Navigating through metaproteomics data: a logbook of database searching

RESEARCH ARTICLE

Navigating through metaproteomics data: A logbook of database searching

Thilo Muth[1], Carolin A. Kolmeder[2*], Jarkko Salojärvi[2], Salla Keskitalo[3], Markku Varjosalo[3], Froukje J. Verdam[4], Sander S. Rensen[4], Udo Reichl[1,5], Willem M. de Vos[2,6,7], Erdmann Rapp[1**] and Lennart Martens[8,9]*

[1] Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany
[2] Department of Veterinary Biosciences, University of Helsinki, Helsinki, Finland
[3] Institute of Biotechnology, University of Helsinki, Helsinki, Finland
[4] Department of General Surgery, NUTRIM, Maastricht University Medical Center, Maastricht, The Netherlands
[5] Otto-von-Guericke University, Bioprocess Engineering, Magdeburg, Germany
[6] Department of Bacteriology and Immunology, University of Helsinki, Helsinki, Finland
[7] Laboratory of Microbiology, Wageningen University, Wageningen, The Netherlands
[8] Department of Biochemistry, Ghent University, Ghent, Belgium
[9] Department of Medical Protein Research, VIB, Ghent, Belgium

Received: November 28, 2014
Revised: February 13, 2015
Accepted: March 6, 2015

Metaproteomic research involves various computational challenges during the identification of fragmentation spectra acquired from the proteome of a complex microbiome. These issues are manifold and range from the construction of customized sequence databases, the optimal setting of search parameters to limitations in the identification search algorithms themselves. In order to assess the importance of these individual factors, we studied the effect of strategies to combine different search algorithms, explored the influence of chosen database search settings, and investigated the impact of the size of the protein sequence database used for identification. Furthermore, we applied de novo sequencing as a complementary approach to classic database searching. All evaluations were performed on a human intestinal metaproteome dataset. *Pyrococcus furiosus* proteome data were used to contrast database searching of metaproteomic data to a classic proteomic experiment. Searching against subsets of metaproteome databases and the use of multiple search engines increased the number of identifications. The integration of *P. furiosus* sequences in a metaproteomic sequence database showcased the limitation of the target-decoy-controlled false discovery rate approach in combination with large sequence databases. The selection of varying search engine parameters and the application of de novo sequencing represented useful methods to increase the reliability of the results. Based on our findings, we provide recommendations for the data analysis that help researchers to establish or improve analysis workflows in metaproteomics.

Keywords:
Bioinformatics / De novo sequencing / False discovery rate / Metaproteomics / Search parameters

Additional supporting information may be found in the online version of this article at the publisher's web-site

Correspondence: Professor Lennart Martens, Department of Medical Protein Research and Biochemistry, VIB and Faculty of Medicine and Health Sciences, Ghent University, A. Baertsoenkaai 3, B-9000 Ghent, Belgium
E-mail: lennart.martens@ugent.be
Fax: +32-92649484

Abbreviations: **FA**, formic acid; **HIMPdb**, human intestinal metaproteome database; **HS**, high scoring; **LS**, low scoring; **MC**, missed cleavages; **PSM**, peptide-spectrum match; **Pyrodb**, *Pyrococcus furiosus* database; **PyroHIMPdb**, concatenated *pyrococcus furiosus* and human intestinal metaproteome database; **RMIC**, relative matched ion count

*Both authors contributed equally and share the first authorship.
**Additional corresponding author: Dr. Erdmann Rapp,
E-mail: rapp@mpi-magdeburg.mpg.de
Colour Online: See the article online to view Figs. 1, 2 and 3 in colour.

1 Introduction

Over the past few years, progress in instrumentation as well as methodological improvements have greatly facilitated the automated analysis of high-throughput MS-based proteomic data. In particular, this concerns the matching of mass spectra to peptides from an in-silico digested protein sequence database [1]. In addition, de novo sequencing approaches have become an alternative option for assigning peptide sequences to their fragmentation mass spectra (MS/MS) [2]. Both methods result in peptide-spectrum matches (PSMs) with algorithm-specific scores reflecting the identification quality. These scores are subsequently processed by built-in statistical routines of the data analysis software or by external validation algorithms to estimate the probability of these PSMs to be correct [3]. The final outcome of a high-throughput proteomics experiment therefore not only depends on the quality of the sample processing and MS analysis steps, but to a large extent also on the peptide and protein identification algorithms used and the parameters employed in the data analysis [4–6]. The first and most important requirement is that the sequences of the analyzed peptides have to be contained in the protein sequence database used, or have to be suggested by de novo sequencing. Second, from the wide range of available parameters that can be modified to influence the search results, the user needs to select the appropriate set [7]. Third, the results need to be validated by reliable statistics using an acceptable false discovery rate (FDR) threshold. In classic proteomic experiments, where proteins from a specific tissue or organism are analyzed, these issues are particularly pronounced when uncommon splice variants and unexpected PTMs are present in the samples. However, when analyses are performed on environmental samples comprising proteins from many different organisms the issues become much more severe [8].

The first challenge, the necessity to know the correct sequence space to perform a meaningful search, is particularly troublesome in metaproteomics. Only a low proportion of common organisms is fully sequenced. Therefore, currently only in controlled environments, i.e. for model organisms or in systems with low complexity, protein sequence databases can be considered to be reasonably representative. In typical metaproteomic analyses, it is therefore highly likely that entire genomes of several organisms are missing from the protein sequence database, resulting in unidentified high-quality spectra. While metagenomics offers the possibility to approach the real protein content of samples by incorporating the genetic background of the ecosystem, this method may not provide complete coverage of all potential protein sequences as the quality of sequencing, assembly and annotation still have an impact on the obtained metagenome [9]. De novo sequencing, where peptide sequence information is inferred directly from the spectrum without the use of a sequence database, has become another possibility to assign peptide sequences to MS/MS spectra from metaproteomic experiments. However, this particular approach has not been widely used for metaproteomic data so far. Instead, steadily increasing public repositories with metagenomic data of several ecosystems are becoming available that can provide a valuable search space for metaproteomic data [10]. Furthermore, such resources can be modified for specific samples by adding genomes from individual organisms in order to increase the coverage. Therefore, protein sequence databases used in metaproteomics can quickly grow to millions of sequences. Because such enormous databases strongly challenge most peptide identification algorithms, specific strategies have been developed to narrow this search space. A common approach presents the sectioning of the database in order to reduce the number of protein sequences by only retaining taxonomy-specific entries [11]. Another strategy employs searching in two steps: first, the initial database is searched without any FDR limitation, which is then followed by a second search with a stringent FDR threshold against a refined database created by extracting the protein identifications derived from the first search [12]. This method has already been applied recently both in metaproteomics [13] and proteogenomics [14, 15]. In a related iterative approach, protein sequences from a first search are used to generate a second database containing gene differences in order to substantiate the previously identified proteins [16].

The second challenge involves the selection of optimal search parameters and is further complicated by the first challenge. The effect on the search results by varying the maximum number of missed cleavages (MC) or the chosen enzyme for digestion (such as trypsin, chymotrypsin, or semi-tryptic cleavage) has hardly been studied for large protein sequence databases encountered in metaproteomics. Another key issue in search parameter selection is the choice of considered PTMs, which is already a challenging task in less complex systems [7]. In metaproteomics, it is even more difficult due to limited a priori information and intricate predictions about expected modifications in the studied system [17]. A clear view on the impact of these parameters on the search outcome has so far been difficult to obtain, since different studies tend to use individual parameter settings, which strongly impairs a comparison of the results.

The third challenge, filtering results at a reliable and acceptable FDR level, also provides specific challenges for metaproteomics research. First of all, despite tremendous progress in proteome bioinformatics over the past decade, currently available search algorithms have been mainly developed and tested for single organism data. The performance characteristics of these algorithms on metaproteomic data therefore are largely unknown. And while it is known that different database search engines can provide complementary results [18–20] and that the use of multiple algorithms may therefore be particularly appropriate in the analysis of complex systems, most studies employ a single search engine only. Second, although FDR estimation in metaproteomics typically relies on the target-decoy approach used in traditional proteomics [21], various studies indicate that this method is not perfectly suited for all cases [5, 22, 23].

