Identification of multi-omics biomarkers of preeclampsia

Tien Dung Pham[1], Ellen Menkhorst[2], Eva Dimitriadis[2], Kim-Anh Lê Cao[1]

[1] Melbourne Integrative Genomics & School of Mathematics & Statistics, University of Melbourne, Australia
[2] Gynaecology Research Centre, The Royal Women's Hospital, Australia

THE UNIVERSITY OF MELBOURNE

Interactive results
rebrand.ly/omics

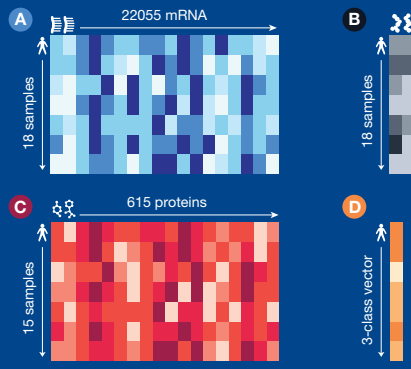
Background

- Preeclampsia is a pregnancy complication characterized by high blood pressure and can develop into more serious eclampsia seizures [1].
- Although past studies have identified some genetic pathways involving in preeclampsia development using single cell sequencing, few have focused on the biological differences between early and late onset of the condition in the trimester.
- Multivariate non-parametric computational methods can now integrate transcriptome with other omics data such as proteome. Such methods allow the discovery of correlated biomarkers across single-cell omics layers and uncover a complete picture of the condition as a dynamic biological system.
- This project aims to enrich results from edgeR Differential Expression (DE) Analysis [2] with methods based on Projection to Latent Structures (PLS) [3] to integrate transcriptome and proteome data and identify relevant biomarkers in different preeclampsia conditions.

An interactive result web page is available via the QR code above.

Data

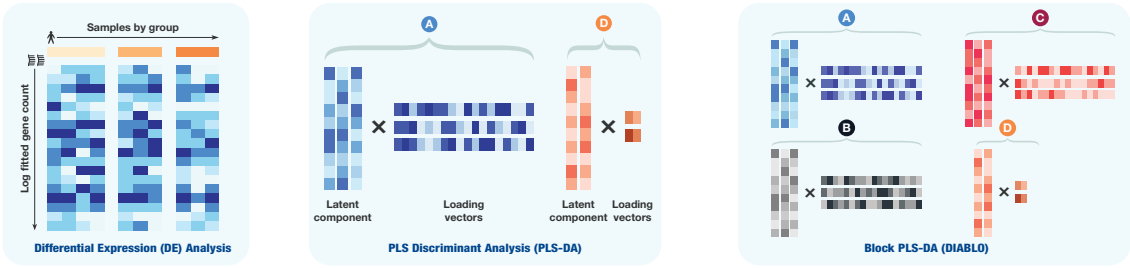
Illumina Novaseq and shotgun proteomics were used to sequence chorionic villus sampling from a cohort of 18 women to obtain (A) mRNA, (B) short RNA, and (C) proteomics count datasets. Of the samples, 8 had a healthy pregnancy, 4 experienced early onset (preterm) preeclampsia, and 6 experienced late onset (term) preeclampsia. This information was encoded in a class vector (D). Genes with low counts were removed and the datasets were then normalised using the TMM method [2].



Reference

[1] Redman, C. (2005). "Latest Advances in Understanding Preeclampsia". Science, 308(5728), 1592-1594. doi: 10.1126/science.1111726
[2] Robinson MD, McCarthy DJ, Smyth GK (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics, 26(1), 139-140.
[3] Rohart et al. (2017) "mixOmics: An R Package for 'omics Feature Selection and Multiple Data Integration." PLOS Computational Biology, 13(11).

Methods



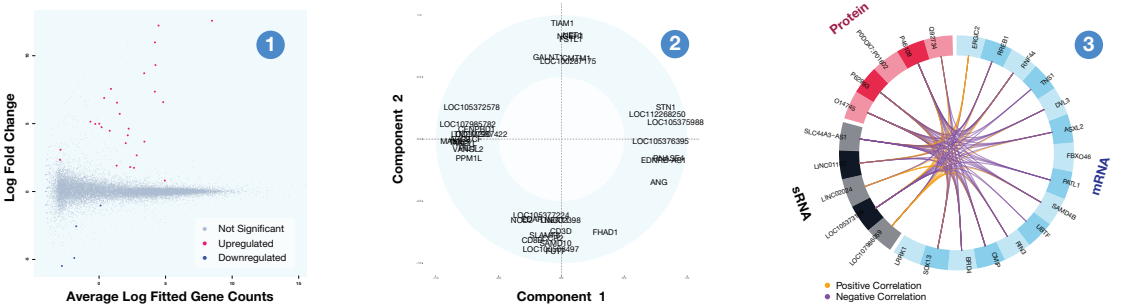
DE Analysis (Univariate)

- edgeR pipeline [2] fits a negative binomial distribution for each gene in (A) and (B).
- Pairwise comparisons between the three groups are then performed with a likelihood ratio test and adjusted for multiple testing using the Benjamini - Hochberg method.
- Significant genes are labelled upregulated if the log fold change is positive and downregulated otherwise.

Projection to Latent Structures - Discriminant Analysis (Multivariate)

- PLS-DA [3] decomposes the class vector (D) and each of (A), (B) and (C) into linear combinations of a latent component and a loading vector, using matrix factorisation, such that the covariance between the components is maximised.
- The number of components to decompose into and the number of variables per component were tuned by treating the output as a classification problem and cross-validating using leave-one-out schema.
- Its multi-omics variant, DIABLO [3], integrates all (A), (B), (C) and (D) in maximising covariance.

Results



(1) Mean-Difference plot highlights 31 significant differentially expressed mRNAs identified from edgeR, classified into upregulated genes and downregulated genes according to log fold change level.

(2) PLS-DA correlation plot shows that the identified mRNAs in each of first 2 latent components tend to have strong correlation (the interior and exterior circle corresponds to correlation levels of 0.5 and 1 respectively) and clustered well into separate subsets.

(3) DIABLO circos plot identifies a set of strongly correlated 15 mRNAs, 5 sRNAs and 5 proteins that maximises covariance. Correlations between omics are predominantly negative, implying potential patterns of expression in preeclampsia conditions.

Summary

- For single omics analysis, edgeR DE analysis and PLS-DA both result in similar number of mRNA markers, but edgeR exhibits lower sensitivity in identifying lowly expressed sRNAs and proteins.
- edgeR is a parametric univariate method, which does not extend well to intrinsically sparse datasets such as small RNA and proteomics. PLS-based tuning ensures that these lowly expressed omics are not undermined in the multivariate methods.
- Given the superior performance over edgeR in identifying novel sRNA and protein markers, this project illustrates how multi-omics methods can enhance data mining and hypothesis generation in the RNA-seq workflow.
- Small sample size results in low predictive generalisability and precludes removal of outliers. Future work should include sample size analysis to improve the statistical power of both methods.