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

Matthys G Potgieter[1,2], Andrew JM Nel[2,] David L. Tabb[2-5], Suereta Fortuin[2], Shaun Garnett[2], Jerome M. Wendoh[6], Jonathan M Blackburn[2], Nicola J Mulder[1]

1. Computational Biology Division, Department of Integrative Biomedical Sciences, IDM, University of Cape Town Cape Town, South Africa.
2. Division of Chemical and Systems Biology, Department of Integrative Biomedical Sciences, IDM, University of Cape Town Cape Town, South Africa.
3. Division of Molecular Biology and Human Genetics, Department of Biomedical Sciences, Stellenbosch University, Cape Town, South Africa
4. Bioinformatics Unit, South African Tuberculosis Bioinformatics Initiative, Stellenbosch University, Cape Town, South Africa
5. The African Microbiome Institute, Stellenbosch University, Stellenbosch, South Africa
6. Division of Immunology, Department of Pathology, IDM, University of Cape Town Cape Town, South Africa.

Abstract

The characterization of complex mass spectrometry data obtained from metaproteomics or clinical studies presents unique challenges and potential insights in fields as diverse as the pathogenesis of human disease, the metabolic interactions of complex microbial ecosystems involved in agriculture, or climate change. However, accurate peptide identification requires representative sequence databases, which typically rely on prior knowledge or matched genome sequencing, and are often error-prone. We present a novel software pipeline to directly estimate the proteins and species present in complex mass spectrometry samples at the level of expressed proteomes, using *de novo* sequence tag matching and probabilistic optimization of very large sequence databases prior to target-decoy search. We validated our pipeline 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, and novel identifications. We showed that using the entire release of **UniProt** we were able to identify a similar taxonomic distribution compared to a matched metagenome database, with improved identifications of certain taxa. Using **MetaNovo** to analyze a set of single-organism human neuroblastoma cell-line samples (SH-SY5Y) against **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. The **MetaNovo** software is available from github[1] or can be run as a standalone Docker container available from the Docker Hub[2].

[1] https://github.com/uct-cbio/proteomics-pipelines
[2] 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 transmission of human diseases. Parallels have been drawn between the role of soil microbiota in the biosphere with 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 pervasive role in the complex determinants of human well-being, with potential disease implications should the delicate balance be disrupted.

Genome and transcriptome sequencing approaches allow for functional and taxonomic characterization of the genes and organisms involved in complex microbial communities, yet evidence for gene transcription does not always imply protein translation (Castellana and Bafna, 2010). On the other hand, metaproteomic and clinical proteomic approaches allow researchers to obtain a direct snapshot of all the expressed proteomes present in a sample, providing a direct window into the metabolic components of the complex pathways and organism interactions involved.

The ideal database for mass spectrometry peptide identification is both comprehensive and exclusive of absent proteins. However, metaproteome datasets offer unique challenges to sensitive MS/MS peptide identification. While matched metagenome sequencing is the gold standard to create sequence databases to search metaproteomics mass spectrometry data against, such data is not always available, and does not guarantee an accurate reflection of the proteomes present. Curated databases for a given microbiome or organism may not be comprehensive or account for all possible sequence polymorphisms or sample contaminants, while using the target-decoy approach to search comprehensive, yet very large, search spaces may lead to less sensitive MS/MS identification (Heyer *et al.*, 2017).

Further, identified peptides may belong to multiple homologous proteins, complicating protein inference and quantification. Challenges faced by adult gut metaproteomics, in particular, 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. 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).

Various tools have been developed to face 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 gene catalogue 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). However, scalable and robust pipelines need to be able to process the rapidly expanding amount of proteomics and genomics information available.

We present a database generation pipeline 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 FDR-controlled protein identification directly at the level of the expressed proteomes of the samples involved, and is 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. The pipeline can be obtained from GitHub[3] or downloaded as a standalone Docker container from the Docker Hub[4].

