Untargeted LC-MS/MS proteomics and CORAVALID data processing for biodiscovery and biomarkers : an efficient workflow

Borderie L[(1)], Monneuse JM[(1)], Metton I[(1)], Skorski G[(1)(2)].

[(1)] **Phylogene, Bernis France.** http://ms.phylogene.com [(2)] Corresponding author (gskorski@phylogene.com)

Overview

•Purpose: Using High resolution mass spectrometry DIA LC-MS/MS experiments in study effects of pathology, stress or compounds , and demonstrating how high resolution/ accurate mass spectrometry is employed to enable wide scale proteomics studies.

•Methods: Epidermis samples from 6 **patients** (3 UV challenged and 3 control subjects) were collected, extracted and trypsinized using a membrane-based processing method (FASP). For generation of a spectral library, all samples were analyzed by DDA methods. DIA analysis was performed on the high resolution mass spectrometer MS to quantitatively map the epidermis proteomes.

•Results: In this study, we show that DIA method for the comprehensive mapping of proteome enables large scale proteomics studies with high throughput (less than 14 days for 12 biological sample measurements) and good reproducibility (media of CV% is 12 %). Our proprietary CORAVALID data processing allows a perfect fit with biological effect induced in this example. It reveals the most biological/mechanistic meaning from biological processes, molecular functions, cellular components, metabolic and signaling pathways, interactors, related transcription factors, protein domains involved. **With our method, not only we could determine that UV stress was active on the classical pathways, but also that the respiratory chain function was modified.**

Introduction

•Until now, the approach of choice for biomarker development was using various kinds of targeted methods. In case of proteomics MS, they allowed to bypass some protein interferences and were more sensitive than global approaches, and especially multiplex. While effective, their development is long and expensive, they produce limited data and interpretation can be biased, as they could be missing important parts of the cell puzzle. Also, early targeting is at the risk of development failure and dramatic associated costs.

•Recently, upgrades of hardware for global approaches (nano LC, High Resolution MS) lifted sensitivity/specificity/cycle time limitations. Also, data independent acquisition (DIA) strategies provides a comprehensive and reproducible data collection for large-scale quantitative proteomics experiments. The analysis efficiency may also be improved by protein fractionation using devoted sample treatment (subcellular fractionation or PTMs enrichment) which allows identification of more unique proteins with a broader dynamic range. Furthermore, advances in data science allow for faster, more in-depth and accurate data processing, which was complicated before. Information gain is higher as all cell mechanisms are considered, improving the understanding of the cells.

In this regard, we present an improved workflow for biodiscovery and biomarkers: Relative quantitation by Label-free DIA Mass Spectrometry coupled to nano-Liquid Chromatography together with our CORAVALID™ data analysis process.