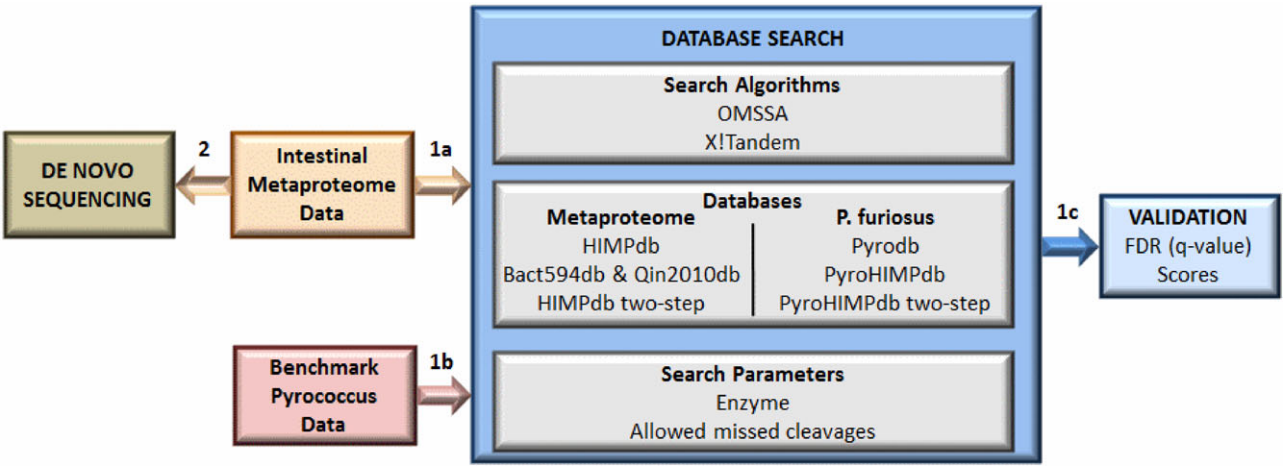

Figure 1. Experimental setup. MS/MS spectra of the human intestinal metaproteome dataset were inferred to peptide sequences by database searching (1a), with various different search settings, as well as de novo sequencing (2). Database search of *Pyrococcus furiosus* spectra was performed equivalently to the metagenome data searches (1b). Both HIMPdb and *P. furiosus* search identifications are validated by their scores and *q*-values via the target-decoy approach (1c).

Consequently, if the FDR estimation is actually influenced by certain unwanted factors, a false impression of error control may be created when results are presented for a definite FDR threshold, usually set at 1% [24, 25] or 5% [26, 27].

In order to obtain a better insight into these three challenges, we analyzed ten metaproteomic samples deriving from human fecal material and evaluated the effect on the identification rate by using two search engines against a targeted protein database. In addition, de novo sequencing was applied to complement the database searches. Variation in search parameters (enzyme, number of MC) and the behavior of the FDR with respect to the database size was tested via benchmark experiments on three out of the ten samples and additional *Pyrococcus furiosus* data [28].

2 Materials and methods

2.1 Study setup

As detailed below, two datasets were used for database searching, one comprising ten human fecal metaproteomes and one of *P. furiosus* proteins (Fig. 1). The search algorithms X!Tandem and OMSSA were applied and different settings of the search parameters (i.e. allowed MC and enzyme selection) were used for database searching. In addition, searches against protein sequence databases of varying size were performed. The PSM scores were obtained from the individual search algorithms and FDR estimations based on *q*-values [29] were used for search results combination. De novo sequencing was applied for the human metaproteomic dataset. Relative matched ion counts were used as rescoring method to compare the performance of database searching and de novo sequencing.

2.2 Metaproteomic dataset

2.2.1 Study cohort and sample preparation

As test metaproteomic dataset we used ten metaproteomes of adults (P1, P3, P8, P11, P17, P23, P27, P28, P31, P34) who took part in a larger study as detailed in [30]. This study was approved on July 12, 2011 by the Medical Ethics Committee of the Atrium Medical Center (Heerlen, the Netherlands, registration number NL30502.096.09), and conducted according to the revised version of the Declaration of Helsinki (October 2008, Seoul). Informed consent in writing was obtained from each subject individually. The biological findings on the metaproteomes will be published separately. Aliquots from the same fecal samples were used, which were analyzed with a phylogenetic microarray [30]. The metaproteomes were prepared as follows: proteins from frozen fecal material were extracted and fractioned as described in a previous study [27]. In brief, cells were lysed by bead beating in PBS. For bead beating, a FastPrep 24 (MP Biomedicals) with cooling device was used, which required 5-min cooling steps on ice in between the five cycles of bead beating. From the resulting supernatants, beads were removed by low-speed centrifugation and cell debris by low-speed centrifugation. Proteins were separated on a 4–12% NUPAGE Bis-Tris gel (Invitrogen) and the 37 and 75 kDa bands of a prestained protein marker (Precision Plus™ Dual Color, BioRad) was used for cutting a gel region. The 37–75 kDa region was subjected to in-gel protein digestion using trypsin [31]. After digestion, peptides were purified with C18 microspin columns (Harvard Apparatus, USA) according to manufacturer's instructions and redissolved in 30 μL of 0.1% trifluoroacetic acid and 1% ACN in HPLC-grade water.

2.2.2 Liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS)

Reverse-phase liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) analysis was carried out on a nanoHPLC system EASY-nLC II (Thermo Fisher Scientific, Germany) connected to an Orbitrap Elite hybrid mass spectrometer (Thermo Fisher Scientific, Germany) with nanoelectrospray ion source (Thermo Fisher Scientific). The tryptic peptide sample mixture was automatically loaded from autosampler into a C18-packed precolumn (EASY-Column™ 2 cm × 100 μm, 5 μm, 120 Å, Thermo Scientific) at a flow rate of 1 μL/min in 10 μL volume of buffer A (1% ACN and 0.1% formic acid in HPLC-grade water). Peptides were transferred onward to C18-packed analytical column (EASY-Column™ 10 cm × 75 μm, 3 μm, 120 Å, Thermo Scientific) and separated starting at 5% Buffer B (98% ACN and 0.1% formic acid in HPLC-grade water) for 5 min followed by a 120-min linear gradient from 5 to 35% of buffer B at the flow rate of 300 nL/min. Gradient was followed by 5-min gradient from 35 to 80% of B, 1-min gradient from 80 to 100% of B and 9-min column wash with 100% B at the constant flow rate of 300 nL/min.

Per LC-MS/MS run, 4 μL of sample were injected and analyzed. Full MS scan was acquired in positive ion mode with a resolution of 60 000 at normal mass range in the orbitrap analyzer. The method was set to fragment the 20 most intense precursor ions with CID (energy 35).

Data were acquired using profile mode for survey scan (MS) carried out in the high-resolution orbitrap mass analyzer and centroid mode for fragment ion scan (MS/MS) carried out in the linear iontrap mass analyzer. Peak picking was calculated by Thermo Proteome Discoverer 1.4.1.14 and output to MASCOT Generic Format (MGF), converted with default settings on the Spectrum Selector node for the conversion workflow (Mass range settings: precursor mass range: 350–5000 Da). The MGF files are available at ftp://MSV000079035@massive.ucsd.edu.

2.3 *P. furiosus* dataset

For the benchmark experiments, we took 14 467 MS/MS spectra obtained from *P. furiosus* proteins as described in detail in [28], as representative proteomics sample. *Pyrococcus furiosus* presents a hyperthermophilic archaeon known for its unique biochemistry [32, 33].

2.4 Database searching

For protein database searching, we used X!Tandem (version 2013.02.01) [34] and OMSSA (version 2.1.8) [35] as search engines with the following parameters: trypsin was used as default enzyme and up to two MC were allowed. Carbamidomethylation of cysteine was chosen as fixed, and oxidation of methionine as variable modification. The fragment ion tolerance was set to 0.4 Da and the precursor tolerance to 0.03 Da. For some explicitly mentioned experiments, we modified the default parameters MC and enzyme. Depending on the experiment, one out of a total of six different protein sequence databases was used (Table 1). The main protein sequence database used was the Human Intestinal Metaproteome database (HIMPdb, 6 153 068 protein sequences), which was constructed from different sources such as metagenomes, bacterial genomes, the human genome, as well as plant genomes and therefore represents a wide range of expected proteins in fecal samples (Supporting Information Table 1). To study the effect of the database size, two subsets of HIMPdb were taken: a collection of 594 bacterial genomes expected to occur in the human gut (Bact594db, Supporting Information Table 2) and intestinal metagenomes of 124 individuals (Qin2010db) [36]. For *P. furiosus* searches both Pyrodb, a collection of 2139 *P. furiosus*, 7325 *Saccharomyces cerevisiae* and 50 *Homo sapiens* protein sequences (Uniprot/Swissprot), was used as well as a combination of HIMPdb and Pryodb (called PyroHIMPdb).

