Omics Integration and Systems Biology

An Introduction

Ashfaq Ali

NBIS, SciLife lab, Lund University, Sweden

2020/05/14 (updated: 2020-05-19)

Omics Integration and Systems Biology Overview

- **Background and rationale**
  - Definitions: What is *"Omics"* integration and Systems Biology
  - Challenges
- **Omics Integration Techniques**
  - Data driven approaches: relying on measurements at hand
  - Network Approach
  - *Genome scale models and biological networks*
  - *Gene Set Enrichment and Pathway analyses*
- **Important steps in data integration**
  - Data Management and harmonization
  - Data scaling, dimensionality and distribution
  - Visualizations that deliver relevant message

Background and rationale

- **Omics Integration**
  - **What is Omics data Integration:** Ask 5 different people, get 5 different answers
  - May refer to study of **commonalities and differences** between two or more omics data sets
  - May refer to **variation explained beyond single omics**
  - May refer to **knowledge integration** to strengthen evidence obtained on a single or multiple omics
- **Systems Biology**
  - What is Systems Biology: *Ask 5 different people, get 10 different answers*
  - May refer to study biological processes at a holistic level, not always omics or high throughput
  - A system may comprise of a cell, a tissue, an organism or an ecosystem
  - Systems maybe static or dynamic
  - Involves combination *In Silico* and *In-vitro* experiments.

Why Data integration?

- Biological processes do not work alone in isolation
- Model and Address heterogeneity across data sets

Omics measurements

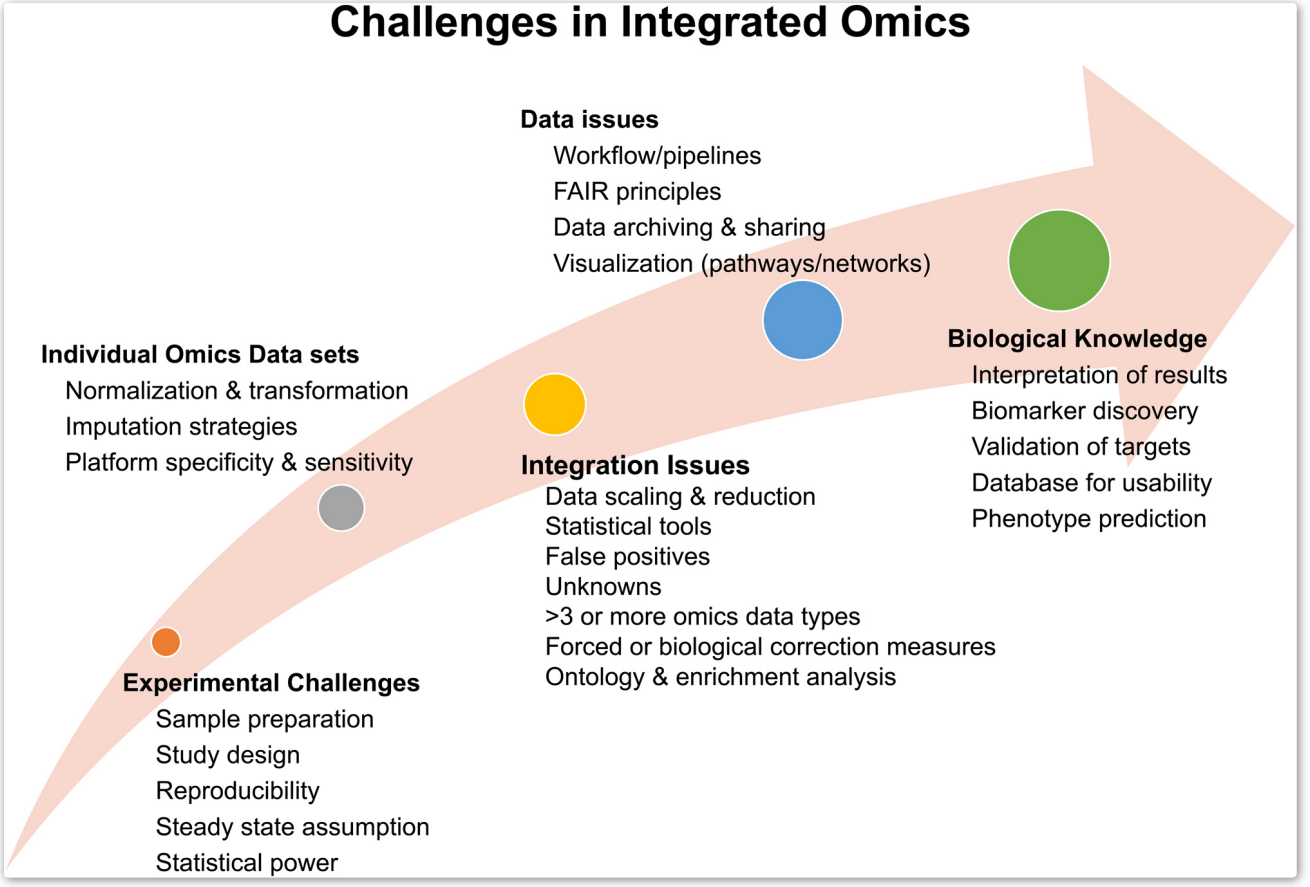
- **Omics measurements are fundamentally different due to technical reasons**
  - Sequencing techniques vs. mass spectrometry based technologies have different dynamic range and measurement space.
  - Itnesities do not always correspond to actual amount in the sample. This makes global normalization tricky.
  - Same samples measure after a period of time does not correspond to same values.
- **Different Omics measure fundamentally different aspects of biology**
  - A classic example is *Phosphoproteomics* where proteins are present in very small amounts but just addition of modification leads to activity.
  - Obviously measuring omics in blood vs. tissue have different interpretations
- **Some times (in fact most of the time) it is best to statistically analyse omics data with respective pipelines and then try to Integrate them.**
  - Add followup *look ups* or statistical test for association/correlations. **Meta Analyses**
  - Rather rely on visualizations that communicate with the audience you are interested in.
- **Avoid strict cut-offs and binning of the data just based on p-values when comparing different Omics**

Omics Data Integration

- **Under developed and continuously filed in bioinformatics**
- Combine data, find shared and distinct information information.
- We learn tools and ways of working with different types of data (Venn diagram's anyone?)
- Interpretation and application depends on the study questions
- More data -> More tools -> More complexity (Redundancy, contradictions)
- Needs across disciplinary communication
- Clever visualizations lead to impact, **poor visualizations = less impact**, No visualization = Hardly any impact. Big impact papers these crap a lot of data into complex figures.
- Many tools and approaches, guideline only for specific purposes. ***More Responsibility on the analyst to document analyses pipelines.***

Challenges in Omics-integration to keep in mind

A comprehensive overview of tools for omics integration (Misra et al).



