Multivariate Methods to detecting co-related trends in data

- Canonical correlation analysis
- Partial least squares
- Co-inertia analysis

- Classical CCA and PLS require n>p. Can apply Penalized CCA and sparse PLS (Comparison of methods Lê Cao et al., (2009). CIA needs no penalization (feature selection)

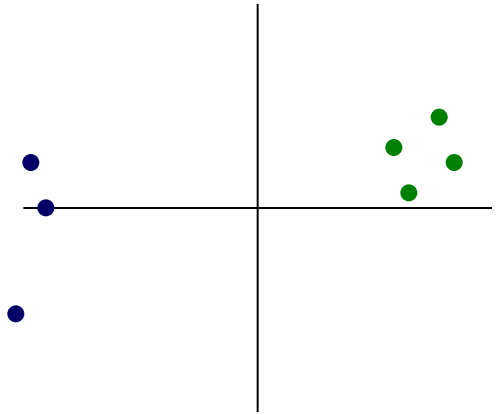
Co-inertia analysis

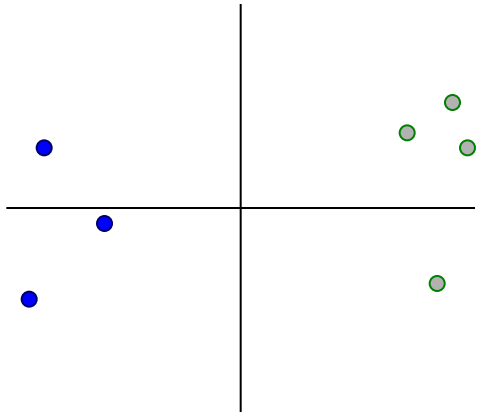
– tables with same number of cases
– when no. of variables >>> cases
– with disparate variables
– Quantitative or qualitative variables
– Can weight cases

Uses PCA, COA or other ordination method

Dolédec, S. & Chessel, D. (1994) *Freshwater Biology*, 31, 277-294.
Thioulouse, J. & Lobry, J.R. (1995) *Computer Applications in the Biosciences*, 11, 321-329
Culhane et al., 2003, Jeffery et al., 2007, Fagan et al., 2006

