WHY DATABASE MATTERS: IMPACT OF SEQUENCE DATABASES ON METAPROTEOME ANALYSIS OF A MOCK MICROBIAL MIXTURE

Parco Scientifico e Tecnologico della Sardegna

PortoConte Ricerche

uniss UNIVERSITÀ DEGLI STUDI DI SASSARI

Alessandro Tanca[a*], Antonio Palomba[a,b], Massimo Deligios[a,b], Tiziana Cubeddu[a,$], Cristina Fraumene[a], Grazia Biosa[a], Daniela Pagnozzi[a], Maria Filippa Addis[a], Sergio Uzzau[a,b]

[a]Porto Conte Ricerche Srl, Tramariglio, Alghero, Italy, [b]Dipartimento di Scienze Biomediche, Università di Sassari, Sassari, Italy [$]Current affiliation: Dipartimento di Medicina Veterinaria, Università di Sassari, Sassari, Italy*
tanca@portocontericerche.it

1

INTRODUCTION

Metaproteomics allows the study of the protein repertoire expressed by complex microbial communities. Two major issues affect metaproteome analysis: first, genome sequence data might be unavailable for most of the species of the microbial community under study; second, a typical environmental sample contains thousands of proteins belonging to up to thousands of different microbial species making both peptide-to-protein and peptide-to-taxa assignments a huge task. In this context, the selection of proper protein databases (DBs) represents a key challenge, especially when dealing with poorly characterized microbiomes. Improvement in bioinformatic approaches, concerning data analysis and interpretation, are still needed. Here we aimed to assess the impact of different protein DBs on the metaproteomic investigation using a lab-assembled microbial mixture.

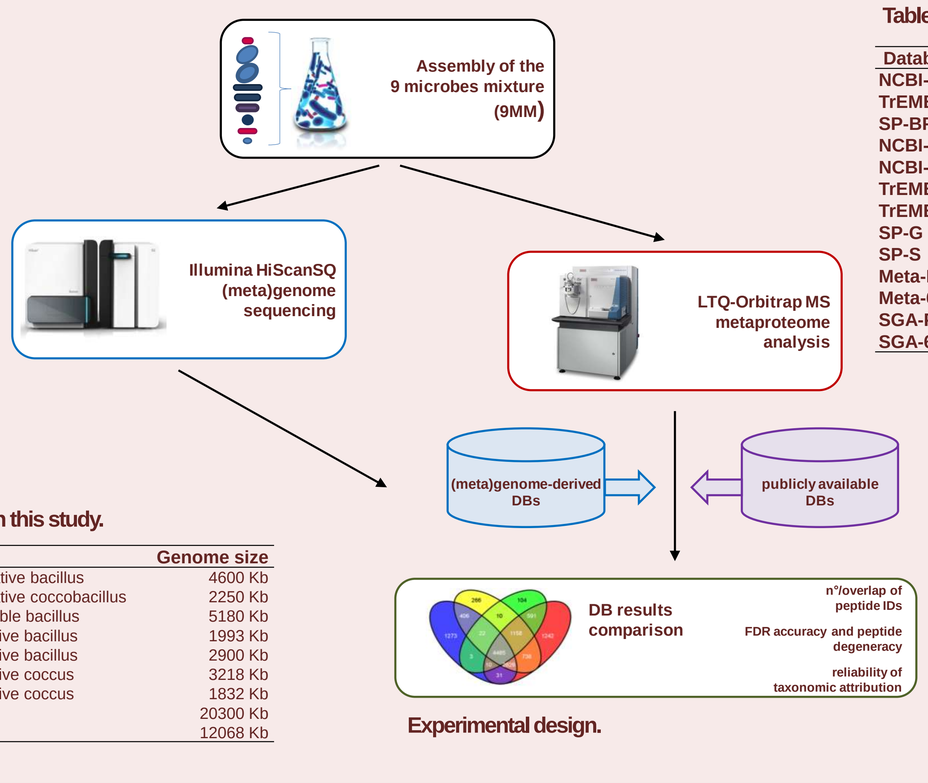
2

EXPERIMENTAL DESIGN

- A microbial mixture (9MM) including seven prokaryotes and two eukaryotes with heterogeneous structural features was assembled (Table 1).
- Genomes extracted from the 9 individual strains and from the 9MM were subjected to Illumina NGS.
- The 9MM metaproteome was analyzed by shotgun LTQ-Orbitrap MS.
- MS data were searched against publicly available and matched experimental DBs (Table 2).
- The information achieved was comparatively evaluated in respect to:
  - ✓number and overlap of peptide identifications;
  - ✓FDR behavior and peptide degeneracy;
  - ✓reliability of taxonomic attribution (using MEGAN and Unipept software).

