

UNTARGETED METAPROTEOMICS IS THE MOST RATIONAL APPROACH FOR FUNCTIONAL MICROBIOTA DISCOVERY & UNDERSTANDING

OPINION PIECE BY GILBERT SKORSKI
CHAIRMAN & CEO, PHYLOGENE

Introduction

The microbiota complexity has been discovered through untargeted sequencing approaches which showed us a complex new world. To understand how a holobiont works, untargeted approaches are available and essential as the traditional way of using targeted techniques to validate a hypothesis will ignore other parallel events which may impact a complex biological event.

Omics Landscape

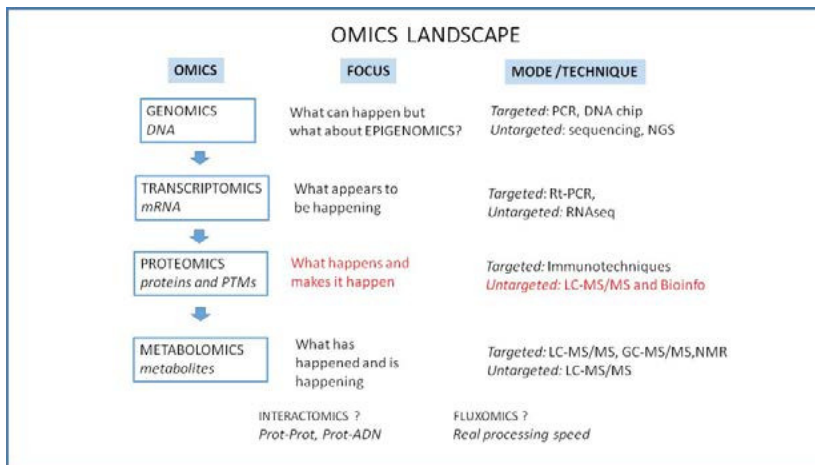
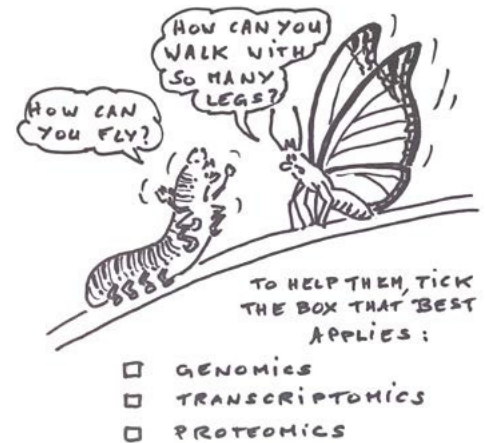


Fig1: Techniques are operational at different steps of the systems biology and provide different levels of information which are targeted specifically on analytes or only assign any signal to analytes through databases.