Coinertia Analysis

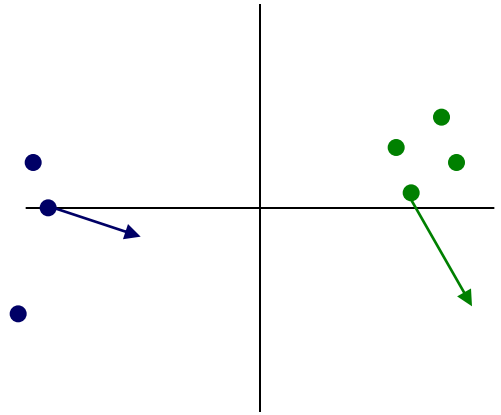




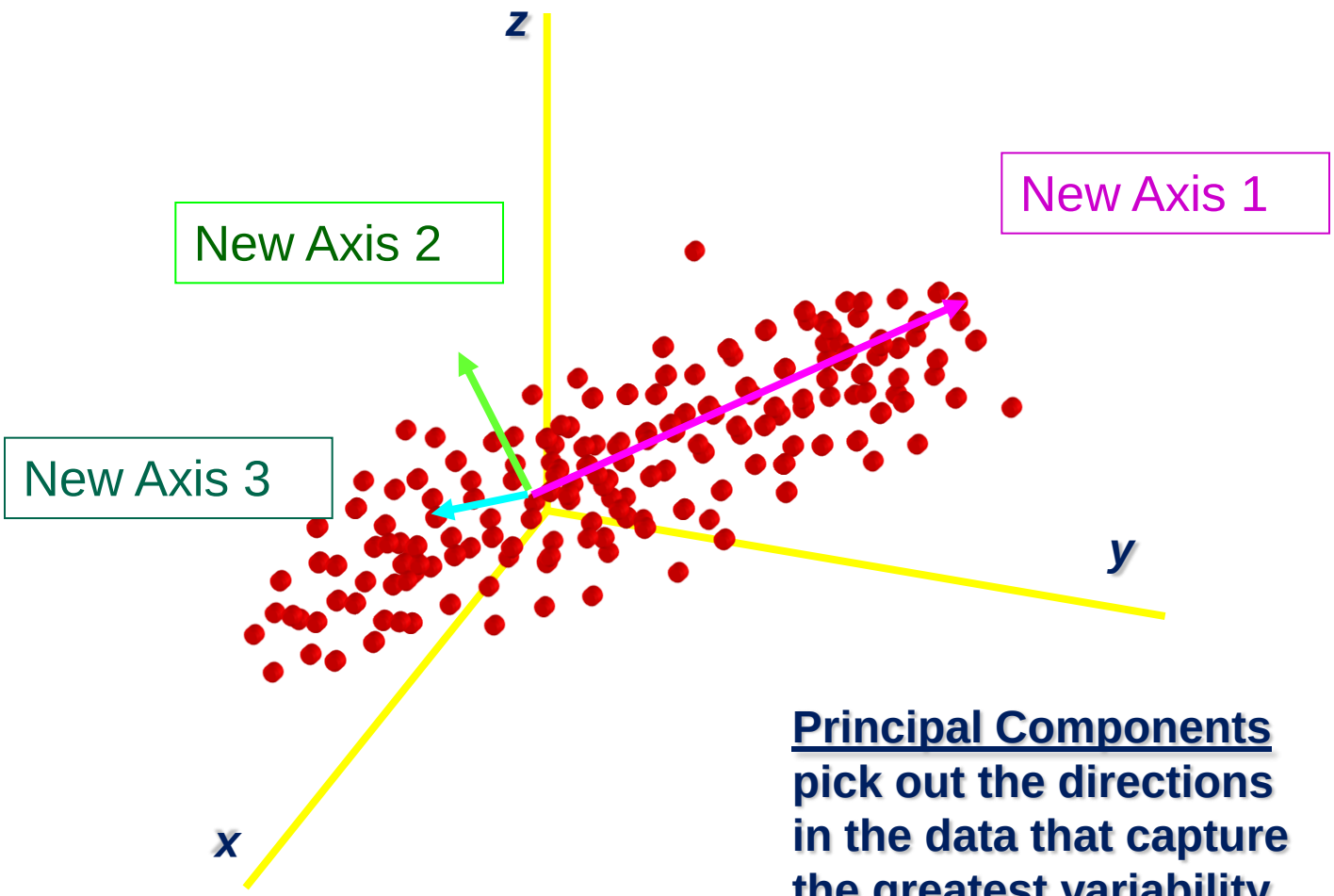
Given A and B datasets.

Find 2 ordinations most similar.

Find successive axes (a_i and b_i) such that
COVARIANCE (a_i, b_i) is maximum



Dimension Reduction (Ordination)



Principal Components
pick out the directions in the data that capture the greatest variability