Table 1. Name and size of the used protein sequence databases

Protein sequence database name	Size[a)]
HIMPdb	6 153 068
Qin2010db[b)]	3 267 604
Bact594db[b)]	1 850 744
Pyrodb	9514
PyroHIMPdb	6 162 582
HIMPdb two-step	90 040 (average)

a) Number of protein sequence entries in the FASTA database.
b) Sectioned database of HIMPdb.

For target-decoy-based FDR estimations for each of the protein databases, a decoy database was generated by reversing the target protein sequences. The target and decoy searches were performed separately.

2.5 Quality control of database search results and combination of search results

Based upon target-decoy searching, we used qVality [29] for *q*-value (minimum FDR) estimation to guarantee consistency across heterogeneous search engines with different scoring techniques. Results were always filtered with 5% FDR, however, in specific cases, as described in the results section with 1 and 10% FDR.

Original search scores were used for quality control of the search results. For X!Tandem, we used the hyperscore as the most objective scoring metric while it relies only on the assigned fragment ions [37]. For OMSSA, the score was transformed by taking minus ten times the ten-base logarithm of the OMSSA *e*-value [38]. This provided us with a comparable scale for the score of both search engines.

X!Tandem and OMSSA search results were combined by taking the individual spectrum identifications filtered by their *q*-values. For peptide identification, the union set of both search algorithms was retained.

2.6 Reporting of search results

The report of the results was limited to the number of identified spectra and distinct peptides, as the total number of proteins cannot be accurately calculated in a metaproteomics experiment. In particular, one peptide often maps to multiple proteins of different organisms resulting in the complicated and essentially unsolvable protein inference problem [8, 39, 40].

2.7 Two-step searching

For samples P1, P23, and P34, the unfiltered protein identifications from the HIMPdb search were collected. The corresponding protein sequences were extracted from the original HIMPdb to create a refined protein sequence database, which was used for a second search [12]. In the following, this approach is called two-step searching.

2.8 De novo sequencing

For de novo sequencing, we used the software DeNovoGUI (version 1.2.0) [41] that provides a graphical user interface and parallelization of the PepNovo+ algorithm [42] using the same tolerances and PTM parameters as described before. The default fragmentation model (CID_IT_TRYP) was used in PepNovo+, accounting for CID fragmentation and tryptic cleavage. The maximum number of peptide solutions was set to 20. The PepNovo+ algorithm was used in multithreaded mode by using four compute cores. Additionally, the de novo peptide suggestions were filtered by a PepNovo+ score threshold above 100 for high-quality identifications and between 50 and 100 for low-quality identifications.

2.9 Rescoring

To evaluate the quality of the identified spectrum, the relative matched ion count (RMIC) score was calculated by the intensities of the matched fragment ions (a/b/c, x/y/z, y-NH_3, y-$H_2$0,b-NH_3, b-$H_2$0, precursor MH, MH-NH_3, and MH-H_2O) of the peptide divided by the total ion current of the related spectrum; an RMIC of 0.5 thus means that 50% of the spectral peak intensity can be explained by fragment ion peaks (within a window of 0.5 Da for each peak). The RMIC score served as a basic and rapid quality measurement independent of the applied peptide identification strategy, as it relies only on the given spectrum and peptide information. Consequently, the rescoring of peptide identifications was consistent for both de novo sequencing and database search results.

3 Results

To explore different spectrum identification strategies in metaproteomic data analysis, we carried out both protein database searching as well as de novo sequencing (Fig. 1). We focused on database searching as it represents the most frequent approach in metaproteomics. We therefore analyzed different setups to untangle the influence of three major components: the search algorithm, the search database, and the search parameters. *Pyrococcus furiosus* data [28] were used as a benchmark to contrast the impact of the evaluated parameters in a metaproteomic to a classical proteomic experiment. In order to evaluate the impact on the quantity and quality of peptide identification results when searching MS/MS spectra against a large database, we chose the technique of integrating *P. furiosus* sequences into a metaproteomic sequence database.

3.1 Database searching

3.1.1 Combination of two search algorithms for peptide identification in human intestinal metaproteomic data

To investigate whether metaproteomic data analysis may benefit from combining multiple search engines, we used two popular and noncommercial database search algorithms (X!Tandem and OMSSA) to perform peptide to spectrum matching of MS/MS data from ten human fecal samples (comprising 317 375 MS/MS spectra in total) against a customized protein sequence database (HIMPdb). On average, 30% of the spectra and 7322 peptides per sample were identified when combining the results from both search algorithms at 5% FDR (Table 2).

X!Tandem provided significantly more identified spectra than OMSSA at 5% FDR and a substantial amount was mutually exclusive: 25% were identified only by X!Tandem and 11% only by OMSSA. On average, the contribution of exclusive peptides was 23% (X!Tandem) and 16% (OMSSA), respectively. As expected, the application of a more stringent 1% FDR criterion resulted in a decrease of the average spectrum identification rate (21%, Supporting Information Table 3). Consequently, the average number of peptides dropped to 5476 per sample (Supporting Information Table 4).

3.1.2 Effect of database size on quality and quantity of search hits

The nature of metaproteomic samples, i.e. the likely presence of hundreds of different organisms, requires the protein

Table 2. Number of measured spectra, identification rate, and number of peptides obtained from the combination of the search algorithms X!Tandem and OMSSA for ten different samples (FDR < 5%)

Sample	Total	ID rate (%)	Peptides	Excl. Spectrum ID (%)		Excl. Peptide ID (%)	
				X!Tandem	OMSSA	X!Tandem	OMSSA
P1	35 179	31.7	8473	21.5	11.4	19.8	16.4
P3	26 560	26.0	5624	24.0	11.2	22.9	15.3
P8	31 891	31.6	7640	19.1	11.2	17.5	17.1
P11	31 744	26.1	6295	30.4	12.0	26.2	16.5
P17	32 203	31.7	8082	19.5	11.5	18.9	16.3
P23	34 050	33.2	8255	35.8	8.6	30.4	13.7
P27	27 339	24.8	5266	24.8	11.2	22.6	14.9
P28	32 037	30.9	7273	25.9	10.4	23.5	14.9
P31	35 848	34.6	9084	30.1	9.9	25.9	15.4
P34	30 524	31.1	7231	20.1	12.1	19.0	17.1
Average	31 737	30.2	7322	25.3	10.8	22.7	15.8

The exclusive contributions by each of the algorithms are given both for spectrum and peptide identifications.

sequence database to be different to the ones used in a classic proteomics experiment. Therefore, a whole collection of translated genomes and metagenomes is typically collated in order to represent as many potential target proteins in the samples as possible. While a few specific search strategies for metaproteomic data have been proposed based on customization of the search database or iterative adaptation of the search space [11, 12, 16], a systematic study of the effects of the search database on the results has not yet been carried out.

In the following, we address the effect of the database size on the search result of three samples in the metaproteomic dataset (P1, P23, P34) by using the targeted HIMPdb, containing potential protein sequences for the studied samples, and subsets of this database, Bact594db and Qin2010db (Table 1). We took those particular sectioned databases as the highest number of peptides could be obtained from them specifically (Supporting Information Table 5).

For all three samples, HIMPdb two-step searching yielded by far the highest number of PSMs and peptides (Fig. 2A, Supporting Information Table 6). HIMPdb and Qin2010db provided a similar amount, whereas Bact594db resulted in the fewest PSMs and peptides. Both for the Bact594db and the Qin2010db searches, the number of database-specific PSMs and peptides was higher than the amount of corresponding identifications from the HIMPdb searches (Fig. 2B and C, Supporting Information Tables 7 and 8). This effect was relatively stronger for the Bact594db (30% of HIMPdb size) search than for the Qin2010db (53% of HIMPdb size) search. The increase in PSMs and peptides was larger at 1% FDR than at 5% FDR, and for PSMs slightly weaker than for peptides.