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

Materials and Methods

MetaNovo database generation workflow

Generation of *de novo* sequence tags. All *Thermo* raw 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. A default 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] Fragment and precursor tolerance of 0.02 Da was selected, with a fixed modification '*Carbamidomethylation of C'* and variable modifications '*Oxidation of M'* and *'Acetylation of protein N-term'*. The output of **DirecTag** was parsed with a custom **Python** script and all sequence tags for each MS/MS across replicates were stored in an **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 all available cores in parallel.

[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.

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.

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 MSMS 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 normalized to a maximum value of 1 to obtain a protein-level posterior probability based on estimated relative protein abundance across replicates. 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 MSMS scan id compared to the set of proteins above it's position in the list. The organism name was obtained from the **UniProt** FASTA headers OS entry, and summed NSAF values aggregated by organism and normalized to obtain an organism level posterior probability. Following this, the proteins in the original set were annotated with their respective calculated organism posterior probability. The original set of identified proteins is re-ranked based on the joint probability of the organism and protein probability values, facilitating the inclusion of low abundance proteins with a higher estimated relative abundance at the organism level. Following the second ranking step, the database is re-filtered to obtain a minimal subset of proteins that can explain all mapped *de novo* sequence tags.

MLI sample metagenomic and proteomics data. Metagenome and proteomics data of 8 MLI samples from adolescent volunteers obtained during colonoscopy were downloaded from PRIDE[6] 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**[7] 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.

---

[6] http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD003528
[7] 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 (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 filter. 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 by a 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. 16 Human neuroblastoma (*SH-SY5Y)* cell-line samples 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 central nervous system samples were searched against the human Reference proteome and **MetaNovo** database 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.

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

Protein extraction and LC MSMS analysis. Four volunteer adult stool samples were taken over a week from each of 2 timepoints 3 months apart, representing known differences in diet and clinical symptoms. The first time point samples were taken during a low-carbohydrate diet, while the second time point samples were taken during a self-reported normal diet. A tropical travel history was 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 is 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/UniProt comparable to a matched metagenome approach in human MLI samples

Protein and peptide identifications. The **MetaNovo/UniProt** pipeline resulted 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.*[8]

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/UniProt**. In total, the **MetaNovo/UniProt** analysis yielded 45 distinct phyla, while the **MetaProt-IQ/IGC** run yielded 14. At the species level, **MetaNovo/UniProt** yielded 405 distinct species, followed by **MetaPro-IQ/IGC** with 123 distinct species identified. *See Figure 1C.*

[8]An attempt by the author 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.