Cross platform comparison

Circles ●

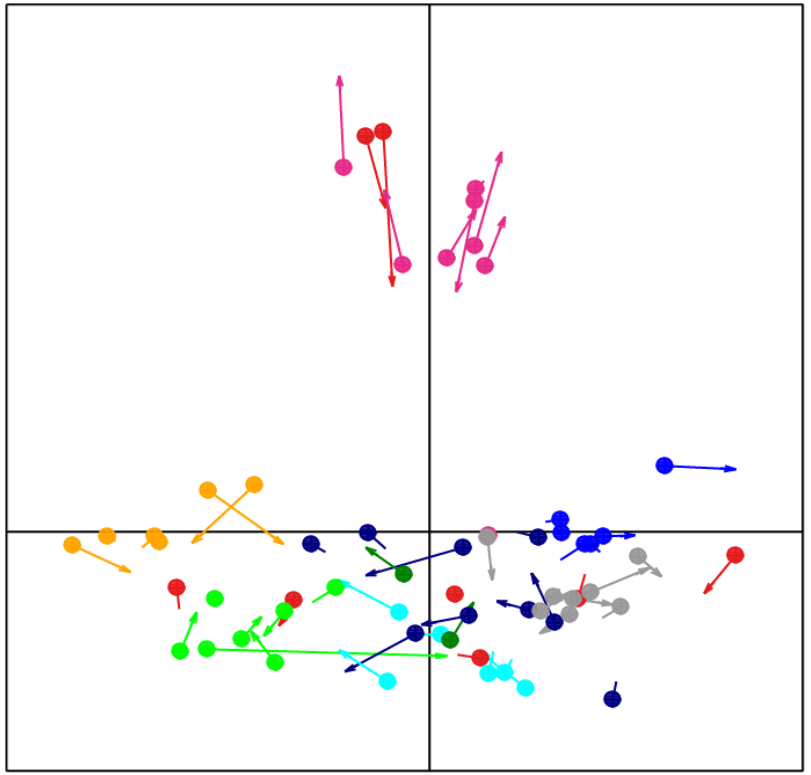
Spotted cDNA arrays

Arrows →

Affymetrix

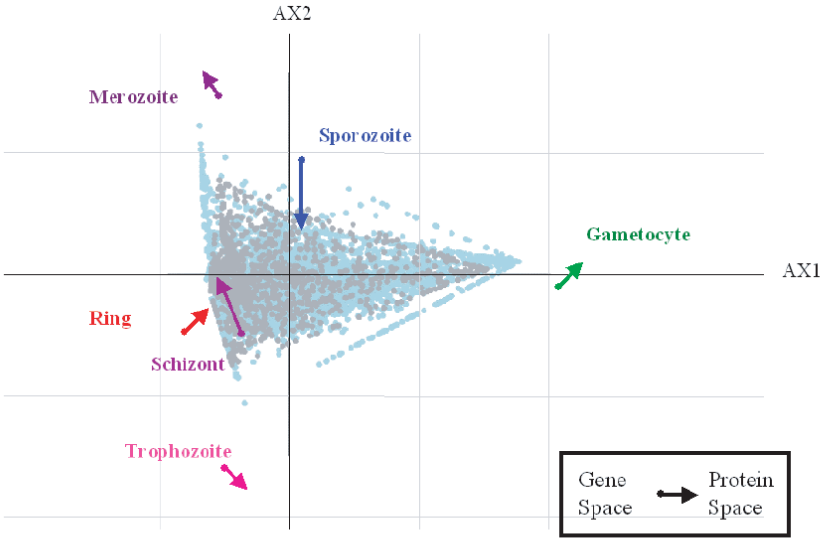
Circles, arrows are joined by line.

Length of line is $\propto$ to divergence between gene expression profiles.

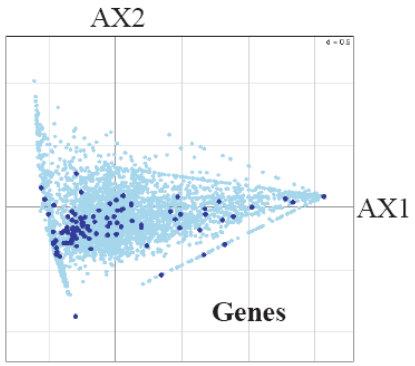


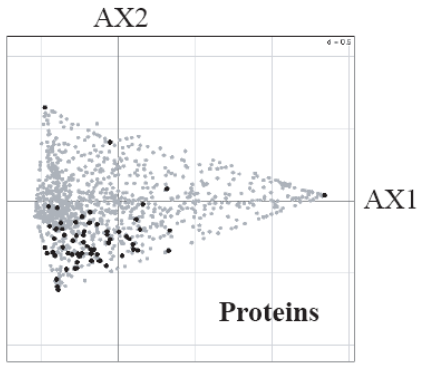
Project GO terms on Genes & Proteins space

Sample with variables (tri-plot)

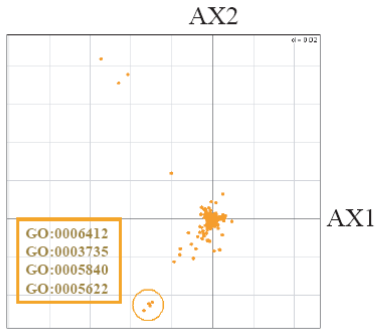


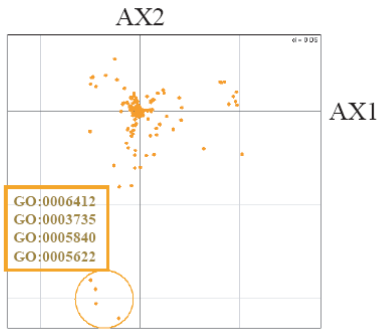
Variables





GO Terms





Axis 1 (horizontal) Accounts for 24.6% variance. Splits sexual & asexual life stages