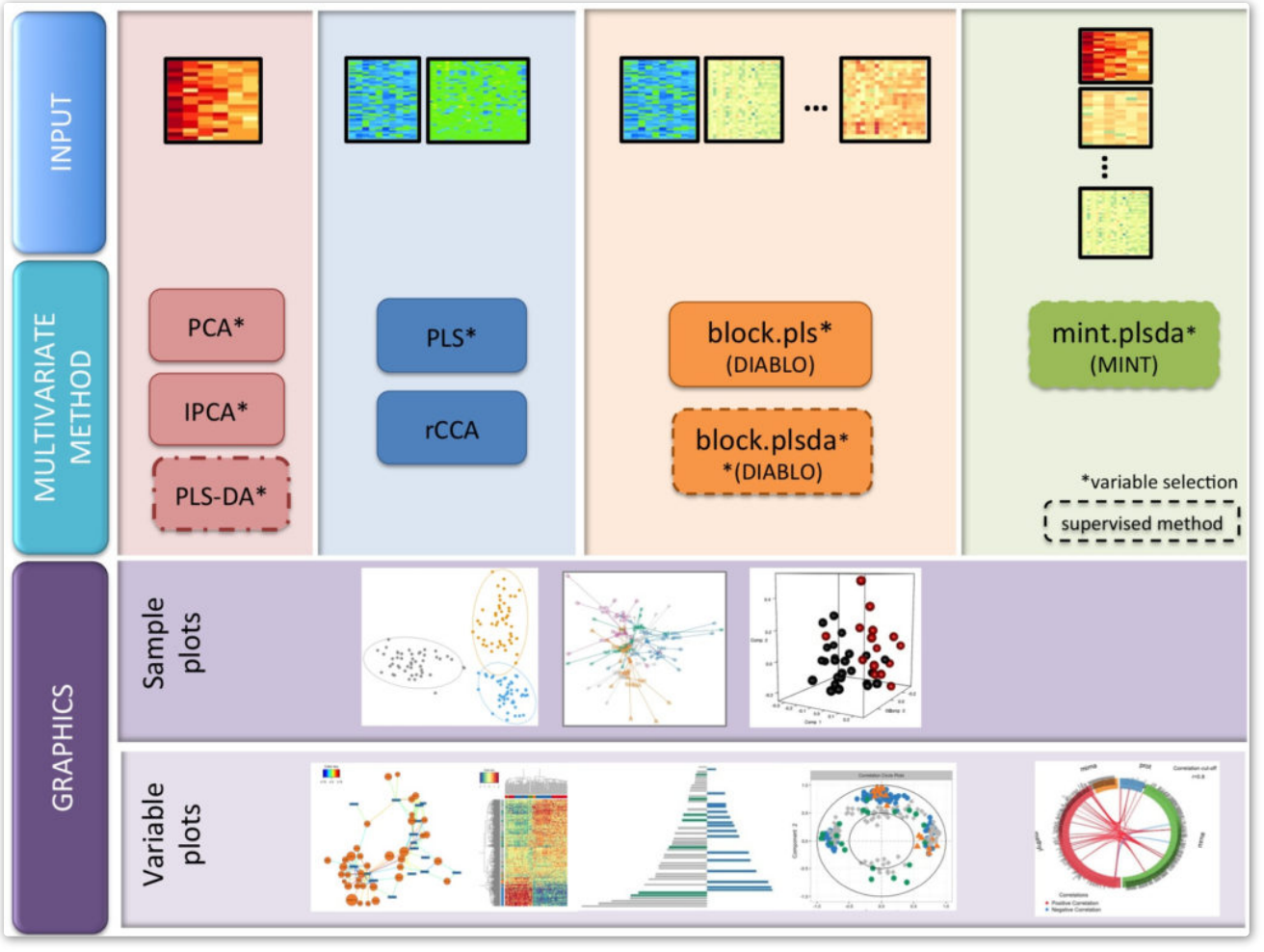
Reproducibility in a science at juvenile stage.

Omics Integration Techniques: Data Driven

- **Unsupervised analysis:** no known sample groups and is exploratory
    - Principal Component Analysis (PCA) for dimensionality reduction
    - Projection to Latent Structures (PLS) to model multiple phenotypes
    - Canonical Correlation Analysis (CCA) to find linear combinations of maximum correlations
- **Supervised analysis:** Class membership to discriminate sample groups and perform prediction
    - PLS Discriminant Analysis (PLS-DA),
    - Data Integration Analysis for Biomarker discovery using a Latent cOmponents **DIABLO**
    - MINT (F Rohart et al. 2017).
- A show case example applying mixomics and biological network integration.

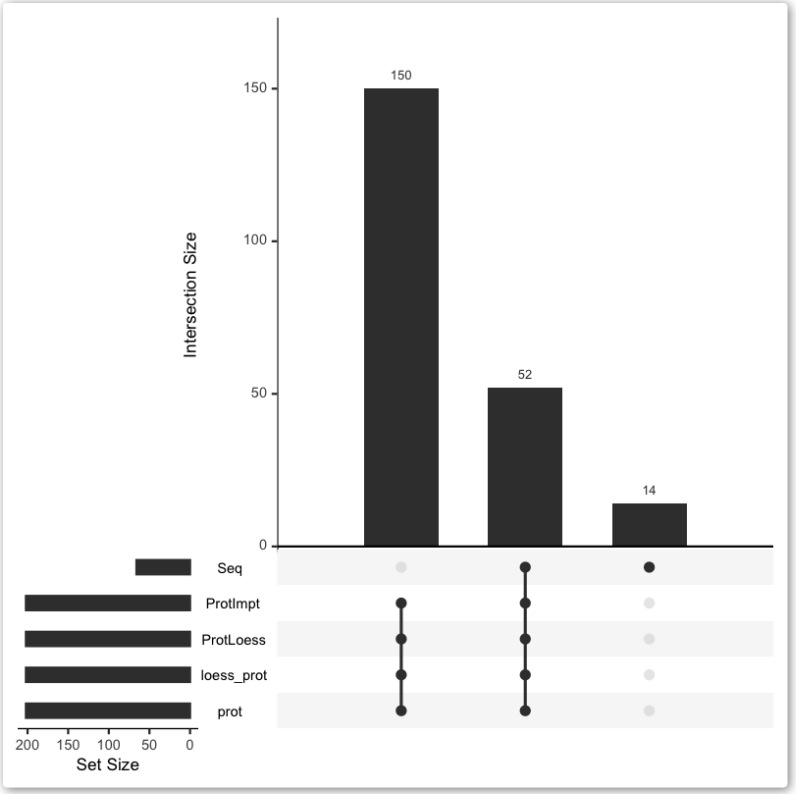
Omics Integration Techniques: Data Driven

- **R Package mixOmics** implemented methods



Omics Integration Techniques: DIABLO Example

A ProteoGenomics Study Example from Diffuse large B-cell lymphoma (DLBCL)

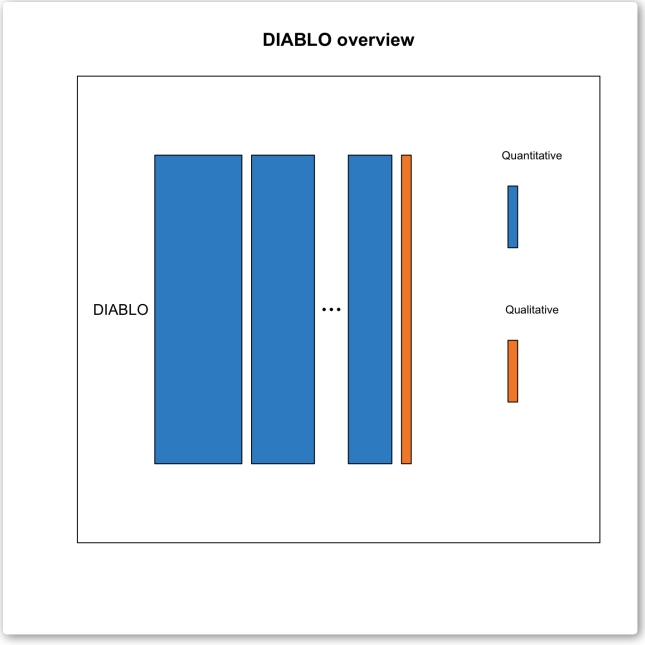


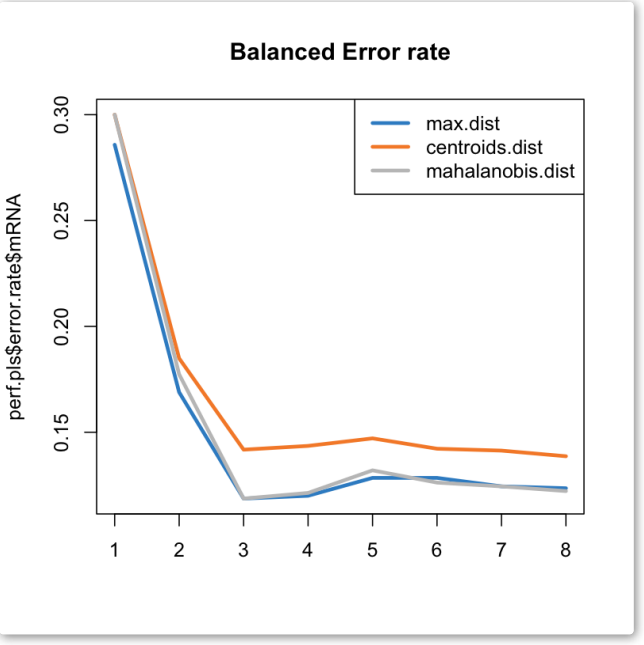
- The study aims to find predictive markers for Immunotherapy response
- Measured ~6000 proteins (TMT) in ~200 samples
- RNA-Seq ~ 20 000 genes in 66 samples
- About 50 overlapping samples