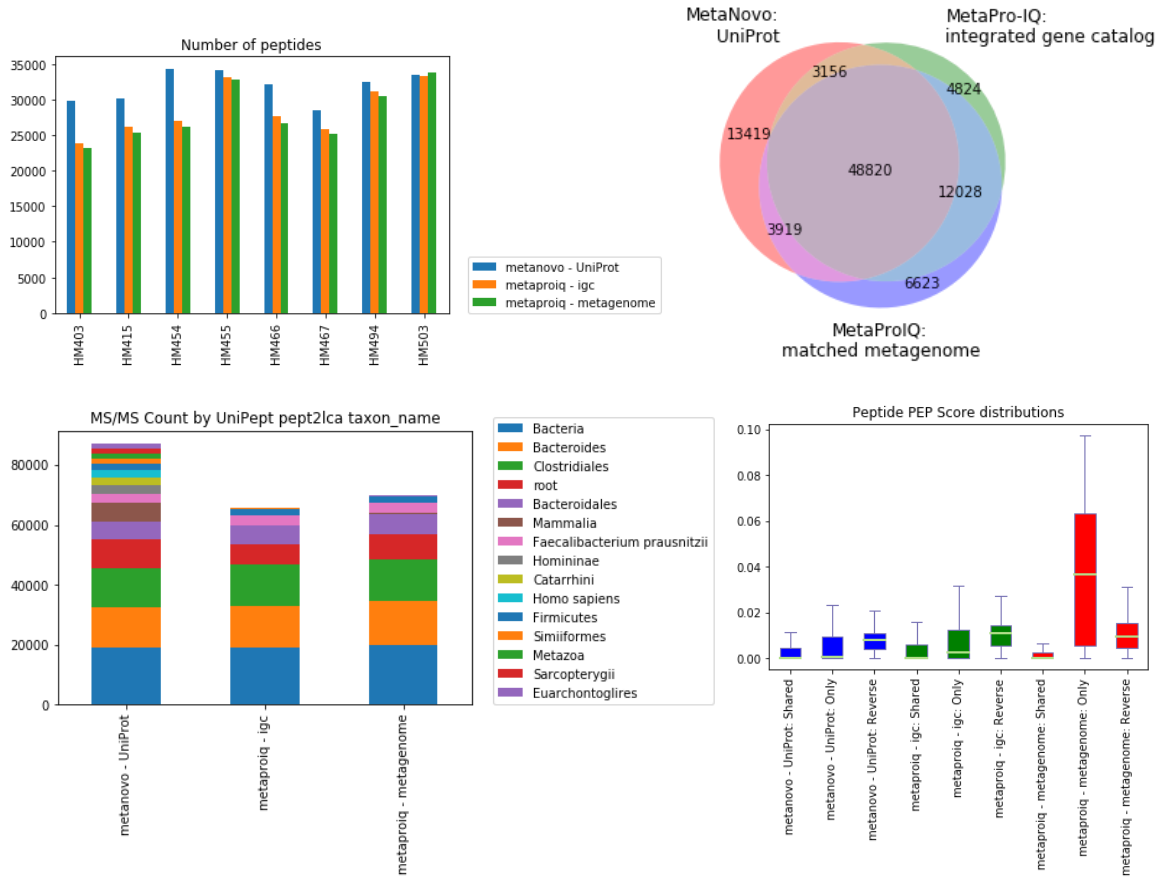

Figure 1. A. Barchart of peptide identifications. The identification rates of **MetaNovo/UniProt** are comparable to the previously published results of **MetaPro-IQ** using matched metagenome and integrated gene catalog databases. **B.** Venn diagram showing large overlap in identified sequences using different approaches, with the highest number of sequences identified using **MetaNovo/UniProt. C.** Spectral counts by *UniPept lowest common ancestor* showed similar taxonomic distributions obtained from different approaches. **D.** Peptides uniquely identified by **MetaNovo/UniProt** 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.

MetaNovo approach comparable to Reference proteome in human neuroblastoma cell-line analysis, novel polymorphisms detected

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/UniProt**, 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.*

Novel polymorphism identifications from orthologous sequences. Peptides identified using the **MetaNovo/UniProt** 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. Taxonomic analysis was used to narrow down the list to only 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 (Krishna *et al.*, 2014), with 9 of these peptides identified at the proteomics level by the same authors. *See Supplementary Table S1.*

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, even though the variant positions do not correspond to annotated sequence variations in *Homo sapiens*. Although this peptide identification may be a true variant in *Homo sapiens*, it may also be a *Bos* peptide that was included as a contaminant from bovine *Trypsin*, or even a false positive hit. It is reasonable that caution needs to be applied to identified sequences that may be caused by contamination during sample preparation.

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 MSMS level by the same authors (dbSNP:rs9973235). Thus, although contamination can not be excluded, a higher level of confidence can be placed in this identification as a true human variant.