Table 1. Microorganisms used in this study.

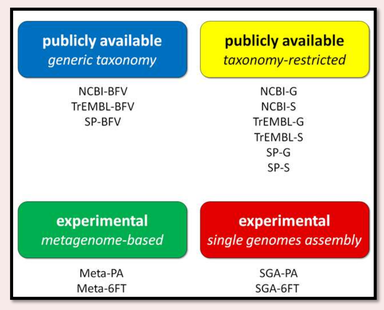
Species	Cell type	Genome size
Escherichia coli	Gram-negative bacillus	4600 Kb
Pasteurella multocida	Gram-negative coccobacillus	2250 Kb
Brevibacillus laterosporus	Gram-variable bacillus	5180 Kb
Lactobacillus acidophilus	Gram-positive bacillus	1993 Kb
Lactobacillus casei	Gram-positive bacillus	2900 Kb
Enterococcus faecalis	Gram-positive coccus	3218 Kb
Pediococcus pentosaceus	Gram-positive coccus	1832 Kb
Rhodotorula glutinis	Yeast	20300 Kb
Saccharomyces cerevisiae	Yeast	12068 Kb



Experimental design.

Table 2. Database used for peptide identification from MS spectra.

Database acronym	Original database	Taxonomy/ Processing
NCBI-BFV	NCBI	Bacteria, Fungi, Viruses
TrEMBL-BFV	UniProtKB/ TrEMBL	Bacteria, Fungi, Viruses
SP-BFV	UniProtKB/ Swiss-Prot	Bacteria, Fungi, Viruses
NCBI-G	NCBI	8 selected genera
NCBI-S	NCBI	9 selected species
TrEMBL-G	UniProtKB/ TrEMBL	8 selected genera
TrEMBL-S	UniProtKB/ TrEMBL	9 selected species
SP-G	UniProtKB/ Swiss-Prot	8 selected genera
SP-S	UniProtKB/ Swiss-Prot	9 selected species
Meta-PA	Matched metagenome	CDS prediction + TrEMBL annotation
Meta-6FT	Matched metagenome	six-frame translation
SGA-PA	Single genomes assembly	CDS prediction + TrEMBL annotation
SGA-6FT	Single genomes assembly	six-frame translation

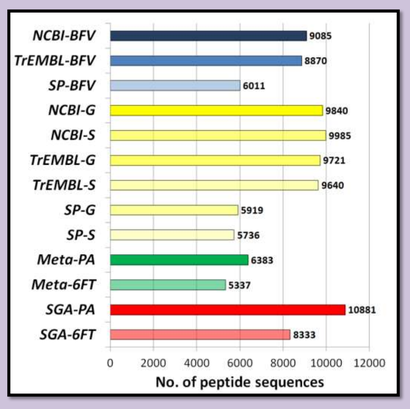


Database classes examined.

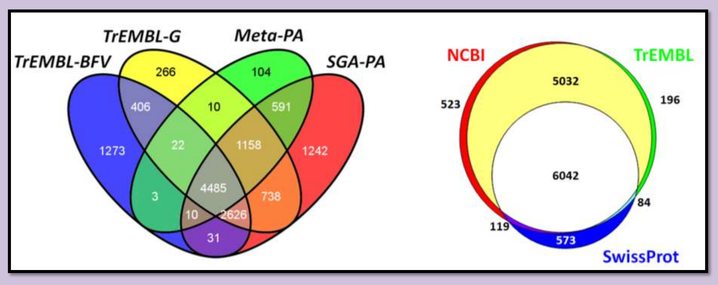
3

COMPARISON OF METAPROTEOMIC DATA OBTAINED USING DIFFERENT PROTEIN DBS

- SGA-PA led to the identification of the highest number of peptides and PSMs, while SwissProt-based DBs of the lowest.
- Taxonomy-restricted DBs from NCBI and TrEMBL performed better than the corresponding DBs with wider taxonomy.
- Genus/species-specific DBs provided up to 17% more identifications compared to DBs with generic taxonomy.
- Only 35% of peptides were common to all database classes.
- Approximately half of the peptides were common to NCBI, TrEMBL and SwissProt.
- About 8% of SwissProt peptide sequences (5% of the total) were not identified in the other DBs.



Comparison of metaproteomic data obtained with different DBs. Number of peptide sequences identified in the 9MM using different sequence DBs (FDR<1%).