Steps in data Integration

- Transform protein ratio data to intensities
- Remove Batch effects in the data
- Transform count RNA-seq data (voom)
- Scale the data
- Extract overlapping samples from omics
- Identify explained variance and optimal number of PCs
- Identify number of features fro each PC and each data set
- Model important discriminating features

DIABLO Application

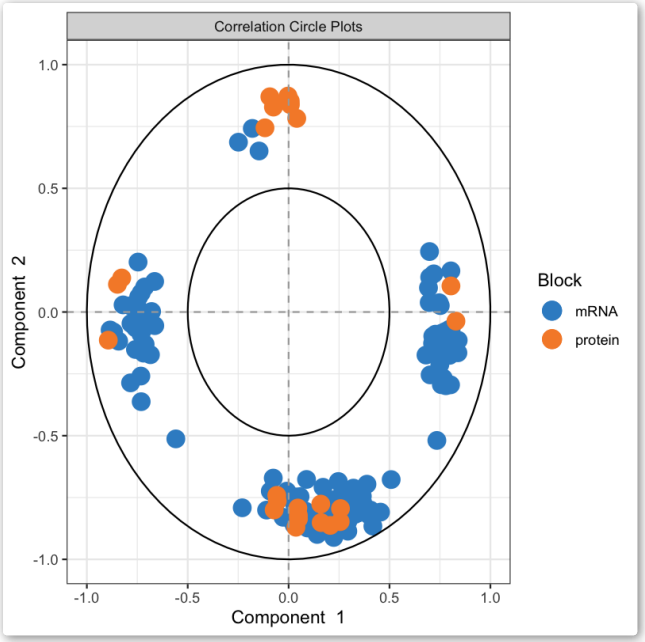


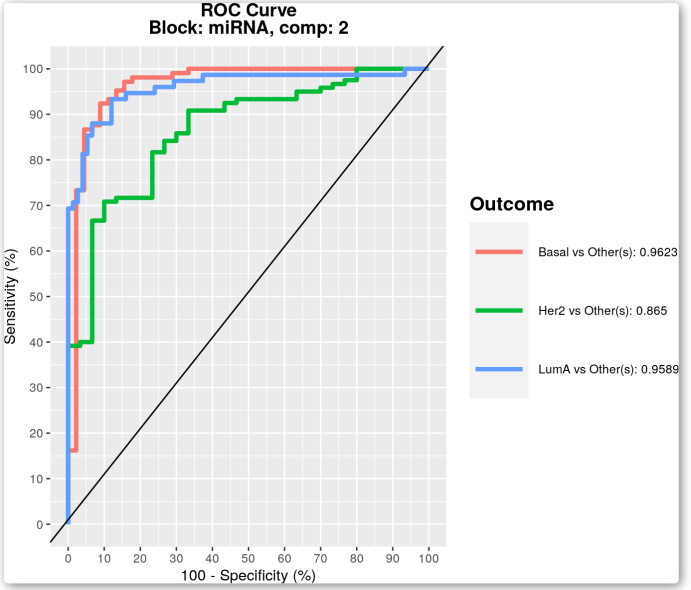


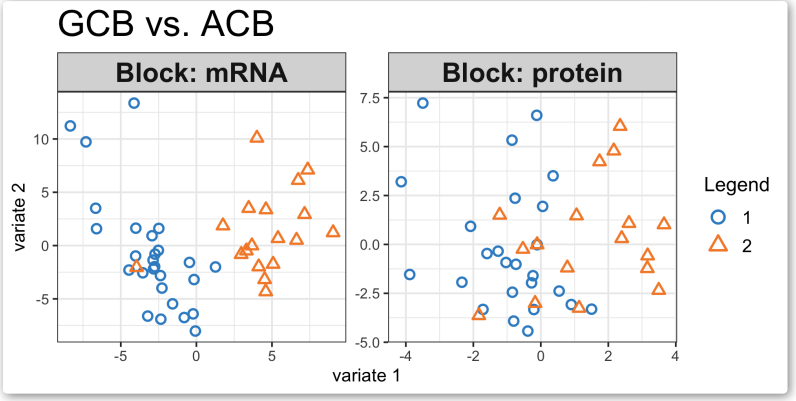
- Dimensionality reduction to extract latent variables
- Association of latent variables with phenotype

- The first two components capture most of the variation in the data here
- One needs to test various combinations of features for each PC and data set

Application of DIABLO to Large B cell Lymphoma data



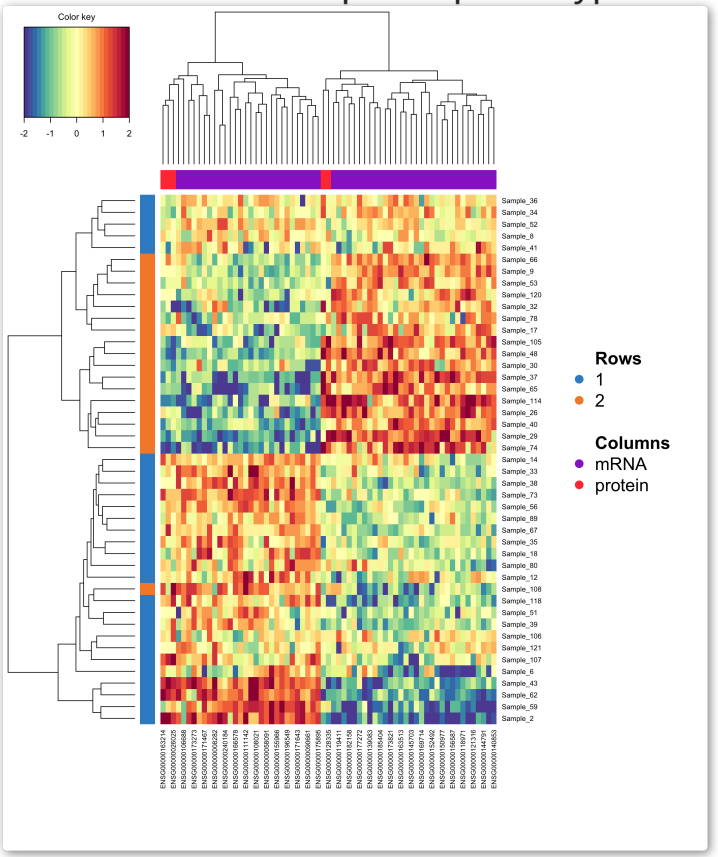




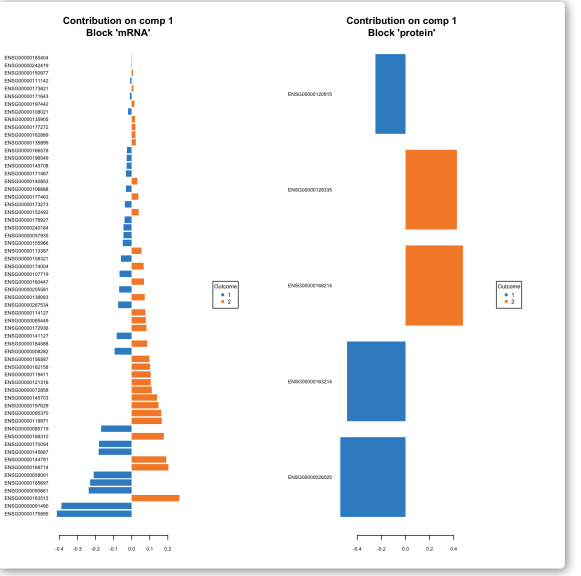
- Model allows ROC curves and prediction analyses
  - Dummy example

Most imporant variable from the integrated analyses

- Heat map of selected RNA and proteins and their relationship with phenotype

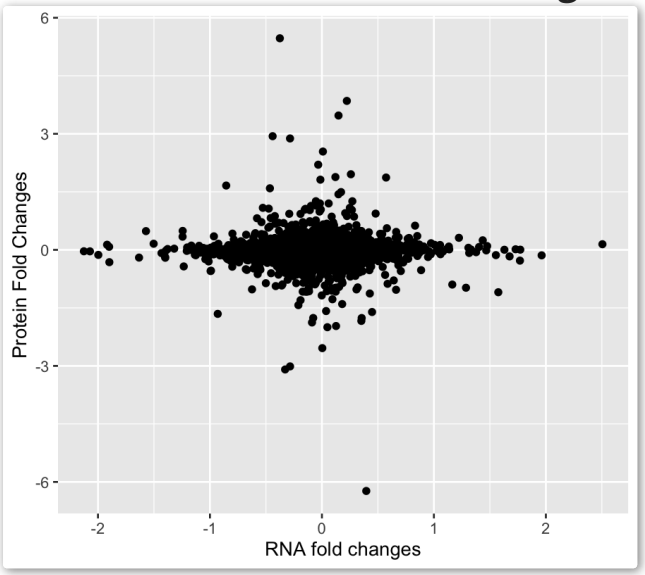


- Weights of individual variables in disciminant analyses
- Here RNA-seq data provided more features relevant for this phenotype