These results demonstrate that a high amount of PSMs and peptides were exclusively found when searching against subsets of HIMPdb.

Next, we investigated two-step searching in more detail: the first step involved searching HIMPdb without applying any filtering. This resulted in an average of 90 040 protein sequences, representing 1.5% of HIMPdb (HIMPdb two-step; see Table 1). At 1 and 5% FDR, the applied method more than doubled the amount of PSMs and peptides in comparison to HIMPdb searching (Fig. 2A). The number of identified peptides was higher than the amount of the identified spectra at 5% FDR. This can be explained by the fact that in several cases multiple peptides were suggested for one spectrum by the database search algorithms.

For sample P1, we also performed a rescoring of HIMPdb and HIMPdb two-step searching results by taking the matched fragment ions of the original MS/MS spectra into account: The RMIC distribution was shifted to the left for the HIMPdb two-step searching results, indicating lower scores derived from both search engines (Supporting Information Fig. 1). The absolute number of PSMs was higher for HIMPdb two-step searching. OMSSA provided a shifted RMIC to the right in comparison to X!Tandem for both HIMPdb and HIMPdb two-step searching.

3.1.3 Effect of search parameters on quality and quantity of search hits

To investigate the influence of the search engine parameters on the number of identifications, additional experiments on the samples P1, P23, and P34, the same as used for testing the effect of the database size, were performed.

First, we modified the maximum limit of allowed MC, in a range from zero to a maximum value of three. Depending on FDR or sample number, the highest number of PSMs and peptides could be found for zero or one MC (Supporting Information Tables 9–14). Studying this in more detail for sample P1 showed that at 1% FDR a total of 18% hits for zero allowed MC was not found with the 1–3 MC settings, however, at 10% FDR, this value decreased to 0.3% (data not shown).

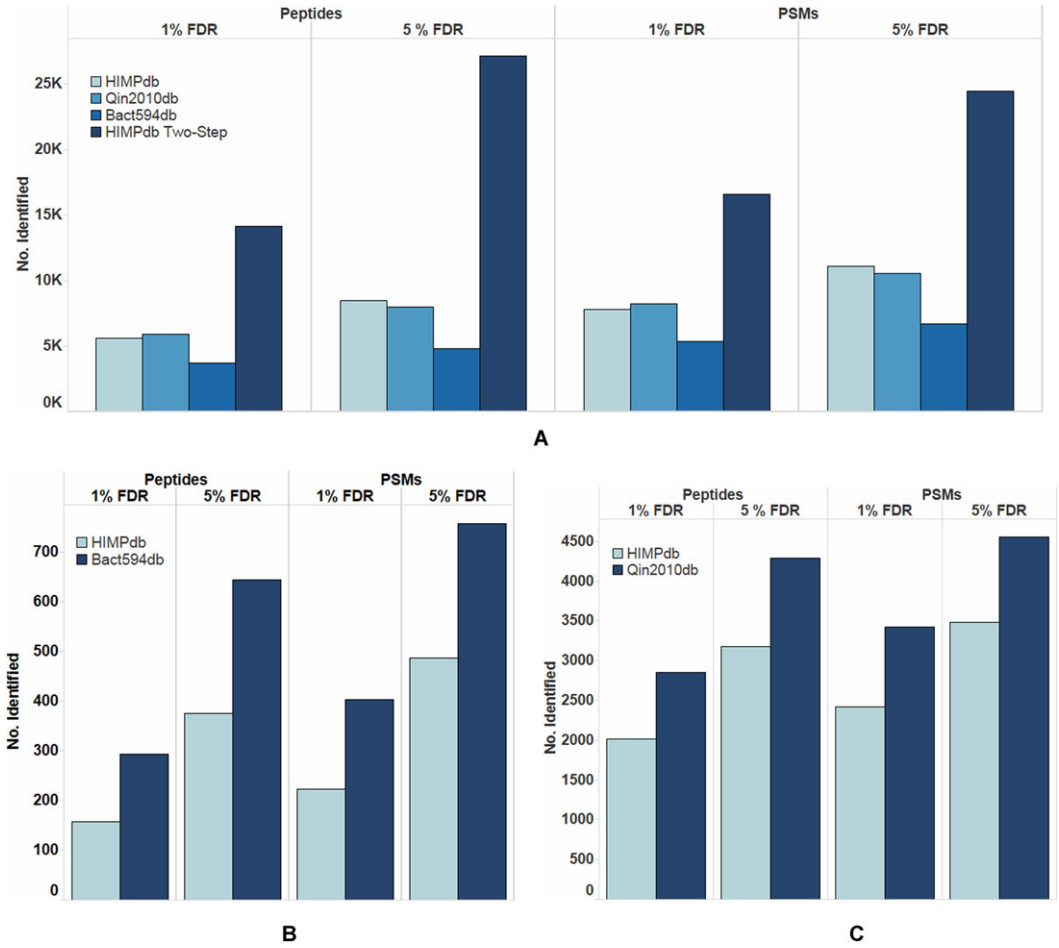

Figure 2. Comparison of results from searching the human intestinal metaproteome sample P1 against a collection of databases. (A) Number of PSMs and peptides for HIMPdb, Bact594db, Qin2010db, and HIMPdb two-step searching method from sample P1. (B) Bact594-specific identifications from HIMPdb and Bact594db database searches. (C) Qin2010-specific identifications from HIMPdb and Qin2010db database searches.

Second, we analyzed the impact of the different cleavage enzyme used. This is appropriate since the analyzed proteins are derived from the human intestinal tract—a protease rich environment—that can easily lead to the presence of various non-tryptic protein fragments. Therefore, we repeated the searches for the samples P1, P23, and P34 using a semi-tryptic cleavage (allowing one terminus to be non-tryptic, whereas the other terminus remains tryptic), and two other common intestinal proteases: chymotrypsin (pancreatic enzyme) and pepsin A (gastric enzyme) for cleavage. At 5% FDR, the tryptic searches resulted in more PSMs and peptides than the semi-tryptic searches (Table 3).

Table 3. Number of PSMs and peptides and percentage of exclusive identifications for tryptic and semi-tryptic search settings for samples P1, P23, and P34 (FDR < 5%)

Sample	Tryptic cleavage				Semi-tryptic cleavage			
	PSMs		Peptides		PSMs		Peptides	
	All	Excl. (%)	All	Excl. (%)	All	Excl. (%)	All	Excl. (%)
P1	11 133	8.1	8473	12.5	10 354	1.1	7959	6.9
P23	11 288	6.1	8255	10.5	10 777	1.7	7976	7.4
P34	9491	8.6	7231	13.1	8743	0.8	6678	5.9
Average	10 637	7.6	7986	12.0	9958	1.2	7538	6.7

Table 4. Number of PSMs and peptides from *Pyrococcus furiosus* searches against three different databases and percentage of *P. furiosus* specific hits (Pyro) at 1 and 5% FDR

Database	1% FDR				5% FDR			
	PSMs		Peptides		PSMs		Peptides	
	ID	Pyro (%)	ID	Pyro (%)	ID	Pyro (%)	ID	Pyro (%)
Pyrodb	10 428	100.0	6032	100.0	11 517	100.0	6884	100.0
PyroHIMPdb	5330	99.4	2951	98.5	7035	97.0	4074	92.8
PyroHIMPdb two-step	9462	91.5	6081	80.2	10 442	90.4	9064	59.2

The semi-tryptic searches identified an exclusive percentage of around 1% PSMs and 7% peptides on average compared to the tryptic searches. In contrast, at 1% FDR, the total number of PSMs and peptides was higher for the semi-tryptic searches (Supporting Information Table 15). However, the running times for the semi-tryptic searches with both X!Tandem and OMSSA increased fivefold on average (data not shown). Choosing chymotrypsin or pepsin A as cleavage enzyme resulted only in about 100 PSMs and peptides (data not shown).

3.1.4 Classic proteomic control—*P. furiosus* benchmark experiment

Next, we performed a benchmark experiment with 14 467 MS/MS spectra from the proteome of *P. furiosus* [28], which were searched against a targeted Pyrodb database containing 9514 *P. furiosus* and untargeted control proteins and an untargeted merged database of the aforementioned *HIMPdb* and *P. furiosus* database (PyroHIMPdb). A two-step search was performed against this database as well (PyroHIMPdb two-step). Similar to the metaproteomic data, the effect of search algorithm database size, and search parameters were studied (Fig. 1).

