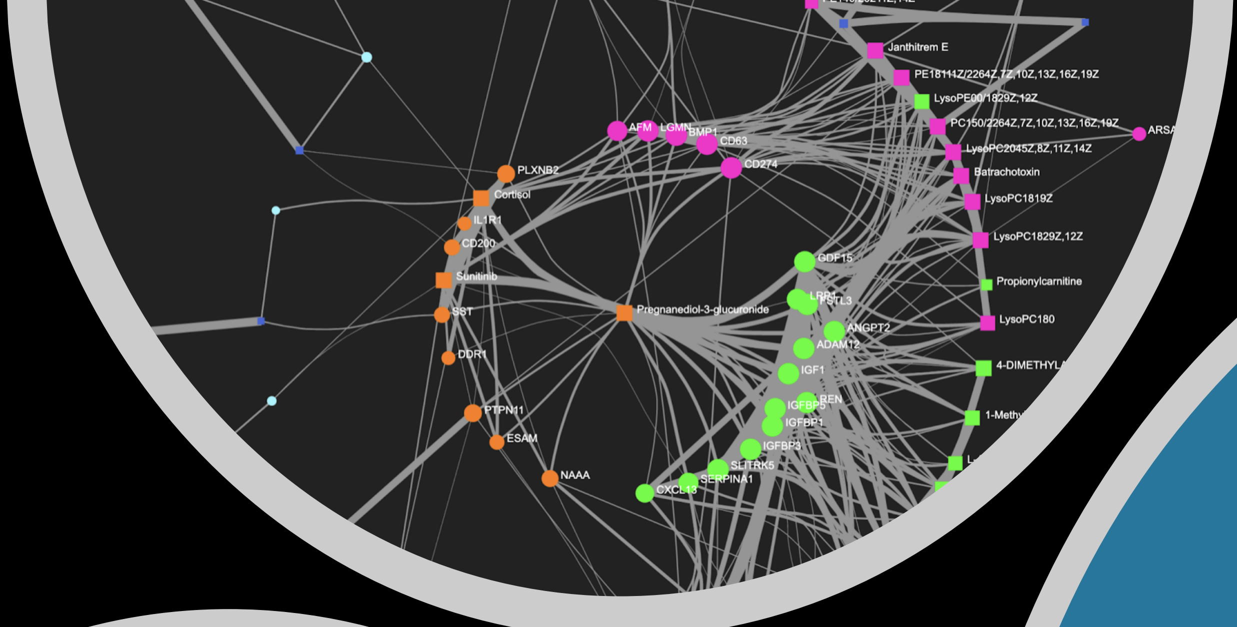
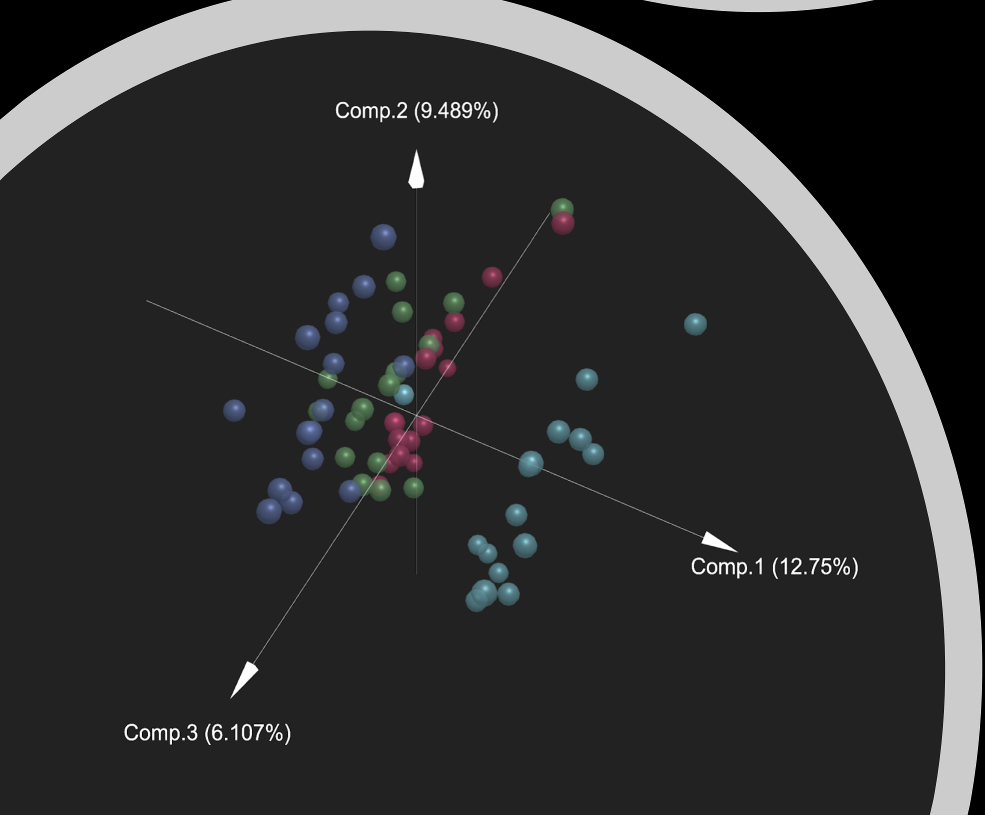
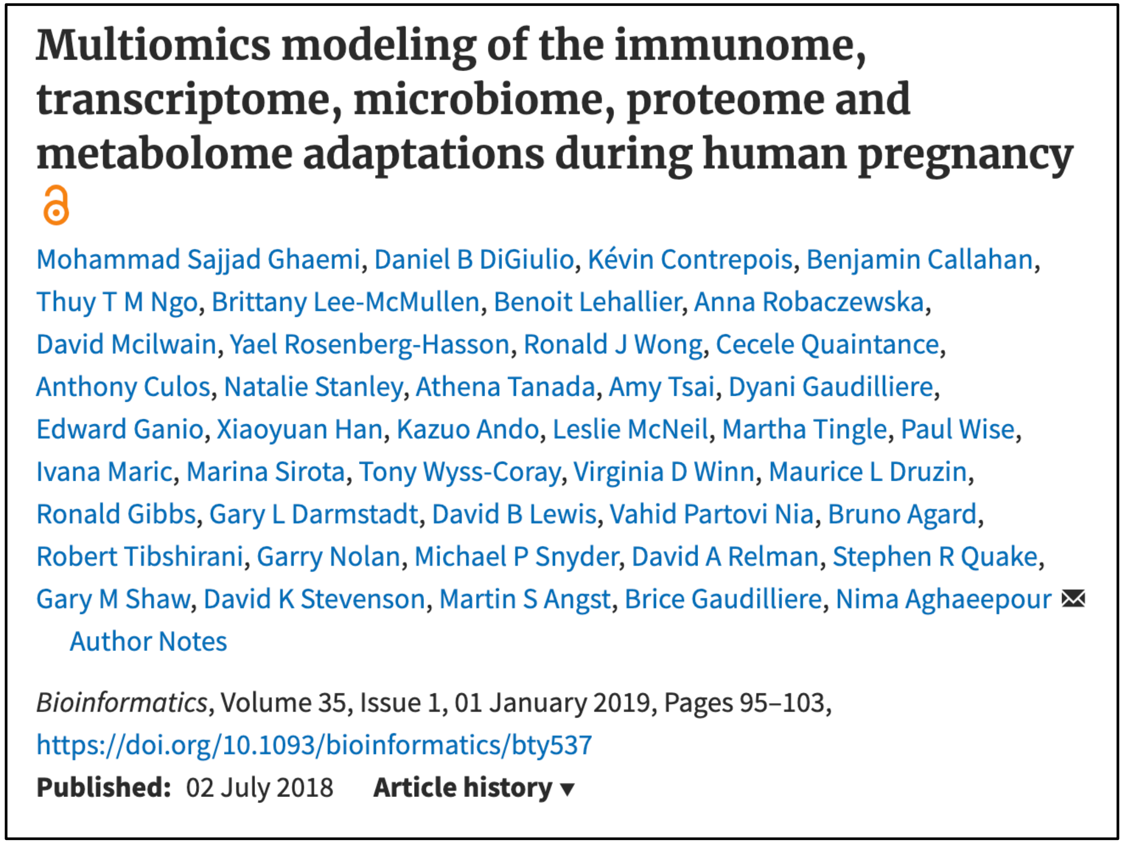