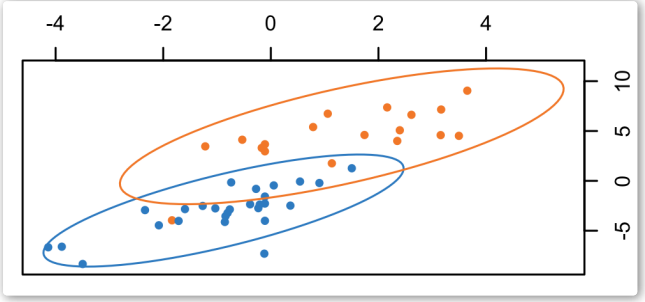


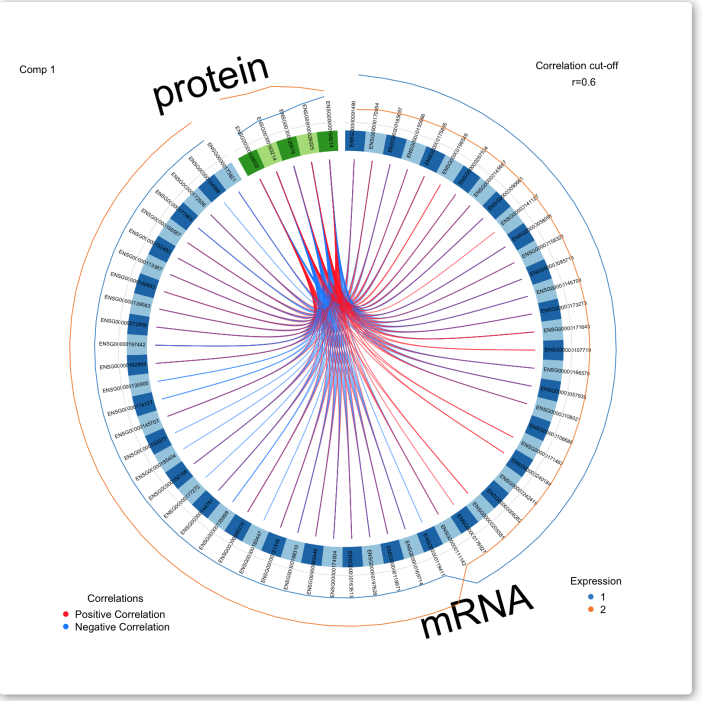
Correlation across RNA and protein data

Global correlation at fold changes level



Correlation for extracted variables DIABLO

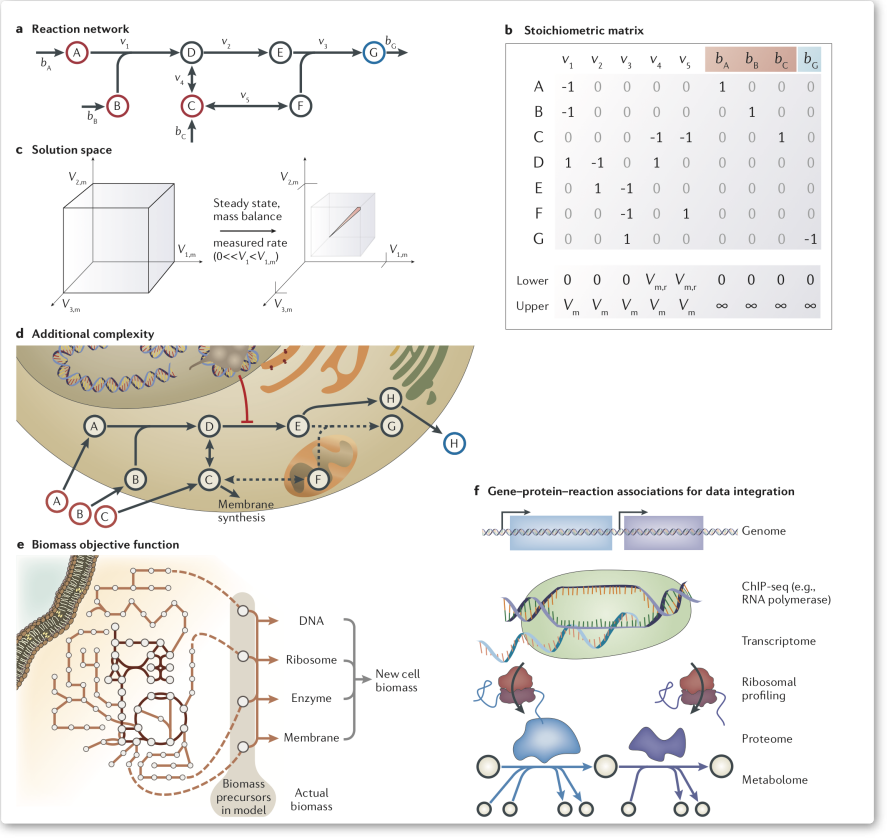




- Diablo model extracts variables most consistent across omics

What are Genome scale metabolic models (GEMs)

- Genome-scale metabolic models (GEMs) are mathematical reconstructions of the metabolic networks with all known metabolic reactions
- In some cases, GEMs could represent the whole tissue or body of a multicellular organism.
- In these metabolic networks, the gene-protein-reaction (GPR) relationships are annotated.
- All the reactions in GEMs are mass-balanced, ensuring stoichiometric balance.



Fundamentals of Genome Scale Metabolic Modelling

Metabolic reconstruction:

A carefully curated and biochemically validated knowledge base in which all known chemical reactions for an organism are detailed and cataloged.

Genome-scale model A condition-specific, mathematically described, commutable derivative of a metabolic reconstruction, containing comprehensive knowledge of metabolism.

Solution space: The feasible region satisfying a set of constraints. In constraint-based reconstruction and analysis (COBRA) models, this represents the feasible flux values for all of the reactions in the model.

