Enhancing protein biomarker quantification in Cervicovaginal Fluid through Data Independent Acquisition proteomics approach

Hernandez M,[1] Saldivia P,[1] Nourdin G,[2] Stecher J,[2] Vargas C,[1] Koch E[1,2*]
1.biotechnology division, MELISA Institute, San Pedro de la Paz Chile
2. Fundacion de Investigacion San Ramon (FISAR), Concepcion, Chile
mail: ekoch@melisainstitute.org

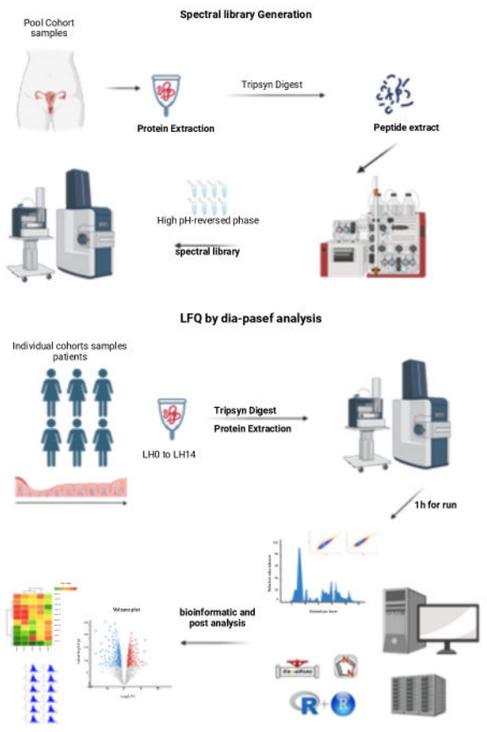
INTRODUCTION

Cervicovaginal Fluid (CVF) is a promising source of protein biomarkers, given its role in the homeostasis and immunity of the female lower genital tract. Comprising vaginal fluids, cervical mucus, and endometrial and oviductal fluids, CVF can be collected with minimal invasiveness, making it ideal for searching protein biomarkers of female reproductive diseases. However, its complexity presents a technical challenge for proteomic profiling the entire protein composition for biomarker discovery.

The objective of this work is to generate a robust and reproducible nLC-MS/MS method via dia-PASEF for the large-scale study of patient cohorts in cervicovaginal fluid. This is reflected in a comparative analysis using deep proteomics methodologies by offline fractionation, where a large number of missing values were observed, generating uncertainty at the level of relative quantification by LFQ.

METHODS

Six female patients were analyzed over seven days of their cycle, with ovulation determined by a surge in luteinizing hormone (LH) from day zero (LH0) to six (LH6). Each protein sample was digested. DP samples were fractionated and analyzed by TimsTOF in DDA mode, isolating the ten most abundant ions. Protein identification used MSFragger against a Uniprot *Homo sapiens* database. DIA samples were analyzed using a spectral library built from DP data with DIA-NN.



Workflow for spectral library generation and LFQ dia-pasef individual analysis

RESULTS

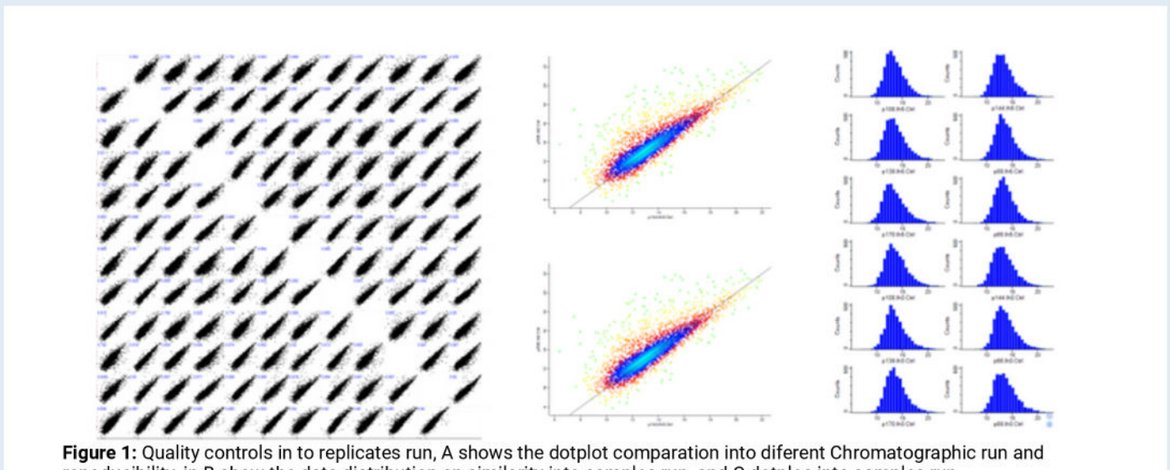


Figure 1: Quality controls in to replicates run, A shows the dotplot comparation into diferent Chromatographic run and repoducibility, in B show the data distribution an similarity into samples run, and C dotplos into samples run.