Knowledge Depends on Existing Analytical Techniques

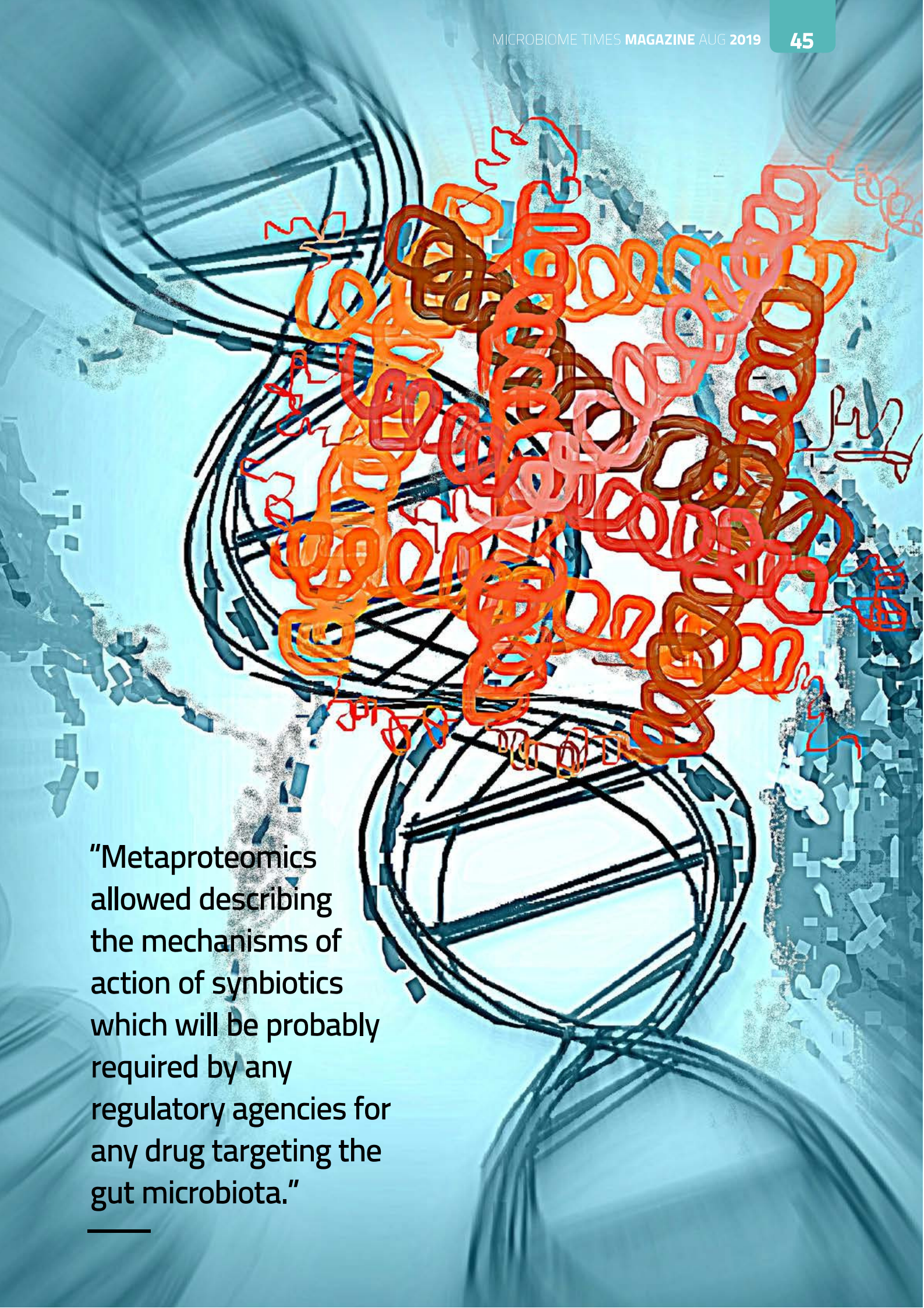
PCR was discovered around 1985 and allowed to detect specifically DNA strands. Since the years 1990, DNA sequencing techniques opened the door to untargeted approaches of genes. Now NGS allows sequencing a genome at low cost.



For over a century, mass spectrometry allowed detection of small chemical molecules. For 15 years now, mass spectrometry evolution allows to analyze and quantify proteins. The proteogenomics approach through the genes knowledge and with LC-MS/MS and databases evolutions allows identifying and quantifying proteins in an untargeted mode. Improvements in the mass accuracy, resolution, and sensitivity of MS instruments are enabling the rapid and reliable detection, identification, and quantification of proteins in complex mixtures.

These techniques are opening the door to improved methods for discovering disease-specific biomarkers with the potential to support early disease detection and even individualized therapies.

In an untargeted way (shotgun NGS) and quantitatively, a gene gives the information that the translated protein can structure or work in the cell, but we now know that epigenetics mechanisms may interfere or inactivate gene. Genomics stays reasonably at the informational level. (fig.1)



“Metaproteomics allowed describing the mechanisms of action of synbiotics which will be probably required by any regulatory agencies for any drug targeting the gut microbiota.”

The transcript (mRNA), RNAseq at the quantitative shotgun level, tells us a protein may be translated but if and when remains an open question.

Only LC-MS/MS proteomics tells us which protein is there

Of course, no technique is able in an untargeted way to tell us if the protein is active and at what speed it produces metabolites, even if the post-translational modifications (PTMs) can be now reached by mass spectrometry (1). Also, no untargeted technique is able to know which protein interact with which other molecule.

Metabolites are produced by proteins and may be metabolized by other proteins. Metabolites may be tracked using LC for the liquid part, GC for the volatile part coupled with MS/MS or NMR but it works rather in a targeted way as databases are poor up to now (2, 3): of course, there is no structural correspondence between the gene and the metabolite as there is between the gene and the protein.