At 5% FDR, 11 517 PSMs and 6884 peptides were identified when searching against Pyrodb (Table 4). However, searching the same spectra against PyroHIMPdb resulted in a drop in the number of PSMs and peptides (Fig. 3A). Combining OMSSA and X!Tandem increased the number of PSMs to a higher extent for PyroHIMPdb than for Pyrodb. While target and decoy distributions in Pyrodb searches could be clearly differentiated for both search engines, the searches against the PyroHIMPdb showed a broader overlap of target and decoy hit distributions for both X!Tandem (Fig. 3B) and OMSSA (Fig. 3C). For a fair benchmark comparison, we applied two-step searching also against PyroHIMPdb that contains few *P. furiosus* sequences among many unrelated sequences (the proteins found in human intestinal microbiota).

For both search engines, the distributions of decoy PSMs from Pyrodb and PyroHIMPdb two-step searching were in a comparable score range (Fig. 3B and C). However, the decoy PSM distribution resulting from PyroHIMPdb searching was broader and also shifted to the right, which explains an increased score threshold for FDR estimations. These distributions explain the steep gain in PSMs and peptides with PyroHIMPdb two-step searching compared to HIMPdb both at 1 and 5% FDR but also the relative decrease of *P. furiosus* specific peptides, which was at 5% FDR only 59.2% of the overall identified peptides (Table 4). In addition, two-step searching could not recover as many PSMs as the Pyrodb search.

As a direct comparison for metaproteomic data, we further evaluated the search parameter selection on the *P. furiosus* dataset. At all FDR levels, the number of PSMs and peptides increased with the number of allowed MC until a value of two MC and dropped slightly for a value of three MC (Supporting Information Tables 16 and 17).

We also searched the *P. furiosus* dataset with a semi-tryptic enzyme selection. This search resulted in less PSMs and peptides than the tryptic search (Supporting Information Tables 16 and 17).

3.2 De novo sequencing

De novo sequencing has the indisputable advantage of being independent of a protein sequence database. Since public databases often fail to fully cover the expected proteome in metaproteomic analyses, de novo sequencing can be useful by inferring peptide sequences directly from MS/MS spectra. Thus, we performed de novo sequencing on the ten metaproteomes using DeNovoGUI [41] based on the PepNovo+ algorithm [42], and thereby obtained de novo peptide suggestions for each spectrum. For de novo sequencing, we employed the same parameters as for database searching to ensure an even comparison of both methods. With a strict PepNovo+ threshold score of 100 or more, on average 23% of the spectra were identified with de novo sequencing (Table 5). Lowering the score threshold to 50 increased the identification ratio to over 60% implying that this number contains a higher percentage of unreliable identifications. The resulting de novo sequencing peptides (PepNovo+ score above threshold 100) were matched against (i) peptides obtained by database searching against HIMPdb at 5% FDR (X!Tandem and OMSSA), and (ii) peptides derived from an in silico digested HIMPdb protein database. On average, 23% of all peptides from database searching were also found via de novo sequencing. This ratio increased to 25% when no score threshold was applied for de novo sequencing. When the de novo sequencing peptides

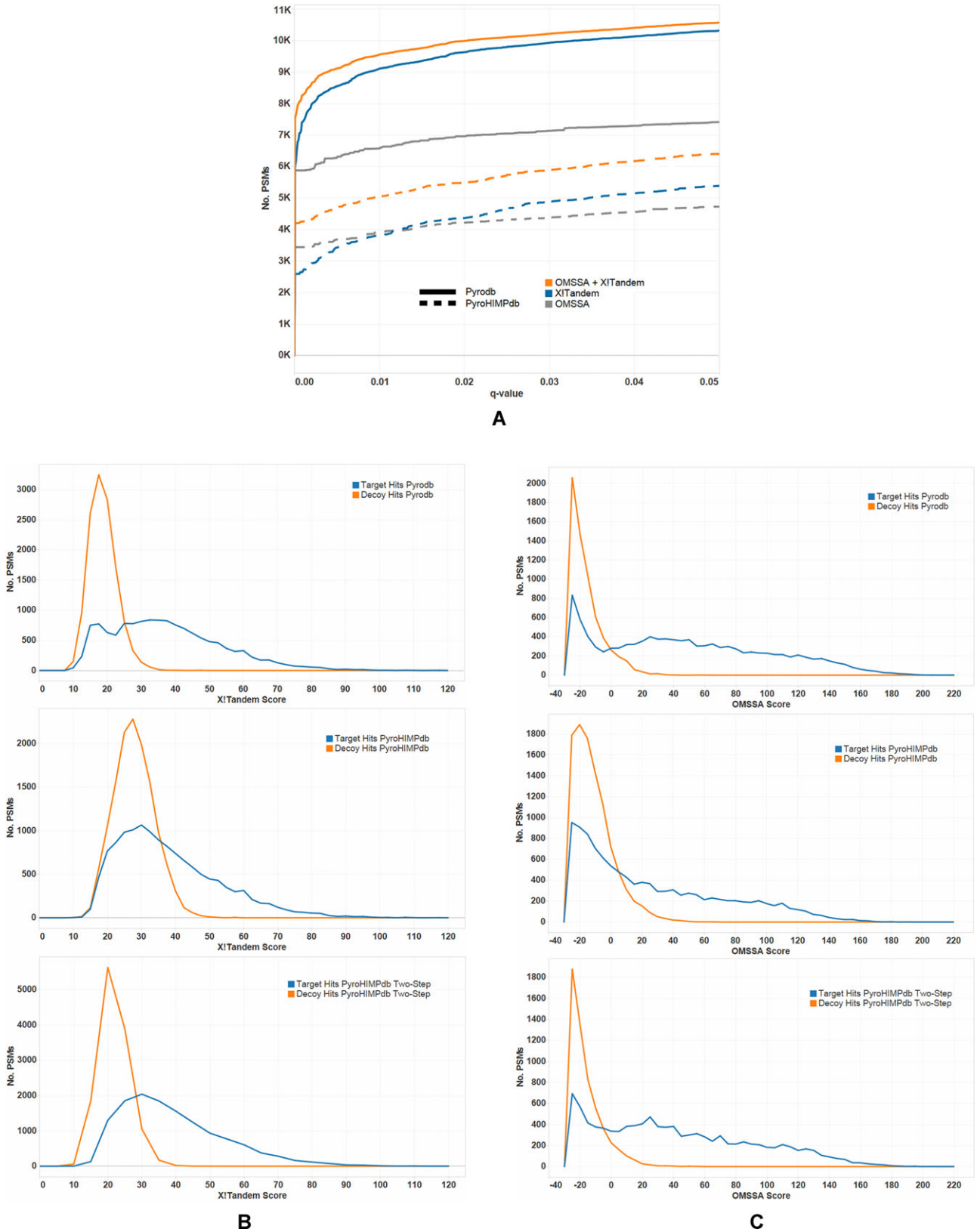

Figure 3. Comparison of results from searching the *Pyrococcus furiosus* sample against a collection of databases. (A) Number of PSMs (X!Tandem and OMSSA algorithm) in Pyrodb and PyroHIMP database searches as a function of the respective *q*-value (minimum FDR). Distribution of target and decoy PSMs identified by (B) X!Tandem and (C) OMSSA according to their score for the *P. furiosus* sample (upper panel: Pyrodb, middle panel: PyroHIMPdb, lower panel: PyroHIMPdb two-step).

Table 5. de novo sequencing results for ten different samples (P1–P34)

Sample	Spec. Ids$^{S>100}$	Spec. Ids $^{S>50}$	Rec. Pept. $^{S>100}$ (%*)	Rec. Pept. ^{ALL} (%*)	In silico Pept.$^{S>100}$	In silico Pept. $^{S>50}$
P1	8121	22 110	1822	1999	1723	5881
P3	6196	16 617	1628	1791	1345	4960
P8	7207	18 933	1758	1909	1567	5396
P11	6766	18 612	1510	1638	1406	5076
P17	7527	19 900	1775	1911	1691	5516
P23	7952	21 743	1814	1948	1666	5480
P27	6281	16 378	1437	1529	1330	4694
P28	7217	19 927	1591	1733	1491	5085
P31	8699	23 561	1893	2056	1807	5839
P34	6677	18 360	1665	1802	1539	4940
Average	7264 (23%)	19 614 (62%)	1689 (23%)	1831 (25%)	1557	5287

Spec. Ids: Number of identified spectra.
S > 100: PepNovo+ score above 100.
S > 50: PepNovo+ score above 50.
ALL: Without PepNovo+ score limitation.
Rec. Pept.: Number of recovered peptides, i.e. peptides identified both with de novo sequencing and HIMPdb search at 5% FDR.
In silico Pept.: In silico peptides of the HIMPdb matched against peptides identified with de novo sequencing.

with a PepNovo+ score above 100 were matched against an in silico digested HIMPdb protein database, an average of 1557 peptides could be found. A less stringent PepNovo+ score threshold of above 50 resulted in a significantly higher amount of 5287 de novo sequencing peptides matched successfully against the in silico digested HIMPdb.