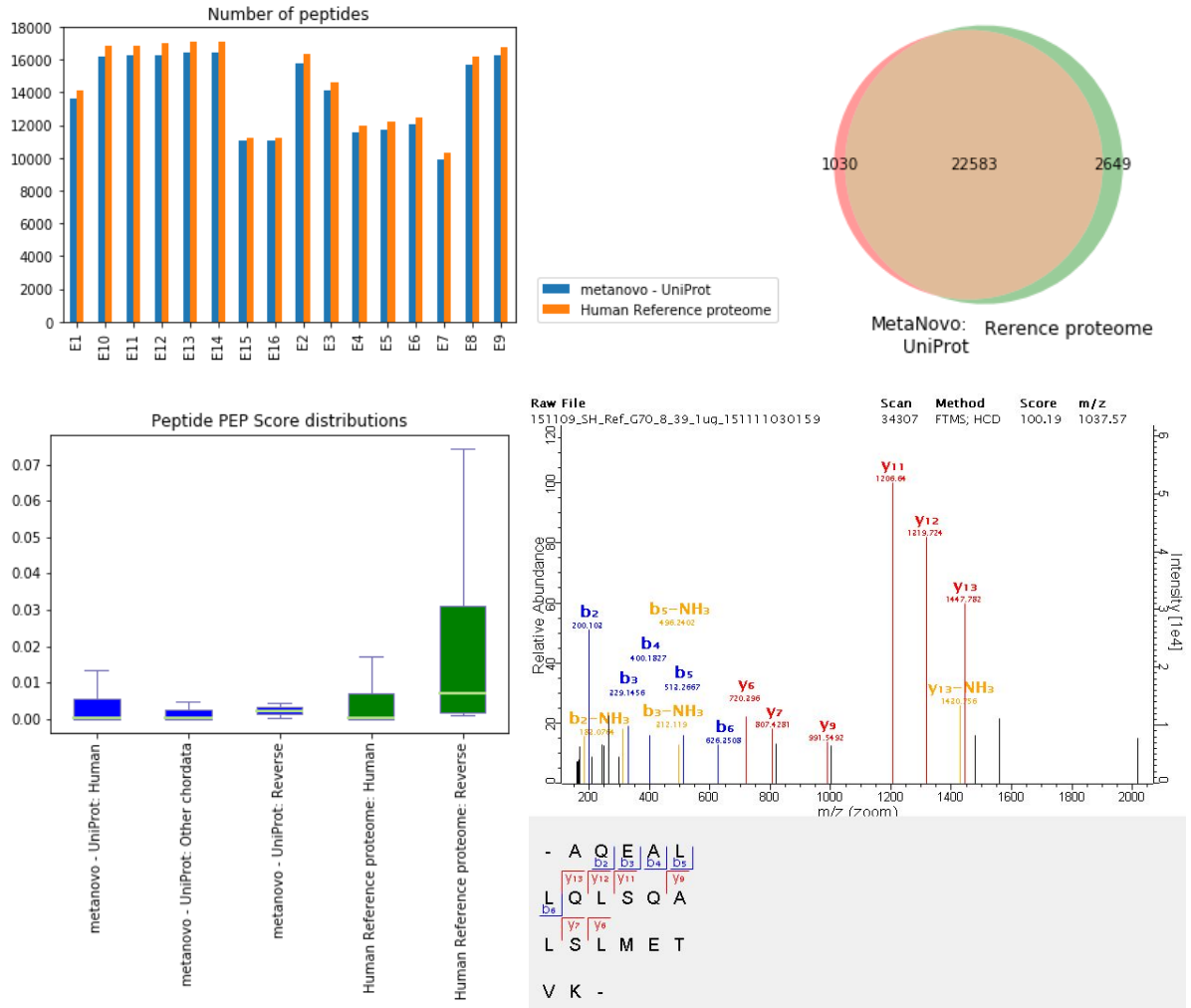

Figure 2. **A.** Bar chart comparing the peptide identification results of **MetaNovo/UniProt** with the human Reference proteome on human neuroblastoma cell-line samples. Comparable identification rates are evident. **B.** Venn diagram illustrating peptide sequence identification overlap between **MetaNovo/UniProt** 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. **D.** An identified non-canonical peptide sequence from **MetaNovo/UniProt** 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.