# Case Study: Multi-omics Analysis of Pregnancy

# Introduction



**Multiomics modeling of the immunome, transcriptome, microbiome, proteome and metabolome adaptations during human pregnancy**

Mohammad Sajjad Ghaemi, Daniel B DiGiulio, Kévin Contrepois, Benjamin Callahan, Thuy T M Ngo, Brittany Lee-McMullen, Benoit Lehallier, Anna Robaczewska, David Mcilwain, Yael Rosenberg-Hasson, Ronald J Wong, Cecele Quaintance, Anthony Culos, Natalie Stanley, Athena Tanada, Amy Tsai, Dyani Gaudilliere, Edward Ganio, Xiaoyuan Han, Kazuo Ando, Leslie McNeil, Martha Tingle, Paul Wise, Ivana Maric, Marina Sirota, Tony Wyss-Coray, Virginia D Winn, Maurice L Druzin, Ronald Gibbs, Gary L Darmstadt, David B Lewis, Vahid Partovi Nia, Bruno Agard, Robert Tibshirani, Garry Nolan, Michael P Snyder, David A Relman, Stephen R Quake, Gary M Shaw, David K Stevenson, Martin S Angst, Brice Gaudilliere, Nima Aghaeepour

Author Notes

*Bioinformatics*, Volume 35, Issue 1, 01 January 2019, Pages 95–103, https://doi.org/10.1093/bioinformatics/bty537

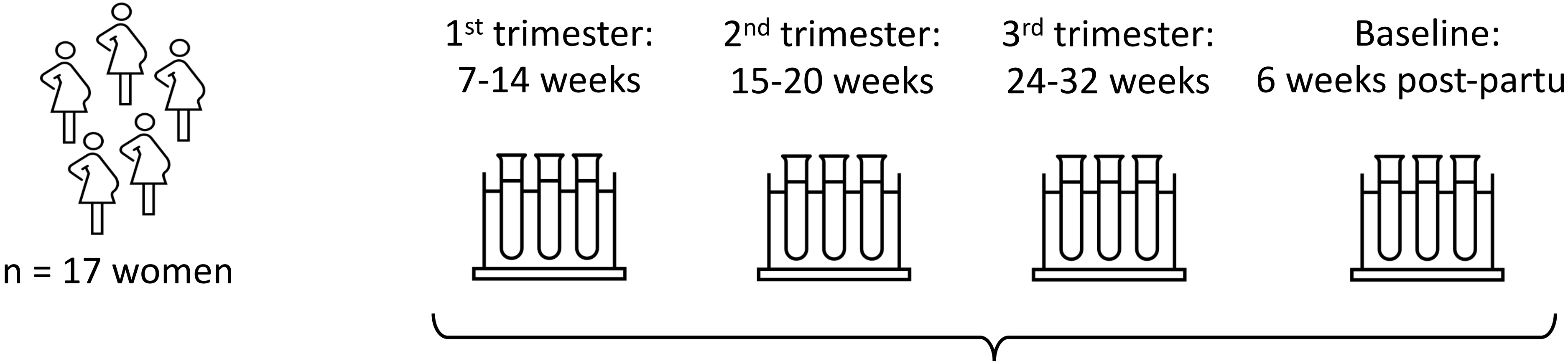
**Published:** 02 July 2018 **Article history** ▾

The goal of the original study was to find molecular markers that predict gestational age during normal pregnancy. If successful, these markers could be measured to predict whether a mother is at risk for a preterm birth. Seven types of ‘omics data were collected from pregnant women at four different time points. Samples were taken during the first, second, and third trimester, as well as six weeks post-partum to get baseline data.

In this case study, we re-analyze the metabolomics and proteomics data to demonstrate how the visual analytics tools in OmicsAnalyst can be used for data-driven multi-omics analysis and interpretation.

# Experimental Design

A total of 68 samples were collected from each ‘omics type (17 women * 4 collection points). All data are available in the supplementary materials of the original publication. We use the metabolomics and proteomics data for this case study (available as Example Data on OmicsAnalyst).



n = 17 women

1st trimester:
7-14 weeks

2nd trimester:
15-20 weeks

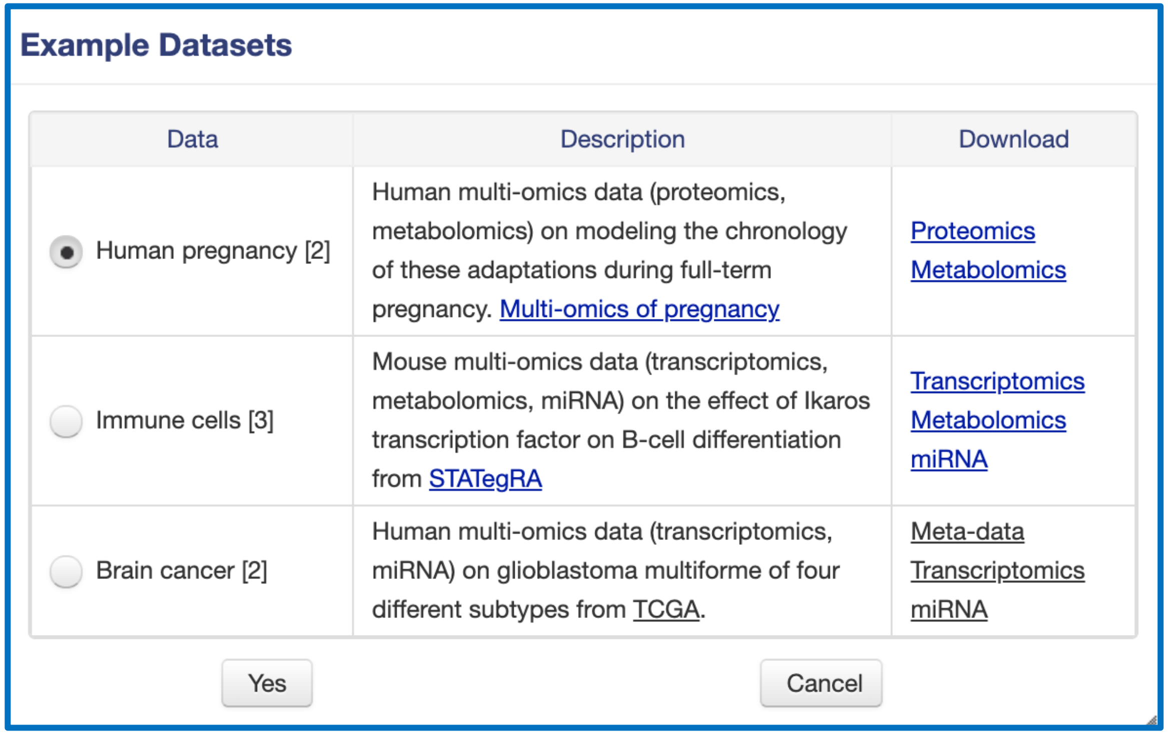
3rd trimester:
24-32 weeks

Baseline:
6 weeks post-partum

7 types of ‘omics data collected: cell-free RNA from plasma, cytokine levels from plasma, cytokine levels from serum, microbiome from vaginal swabs/stool/saliva/gum swabs , single-cell characterization of whole blood, **metabolomics from plasma**, and **proteomics from plasma**.

# Data Processing

All pre-processing steps prior to the visual analytics can be reproduced using the default settings loaded along with the example data.