Shotgun sequencing allows finding all the known genes in a sample, genes which account for around 2% of DNA, the remaining being the 'junk DNA' which is now known to be of primary importance. Based on genes sequences databases, LC-MS/MS proteomics can easily identify the functional proteome and may also reach the signaling proteins using sample pre-fractionation. Proteomics has also the 'dark proteome' which accounts for around half of LC-MS/MS spectra and corresponds to disordered protein regions but which seems to be essential for the proteome activity disordered regions.

Up to now, metabolomics remains limited in an untargeted mode as only 10% of spectra obtained in mass spectrometry can be identified (2).

Although transcriptomics, proteomics, and metabolomics all generally measure the products of gene expression, one should not necessarily expect exact quantitative correlation among them. Measuring the level of a transcript reflects the production rate of its protein product, but does not accurately predict the concentration or stability of the protein product. In fact, the correlation of mRNA abundances with their corresponding protein abundances, while reasonable for some core metabolic processes in some microbial systems, in general is poor or non-existent in most biological systems examined to date, suggesting proteomics data is likely more indicative of biological phenotype than mRNA. (4, 5)

What is LC-MS/MS untargeted quantitative metaproteomics?

Proteomics means the study of expressed proteins in a cell, a tissue, an organ or an organism at definite time and conditions. The proteome is the entire set of these proteins. Metaproteomics refers to all the proteins of the ecosystem, mainly the host and its microbiotas, named holobiont. An LC-MS/MS instrument analyses continuously the peptide fractions obtained from HPLC after enzymatic digestion of a protein extract. The mass spectrometer does every second an MS spectrum followed by further MS/MS fractionation on the most intense components. This information generated on the different proteins segments is compared to mass maps and spectra available in databases and it allows the identification of

proteins present in samples. This identification is made by using proteogenomics (6) approach which makes the bridge between genes and proteins. The peptide pikes area ratios gives a relative quantification.

Relative quantitative proteomics or metaproteomics refers to the analysis of two or more groups of samples by globally and quantitatively comparing their proteomes. Mass spectrometry has the unique ability to measure changes in complex protein mixtures.

As the identification is a process without targeting particular analytes, this is a hypothesis-free approach.

The analytical workflows are now simple and linear, in a first intent without need of 1D nor 2D prefractionation to reach 5000 proteins in a sample. One challenging difference of metaproteomics compared to proteomics remains the size of the identification databases. The human side contains 23000 genes, on which at least 1000 genes per microbial species are filled out. (Fig. 2)

Analytically, LC-MS/MS relative quantification of proteins is now a reliable technique (7) with a dynamic range of more than 5 logs, inter-assay CVs less than 20% (8) and Limit Of Detection at the attomole level. The main variability is brought, as for most techniques, by the sample protein extraction (9) which is minimized in relative quantification of same kind of samples.

Associated Bioinformatics

The current proteomics LC-MS/MS output data reach easily more than

5000 identified proteins. More than 10000 proteins can be reached easily in metaproteomics. The relative quantification of two conditions reveals tens to hundreds of under- or overexpressed proteins.

This requires heavy data processing pipelines (10) which will analyze all impacted proteins in correspondence with the metabolic pathways in which they are involved. It is so possible to understand the biological events induced by the disease or the dysbiosis.

Usually, a taxonomic analysis, a functional analysis by taxon (Homo sapiens, Fungi, Bacteria and Archae) and inter-functions correlations can be produced which give access to links between signaling/metabolic pathways, potential association of functions to particular taxons, potential interspecific relationships between microorganism and between microorganism and human (11). As a synthesis, the mechanism of action can be formulated (12).

Multi-omics as Complementarity Factors

Although untargeted LC-MS/MS proteomics data provides information about protein production, it cannot accurately predict a protein's activity or functional state. To achieve a more complete understanding of metabolic activities, it is interesting to integrate -omics data. It is important to realize that while DNA and proteins are relatively stable, transcripts and metabolites often have very short half-lives; thus, there are dramatic temporal information differences in these omics measurements. By integrating these large-scale datasets, cellular metabolism can be examined at an upgraded information level. For complex samples such as gut microbiomes, this integrated omics information has potential to provide a detailed molecular view at a higher resolution (4). For example, having the information of major impacted taxons after a 16srDNA profiling will focus the metaproteomics database queries on these taxons.