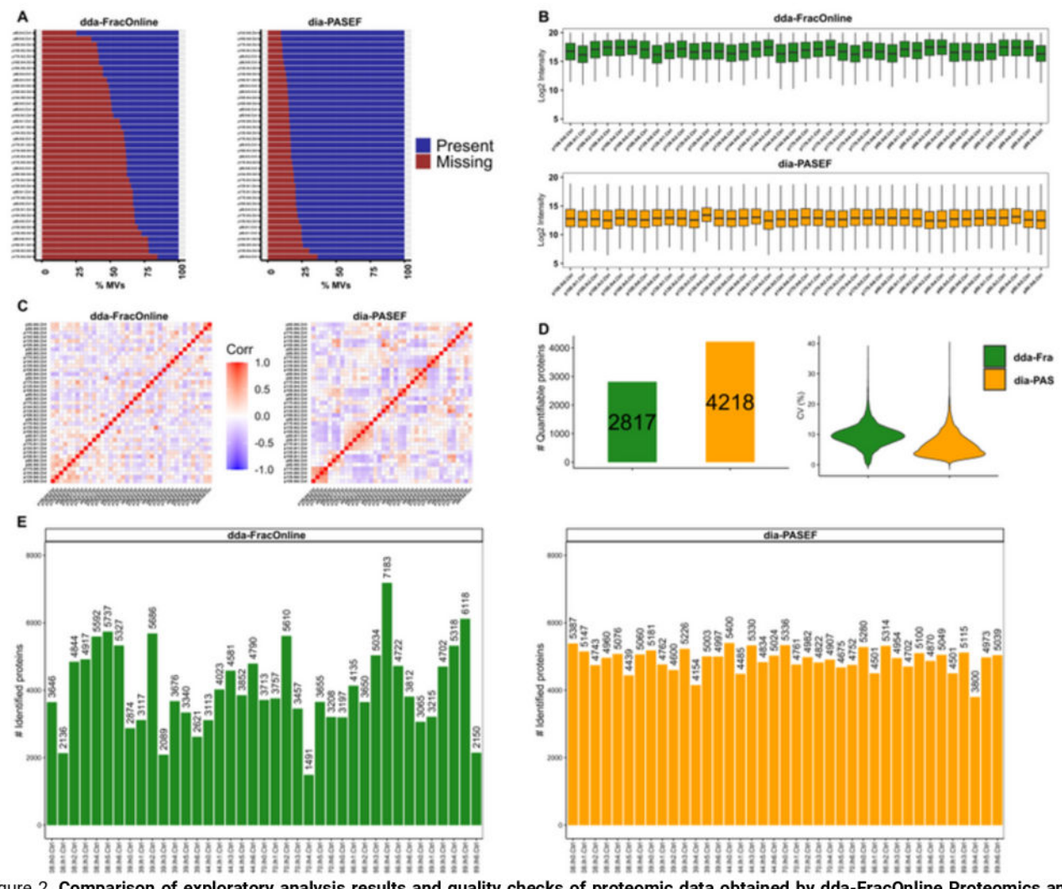


Figure 2. **Comparison of exploratory analysis results and quality checks of proteomic data obtained by dda-FracOnline Proteomics and dia-PASEF.** This analysis included a dataset from six patients over seven days, from LH0 to LH6. A) Missing Values Bar Chart: showed the percentage of missing values (MVs) in each sample, using blue for present values and red for absent values. B) Box Plots for Raw Data Variability: visualized the variability between samples, providing a direct representation of the dispersion in the raw data. C) Heat Map of Correlation between Chromatographic Runs: depicted the correlation between runs for each sample, where red and blue colors indicated positive and negative correlations, respectively. D) Visualization of Quantifiable Proteins and Variability: Quantifiable Protein Bar Chart: showed the number of proteins identified in each sample, differentiated by color: green for dda-FracOnline and orange for dia-PASEF. Coefficient of Variation (%CV) Violin Plot: complemented the previous graph by showing the variability of the measurements for each analytical strategy. E) Number of Protein Identifications per Sample: each bar in the graph represented a different sample and the number of proteins identified, with color coding distinguishing between dda-FracOnline (green) and dia-PASEF (orange).