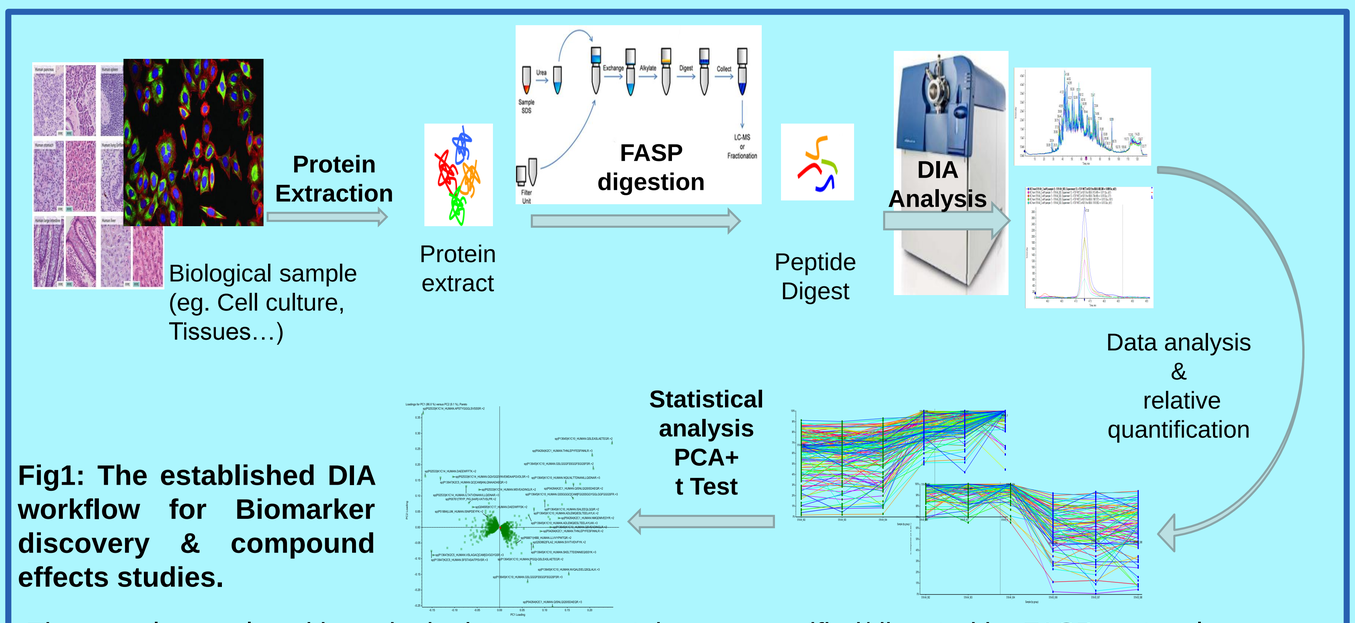
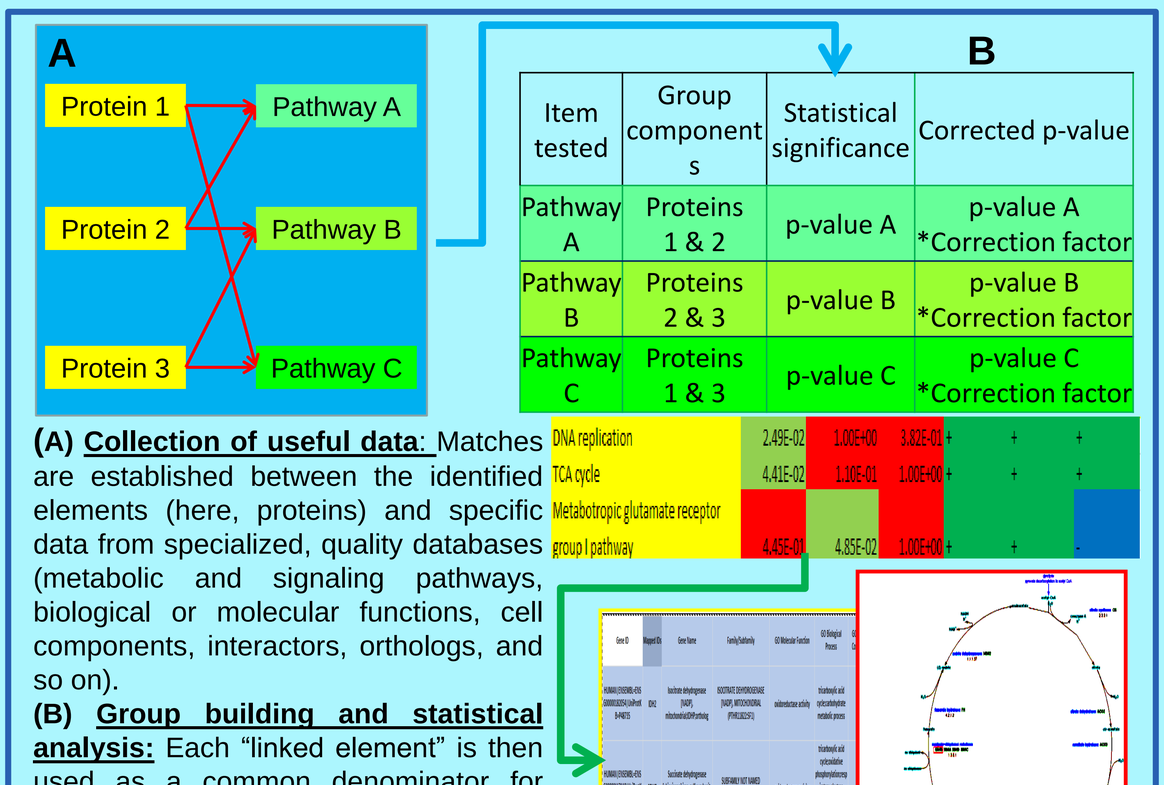


Fig1: The established DIA workflow for Biomarker discovery & compound effects studies.

The samples are lysed by anionic detergent, proteins were purified/digested by FASP protocol. Resulting peptides are analysed with LC-MS/MS. Aspectral library is first generated using DDA runs accumulation, then quantification is performed using DIA. Samples are compared using statistical data analysis methods (PCA, T-test).



(A) Collection of useful data: Matches are established between the identified elements (here, proteins) and specific data from specialized, quality databases (metabolic and signaling pathways, biological or molecular functions, cell components, interactors, orthologs, and so on).

(B) Group building and statistical analysis: Each "linked element" is then used as a common denominator for building groups of "associated proteins", which are then statistically tested for significance, allowing confirmation of the fact that such group existence is not random**.** As numerous tests are realized we also use multiple testing statistical corrections.

(C) Data Interpretation: Data is thoroughly analyzed and synthetized for obtain condensed and meaningful information.