Axis 2 (vertical) 4.8% variance. Splits invasive stages (Merozoite and Sporozoite stages which invade red blood)

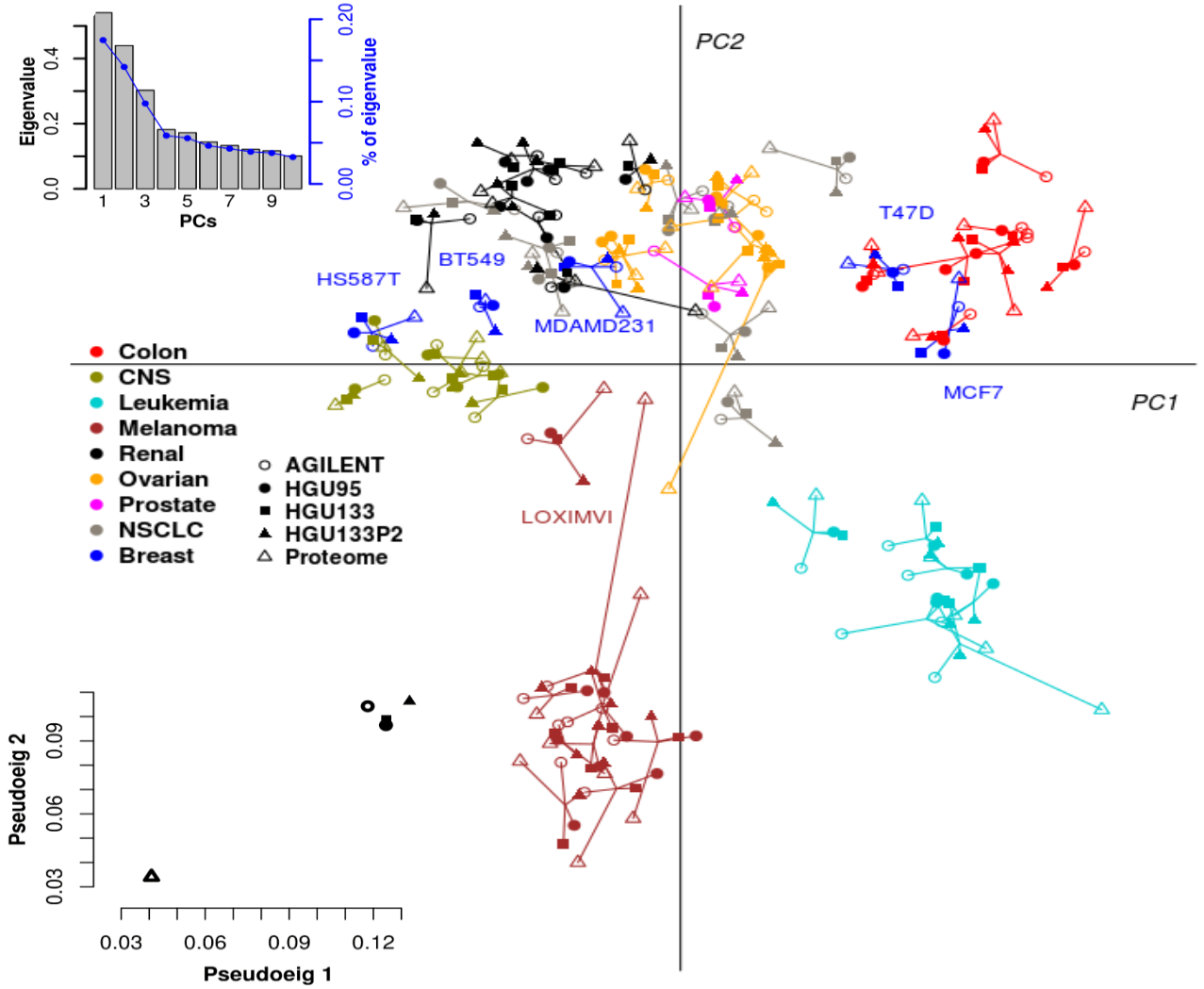
Fagan A, **Culhane AC**, Higgins DG. (2007) A Multivariate Analysis approach to the Integration of Proteomic and Gene Expression Data. *Proteomics.* 7(13):2162-71.