Using MetaNovo to identify temporal patterns of diet, disease and microbial diversity in the adult gut microbiome.

Peptide and protein identifications. Using **MetaNovo/UniProt** to analyze 8 adult stool samples from 2 different timepoints, 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 timepoints taken 3 months apart. *See Supplementary Figure S2.*

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 higher median entropy measure (2.62) than the second time-point (2.51) indicating a higher diversity of identified taxa. *Kruskal-Wallis* analysis of variance showed a significant difference between the diversity of the two groups (p-value 0.02). *See Supplementary Table S3.*

Differential abundance analysis using UniPept *pept2lca*. Using **UniPept** *pept2lca*, 373 distinct lowest common ancestor ids were identified. Using summed precursor ion intensities with LIMMA moderated t-test, 50 were significantly different (qvalue < 0.05) between the timepoints. The top 5 differential lowest common ancestor groups between the two first and second time points were *Trichinella papuae* (log2FC 8.30, adj. p-value 0.00014), *Ruminococcus flavefaciens* (log2FC 10.12, adj. p-value 0.00014), *Hyphomonas* (log2FC 6.067, adj. p-value 0.0002), *Hypoxylon* (Iog2FC 9.99, ad.p-value 0.0002) and *Aspergillus* (log2FC 5.80, adj. p-value 0.0006). Other significantly different taxa include animal components *Bos*, *Bovidae, Ovis aries,* and *Pecora* that were more abundant in the first time point, and plant components *Malvaceae, Musa* and *Cocosae* and *Arecaceae* that were more abundant in the second. *Firmicutes* such as *Roseburia intestinalis* were also more abundant in

the second time point, likely a response to an increased proportion of complex carbohydrates in the gut. *See Figure 3.*