**Example Datasets**

| Data | Description | Download |
|---|---|---|
| ● Human pregnancy [2] | Human multi-omics data (proteomics, metabolomics) on modeling the chronology of these adaptations during full-term pregnancy. Multi-omics of pregnancy | Proteomics<br>Metabolomics |
| ○ Immune cells [3] | Mouse multi-omics data (transcriptomics, metabolomics, miRNA) on the effect of Ikaros transcription factor on B-cell differentiation from STATegRA | Transcriptomics<br>Metabolomics<br>miRNA |
| ○ Brain cancer [2] | Human multi-omics data (transcriptomics, miRNA) on glioblastoma multiforme of four different subtypes from TCGA. | Meta-data<br>Transcriptomics<br>miRNA |

## Annotation

**Value type:** Continuous
**Data type:** Metabolomics; Proteomics
**Species:** H. Sapiens
**ID type:** Common Name; Official Gene Symbol

Specifying this option correctly is important because different normalization options should be used for continuous vs. reads count data.

## Missing values

This section is not important for this case study since our data do not contain missing values, but we leave the defaults. OmicsAnalyst enforces estimation of missing values because many of the downstream methods cannot handle missing values.

## Filtering

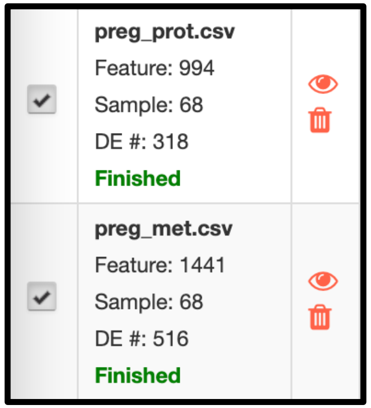
Same settings for both data:
**Variance filter:** 15
**Abundance filter:** 2 – Relative

Uninteresting (low variance) and unreliable (low abundance) features can be removed to increase statistical power and reduce noise in the results. Here, we choose low thresholds since our datasets are relatively small already (< 1500 features).

## Comparison

**Metadata of interest:** CLASS
**Data transformation:** None (already roughly normal distribution)
**Method:** T-tests/ANOVA
**Fold-change cutoff:** 1.0 |log2FC|
**P-value (FDR) cutoff:** 0.005

Cutoffs chosen to give roughly ~30% differential features in both datasets



☑ **preg_prot.csv**
Feature: 994
Sample: 68
DE #: 318
**Finished**

☑ **preg_met.csv**
Feature: 1441
Sample: 68
DE #: 516
**Finished**

# QA/QC and Scaling

The page after Data Processing provides some summary figures (density, PCA, and tSNE plots) so we can do some final QA/QC checks to make sure the processing and normalization have produced reasonable results, and that there are no extreme outliers. In addition, it is generally a good idea to re-scale each dataset so that they have roughly similar overall distributions and are more comparable to each other. Below we see the big difference that auto-scaling can make (Figure 1: before; Figure 2: after). Next, we proceed to the Method Selection page.

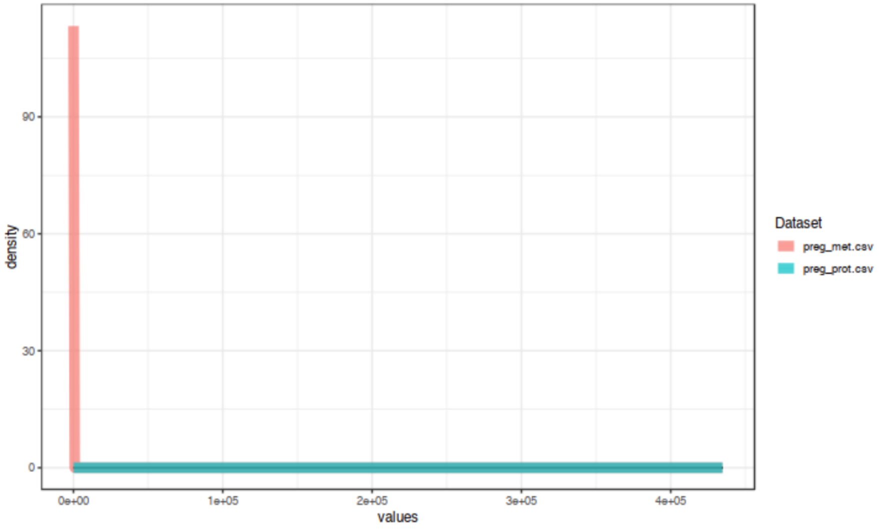


**Figure 1**

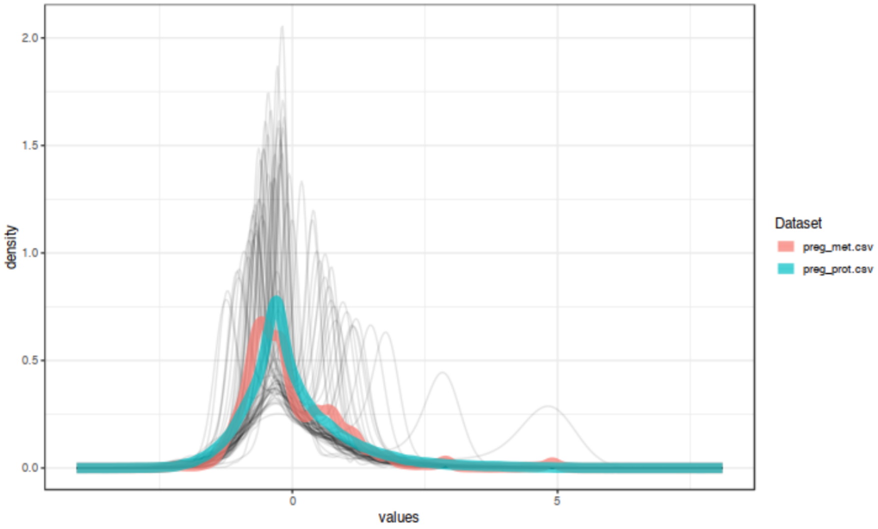
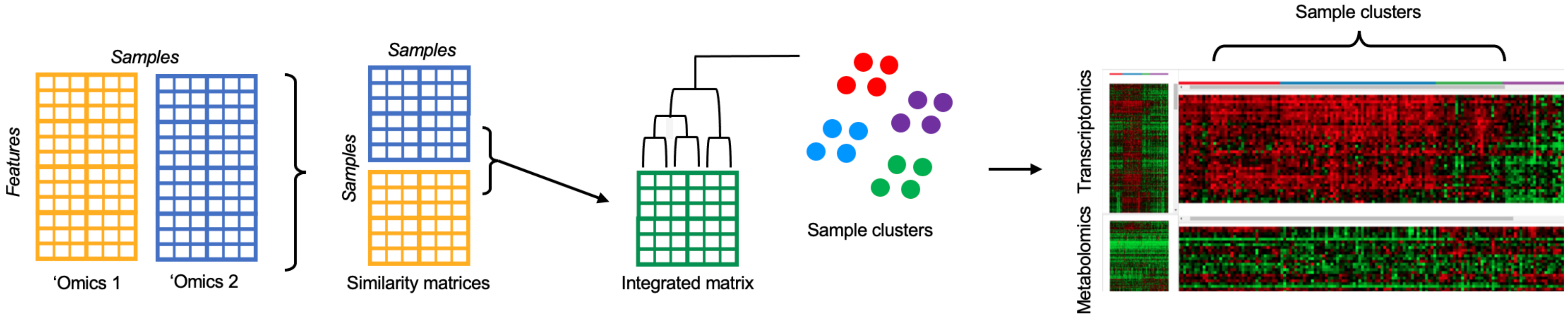


**Figure 2**

# Cluster Heatmap Analysis

In this case study, we start with the heatmap analysis track in "Free Exploration" mode to get a feel for single-omics patterns before moving on to integrating multiple datasets. The dual heatmap provides a quick and easy way to visually detect patterns present in each 'omics dataset and assess whether they are shared across datasets.