Flux distributions

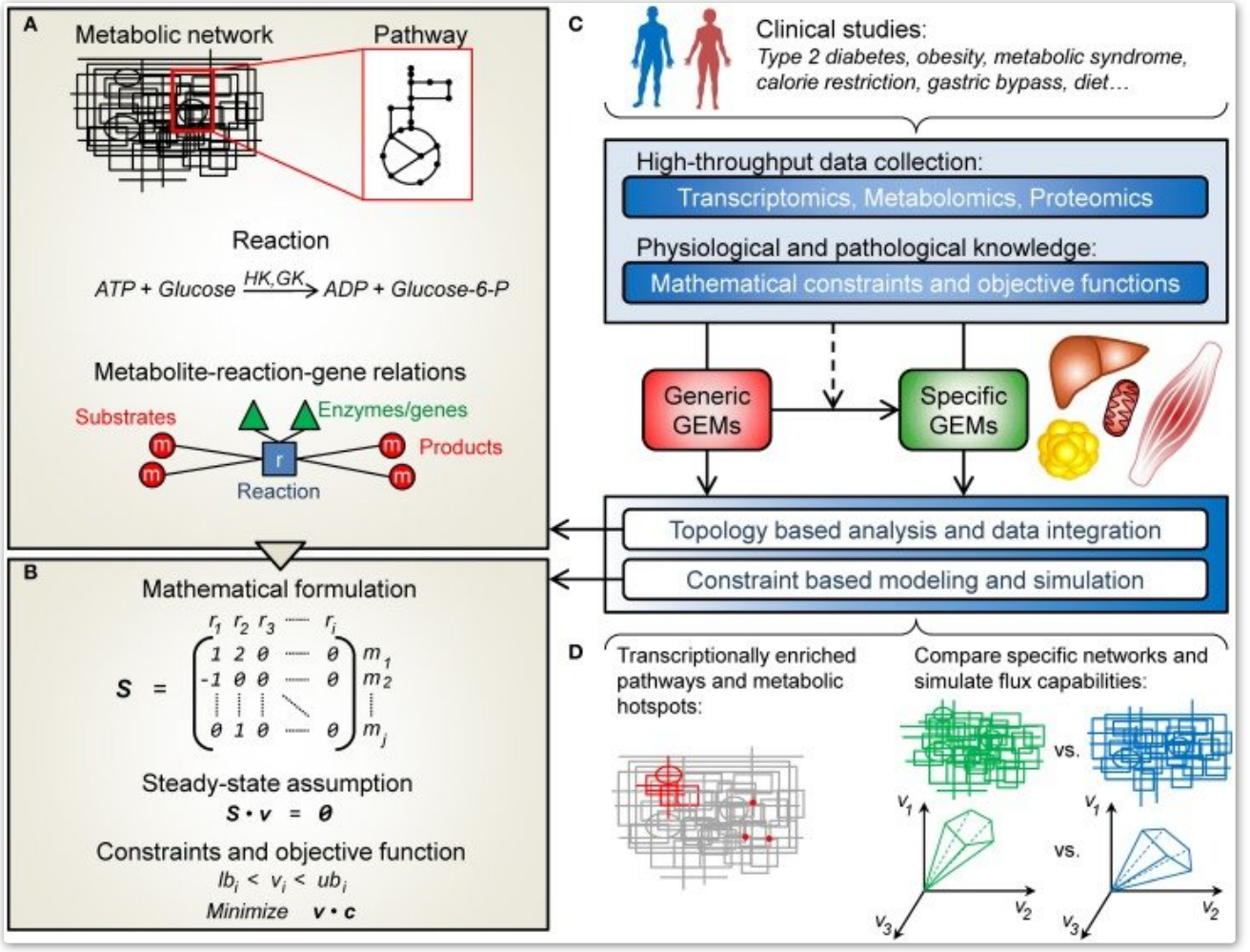
A set of steady-state fluxes for all of the reactions in a metabolic network.

Why GEMs in Omics integration

- (GEMs): a valuable systems biology platforms for
  - Model-guided data analysis of large omics datasets
  - Provide cellular context to the data
  - Allow the integration of diverse omics data
  - Directly link metabolites to enzymes.
  - Elucidate how changes in one component affect other pathways and cell phenotype
  - Growth, cell energetic
  - Pathway fluxes
  - Biosynthesis of cell components, byproduct secretion, etc.
  - These systems biology models can provide a mechanistic link from **genotype to phenotype**

Opdam *etal* *Cell Systems 4, 318–329, March 22, 2017*

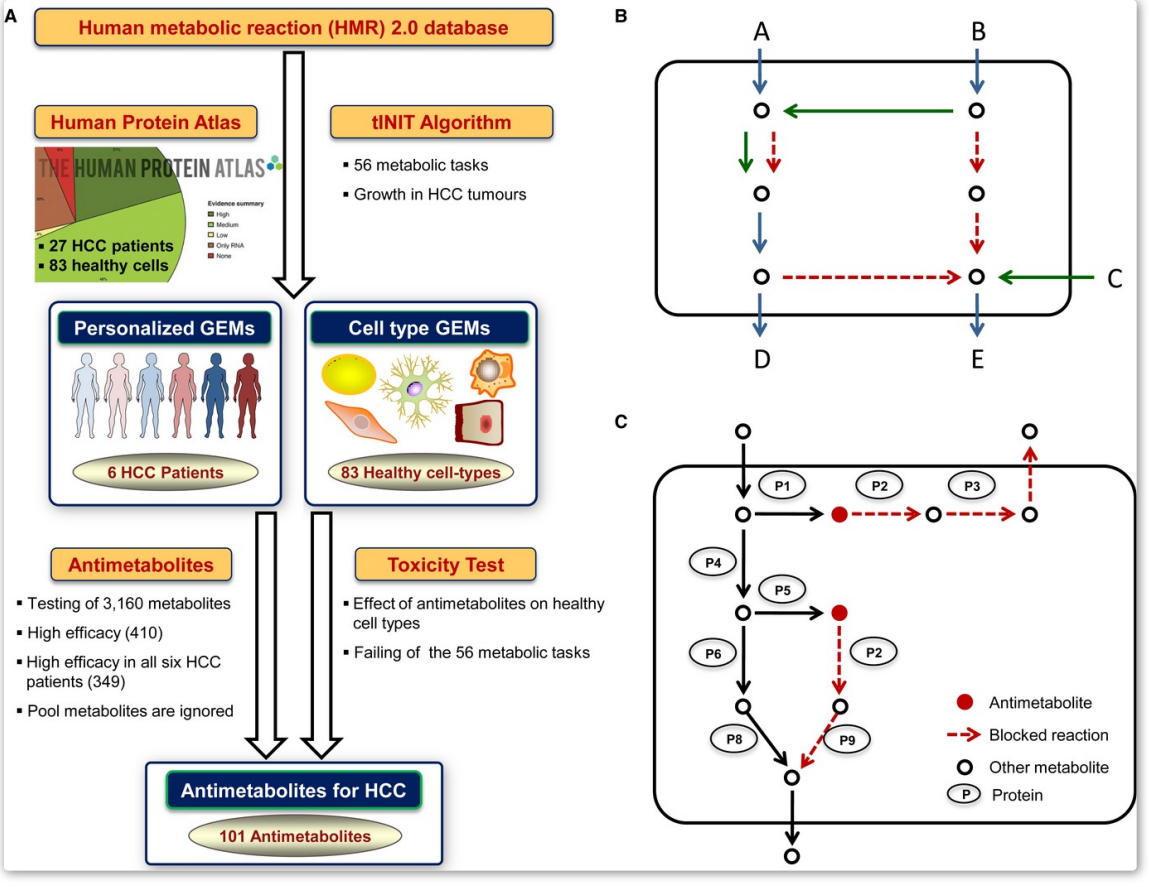
An overview of human genome scale metabolic models (GEMs)



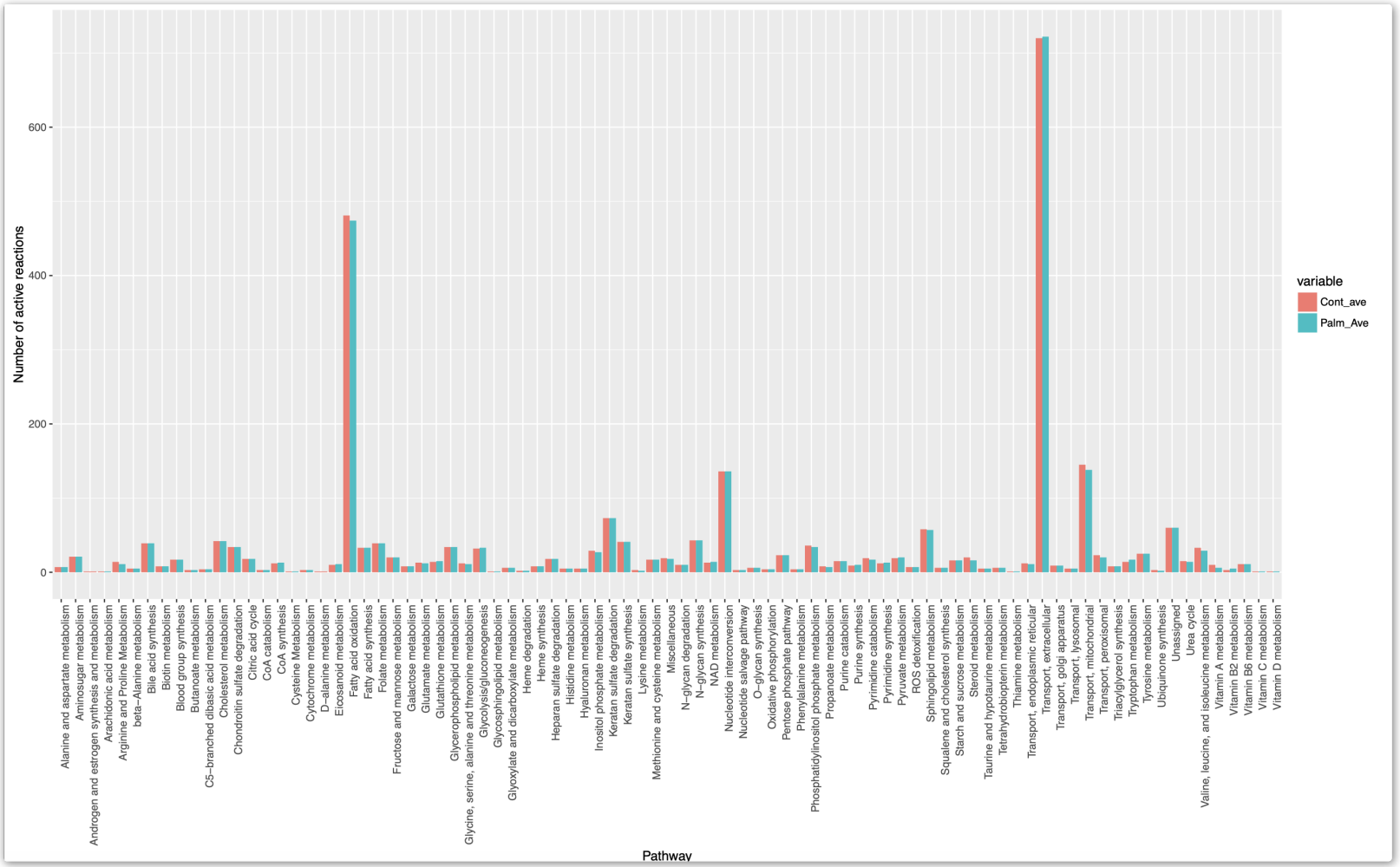
Model Extraction Algorithms and analyses