In order to investigate the reliability of the identifications in more detail, we focused on the samples P1, P23, and P34 and compared the resulting peptides from de novo sequencing with the ones from database searching by their respective scores (Supporting Information Fig. 2). As the individual scoring methods are hardly comparable, we classified the results into two elementary categories of low and high scoring peptides. Therefore, we applied a threshold for de novo sequencing peptides scoring above 100 (PepNovo+ score), and another one for database searching peptides scoring above 40 (Hyperscore for X!Tandem and -10*$\log_{10}$ (*e*-value) for OMSSA). Peptides scoring above those defined thresholds were called high scoring identifications, otherwise low scoring (LS) identifications.

Most peptides overlapping between HIMPdb searching and de novo sequencing were high scoring identifications (Supporting Information Fig. 2). Conversely, the number of LS identifications was increased when comparing the overlapping peptides from HIMPdb two-step database searching and de novo sequencing. As already revealed in previous experiments, OMSSA performed better by resulting in less LS identifications than X!Tandem for both database searching setups. The number of LS identifications was increased for HIMPdb two-step searching in comparison to HIMPdb, which could be confirmed by further rescoring on sample P1 (Supporting Information Fig. 3). Furthermore, more stringent FDR thresholds resulted in higher PepNovo+ scores for overlapping peptides from database searching and de novo sequencing (Supporting Information Fig. 4). HIMPdb two-step searching revealed even more overlap with de novo sequencing results on the peptide level, again with an augmented number of LS identifications.

4 Discussion

Modern mass spectrometric methods allow for the high-throughput analysis of metaproteomic samples. While the resulting spectrum data are processed by automated protein and peptide identification algorithms, the use of algorithms requires the selection of many parameters by the user. As the effects of these settings on the number of identifications are currently poorly understood in the context of metaproteomic analyses, we evaluated the effects of several of these parameters for a human intestinal metaproteome and a benchmark experiment with *P. furiosus* data.

4.1 The benefit of combining results from two database search engines and obstacles found when varying the major search parameters

The effect of combining the results of different search engines has been described before in a metaproteomic study without reporting the achieved gain [43]. Combining the results from the search algorithms X!Tandem and OMSSA lead in our case to an average spectrum identification ratio of 30%. This seems to be a rather high value for a metaproteomics experiment. For instance, in a mouse metaproteome study [44] only 5% of the MS/MS spectra were identified, and in another human intestinal metaproteome study up to 17% [9]. As the database search engines X!Tandem and OMSSA complemented each other by 11 and 25% on the spectrum identification level, applying both of these algorithms and combining the results is justified. This is also in line with the findings in

a recent study with human cell line data where various search algorithms were tested individually and in combination: both X!Tandem and OMSSA showed a high performance and provided a complementarity of around 12% [20]. Searching the *P. furiosus* data against the PyroHIMPdb further illustrated the benefit of using two search algorithms, as it resulted in a significantly higher number of *P. furiosus* PSMs. While the number of MC was initially set to two due to common standards for database searching [45,46], we then investigated the effect of chosen MC value on the number of PSMs and peptides with a subset of three samples. The number of PSMs and peptides slightly decreased when increasing the MC parameter. However, in the *P. furiosus* benchmark experiment, the number of identifications increased with an increasing MC value. A possible biological explanation is that *P. furiosus* represents a hyperthermophile, which may synthetize proteins resistant to enzymatic degradation resulting in a high number of peptides for an increased MC parameter.

According to the literature, up to four MC have been accepted in metaproteomic studies (for example [47]). We found the number of PSMs and peptides increasing when choosing a lower MC parameter. If the sample requires a higher MC parameter, it is recommended to combine the results of at least two database searches with different MC values.

Furthermore, we evaluated the choice of the cleavage enzyme by performing both tryptic and semi-tryptic searches. Each of them resulted in unique PSMs and peptide identifications, while the highest number of identifications was found with tryptic cleavage at 5% FDR. When analyzing fecal samples with emphasis on host proteins, a semi-tryptic cleavage parameter led to a nearly equal amount of semi-tryptic and tryptic peptides [48]. If computer resources allow semi-tryptic searches, these could be used to supplement tryptic search results. However, according to our data, semi-tryptic search is not recommended to be used as the only enzyme option. Chymotrypsin and pepsin A—a stomach enzyme and fragments may not persist until the anus—as additional enzymes did not result in a statistically relevant amount of PSMs and peptides. This can be explained as tryptic peptides are highly abundant in human gut samples.

On top of the issue with large protein sequence databases in metaproteomics, it should be noted that performing searches with non-default parameters leads to an increased search space and less identifications: for example, an MC parameter of more than one resulted in less identified PSMs and peptides compared to conventional MC settings.

4.2 Pitfalls with database searching—de novo sequencing as alternative for metaproteomic data analysis?

The effect of database size on the number of identifications in metaproteomic data had been addressed before. For example, searching against subsets of a large database increased the number of PSMs and peptides when analyzing a mixture of nine different microbial strains [11]. An increase of PSMs and peptides was found when fecal material was studied with the focus on human proteins [48]. Furthermore, it has been noted that lowering the FDR threshold leads to more PSMs and peptides, but the results have not been discussed in detail [11].

Sectioning the HIMPdb to search against subsets had a clear effect and increased the number of resulting PSMs and peptides. Notably, lowering the FDR threshold in large database searches (HIMPdb) recovered identifications which were only found in subset database searches (Bact594db, Qin2010db). This observation reveals the problem that a stringent FDR threshold may exclude valuable identifications, but lowering the FDR increases the number of false positives in the results. In contrast to pure culture proteomics, it is even more difficult to judge, which peptides might be really present in a metaproteomic sample as it is hardly feasible to exclude results based on biological knowledge.

The *P. furiosus* benchmark experiments could explain how the decreased number of identifications in large database searches is resulting from the target-decoy approach used for FDR estimation: The target and decoy identification distributions derived from large database searches (PyroHIMPdb) overlap significantly. As the ratio of decoy to target identifications is increased tremendously when searching against large databases, the target-decoy-based FDR estimation is biased and valuable identifications may be missed out. On the other hand, a too small database search may result in an overestimation of identifications and should therefore be avoided. However, the strategy of reasonably downsizing or sectioning the database is useful as valuable identifications may be excluded when using the target-decoy approach for large databases.

The two-step searching has to be applied with a stringent FDR filtering, as the number of false-positive identifications tends to increase dramatically as shown in the *P. furiosus* benchmark experiment. This finding presents the other side of the coin: the ratio of decoy to target identifications is decreased and the FDR estimation could therefore be biased. A similar phenomenon was observed when setting a stringent parent ion tolerance for MASCOT database searches [49]. Moreover, various studies point out that the target-decoy approach, despite its popularity in proteomics, has to be treated with utmost caution to avoid misguided FDR estimations [50–52].