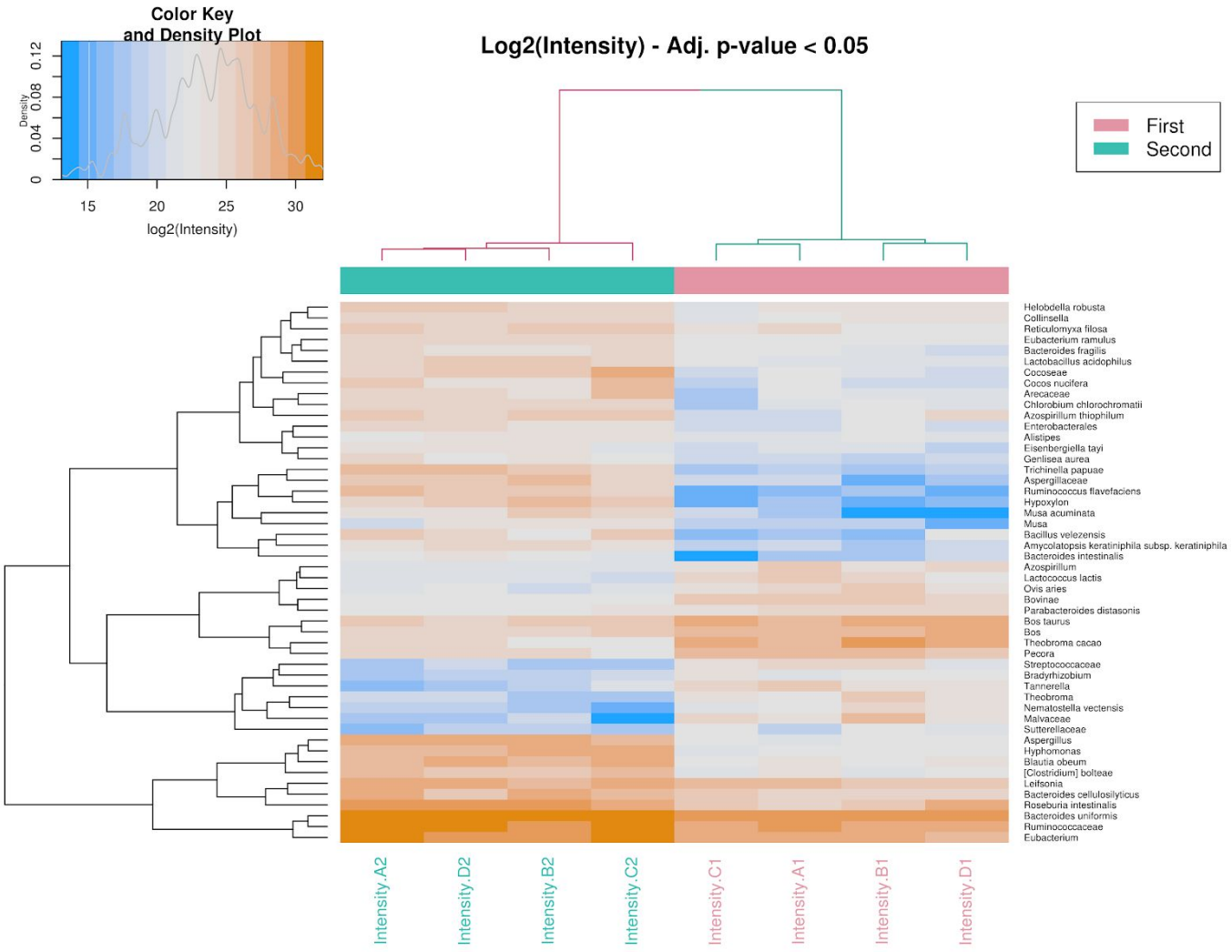

Figure 3: Heatmap of normalized precursor ion intensity of identified peptides from 8 faecal samples. Hierarchical clustering showed a clear grouping by timepoints. Mean precursor ion intensity for each lowest common ancestor identified using the *UniPept pept2lca* tool. At least two peptides per taxonomic assignment were required, and the log2 transformed summed precursor ion intensity calculated per replicate and taxon. The heatmap above shows 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 at the time - a low carbohydrate diet in time point 1, and a normal diet in time point 2. For example, *Bos taurus, Pecora, Ovis aries* assignments are more abundant in *time point 1*, while plant taxa such as *Cocosae,* and *Firmicutes* such as *Roseburia intestinalis*, are more abundant in *time point 2*. *Theobroma cacao* (chocolate) was 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 S4.*

Perturbed KEGG pathways. 5 KEGG terms were differentially perturbed based on *R* GAGE analysis of ion intensity of the constituent proteins identified using *InterProScan* (qvalue < 0.05). GAGE used 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 0.003, set size 34), *Glyoxylate and dicarboxylate metabolism* (*Metabolism; Carbohydrate metabolism*, q value 0.011, set size 71), *Arginine biosynthesis* (*Metabolism; Amino acid metabolism*, q value 0.015, set size 71), *Pentose phosphate pathway* (*Metabolism; Carbohydrate metabolism*, q value 0.026, set size 94), *Alanine, aspartate and glutamate metabolism* (*Metabolism; Amino acid metabolism*, q value 0.048, set size 73) were all upwardly perturbed in the second time point. All perturbed KEGG pathways belonged to the *Metabolism* class, with 2 *Carbohydrate metabolism* and 2 *Amino acid 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 second time point.

Differentially perturbed InterPro terms. 23 InterPro domains were enriched in the upwards direction in the second time point. *Bifunctional inhibitor/plant lipid transfer protein/seed storage helical domain* (IPR016140, q value 0.0, set size 34), *Bifunctional inhibitor/plant lipid transfer protein/seed storage helical domain superfamily* (IPR036312, q value 0.0, set size 36), R*mlC-like cupin domain superfamily* (IPR011051, q value 0.0, set size 103), *RmlC-like jelly roll fold* (IPR014710, q value 0.0, set size 104), *11-S seed storage protein, conserved site* (IPR022379, q value 0.0, set size 36), and *11-S seed storage protein, plant* (IPR006044, q value 0.0, set size 47) describe the top 6 IPR terms, and all of them have

been linked with plant seed storage - consistent with an increase in carbohydrate consumption from grain sources in the second time point.