> 2 datasets: MCIA

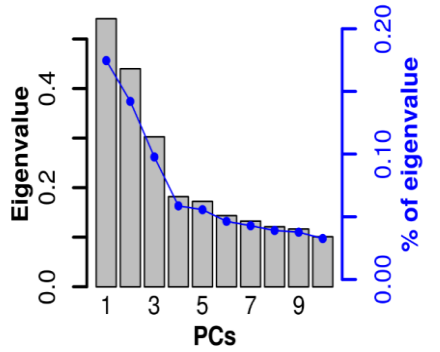
- Application of CIA to two datasets has limited application in integration of ‘omics data for pathway discover.
- Extension – multiple coinertia analysis (MICA)
- Max correlation to psudeo axis

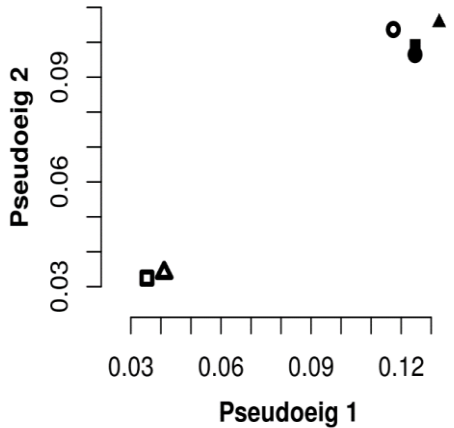
Meng C, Kuster B, **Culhane AC** and Moghaddas Gholami A (2014) A multivariate approach to the integration of multi-omics datasets. BMC Bioinformatics

MCIA of 5 data sets (NCI60)



MCIA of 6 data sets (NCI60)