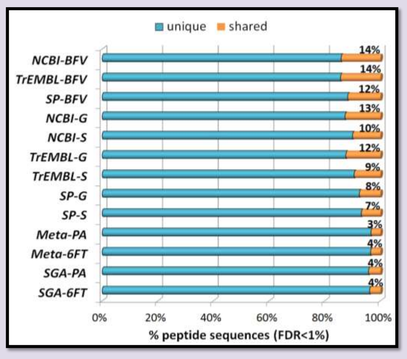


Overlap of metaproteomic results obtained with different DBs. Left, Venn diagram illustrating the peptide distribution among four different DB classes. Right, Venn diagram illustrating the peptide distribution among all NCBI-, TrEMBL- and SwissProt-based DBs used in this study.

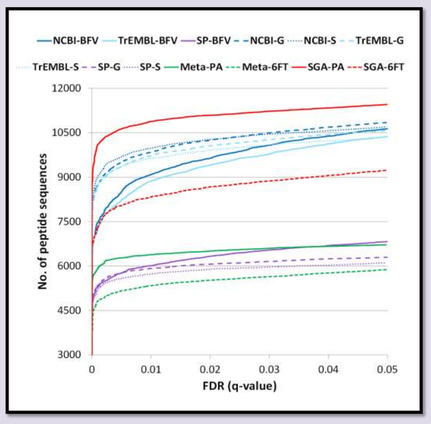
4

EVALUATION OF FDR BEHAVIOR AND PEPTIDE DEGENERACY

- Experimental DBs exhibited significantly lower percentages of shared peptides when compared to publicly available DBs.
- The peptide degeneracy decreased according to the following order: NCBI>TrEMBL>SwissProt and BFV>G>S.
- q-value curves of all publicly available DBs with generic taxonomy, kept on rising much longer compared to the remaining DBs.



Evaluation of peptide degeneracy using different DBs. Peptides identified with each DB at FDR<1%.



Evaluation of FDR behavior degeneracy using different DBs. Diagram plotting the number of peptides (left) and PSMs (right) identified with each database as a function of FDR thresholds based on the Percolator q-values.

5

RELIABILITY OF TAXONOMIC ATTRIBUTION BY UNIPEPT AND MEGAN

- Unipept demonstrated a higher reliability, since the average percentage of incorrect attributions was 3%, 5% and 9% (at the family, genus and species level) compared to respective percentages of 7%, 17% and 32% with MEGAN.
- Meta-PA provided the most specific results, due to the lowest rate of misassignments, whereas NCBI-BFV and TrEMBL-BFV performed worse in this respect.
- No yeast-specific peptides could be identified using Meta-PA, because of the total lack of eukaryotic sequences in this DB.
- The use of a threshold corresponding to 0.5% of the total number of taxon-specific peptides reduces false positive attributions.

Reliability of taxonomic attribution using Unipept and MEGAN. Taxonomic distribution of family specific peptides identified with different DBs, according to Unipept (left) or MEGAN (right) LCA analysis. Red rectangles illustrate misassignments, with indication of their percentage for each DB. Bacterial taxa are represented by shades of blue, whereas yeast taxa by shades of green.

6

CONCLUSIONS

- The parallel use of multiple DBs has to be encouraged, as different DB types can lead to highly complementary results.
- The use of iterative metaproteomic searches with DBs of decreasing size can be key to achieve a wider metaproteome coverage.
- Metagenomics can help investigate less characterized species.
- Software enabling LCA analysis of metaproteome data can provide reliable results even at the species level, but proper filters with specific thresholds have to be set to reduce false positive attributions.

7

REFERENCES

- Huson DH, Auch AF, Qi J, *et al.* (2007) **MEGAN analysis of metagenomic data.** Genome Res 17: 377-386.
- Mesuere B, Devreese B, Debyser G, *et al.* (2012) **Unipept: tryptic Peptide-based biodiversity analysis of metaproteome samples.** J Proteome Res 11: 5773-5780.
- Muth T, Benndorf D, Reichl U, *et al.* (2013) **Searching for a needle in a stack of needles: challenges in metaproteomics data analysis.** Mol Biosyst 9: 578.
- Tanca A, Palomba A, Deligios M, *et al.* (2013) **Evaluating the Impact of Different Sequence Databases on Metaproteome Analysis: Insights from a Lab-Assembled Microbial Mixture.** PLoS ONE 8(12): e82981.

In this web page you will find the links for downloading this and the other posters from our research group, as well as the slides of our oral presentation

http://www.portocontericerche.it/eventi/porto-conte-ricerche-al-congresso-mondiale-hupo