53 InterPro terms were downregulated in the second time point with q value < 0.05. *Immunoglobulin-like domain superfamily* (IPR036179, q value 0.0, set size 224), *Immunoglobulin-like domain* (q value 0.0, set size 214), *Immunoglobulin V-set domain* (IPR013106, q value 0.0, set size 203), *Immunoglobulin subtype* (IPR003599, q value 0.0, set size 174) and *Immunoglobulin-like fold* (IPR013783, q value 0.0, set size 311) were the top five downregulated IPR 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 *time point 2* 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).

Gut pathogen detection using mass spectrometry. 7 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 second 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.*

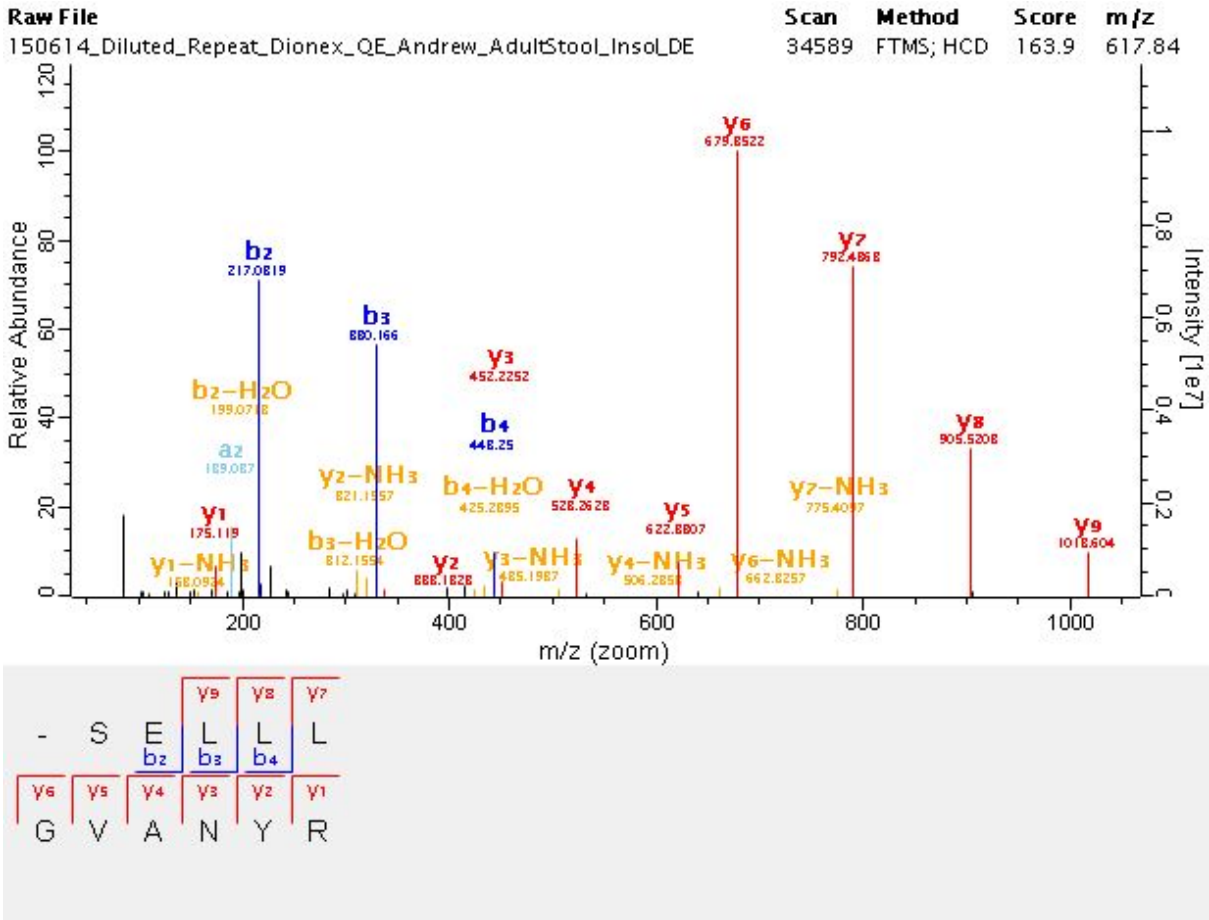

Figure 4. *SELLLGVANYR* (PEP score 4.535300e-07) was significantly more abundant in the second time point (adj.p-value 0.0028). BLAST analysis yielded a match to *Vicilin-like antimicrobial peptides 2-1* of *Trichinella papuae* with 100.0 % identity. The presence of this pathogen and associated peptide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

Conclusions

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 (*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 MSMS percentage assignment between the **MetaNovo** database derived from **UniProt** and the human Reference proteome was found in the analysis of the human central nervous system samples**.** Further, 85.99 percent 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.

We have demonstrated that it is possible to quantify taxa that would not have been identified using 16S rRNA sequencing, allowing for simultaneous characterization of the microbiome 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. Further, the identification of a *helminth* pathogen in the second time point corresponding with decreased alpha diversity is consistent with the well-known association between pathology and decreased gut microbial diversity.

Peptide assignments by **UniPept** of peptides to organisms that are not likely to be present in either the adult gut metaproteome or in the central nervous system 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 peptides also previously identified at the MSMS level.

In conclusion, **MetaNovo** makes characterizing complex mass spectrometry data with an extremely large search space feasible, allowing for novel peptide and polymorphism identification when genomic information is not available, and improving the accuracy and affordability of current clinical proteomic and metaproteomic analysis methods. 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, such as the detection of *helminth* peptides in adult stool samples, or rare clinically relevant sequence polymorphisms in tumour samples.