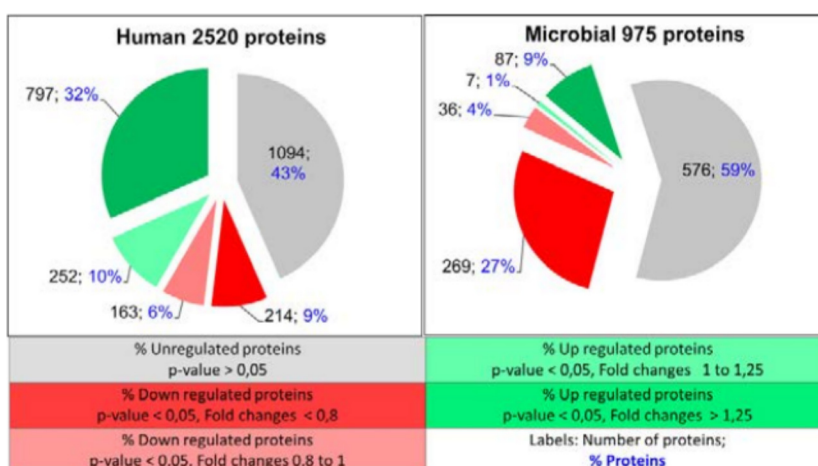


Fig 2. Proteins distribution for relative quantification Male/Female on skin: According to fold changes and p-values, statistical variation of abundances was shown for 1011 human proteins and 356 microbial proteins (including fungi and bacteria). Monneuse and all. ESDR 2017

Understanding the effects on host and microbiotas

Metagenomics studies initiated with the Human Microbiome or MetaHit projects brought huge knowledge about the individual variability of the microbiotas, the importance of the diversity and quantity of the gut microbiome at the taxonomic level, and relationship between a taxonomic state of a microbiota and diseases. Anyway a huge gap remains to understand the working of the holobiont such as the relations between microbiota and host, the impact of a diet, a probiotics, or a disease. The taxon, genus and species impacted by any dysbiosis and the 2 millions of identified genes in the gut microbiome does not tell much on what really occurs.

Already in 2012, the Human Microbiome Project (13) mentioned that the taxon variability was not in relation with the relative stability of functions at the genomic level.

Metagenomic studies have shown that gut microbiotas share a stable set of core functions, in spite of a large inter-individual structural/compositional variability. However, since sequenced genes are not necessarily expressed, metagenomics cannot provide reliable information on which microbial functional traits are actually changing in response to stimuli from host metabolism, immunity, neurobiology, diet, or other environmental factors which induce a substrate change. (12)

But this type of information can be gathered by functional metaproteomics, which displays higher sensitivity to perturbation and may therefore better reflect host-microbiome interactions (14) and mechanisms of action (15).

This way, several correlations were

identified between human and bacterial proteins (16) as both human and microbiota proteins are followed.

Using this approach on stool samples it is even possible to functionally relate human proteins to bacterial extracellular vesicles and to identify the peptides of interest reflecting taxonomic and/or pathway changes, which enables further biomarker discovery (17).

In another study on prebiotics and probiotics (synbiotics) diet in a obese mice model, it was shown first that high fat diet could induce a striking change in the functional activities of the gut ecosystem, which is characterized mainly by an increase in cell motility, and amino acid, carbohydrate and lipid metabolisms as well as a decrease in energy and nucleotide metabolisms. The symbiotic diet significantly reversed 11 KEGG pathways that are involved in carbohydrate, amino acid, and energy metabolisms, biosynthesis of other secondary metabolites, transcription, translation, replication and repair, as well as transport and catabolism (18). Metaproteomics allowed describing the mechanisms of action of synbiotics which will be probably required by any regulatory agencies for any drug targeting the gut microbiota.

In Crohn's disease, metaproteomics and dedicated bioinformatics were used to follow a patient in 4.5 years (11) or a cohort of patients after resection surgery for 1 year (19). Almost all the functions observed across all individuals were observed in multiple phyla, these functions are not specific to any one phylum, genus, or species. There is a clear persistence of conserved metabolic functions across time and individuals. Finally, the gut microbiome's metabolism is not driven by a set of discrete linear pathways but a web of interconnected reactions facilitated by a network of enzymes that connect multiple molecules across multiple pathways (19).