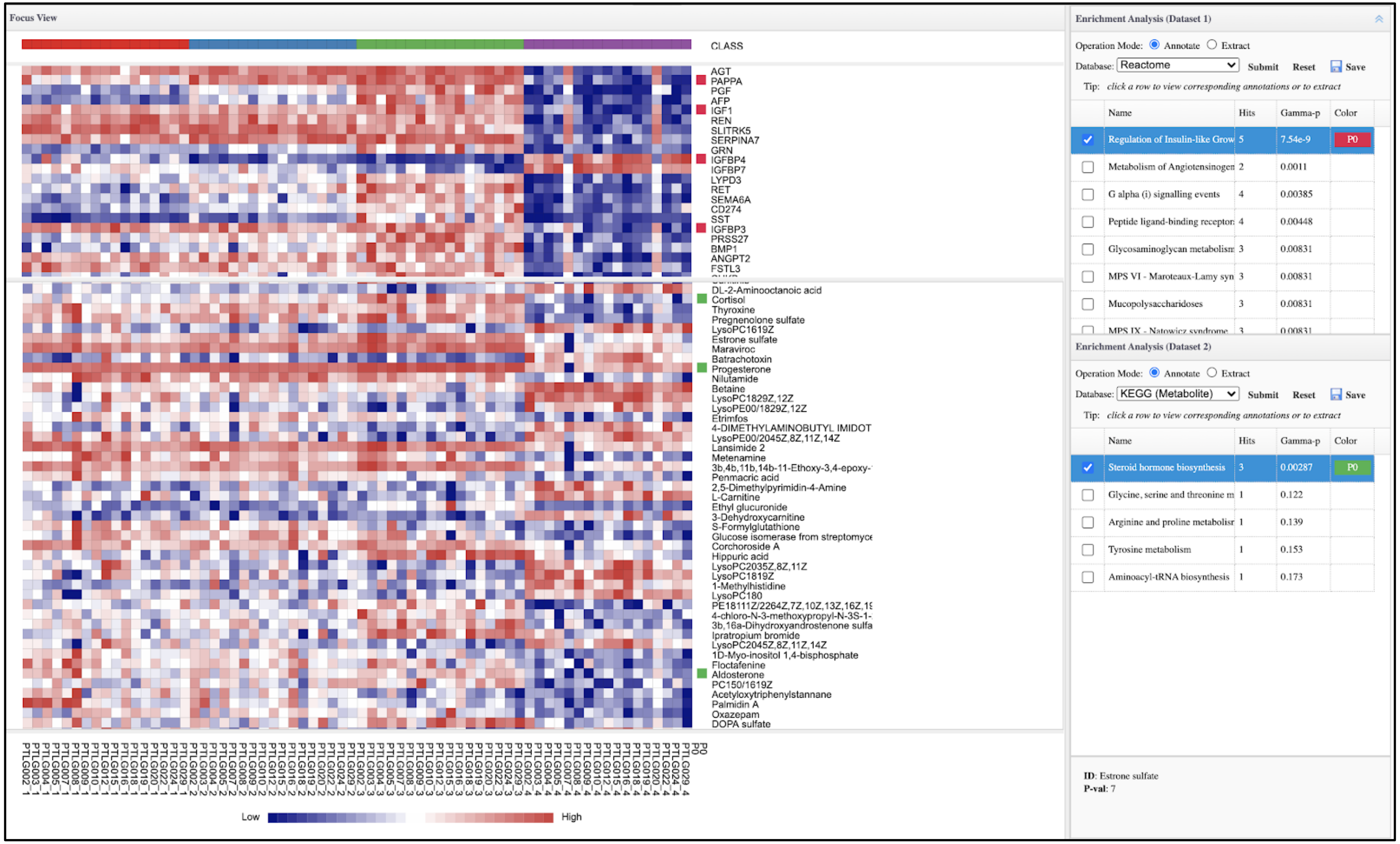
Compute similarity matrices for each 'omics

Integrate individual similarity matrices

Perform multi-view cluster detection

Interactive exploration of side-by-side heatmaps

We start by looking at heatmaps of the top 50 differential features from each ‘omics data. The top enriched pathway (Reactome) for proteomics is “Regulation of Insulin-like Growth Factor”, which is a system known to change throughout pregnancy: https://doi.org/10.1016/j.jcma.2013.07.004. The top enriched KEGG pathway for metabolites is “Steroid hormone biosynthesis”, which also makes sense since many hormones are increased during pregnancy. The heatmap shows that features from both pathways have lower levels in baseline samples compared to samples collected during pregnancy.



Enrichment Analysis (Dataset 1)

Operation Mode: ● Annotate ○ Extract

Database: Reactome Submit Reset Save

Tip: *click a row to view corresponding annotations or to extract*

| | Name | Hits | Gamma-p | Color |
|---|---|---|---|---|
| ☑ | Regulation of Insulin-like Grow | 5 | 7.54e-9 | P0 |
| □ | Metabolism of Angiotensinogen | 2 | 0.0011 | |
| □ | G alpha (i) signalling events | 4 | 0.00385 | |
| □ | Peptide ligand-binding receptor | 4 | 0.00448 | |
| □ | Glycosaminoglycan metabolism | 3 | 0.00831 | |
| □ | MPS VI - Maroteaux-Lamy syn | 3 | 0.00831 | |
| □ | Mucopolysaccharidoses | 3 | 0.00831 | |
| □ | MPS IX - Natowicz syndrome | 3 | 0.00831 | |

Enrichment Analysis (Dataset 2)