- COBRA
- RAVEN

tINIT Algorithm



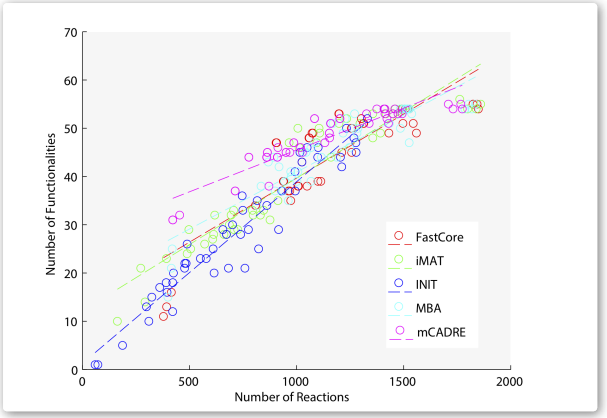
Context Specific Models

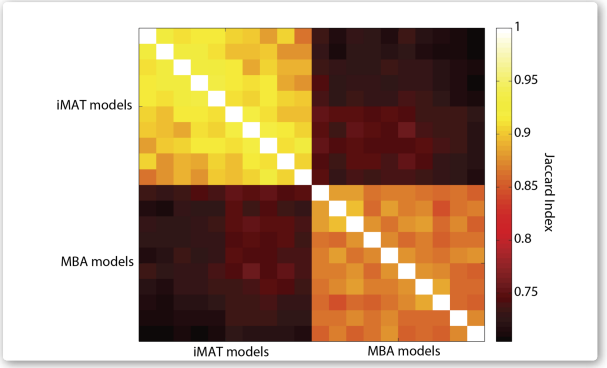


Example: The number of reactions active in given pathways in two conditions.

Performance of diferent algorithms and relevance

- Different parameters and algorithms yielded diverse models
- Extraction method most strongly affected accuracy of gene essentiality predictions





- Essentially the parameters algorithms use to create context specific models explain the largest variation in your data.
- We will only cover Reaction Enrichment in Piano that one can follow up using HMR.

Opdam *etal* *Cell Systems 4, 318–329, March 22, 2017*

Gene Set Enrichment Analyses HMR as a Network and gene set Enrichment Analyses

	ensembl_gene_id	baseMean	log2FoldChange	pvalue	padj
1	ENSG00000000003	490.01721	0.9145204	3.661641e-17	0.00376
2	ENSG00000000419	817.78066	-0.1894651	6.001737e-02	0.04354
3	ENSG00000000457	82.07877	0.3307639	1.207585e-01	0.00005
4	ENSG00000000460	356.07160	-1.8636578	4.096103e-51	0.00025
5	ENSG00000001036	919.60675	-0.3482723	3.922539e-05	0.19231
6	ENSG00000001084	529.59397	-0.6764194	8.192621e-13	0.06244

> Is there a pattern in my list of DEGs?

- Do my DEGs work together?
- Are they involved in a biological process?
- Are they involved in a known pathway?
- Reduce gene lists to biologically interpretable terms
- Pick interesting genes based on function
- Less prone to false-positives on the gene-level
- Interpretation of genome-wide results

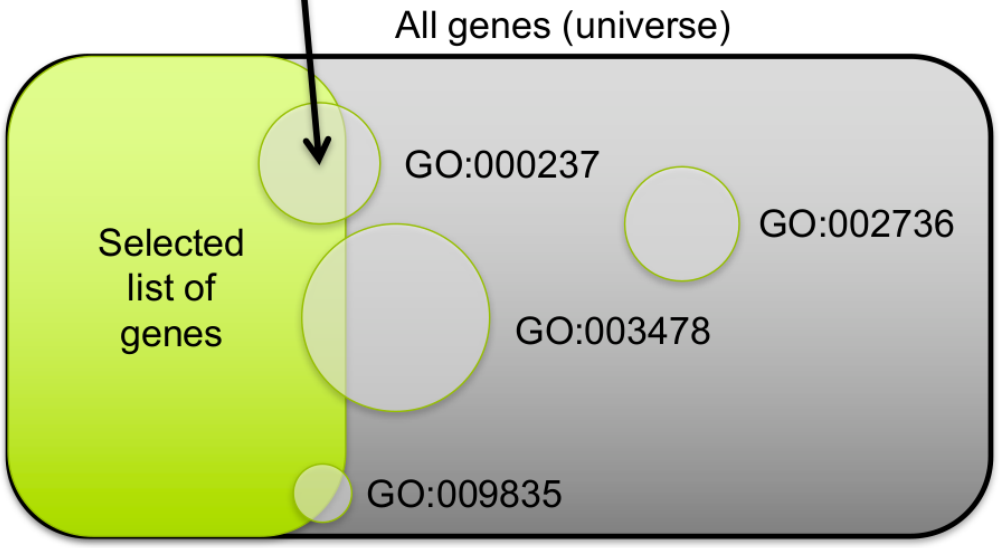
Gene set analyses (GSA)

- Requires cut-off
- Omits any expression metric
- Good to test overlap of signif genes in two comparisons
- Computationally fast

Hypergeometric test (Fisher's exact test)

Is this overlap bigger than expected by random chance?