Fig3: CORAVALID workflow

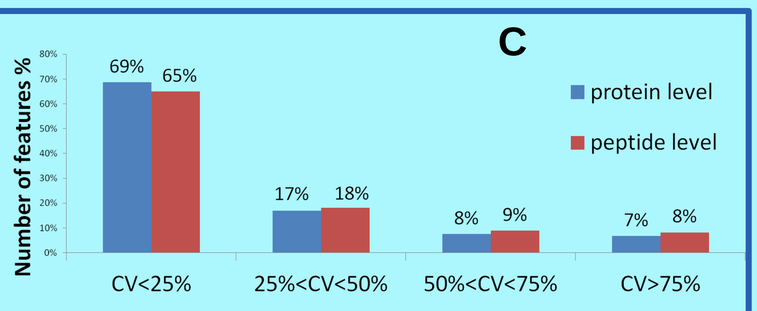
A

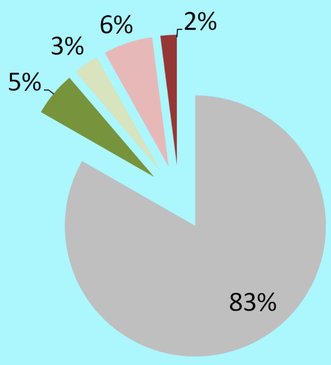
Sample	Protein Level (1% FDR)	Peptides level (1% FDR)
Epidermis Control	1067	7965
Epidermis UV	1095	8199

B

Spectral library generation

Spectral library	Protein Groups	Peptides	Fragment Ions	Sources
Epidermis	1257	7776	38441	12 DDA runs of epidermis proteomes





% Non-regulated proteins, p-value >0,05
% Down-regulated proteins, p-value <0,05, Fold change <0,5
% Down-regulated proteins, p-value <0,05, Fold change between 0,5 to 1
% Up-regulated proteins, p-value <0,05, Fold change between 1 to 2
% Up-regulated proteins, p-value <0,05, Fold change >2

Fig2: Analytical results:
(A) Identification results
(B) Spectral library
(C) Quantitation quality assessment
(D) Quantitation results

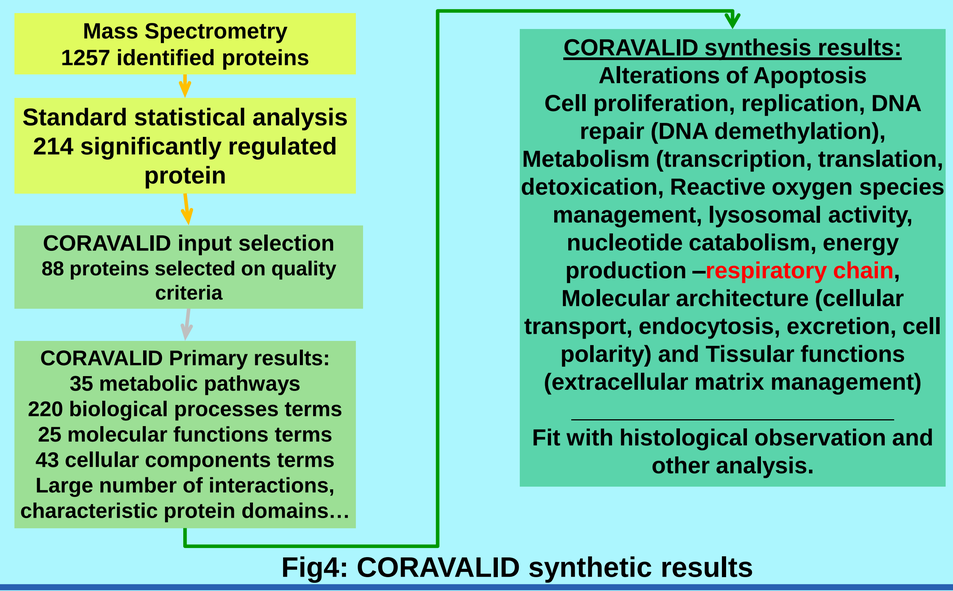


Fig4: CORAVALID synthetic results

Conclusion

Observing protein regulations occurring in our samples allows for a more thorough and clear investigation of the mechanisms at work.
Biasing an experiment by deciding even before the start to use some protein as a marker may encounter some problems:
- A marker identified at gene level may not be a good marker at protein levels, as there are known discrepancies between genomics/transcriptomics and proteomics.
- Protein function may not be documented enough to allow for correct interpretation of over- or under-expression depending on the context, since correlation of a marker to a phenotype is not necessarily explained. This is even worse when the protein context is missing.
- When using a marker alone, differences could not be revealed between phenotype, while a more complete proteomic profile enable to have a finer approach.
- Even when not impairing the main results, not having a global approach may hide concurrent phenomenon explaining the phenotype, thus delaying a lot its discovery. Finding it at the same time may allow to schedule and define experiments more efficiently.

Proteomics Nano LC- MS/MS coupled with CORAVALID™, our expert data processing workflow, allows for a complete view on the effects on the cell and is a very efficient tool for biodiscovery and biomarker analysis.

PHYLOGENE
62 RN 113 -30620 BERNIS – FRANCE http://ms.phylogene.com