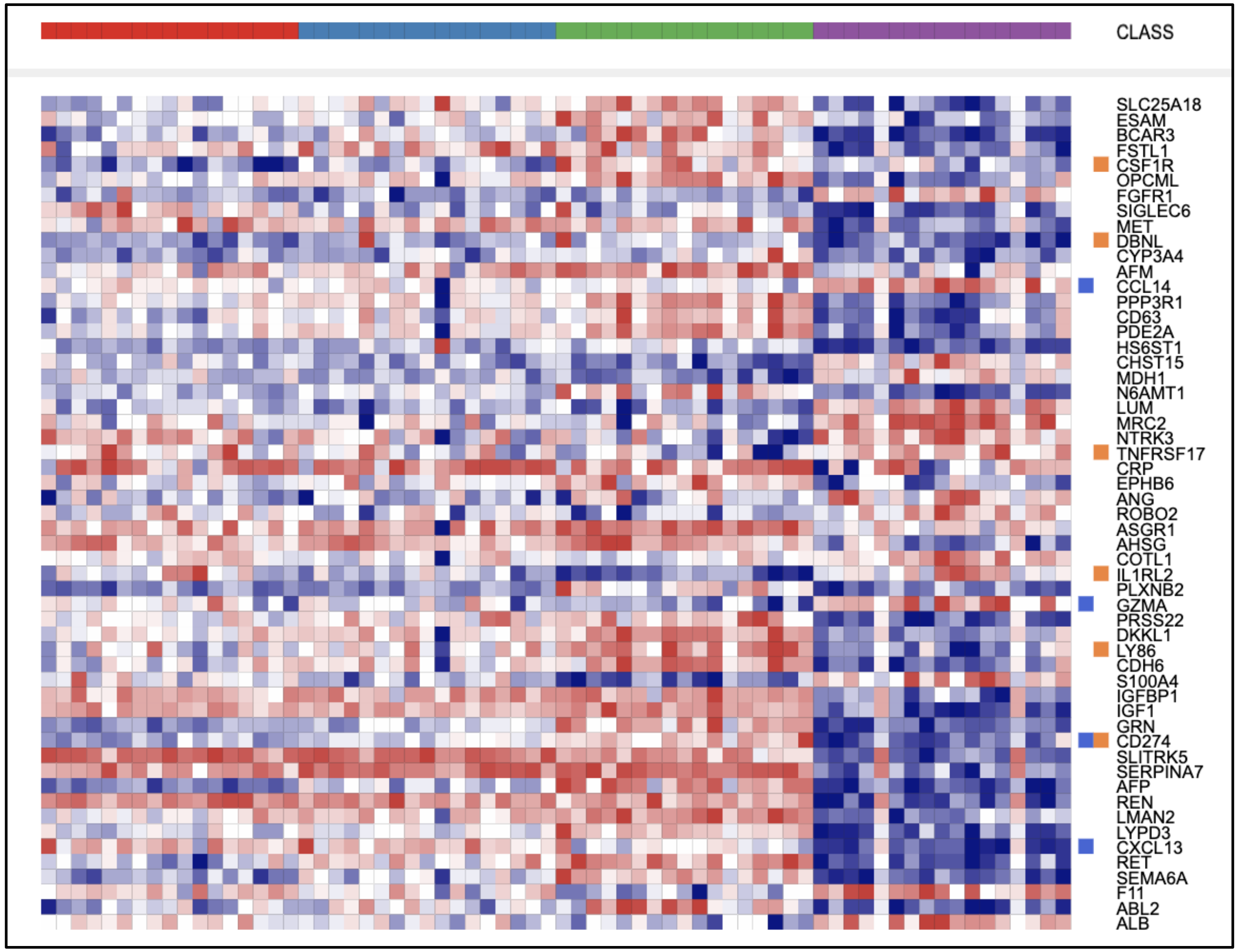
Operation Mode: ● Annotate ○ Extract

Database: KEGG (Metabolite) Submit Reset Save

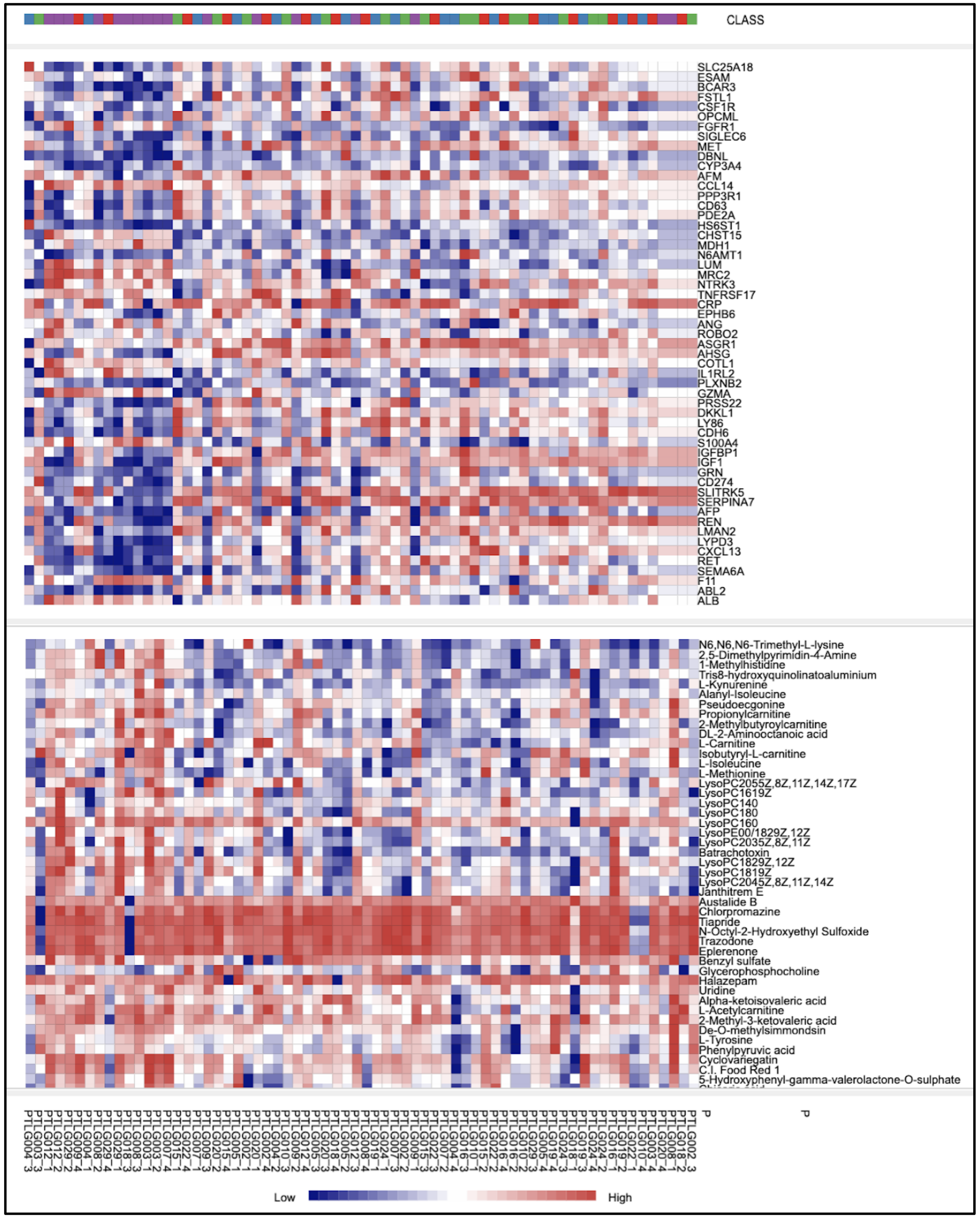
Tip: *click a row to view corresponding annotations or to extract*

| | Name | Hits | Gamma-p | Color |
|---|---|---|---|---|
| ☑ | Steroid hormone biosynthesis | 3 | 0.00287 | P0 |
| □ | Glycine, serine and threonine m | 1 | 0.122 | |
| □ | Arginine and proline metabolisr | 1 | 0.139 | |
| □ | Tyrosine metabolism | 1 | 0.153 | |
| □ | Aminoacyl-tRNA biosynthesis | 1 | 0.173 | |

**ID**: Estrone sulfate
**P-val**: 7



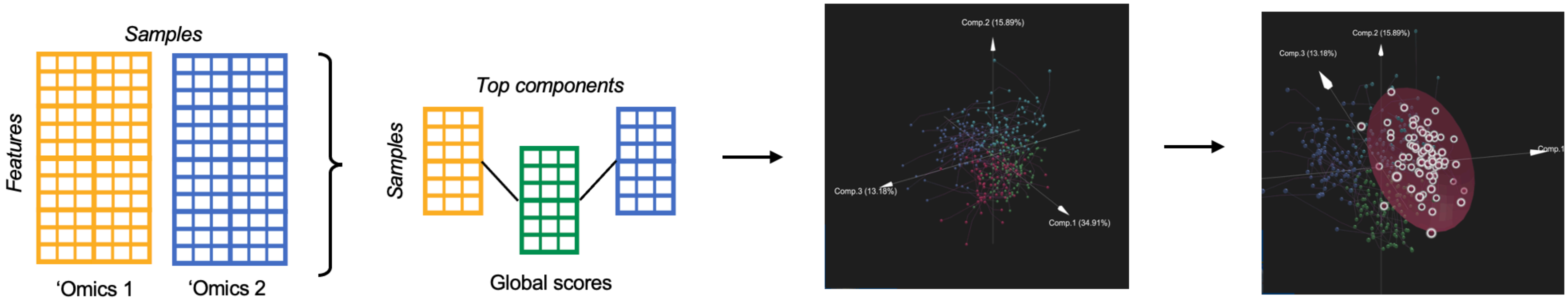
Next, we performed hierarchical clustering of features in both datasets using the "Average linkage" method and found a distinct cluster of proteins with generally lower levels in the baseline samples. The significant enriched pathways (Reactome) with the most hits are both related to the immune system (orange = "Immune system process"; blue = "Immune response"). This makes sense as the immune system is known to change predictably throughout pregnancy: https://pubmed.ncbi.nlm.nih.gov/28864494/.



Finally, we hierarchically clustered samples from both datasets, using the proteomics data as the anchor. We can see a cluster of proteins with lower expression that generally correspond with the baseline samples. This same sample cluster appears to have slightly higher metabolite levels, showing some evidence that these samples form distinct clusters in both datasets.

Looking at the "CLASS" annotation track at the top, we see that the baseline samples roughly cluster together. However, the single 'omics hierarchical clustering does a poor job of separating the different pregnant classes from each other. The green, red, and blue groups are almost perfectly mixed.