	Selected	Not selected
In GO-term	8	2
Not in GO-term	92	19768



GSA input

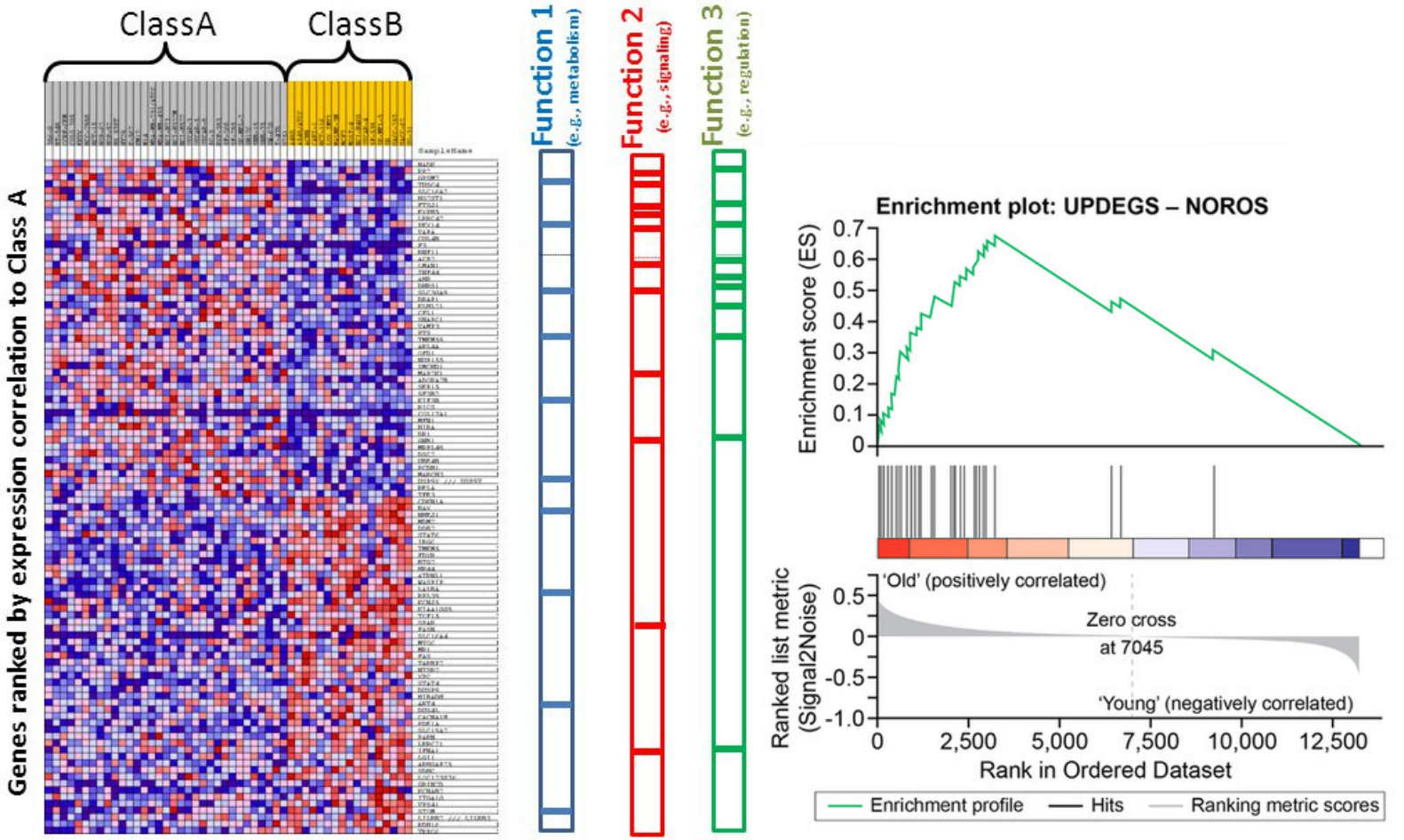
	ensembl_gene_id	baseMean	log2FoldChange	pvalue	padj
1	ENSG00000000003	490.01721	0.9145204	3.661641e-17	0.00376
2	ENSG00000000419	817.78066	-0.1894651	6.001737e-02	0.04354
3	ENSG00000000457	82.07877	0.3307639	1.207585e-01	0.06244
4	ENSG00000000460	356.07160	-1.8636578	4.096103e-51	0.12002
5	ENSG00000001036	919.60675	-0.3482723	3.922539e-05	0.19231
6	ENSG00000001084	529.59397	-0.6764194	8.192621e-13	0.00005

Input set: `ENSG00000000003, ENSG00000000419, ENSG00000001084`

Universe: `ENSG00000000003, ENSG00000000419, ENSG00000000457, ENSG00000000460, ENSG00000001036, ENSG00000001084`

Gene set enrichment analyses (GSEA)

- All genes are used
- Ranked by an expression metric/gene-level statistic



GSEA input

```
  ensembl_gene_id  baseMean log2FoldChange       pvalue    padj
1 ENSG00000000003 490.01721      0.9145204 3.661641e-17 0.00376
2 ENSG00000000419 817.78066     -0.1894651 6.001737e-02 0.04354
3 ENSG00000000457  82.07877      0.3307639 1.207585e-01 0.06244
4 ENSG00000000460 356.07160     -1.8636578 4.096103e-51 0.12002
5 ENSG00000001036 919.60675     -0.3482723 3.922539e-05 0.19231
6 ENSG00000001084 529.59397     -0.6764194 8.192621e-13 0.00005
```

- Input is a set of labelled ranked expression metrics.

```
## ENSG00000000003 ENSG00000000457 ENSG00000000419 ENSG00000001036 ENSG00000001084
##       0.9145204       0.3307639      -0.1894651      -0.3482723      -0.6764194
## ENSG00000000460
##      -1.8636578
```

Available Tools

R packages

topGO, goana, goseq, topKEGG, kegga, enrichR, piano, clusterProfiler, Pathview, fgsea, gprofileR

Online

DAVID, GOrilla, Enrichr, Revigo, Webgestalt, Panther, Tair

Downloadable

GSEA, ErmineJ, Ingenuity Pathway analyses

Considerations

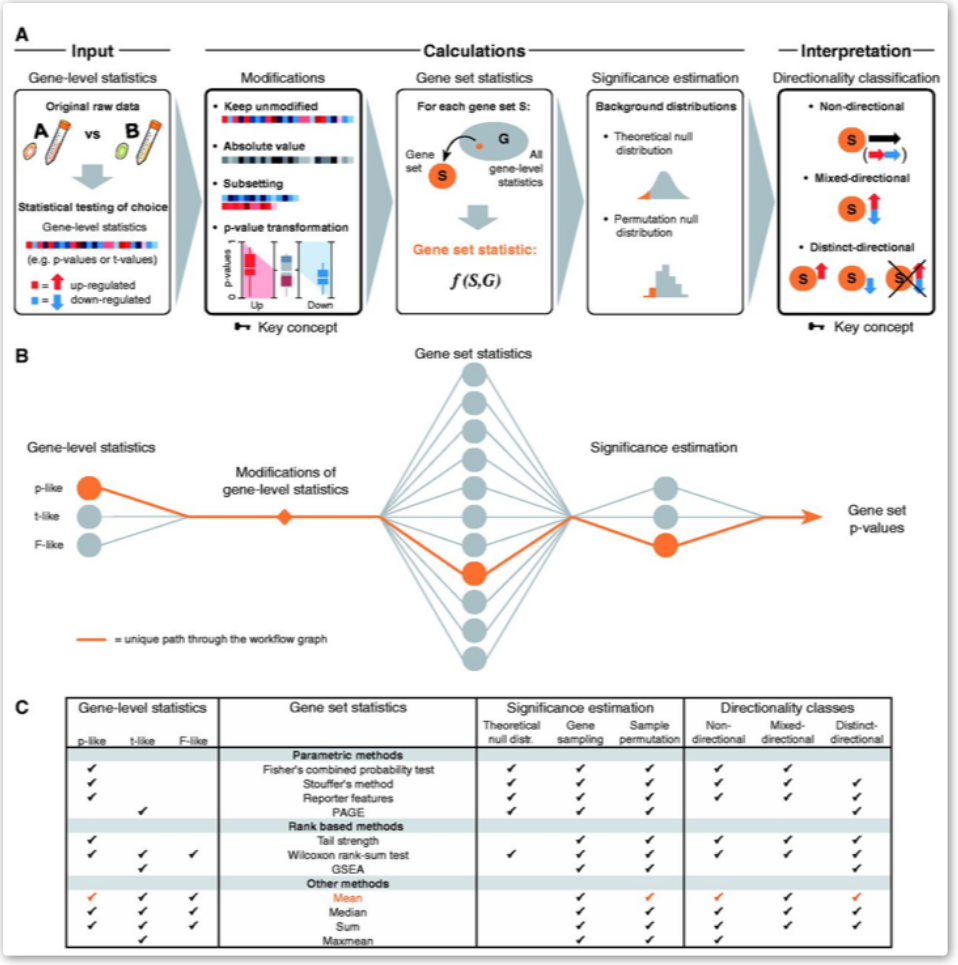
- Pay attention to gene IDs
- Bias in gene sets
- Confusing gene set names
- Consider gene set size
- Adjust for multiple testing
- Large number of highly overlapping gene-sets (representing a similar biological theme) can bias interpretation and take attention from other biological themes that are represented by fewer gene-sets

HMR as a Network and gene set Enrichment Analyses

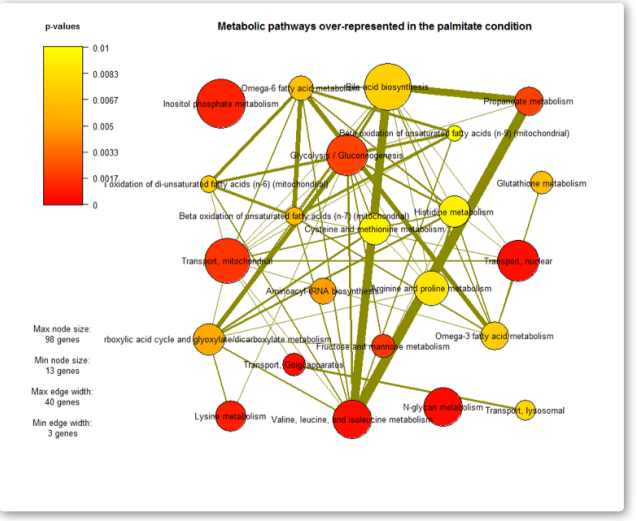
Signalling, Co-expression, metabolic, Protein Protein Interaction, Phosphoproteins etc

- Network representation
  - Gene-metabolite
  - reaction associations
  - Reaction- Pathway associations
- **Piano** Package in R
  - Implements several different methods for gene set enrichment analyses.
  - Provides direct comparison of several methods and allow building a consensus on enrichment analyses.
  - Allow direction specific analyses
  - Provides excellent visualization of result.
  - Easy output/export of results for further analyses.

