatics/btx122

14. Barsnes, H., Vaudel, M., Colaert, N., Helsens, K., Sickmann, A., Berven, F. S., & Martens, L. (2011). compomics-utilities: an open-source Java library for computational proteomics. BMC Bioinformatics, 12(1), 70. https://doi.org/10.1186/1471-2105-12-70
15. Tange, O. (2011). GNU Parallel: The Command-Line Power Tool | USENIX. Retrieved November 6, 2017, from https://www.usenix.org/publications/login/february-2011-volume-36-number-1/gnu-parallel-command-line-power-tool
16. Zybailov, B. L., Florens, L., & Washburn, M. P. (2007). Quantitative shotgun proteomics using a protease with broad specificity and normalized spectral abundance factors. Molecular BioSystems, 3(5), 354. https://doi.org/10.1039/b701483j
17. Krishna, A., Biryukov, M., Trefois, C., Antony, P. M. A., Hussong, R., Lin, J., … May, P. (2014). Systems genomics evaluation of the SH-SY5Y neuroblastoma cell line as a model for Parkinson's disease. BMC Genomics, 15(1), 1154. https://doi.org/10.1186/1471-2164-15-1154
18. Wu, H.-J., & Wu, E. (2012). The role of gut microbiota in immune homeostasis and autoimmunity. Gut Microbes, 3(1), 4–14. https://doi.org/10.4161/gmic.19320
19. Nieman, D. C. (1998). Influence of carbohydrate on the immune response to intensive, prolonged exercise. Exercise Immunology Review, 4, 64–76. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9644095
20. McSorley, H. J., & Maizels, R. M. (2012). Helminth infections and host immune regulation. Clinical Microbiology Reviews, 25(4), 585–608. https://doi.org/10.1128/CMR.05040-11
21. Liguori, G., Lamas, B., Richard, M. L., Brandi, G., da Costa, G., Hoffmann, T. W., … Sokol, H. (2016). Fungal Dysbiosis in Mucosa-associated Microbiota of Crohn's Disease Patients. *Journal of Crohn's & Colitis*, *10*(3), 296–305. https://doi.org/10.1093/ecco-jcc/jjv209

Addendum A: Adult Stool Samples

Proteomic sample preparation. Eight stool samples were collected from the patient under sterile conditions, transported on ice on the day and stored at -20°C. Four of these samples were collected over a week and the remaining over the same period a month later. 1g of each sample was pulverised in the presence of liquid nitrogen using a mortar and pestle. Pulverised sample was transferred into 5 mL of isopropanol and further homogenised by ultrasonication for 5 mins at 30W energy output using an ultrasonic cell disruptor (VirSonic 100, VirTis). Samples were precipitated overnight at 4°C. Precipitates were collected using centrifugation and resuspended in 100 uL. Proteinaceous material was further extracted using 4:1 chloroform: methanol and precipitated with 2 vols of methanol. Precipitates were resuspended in 500 *uL* of RIPA buffer (150 mM NaCl, 50 mM triethylammonium bicarbonate, 1% SDS, 0.5% DCA, pH 8) and boiled at 95ºC for 5 mins. Protein concentration was estimated using the bicinchoninic assay, using BSA as a reference standard. An estimated 200 ug of proteinaceous material was used for both FASP and in-solution protein sample preparation methods.

For the FASP procedure, 200 ug protein was transferred into a 500 µL of Ultracel 30 000 MWCO centrifugal unit (Amicon Ultra, Merck). Protein extracts were buffer-exchanged with three rounds of 200 µL of UT buffer (8 M urea, 0.1 M Tris, pH 8.5). Cysteine bonds were reduced with 100 mM DTT for 1 h at room temperature. Alkylation of reduced cysteine bonds was carried out by incubation in the dark for 20 min in 200 µL of UT buffer containing 0.05 M iodoacetamide (Sigma). Two 200 µL UT buffer exchanges were used to remove the alkylating agent, followed by three buffer exchanges with 100 µL of ABC buffer (50 mM ammonium bicarbonate buffer, pH 8). 40 µL of ABC buffer containing 1:50 ratio of sequencing-grade modified trypsin (Promega) to amount of protein was added to the retentate. Proteolysis was carried out at 37 °C for 18 h in a wet chamber. Three rounds of 40 µL of ABC buffer was used to elute the peptide-rich solution.

For in-solution sample preparation, 200 ug sample was resuspended at a protein concentration of 0.2 ug/mL in denaturation buffer (100 mM Tris-HCL, 6 M Guanidium hydrochloride, 2 M Thiourea, pH 8). Reduction and alkylation was carried out for 1 hour each, using 1 mM DTT and 2.5 mM IAA. 4 vols of ABC was used to dilute the denaturation buffer prior to the addition of Trypsin (same amount as before). Digestion was carried out at 25°C for 18 hrs on a rotator.

Peptide-rich eluate obtained for either FASP or in-solution tryptic digests was desalted using a homemade stage tip containing Empore Octadecyl C18 solid-phase extraction disk (Supelco). Activation, equilibration and peptide wash and elution were all carried out using centrifugation at 5000 g for 5 min. Activation and equilibration of the C18 disk was carried out using three rinses with 80% acetonitrile (ACN), followed by three rinses with 2% ACN, respectively. Peptide-rich solution was loaded onto the disk and centrifuged. Desalting was carried out using three washes of 2% ACN, followed by three washes of 2% ACN containing 0.1% formic acid (Sigma). Elution of desalted peptides into glass capillary tubes was carried out using three rounds of 100 µL of 60% ACN, 0.1% formic acid. Peptides were dried in a vacuum and resuspended in 2% ACN, 0.1% formic acid at 250 ng/µL.

LC/MS/MS analysis. Liquid chromatography separation was done with a home-packed 100 µM ID × 20 mm precolumn connected to a 75 µM × 200 mm analytical column packed with C18 Luna beads (5 µm diameter, 100 Å pore size; Phenomenex 04A-5452). The columns were connected to an Ultimate 3500 RS nano UPLC system (Dionex). 1 µg of desalted peptides was loaded onto the column with starting mobile phase of 2% ACN, 0.1% formic acid. Peptides were eluted with the following gradient of 10 min at 2% ACN, increase to 25% ACN for 115 min, to 35% ACN over 5 min, to 80% ACN over 5 min, followed by a column wash of 85% for 20 min. The flow rate was constant at 300 µL/min. Typical back pressure values during separation were <350 bar.

Mass spectra were acquired with a Q Exactive mass spectrometer in a data-dependent manner, with automatic switching between MS and MS/MS scans using a top-10 method. MS spectra were acquired at a resolution of 70 000 with a target value of 3×10^6 ions or a

maximum integration time of 250 ms. The scan range was limited from 300 to 1750 *m*/*z*. Peptide fragmentation was performed via higher-energy collision dissociation (HCD) with the energy set at 25 NCE. Intensity threshold for ions selection was set at 1.7 *e*4 with charge exclusion of $z = 1$ and $z > 5$. The MS/MS spectra were acquired at a resolution of 17 500, with a target value of 2 × 105 ions or a maximum integration time of 120 ms and the isolation window was set at 4.0 *m*/*z*.