# Dimension Reduction Analysis



**Multi-variate dimension reduction**

**Visualize samples on top 3 components of shared co-variance space**

**Detect and select clusters**

Next, we see how well the sample groups separate when we use multi-variate dimension reduction techniques that both summarize redundant information within single 'omics datasets *and* are correlated across 'omics layers.

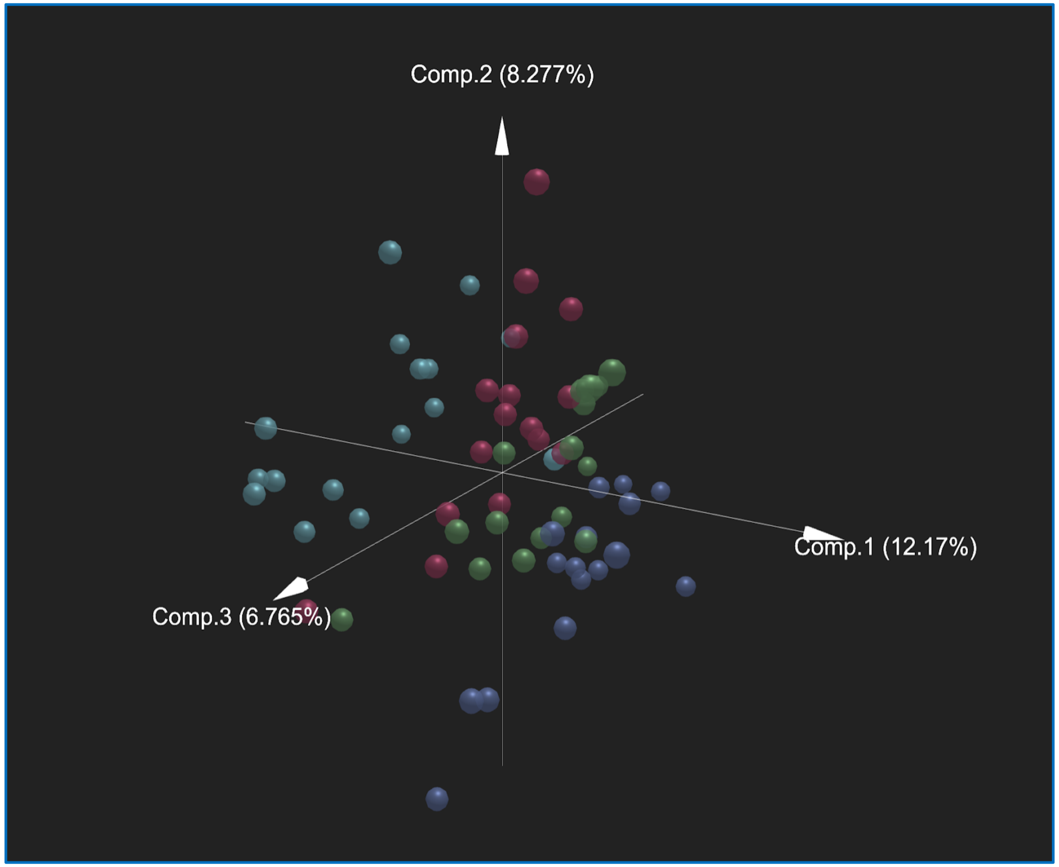


**Figure 1**

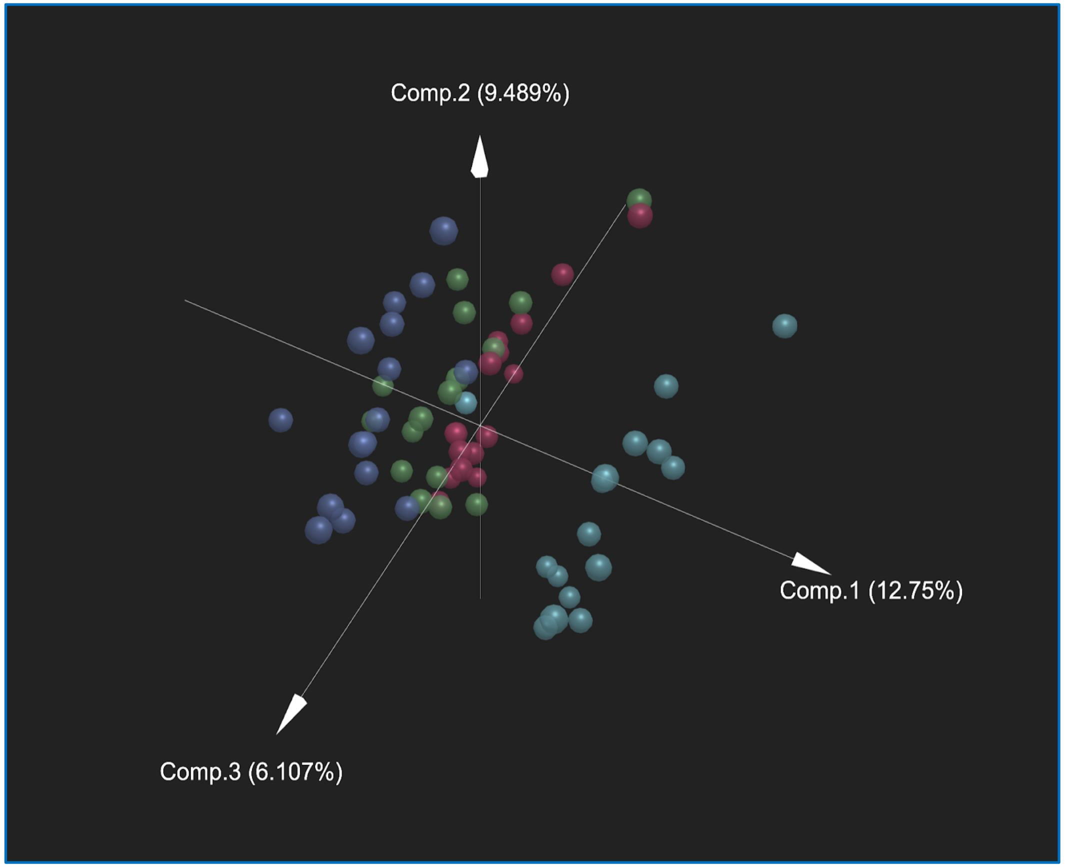
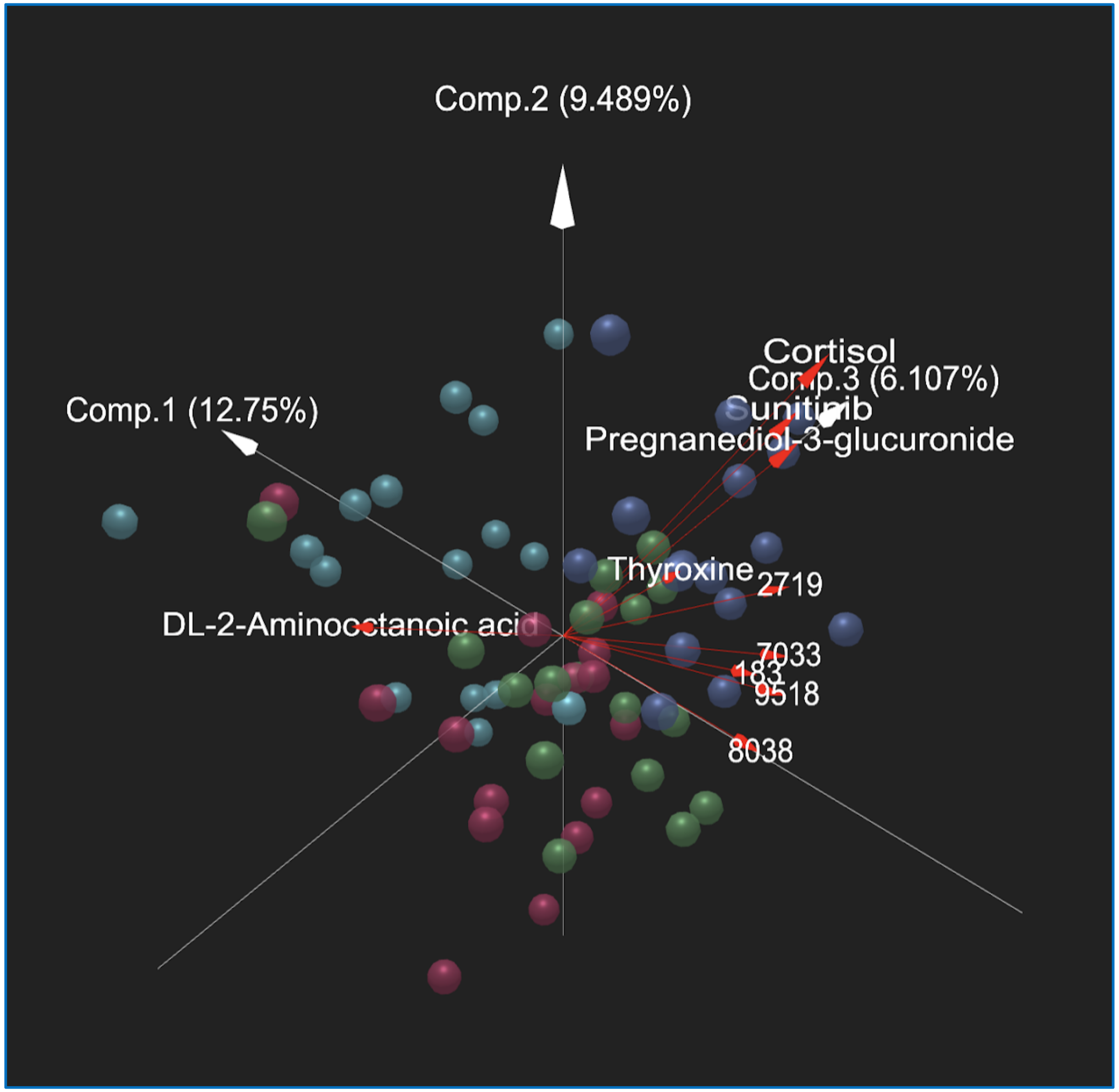