The above-mentioned obstacles concerning peptide identification via database searching and statistical validation on metaproteomics data also suggest looking out for alternative approaches. de novo sequencing, in particular, addresses two issues at the same time: it sidesteps the problem of uncovered sequences and the issue with database size by not relying on a protein sequence database at all. de novo sequencing showed an overlap of only 25% on the peptide level with database searching and therefore hardly qualifies as validation method. However, it can be used as a complementary method to database search engine algorithms, as a

Table 6. Recommendations for data analysis of metaproteomic data

Parameter/analysis type	Recommendation
Search algorithms	Using and combining multiple search algorithms
Sequence database size	As small and dedicated as possible
Enzyme cleavage	Trypsin and semi-tryptic (with high running times)
Maximum missed cleavages	Application between zero and two MC
Two-step searching	To be used with caution and stringent FDR filtering
False discovery rate	5 or 1% (for two-step searching)
Database sectioning	Alternative option with overhead of results combination

significant number of de novo sequencing peptides could be matched back to the original protein sequences. The mapping of de novo sequencing peptides to database search results showed that a PepNovo+ score of 100 corresponds to the results obtained by applying a 5% FDR threshold on peptides from database searching. In addition, the number of matched peptides increased only slightly when no PepNovo+ scoring was applied. On the other hand, mapping de novo sequencing peptides with a PepNovo+ score > 50 against an in silico digest of HIMPdb recovered a much higher number of peptides in comparison to the mapping with a PepNovo+ score > 100. As a consequence, the method of matching de novo sequencing peptides against the in silico digested database may provide valuable peptide identifications hidden in the database search results. In order to improve the reliability of these peptide identifications, a rescoring method, e.g. based on RMIC values, could be performed afterward.

Nowadays, with modern computer architectures, de novo sequencing is a powerful and cost-effective method, and researchers may further benefit from advances in MS technologies in line with algorithm development. At the moment, de novo sequencing already qualifies as complementary approach to common database searching and supports the validation of borderline peptide identifications from database searching. Importantly, while metagenome sequences may be difficult to obtain, this approach allows for the identification of peptides even if the protein sequence information is incomplete or not available [9]. Consequently, certain changes across systems could be determined by frequently identified marker peptides. Furthermore, de novo sequencing may already perform even better than database search algorithms in certain situations. For example, as homologous proteins from different taxonomic origins are present in most metaproteomics samples, database search algorithms tend to favor certain protein identifications by assigning probabilities, while de novo sequencing merely relies on the information present in the spectrum. However, it should be noted that the essential step of mapping peptides to protein sequences is not performed by de novo sequencing but software exists for mapping de novo peptide sequences to proteins [53, 54]. Nevertheless, the protein inference problem is highly challenging in metaproteomics experiments as peptides can be linked to multiple proteins from different organisms [39]. Notably, spectral library searching represents an alternative approach for peptide and protein identification [55, 56]. This method relies on recorded high-quality mass spectra being used as references for matching experimental spectra. Due to the lack of reference libraries for our metaproteomic data, this method could not be evaluated here. However, it is clear that spectral library searching can be readily applied for metaproteomics as soon as such libraries become available.

Another crucial parameter in database searching represents PTMs and has not been addressed by our study. As little is known about expected PTMs in microbial communities, it is nearly impossible to study them with generalized proteomics methods. Therefore, a detailed study on PTMs remains for further evaluation in the future.

We limited our results solely on presenting the number of PSMs and peptides and did not further infer taxonomic or functional information. The next step would be to explore how much of this meta-information can be gained when searching against database subsets and varying the evaluated parameters. Also, the alternative option of de novo sequencing and its results may be interesting for experiments in this context.

Our results have to be regarded with respect to the evaluated datasets which depended on type of sample, sample preparation, LC-MS setup, used search engines, and other data analysis parameters. The range of pipeline solutions for human metaproteome research is reviewed elsewhere [10].

4.3 Conclusion

After a decade of explorative metaproteomic studies applying various peptide identification strategies it is about time to refine our methods (Table 6). In particular, besides the use of different sample preparation strategies [10, 57], heterogeneous data analysis strongly impairs cross-study comparison. We showed here that selection of search algorithm, protein sequence database, and search parameters each affected the quantity and quality of identified peptides pointing to the specific challenges of data analysis in metaproteomics research. The benchmark experiment of concatenating *P. furiosus* sequences to unrelated protein sequences revealed a critical issue with the commonly used target-decoy approach for FDR estimation: valuable peptide and consequently, protein identifications may be missed when the protein search space is artificially increased. On the other hand, a two-step searching method decreased the database size and resulted in an

increased identification rate with the risk of increasing the number of false-positive identifications. In contrast, it is not possible to exclude certain identifications in real metaproteomics experiments without a priori knowledge about the sample composition. Therefore, the phenomenon of finding a compromise between sensitivity and specificity is put to the extreme in large metaproteomics databases. It could well be the case that databases cover a larger fraction of the studied proteomes than present identification ratios are implying. Low identification ratios may therefore also be explained by limits in the sensitivity of the applied search strategy and not only by missing target sequences in the protein sequence database. As protein databases will grow even more rapidly, search algorithms and statistical validation strategies may need a refinement. For example, the target-decoy approach holds pitfalls and should be carefully applied, particularly in metaproteomics and proteogenomics. Some metaproteomics studies yet again neglect FDR estimations by filtering on absolute search algorithm scores. However, a robust validation is advisable due to the number of potential false-positive identifications [58]. In our experiments, we searched separately against target and decoy databases in order not to favor target identifications in a concatenated search which may introduce even more bias in FDR estimation.

Our results highlight that searches against subsets of a large database may be useful to increase the number of identifications. Furthermore, our results indicate that using multiple search engines increases significantly the amount of identifications in metaproteomics workflows. Combining the results of different cleavage enzymes may bring valuable peptide identifications to light and may serve as cross-validation: the assumption would be that peptides identified multiple times with different settings are not obtained purely by chance. However, each additional search increases CPU time and efforts in combining and interpreting the results. In our opinion, the preferred strategy would be to perform a cost–benefit analysis in order to achieve the highest sensitivity and specificity for the search results. Therefore, the most appropriate database size and the optimal setup with respect to time efficiency need to be estimated by pilot experiments before the actual processing and analysis are carried out.

The presented issues in metaproteomic data analysis cannot be ignored and need to be further addressed by analytical scientists in close cooperation with bioinformaticians and statisticians. Finally, the integration of orthogonal information from other research fields such as metagenomics and 16S rRNA sequencing is important to increase the confidence in metaproteomic results.

Outi Immonen and Nahaison Krips are thanked for technical assistance, Airi Palva for generous help with conducting the study, and Mark de Been for sharing the HIMPdb. Funding was provided by the Doctoral Programme in Food Chain and Health, the Finnish Academy of Science (grants 141130, 137389, and 141140), the European Research ouncil (grant 250172-MicrobesInside), the European Commission 7th Framework Programme (grant 262067– PRIME-XS), and Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks").

The authors have declared no conflict of interest.

5 References

[1] Vaudel, M., Sickmann, A., Martens, L., Current methods for global proteome identification. *Expert Rev. Proteomics* 2012, *9*, 519–532.

[2] Allmer, J., Algorithms for the de novo sequencing of peptides from tandem mass spectra. *Expert Rev. Proteomics* 2011, *8*, 645–657.

[3] Nesvizhskii, A. I., A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* 2010, *73*, 2092–2123.

[4] Tharakan, R., Edwards, N., Graham, D. R., Data maximization by multipass analysis of protein mass spectra. *Proteomics* 2010, *10*, 1160–1171.

[5] Colaert, N., Degroeve, S., Helsens, K., Martens, L., Analysis of the resolution limitations of peptide identification algorithms. *J. Proteome Res* 2011, *10*, 5555–5561.

[6] Yadav, A. K., Kumar, D., Dash, D., Learning from decoys to improve the sensitivity and specificity of proteomics database search results. *PLoS One* 2012, *7*, e50651.

[7] Vaudel, M., Burkhart, J. M., Sickmann, A., Martens, L., Zahedi, R. P., Peptide identification quality control. *Proteomics* 2011, *11*, 2105–2114.

[8] Muth, T., Benndorf, D., Reichl, U., Rapp, E., Martens, L., Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. *Mol. Biosyst.* 2013, *9*, 578–585.

[9] Cantarel, B. L., Erickson, A. R., VerBerkmoes, N. C., Erickson, B. K. et al., Strategies for metagenomic-guided whole-community proteomics of complex microbial environments. *PLoS One* 2011, *6*, e27173.

[10] Kolmeder, C. A., de Vos, W. M., Metaproteomics of our microbiome - developing insight in function and activity in man and model systems. *J. Proteomics* 2014, *97*, 3–16.

[11] Tanca, A., Palomba, A., Deligios, M., Cubeddu, T. et al., Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture. *PLoS One* 2013, *8*, e82981.

[12] Jagtap, P., Goslinga, J., Kooren, J. A., McGowan, T. et al., A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics* 2013, *13*, 1352–1357.