Piano workflow



Gene-level statistics			Gene set statistics	Significance estimation			Directionality classes		
p-like	t-like	F-like		Theoretical null distr.	Gene sampling	Sample permutation	Non-directional	Mixed-directional	Distinct-directional
			Parametric methods						
✓			Fisher's combined probability test	✓	✓	✓	✓	✓	
✓			Stouffer's method	✓	✓	✓	✓	✓	✓
✓			Reporter features	✓	✓	✓	✓	✓	✓
	✓		PAGE	✓	✓	✓			✓
			Rank based methods						
✓			Tail strength		✓	✓	✓	✓	✓
✓	✓	✓	Wilcoxon rank-sum test	✓	✓	✓	✓	✓	✓
	✓		GSEA		✓	✓			✓
			Other methods						
✓	✓	✓	Mean		✓	✓	✓	✓	✓
✓	✓	✓	Median		✓	✓	✓	✓	✓
✓	✓	✓	Sum		✓	✓	✓	✓	✓
	✓		Maxmean		✓	✓	✓		

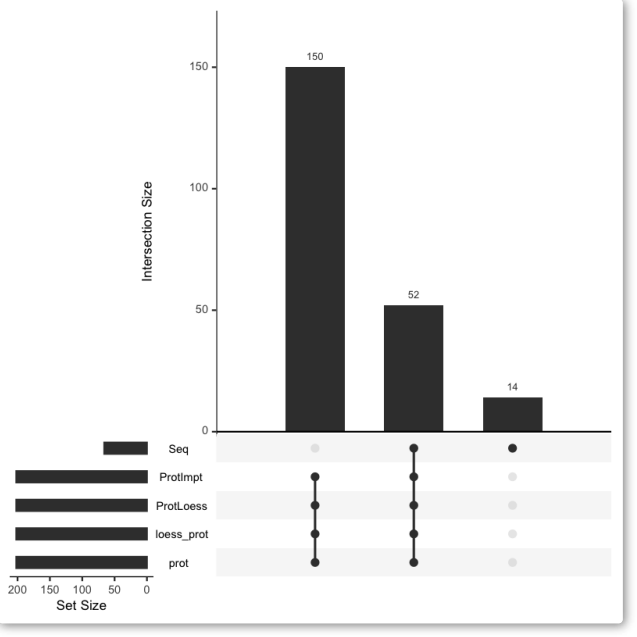


- One may use **Gene Expression**, **proteomics**, **metabolomics** or other data with small tweaks for individual omics level enrichment analyses.
- Compare enrichment results across omics

Leif Wigge's Piano Package in R

Data Management

- Data management is a key aspect in omics integration S4 classes
  - ExpressionSet
  - Summarized Epreriments
  - RaggedExPeriments
  - MultiAssayExperiments
  - Annotations BioMArt, Ensemble
  - BSGenomes
- **Containanerize** your data to avoid irreproducibility hazard
  - **Updates in clinical databases**
  - **Updates in Annotations**
  - **Coding mistakes in matching across datasets**

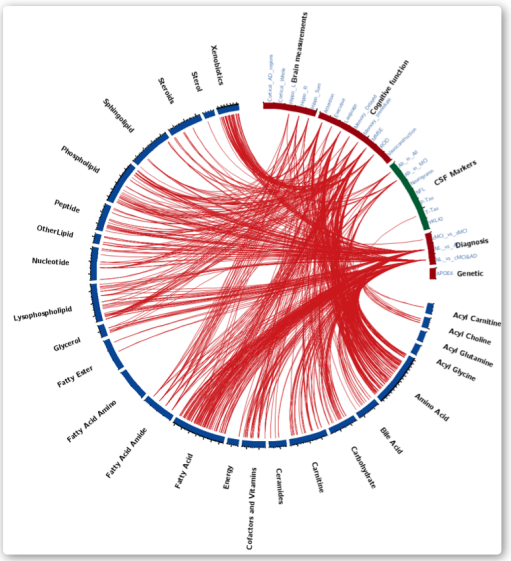


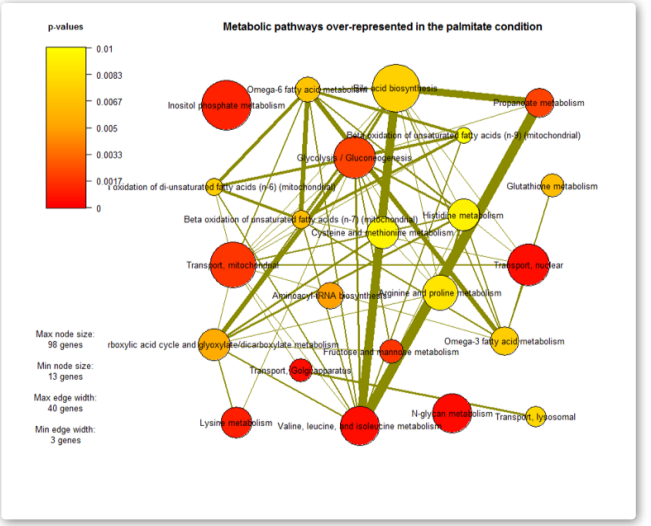
Example

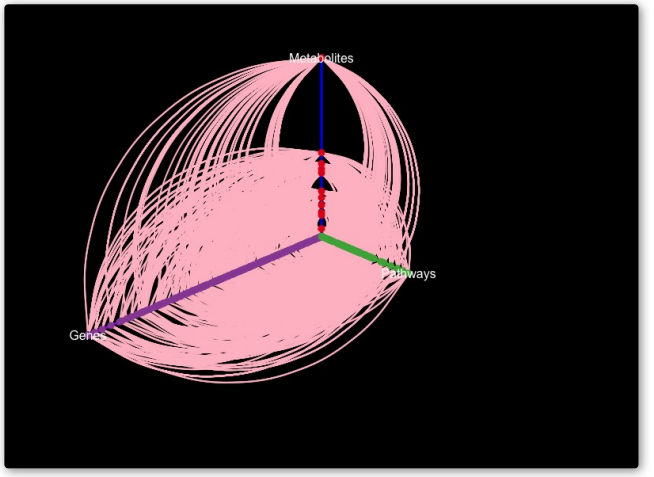
- The sample IDs across data sets
- Sample and row meta data disagree
- Result in loss of costly statitical power

Data Visualization

- Go beyond 2D scatter plots and heat maps
  - Layer your data in **Circular format**
  - Networks
  - Hive plot







Some of the best work goes unnoticed if not visually appealing

Omics Integration NBIS Course

Omics Integration and Systems Biology Course **October 5 - 9 , Lund**

- Apply here, **Deadline for registration: 21/08/2020**
- Topics Covered
  - Data wrangling in omics studies
  - Condition-specific and personalized modeling through Genome-scale Metabolic models based on integration of transcriptomic, proteomic and metabolomic data; Biological network inference, community and topology analysis and visualization
  - Identification of key biological functions and pathways;
  - Identification of potential biomarkers and targetable genes through modeling and biological network analysis;
  - Application of key machine learning methods for multi-omics analysis including deep learning;
  - Multi-omics integration, clustering and dimensionality reduction;
  - Similarity network fusion and Recommender systems;
  - Integrated data visualization techniques

Any Questions

- .....
- .....
- .....

Thank You