**Figure 2**

Figure 1 shows the sample-space 3D scatter plot after performing dimension reduction with the PLS method. We see much better separation between groups compared to univariate hierarchical clustering, even though this method is unsupervised with respect to the sample labels. The ordering also makes sense: light blue = baseline; red = first trimester; green = second trimester; blue = third trimester, thus the samples get progressively further away from baseline as the pregnancy progresses. Figure 2 shows that the separation between the sample groups gets a bit better with DIABLO, the only supervised method.

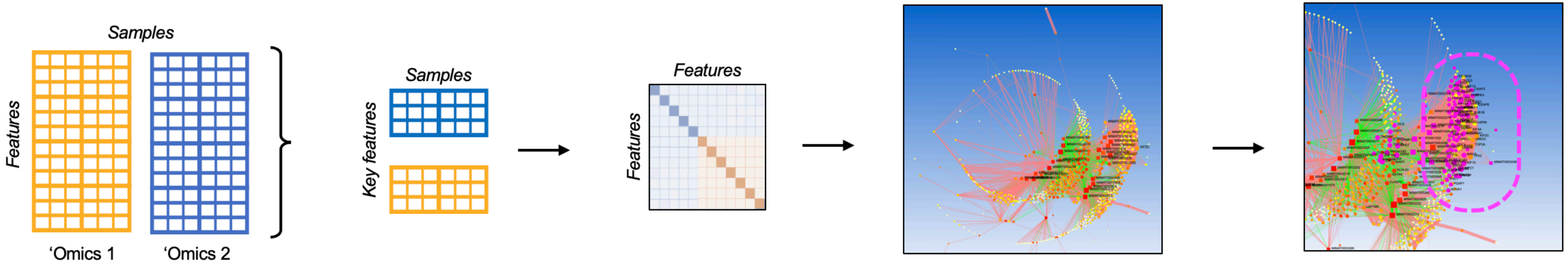


In bi-plot mode of the DIABLO results, we see that many of the key features driving variation along the component axes are clearly related to pregnancy.

Three out of five metabolites are hormones or are derived from hormones that are known to fluctuate throughout pregnancy (thyroxine; pregnanediol-3-glucuronide; cortisol).

Four out of five highlighted proteins are clearly related to pregnancy. The five proteins are ADAM12 (8038), GPC3 (2719), TFF3 (7033), GDF15 (9518), and AGT (183). One (ADAM12) is a commonly used serum marker for pregnancy and two (GDF15, GPC3) correspond to genes that are highly expressed in placenta relative to other tissues (source: NCBI gene profile). A quick search of "angiotensin and pregnancy" (AGT) reveals that this hormone is known to be elevated throughout the course of normal pregnancy (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3275085/). It is the main hormone in control of maintaining blood pressure. The final protein has an unknown function (TFF3).