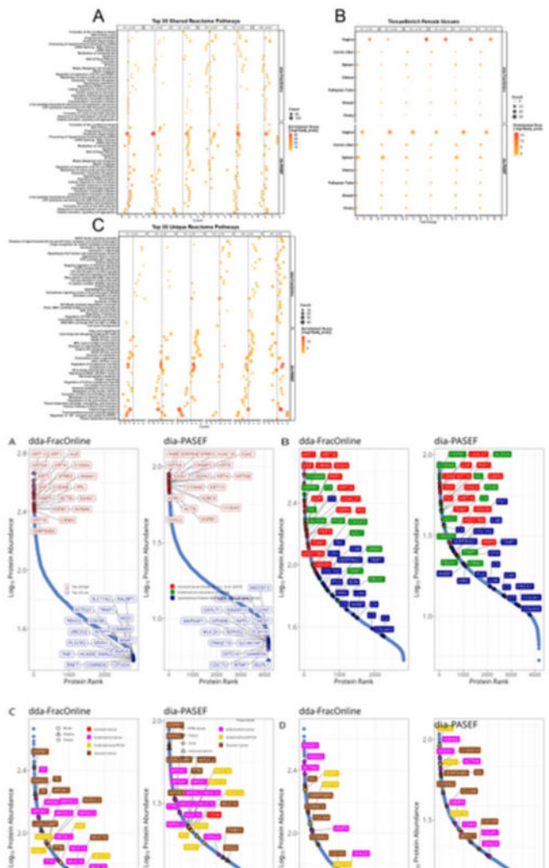
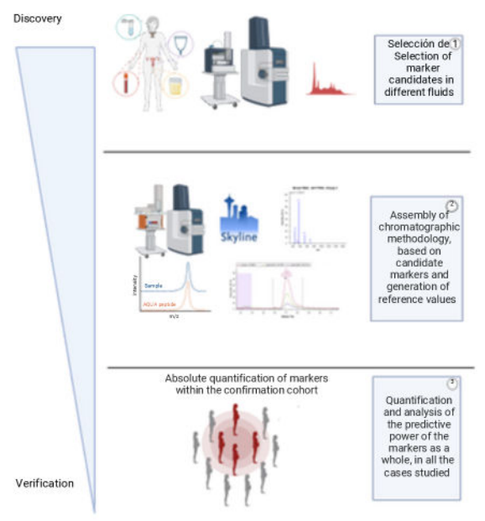


Figure 4. Top 30 most abundant biological pathways from Reactome and enriched female tissues using differentially expressed proteins (DEPs) in longitudinal studies with dda-FracOnline and dia- PASEF methods. Panels A and C show the Reactome enrichment analysis, highlighting the 30 most abundant biological pathways, both shared and unique between dda-FracOnline and dia-PASEF, respectively. Panel B displays the tissue enrichment analysis. Significantly enriched tissues were identified using proteins from the dda-FracOnline and dia-PASEF sets, including only female and immune tissues. For both analyses, the DEPs determined in the longitudinal comparisons within each dda-FracOnline and dia- PASEF set were used. For Reactome, the X-axis represents the activation coefficient (Z-score) of each pathway, while for TissueEnrich, the X-axis represents the fold change in expression of each tissue in the DEP set. The size of the points indicates the protein count (Count), and the color reflects the enrichment score (- log10(p.adjust)). Only female and immune tissues are shown in the TissueEnrich analysis.

Figure 5. Distribution of expression levels of markers associated with female reproductive pathologies identified in dda-FracOnline and dia-PASEF. Each graph displays the abundance of proteins (log10 mean intensity) as a function of the abundance range, ordering the proteins from lowest to highest abundance (Protein Rank) for each set of quantifiable proteins in dda-FracOnline and dia-PASEF. A) The 20 most abundant proteins (in red) and the 20 least abundant proteins (in blue) are shown. B) Markers described in cervical mucus (in red), endometriosis (in green), and spontaneous preterm birth (in blue) are highlighted according to the literature. C) Markers associated with pathologies described in other fluids are highlighted, with different shapes and colors indicating their origin: cervical cancer (in pink), endometrial cancer (in yellow), endometriosis/PCOS (in brown), and ovarian cancer (in dark brown). The shapes used are circle (blood), upward triangle (plasma), and downward triangle (serum). D) Markers with various shapes are highlighted: circle (FFPE tissue, Formalin-Fixed Paraffin-Embedded), triangle (tissue), diamond (urine), and upward triangle (tubal wash).

CONCLUSION

DIA's superior sensitivity and reproducibility make it suitable for biomarker quantification in CVF, enabling accurate analysis of complex samples. This method holds significant potential for early detection of female reproductive complications, particularly in longitudinal retrospective studies. Such validated biomarkers can be used in downstream clinical applications, including laboratory and immunoassay-based techniques. Selecting an appropriate proteomic strategy like DIA can enhance our ability to identify biomarkers for female reproductive pathologies, facilitating earlier interventions and treatments that are more effective. This study highlights the importance of advanced proteomic methods in maximizing CVF information for better clinical outcomes in female reproductive disorders.



Projections

This method will help to create discovery cohorts where marker candidates can be generated, in order to then develop other more targeted tools for absolute quantification such as PRM (Parallel Reaction Monitoring) where we can establish markers and their absolute concentration, in turn determining reference ranges, which allows us to personalize the diagnosis and monitoring of multiple pathologies. It is also important to highlight that this method can also be used for the analysis of other biofluids of clinical interest, which makes it a very versatile tool, with great clinical projections.

ACKNOWLEDGEMENTS

This work was funded by Fisar Chile, under the research fund