Author contributions

MGP - Bioinformatics

AJMN - Bioinformatics, Wet lab

SF - Wet lab

SG - Wet lab

DT - DirecTag configuration, Taxonomic weighting strategy

JB - Corresponding author

NM - 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.

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. 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
11. 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
12. 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
13. 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
14. 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
15. 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
16. 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
17. 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

18. 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
19. 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

Addendum A: Adult Stool Samples

DNA extraction and 16s rRNA gene library preparation. DNA was extracted from stool samples using the PowerSoil®DNA isolation kit (MO BIO Laboratories Inc.) After enzymatic digestion and bead beating, the kit was used to extract DNA which was then quantified using a Qubit® fluorometer (Thermo Fisher Scientific). This was followed by preparation of 16s rRNA libraries for next-generation sequencing on the Illumina HiSeq 2500 platform.

The hypervariable V6 region of the 16S rRNA gene was PCR amplified using conserved universal V6 primers via two steps as described by Arthur *et al*. (Arthur *et al*., 2012). The first step utilized set 1 primers. These barcoded primers were used to carry out PCR using a high-fidelity DNA polymerase (Thermo Scientific™ Phusion™) on extracted DNA. The PCR reaction amplified from 50 ng of DNA template. PCR reactions were held at 94 °C for 3 min followed by 10 amplification cycles using a touchdown protocol with denaturation at 94 °C for 45 s, annealing at 61 °C for 45 s with 1 °C drop in each cycle and an extension at 72°C for 45 s. Amplification continued subsequently with 15 additional cycles using 51 °C as annealing temperature. The PCR was then terminated with a final elongation at 72 °C for 2 min.

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 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 (Ref) and in-solution protein sample preparation methods (Ref).

In brief 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 Guanidine 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 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 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 an Orbitrap 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*.