Conclusion

Due to the proteogenomics shortcut, metaproteomics and dedicated bioinformatics is the most rational and pragmatic approach for discovery and understanding at the functional level. Based now on mature mass spectrometry technique, it provides an unparalleled ratio cost on result which is proposed as a service in skilled labs. Then it can be reinforced with other omics or PTMs investigations, or confirmed by multiplex targeted quantification such as Selected Reaction Monitoring in mass spectrometry to follow a set of biomarkers.

Bibliography

- Olsen JV and all. Status of large-scale analysis of post-translational modifications by mass spectrometry. *Mol Cell Proteomics*. 2013 Dec;12(12):3444-52
- Frankel AE and all. Metagenomic Shotgun Sequencing and Unbiased Metabolomic Profiling Identify Specific Human Gut Microbiota and Metabolites Associated with Immune Checkpoint Therapy Efficacy in Melanoma Patients. *Neoplasia*. 2017 Oct;19(10):848-855
- Hamzeiy H and all. What computational non-targeted mass spectrometry-based metabolomics can gain from shotgun proteomics. *Curr Opin Biotechnol*. 2017 Feb;43:141-146
- Xiong W. and all. Development of an Enhanced Metaproteomic Approach for Deepening the Microbiome Characterization of the Human Infant Gut. *J. Proteome Res*. 2015; 14, 133-141
- Wang J and all. Proteome Profiling Outperforms Transcriptome Profiling for Coexpression Based Gene Function Prediction. *Mol Cell Proteomics*. 2017 Jan; 16(1):121-134
- Low TY and all. Connecting Proteomics to Next-Generation Sequencing: Proteogenomics and Its Current Applications in Biology. *Proteomics*. 2018 Nov 15:e1800235.
- Geyer P and all. Revisiting biomarker discovery by plasma proteomics. *Mol Syst Biol*. 2017; 13: 942
- Perrin R and all. Quantitative Label-Free Proteomics for Discovery of Biomarkers in Cerebrospinal Fluid: Assessment of Technical and Inter-Individual Variation. *PLoS ONE* 2013 8(5): e64314
- Piehowski P. and all. Sources of Technical Variability in Quantitative LC-MS Proteomics: Human Brain Tissue Sample Analysis. *J Proteome Res*. 2013 May 3; 12(5): 2128-2137.
- Hameury S and all. Prediction of skin anti-aging clinical benefits of an association of ingredients from marine and maritime origins: Ex vivo evaluation using a label-free quantitative proteomic and customized data processing approach. *J Cosmet Dermatol*. 2019 Feb;18(1):355-370
- Mills R and all. Evaluating Metagenomic Prediction of the Metaproteome in a 4.5-Year Study of a Patient with Crohn's Disease. *mSystems* 2019 Jan-Feb 4(1): e00337-18
- Starr A. and all. Proteomic and metaproteomic approaches to understand host-microbe interactions. *Anal Chem*. 2018 Jan 2;90(1):86-109.
- The Human Microbiome Project Consortium. Structure, Function and Diversity of the Healthy Human Microbiome. *Nature*. 2012 Jun 14; 486(7402): 207-214
- Tanca A ad all. Potential and active functions in the gut microbiota of a healthy human cohort. *Microbiome* (2017) 5:79
- Zybailov B. and all. Metaproteomics reveals potential mechanisms by which dietary resistant starch supplementation attenuates chronic kidney disease progression in rats. *PLoS One*. 2019 Jan 30;14(1):e0199274.
- Kolmeder CA and all. Faecal Metaproteomic Analysis Reveals a Personalized and Stable Functional Microbiome and Limited Effects of a Probiotic Intervention in Adults. *PLoS One*. 2016 Apr 12;11(4)
- Zhang X. Deep Metaproteomics Approach for the Study of Human Microbiomes. *Anal. Chem*. 2017, 89, 9407-9415
- Xinxin K. Synbiotic-driven improvement of metabolic disturbances is associated with changes in the gut microbiome in diet-induced obese mice *Mol Metab*. 2019 Apr;22:96-109
- Blakeley-Ruiz J. Metaproteomics reveals persistent and phylum-redundant metabolic functional stability in adult human gut microbiomes of Crohn's remission patients despite temporal variations in microbial taxa, genomes, and proteomes. *Microbiome* 2019:7:18