[13] Hansen, S. H., Stensballe, A., Nielsen, P. H., Herbst, F. A., Metaproteomics: evaluation of protein extraction from activated sludge. *Proteomics* 2014, *14*, 2535–2539.

[14] Chapman, B., Bellgard, M., High-throughput parallel proteogenomics: a bacterial case study. *Proteomics* 2014, *14*, 2780–2789.

[15] Jagtap, P. D., Johnson, J. E., Onsongo, G., Sadler, F. W. et al., Flexible and accessible workflows for improved proteogenomic analysis using the Galaxy Framework. *J. Proteome Res.* 2014, *13*, 5898–5908.

[16] Rooijers, K., Kolmeder, C., Juste, C., Doré, J. et al., An iterative workflow for mining the human intestinal metaproteome. *BMC Genom.* 2011, *12*, 6.

[17] Cain, J. A., Solis, N., Cordwell, S. J., Beyond gene expression: the impact of protein post-translational modifications in bacteria. *J. Proteomics* 2014, *97*, 265–286.

[18] Searle, B. C., Turner, M., Nesvizhskii, A. I., Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *J. Proteome Res.* 2008, *7*, 245–253.

[19] Jones, A. R., Siepen, J. A., Hubbard, S. J., Paton, N. W., Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics* 2009, *9*, 1220–1229.

[20] Shteynberg, D., Nesvizhskii, A. I., Moritz, R. L., Deutsch, E. W., Combining results of multiple search engines in proteomics. *Mol. Cell. Proteomics* 2013, *12*, 2383–2393.

[21] Elias, J. E., Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 2007, *4*, 207–214.

[22] Granholm, V., Käll, L., Quality assessments of peptide-spectrum matches in shotgun proteomics. *Proteomics* 2011, *11*, 1086–1093.

[23] Blakeley, P., Overton, I. M., Hubbard, S. J., Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J. Proteome Res.* 2012, *11*, 5221–5234.

[24] Perez-Cobas, A. E., Gosalbes, M. J., Friedrichs, A., Knecht, H. et al., Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. *Gut* 2013, *62*, 1591–1601.

[25] Del Chierico, F., Petrucca, A., Mortera, S. L., Vernocchi, P. et al., A metaproteomic pipeline to identify newborn mouse gut phylotypes. *J. Proteomics* 2014, *97*, 17–26.

[26] Ng, C., DeMaere, M. Z., Williams, T. J., Lauro, F. M. et al., Metaproteogenomic analysis of a dominant green sulfur bacterium from Ace Lake, Antarctica. *ISME J.* 2010, *4*, 1002–1019.

[27] Kolmeder, C. A., de Been, M., Nikkilä, J., Ritamo, I. et al., Comparative metaproteomics and diversity analysis of human intestinal microbiota testifies for its temporal stability and expression of core functions. *Plos One* 2012, *7*, e29913.

[28] Vaudel, M., Burkhart, J. M., Breiter, D., Zahedi, R. P. et al., A complex standard for protein identification, designed by evolution. *J. Proteome Res.* 2012, *11*, 5065–5071.

[29] Käll, L., Storey, J. D., Noble, W. S., Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics* 2008, *24*, i42-i48.

[30] Verdam, F. J., Fuentes, S., de Jonge, C., Zoetendal, E. G. et al., Human intestinal microbiota composition is associated with local and systemic inflammation in obesity. *Obesity (Silver Spring)* 2013, *21*, E607–E615.

[31] Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V., Mann, M., In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* 2006, *1*, 2856–2860.

[32] Verhees, C. H., Kengen, S. W., Tuininga, J. E., Schut, G. J. et al., The unique features of glycolytic pathways in Archaea. *Biochem. J.* 2003, *375*, 231–246.

[33] Ettema, T. J., de Vos, W. M., van der Oost, J., Discovering novel biology by in silico archaeology. *Nat. Rev. Microbiol.* 2005, *3*, 859–869.

[34] Craig, R., Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004, *20*, 1466–1467.

[35] Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L. et al., Open mass spectrometry search algorithm. *J. Proteome Res.* 2004, *3*, 958–964.

[36] Qin, J., Li, R., Raes, J., Arumugam, M. et al., A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010, *464*, 59–65.

[37] Fermin, D., Allen, B. B., Blackwell, T. W., Menon, R. et al., Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol.* 2006, *7*, R35.

[38] Sennels, L., Bukowski-Wills, J. C., Rappsilber, J., Improved results in proteomics by use of local and peptide-class specific false discovery rates. *BMC Bioinformatics* 2009, *10*, 179.

[39] Nesvizhskii, A. I., Aebersold, R., Interpretation of shotgun proteomic data - The protein inference problem. *Mol. Cell. Proteomics* 2005, *4*, 1419–1440.

[40] Martens, L., Hermjakob, H., Proteomics data validation: why all must provide data. *Mol. Biosyst.* 2007, *3*, 518–522.

[41] Muth, T., Weilnböck, L., Rapp, E., Huber, C. G. et al., DeNovoGUI: an open source graphical user interface for de novo sequencing of tandem mass spectra. *J. Proteome Res.* 2014, *13*, 1143–1146.

[42] Frank, A., Pevzner, P., PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* 2005, *77*, 964–973.

[43] Liu, M., Fan, L., Zhong, L., Kjelleberg, S., Thomas, T., Metaproteogenomic analysis of a community of sponge symbionts. *ISME J.* 2012, *6*, 1515–1525.

[44] Daniel, H., Moghaddas Gholami, A., Berry, D., Desmarchelier, C. et al., High-fat diet alters gut microbiota physiology in mice. *ISME J.* 2014, *8*, 295–308.

[45] Ossipova, E., Fenyo, D., Eriksson, J., Optimizing search conditions for the mass fingerprint-based identification of proteins. *Proteomics* 2006, *6*, 2079–2085.

[46] Stead, D. A., Preece, A., Brown, A. J., Universal metrics for quality assessment of protein identifications by mass spectrometry. *Mol. Cell Proteomics* 2006, *5*, 1205–1211.

[47] Sowell, S. M., Abraham, P. E., Shah, M., Verberkmoes, N. C. et al., Environmental proteomics of microbial plankton in a highly productive coastal upwelling system. *ISME J.* 2011, *5*, 856–865.

[48] Lichtman, J. S., Marcobal, A., Sonnenburg, J. L., Elias, J. E., Host-centric proteomics of stool: a novel strategy focused on intestinal responses to the gut microbiota. *Mol. Cell Proteomics* 2013, *12*, 3310–3318.

[49] Cooper, B., The problem with peptide presumption and the downfall of target-decoy false discovery rates. *Anal. Chem.* 2012, *84*, 9663–9667.

[50] Wang, G., Wu, W. W., Zhang, Z., Masilamani, S., Shen, R. F., Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. *Anal. Chem.* 2009, *81*, 146–159.

[51] Gupta, N., Bandeira, N., Keich, U., Pevzner, P. A., Target-decoy approach and false discovery rate: when things may go wrong. *J. Am. Soc. Mass Spectrom.* 2011, *22*, 1111–1120.

[52] Chalkley, R. J., When target-decoy false discovery rate estimations are inaccurate and how to spot instances. *J. Proteome Res.* 2013, *12*, 1062–1064.

[53] Shevchenko, A., Sunyaev, S., Loboda, A., Bork, P. et al., Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal. Chem.* 2001, *73*, 1917–1926.

[54] Leprevost, F. V., Valente, R. H., Lima, D. B., Perales, J. et al., PepExplorer: a similarity-driven tool for analyzing de novo sequencing results. *Mol. Cell Proteomics* 2014, *13*, 2480–2489.

[55] Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K. et al., Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 2007, *7*, 655–667.

[56] Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K. et al., Building consensus spectral libraries for peptide identification in proteomics. *Nat. Methods* 2008, *5*, 873–875.

[57] Bastida, F., Hernandez, T., Garcia, C., Metaproteomics of soils from semiarid environment: functional and phylogenetic information obtained with different protein extraction methods. *J. Proteomics* 2014, *101*, 31–42.

[58] Adamski, M., Blackwell, T., Menon, R., Martens, L. et al., Data management and preliminary data analysis in the pilot phase of the HUPO Plasma Proteome Project. *Proteomics* 2005, *5*, 3246–3261.