# Correlation Network Analysis



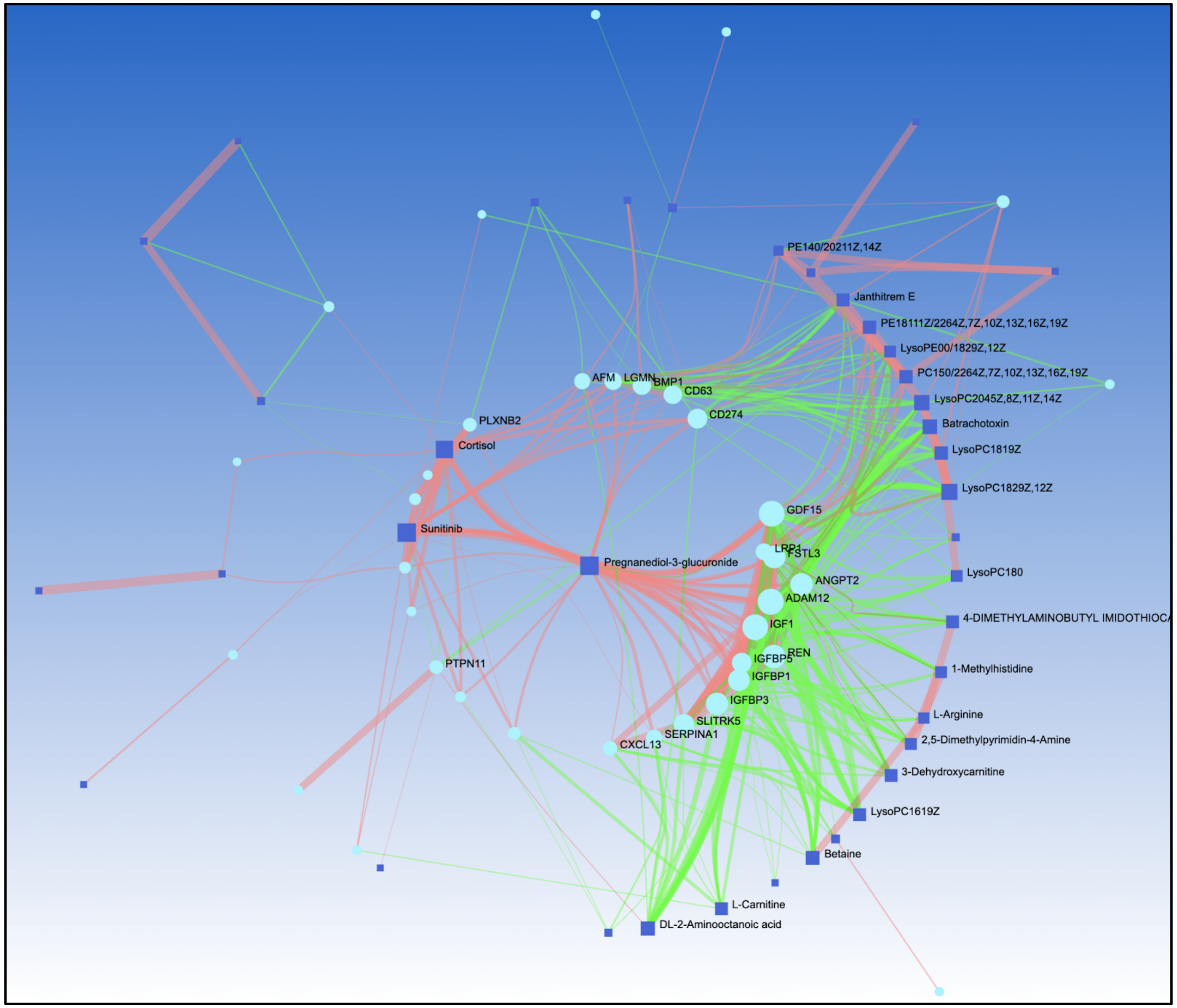
**Select key features**

**Compute similarity matrix across both 'omics**

**Visualize connections between features**

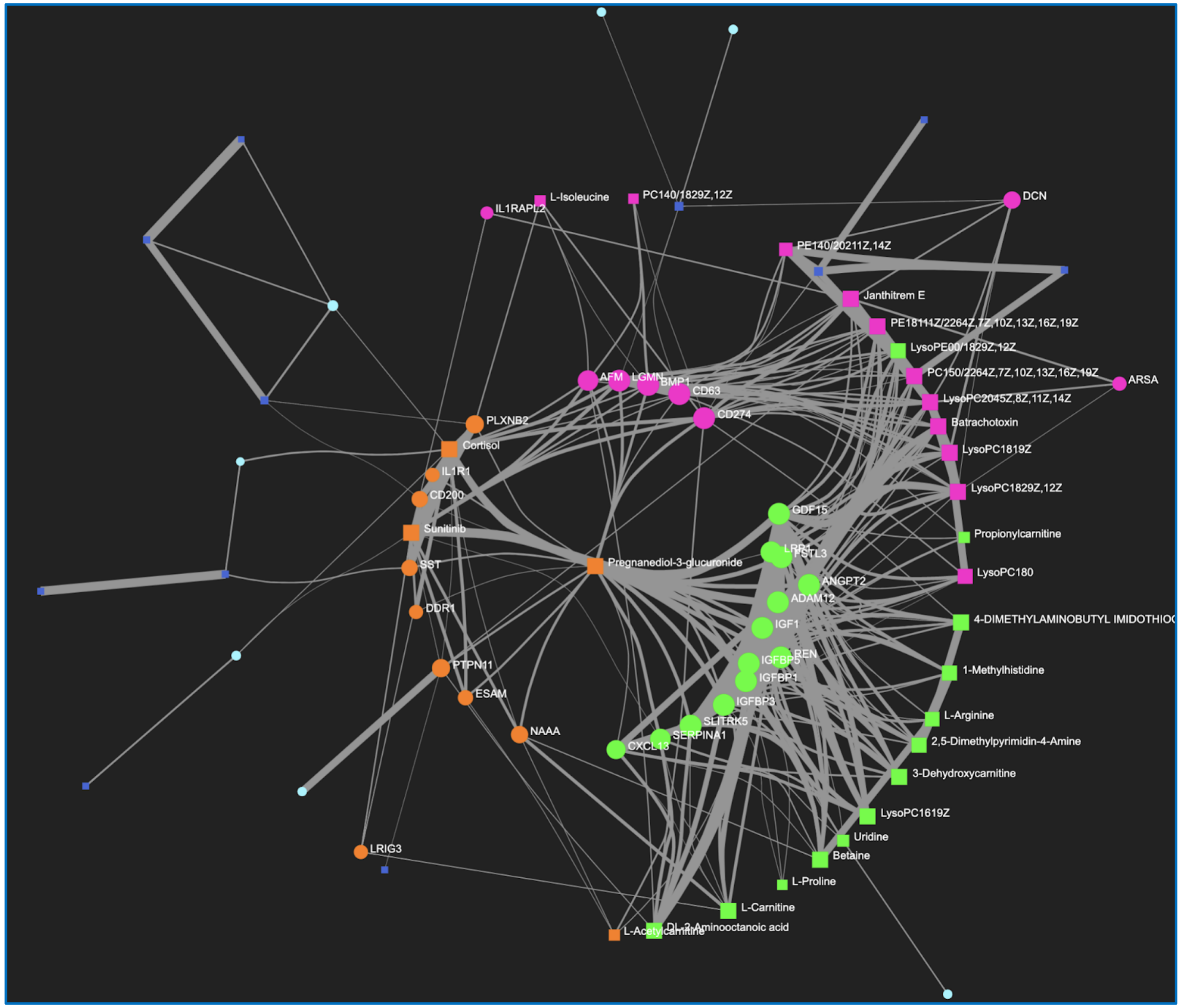
**Detect and select clusters**

Finally, we use correlation network analysis to gain a more detailed understanding of the key features within the top three DIABLO components that showed good separation between sample groups. All plots were made with feature selection from DIABLO components. We selected the top 20 features from each 'omics type for the top three components and allowed both 'omics connections (cross: 0.5; within: 0.8). The following plots are subnetwork 1 (largest one), using the concentric circle layout with edge bundling. To reproduce this plot, double-click "Pregnanediol-3-glucuronide" prior to selecting the method.



Overall, we visually recognize that some of the metabolites (squares), especially those with high degree (larger size), are hormones that were among our 6 top driving metabolites in the DIABLO bi-plot, as was the protein (circles) ADAM12. We see REN and ANGTP2 proteins, which are related to the angiotensin signaling pathway also highlighted in the bi-plot.

On the left side of the plot, we see a mix of proteins and metabolites that are generally all positively correlated with each other (predominantly red connections between features). On the right side, we see that for most part, metabolites are positively correlated with other metabolites, and proteins are positively correlated with other proteins, but protein and metabolite layers are negatively correlated with each other (predominantly green connections between feature layers). This is interesting and deserves further investigation.



Next, we perform module detection (WalkTrap) to find groups of densely connected features. We change the background to black and the edge color to grey to make module highlights stand out more.

Three modules are returned by the WalkTrap algorithm. One of these modules (green) has a significant p-value and so we investigated its components further with enrichment analysis. It is enriched for "Regulation of Insulin-like Growth Factor (IGF) Transport and Uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)" (Reactome pathway; adj p-val = 0.00003). This makes sense: a search of "IGF pregnancy" returns the following study: https://doi.org/10.1016/j.jcma.2013.07.004, which says that there are increases in IGF in blood plasma during pregnancy.

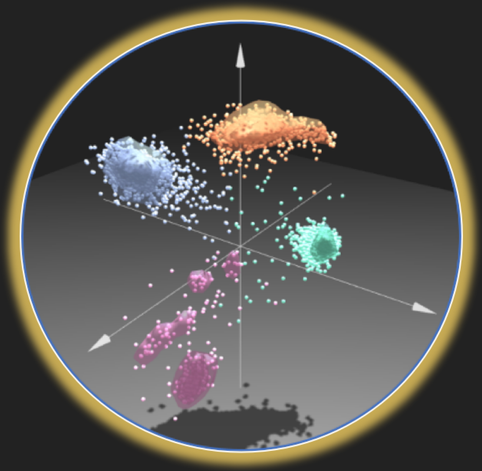
# Conclusions

Overall, there are several key points about multi-omics analysis in OmicsAnalyst:

- The separation between sample groups is limited in single omics. When we cluster samples hierarchically in heatmap, we see that baseline samples cluster together somewhat, but other time points are very mixed.

- The separation between samples is much more pronounced when using dimension reduction of integrated multi-omics data. The separation also matches what we expect: pregnant samples are noticeably different from baseline, and slightly different from each other. There is some separation between pregnancy times, and the separation matches what we expect: as pregnancy progresses, the sample groups get further from baseline.

- The key molecular features driving separation between samples in multi-variate dimension reduction are consistent with the biology of pregnancy. We see this in two ways:
  - Many of the key features highlighted in the bi-plot are known markers for pregnancy
  - These features show up again as topologically important nodes in the correlation network
  - Module analysis of the network returns biological processes that are known to change throughout pregnancy



# The End

For more information, visit the **FAQs, Tutorials, Resources** and **Contact** pages on www.omicsanalyst.ca