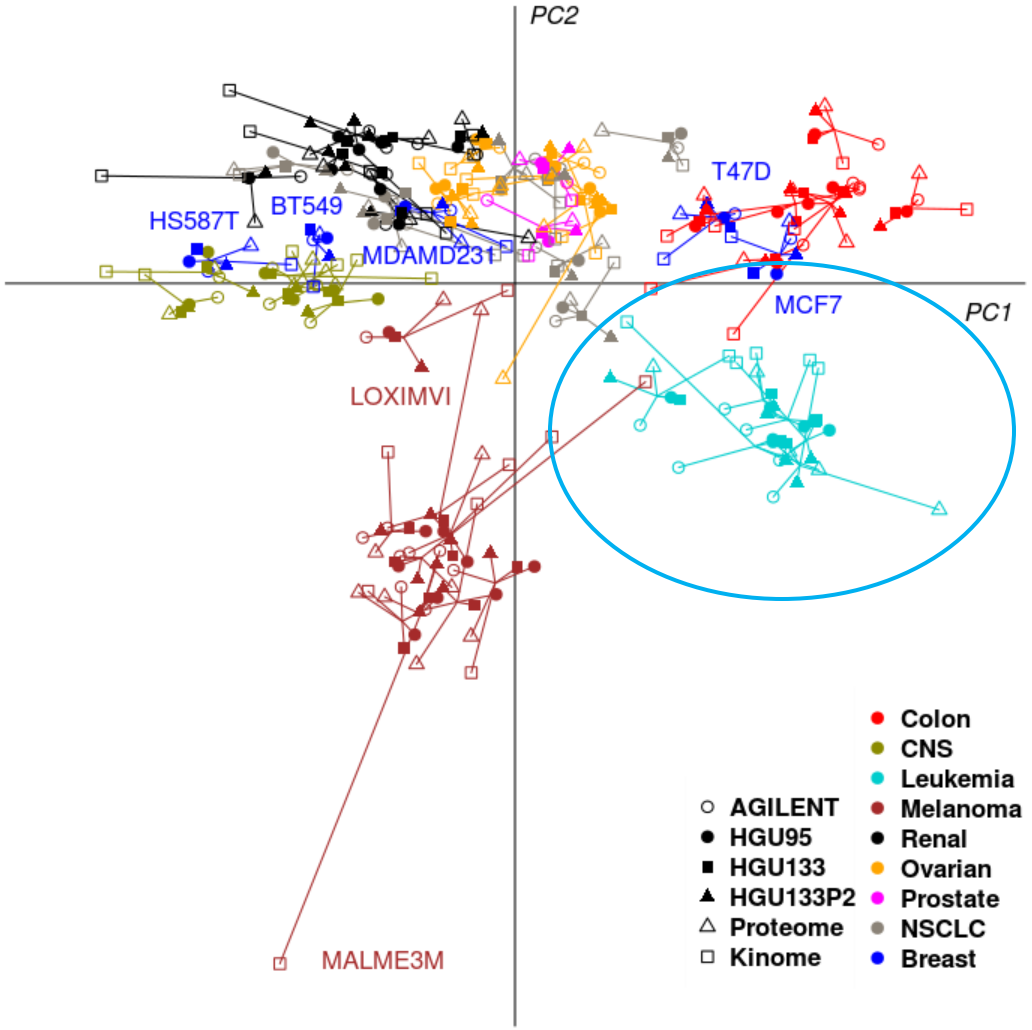
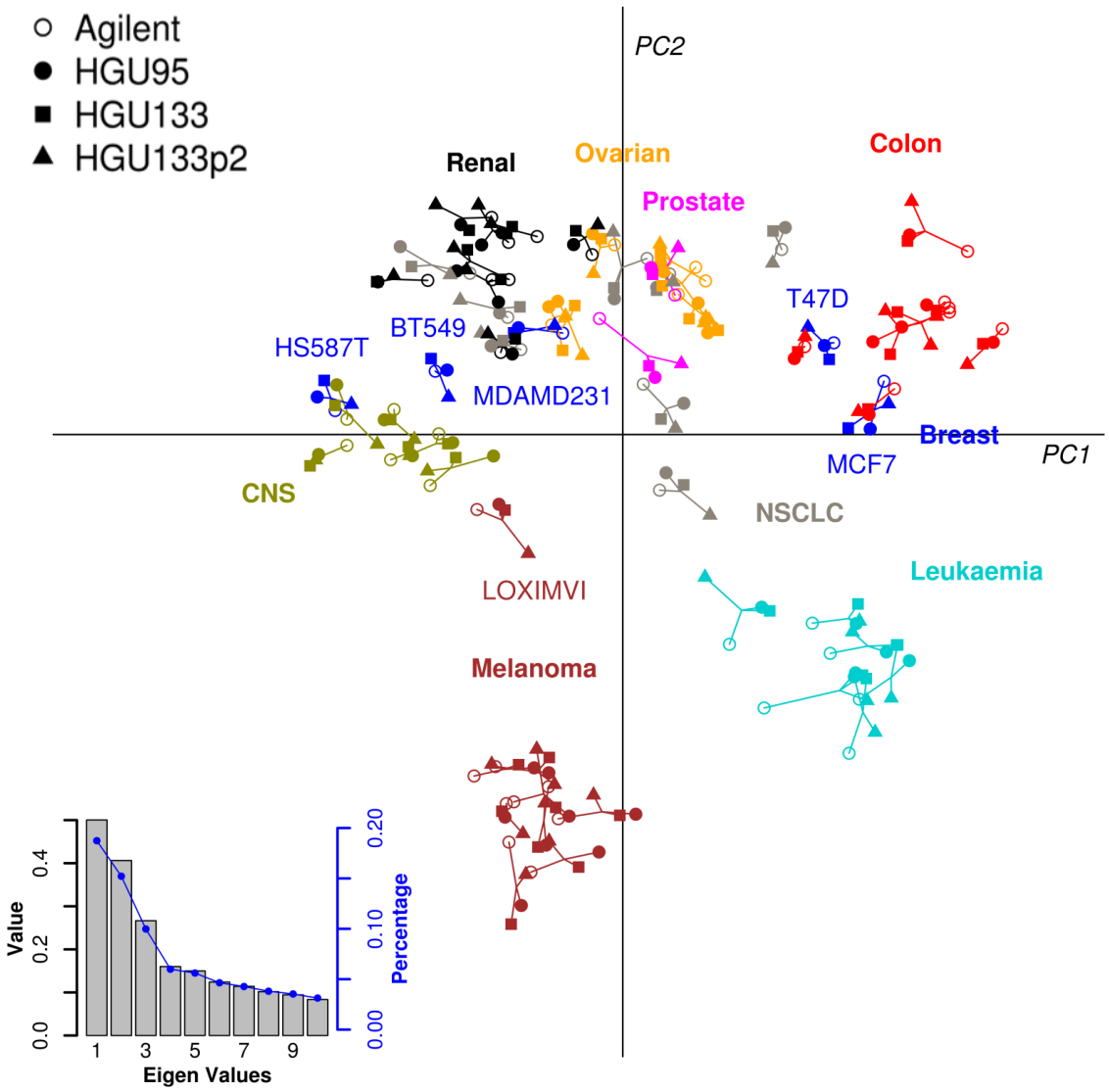
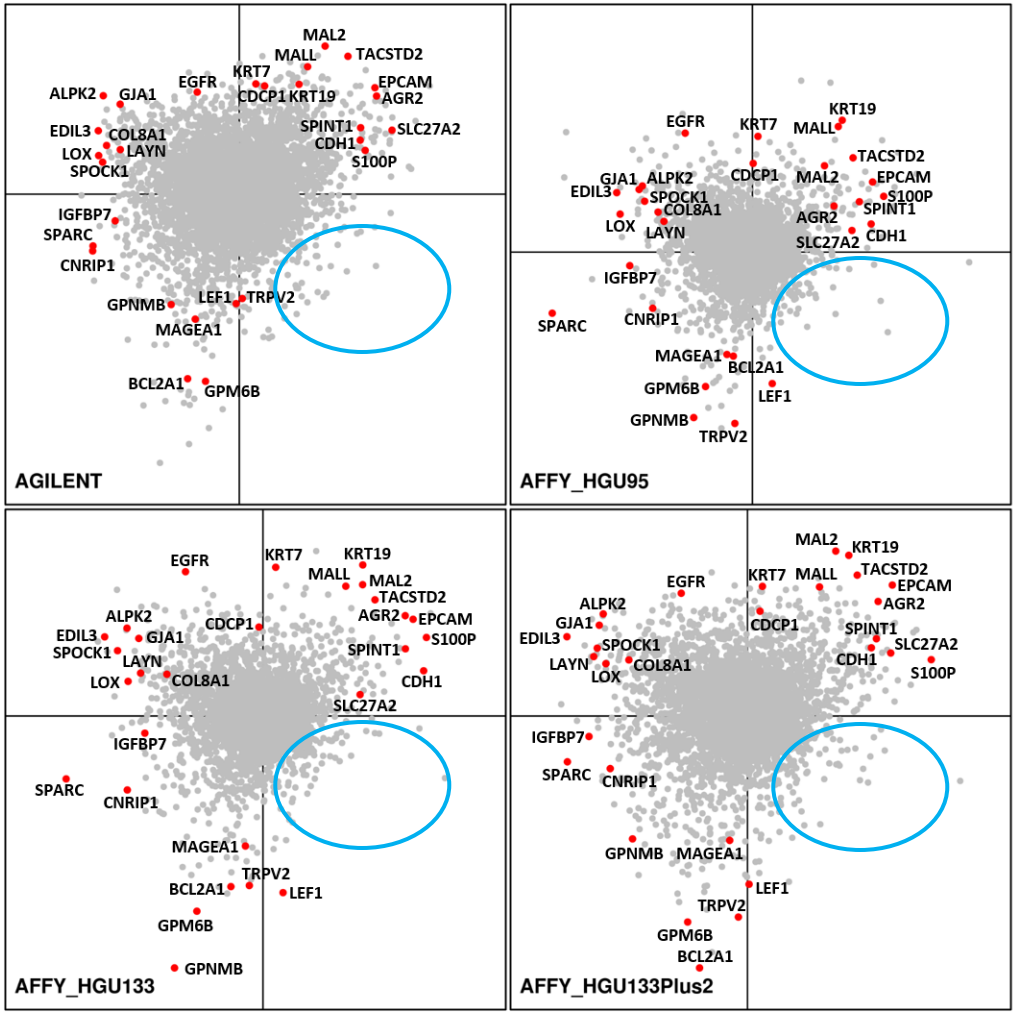


Figure 1 Visualization of the cell lines from multiple datasets. Cell lines are projected on the first two axes. Cell lines from different platforms are distinguished by shapes. Same cell line are linked by edges, the point of intersection is calculated from the synthetic analysis.



Extract features with similar co-ordinates



As features were projected on the same scale, it enables easy pathway analysis of integrated data

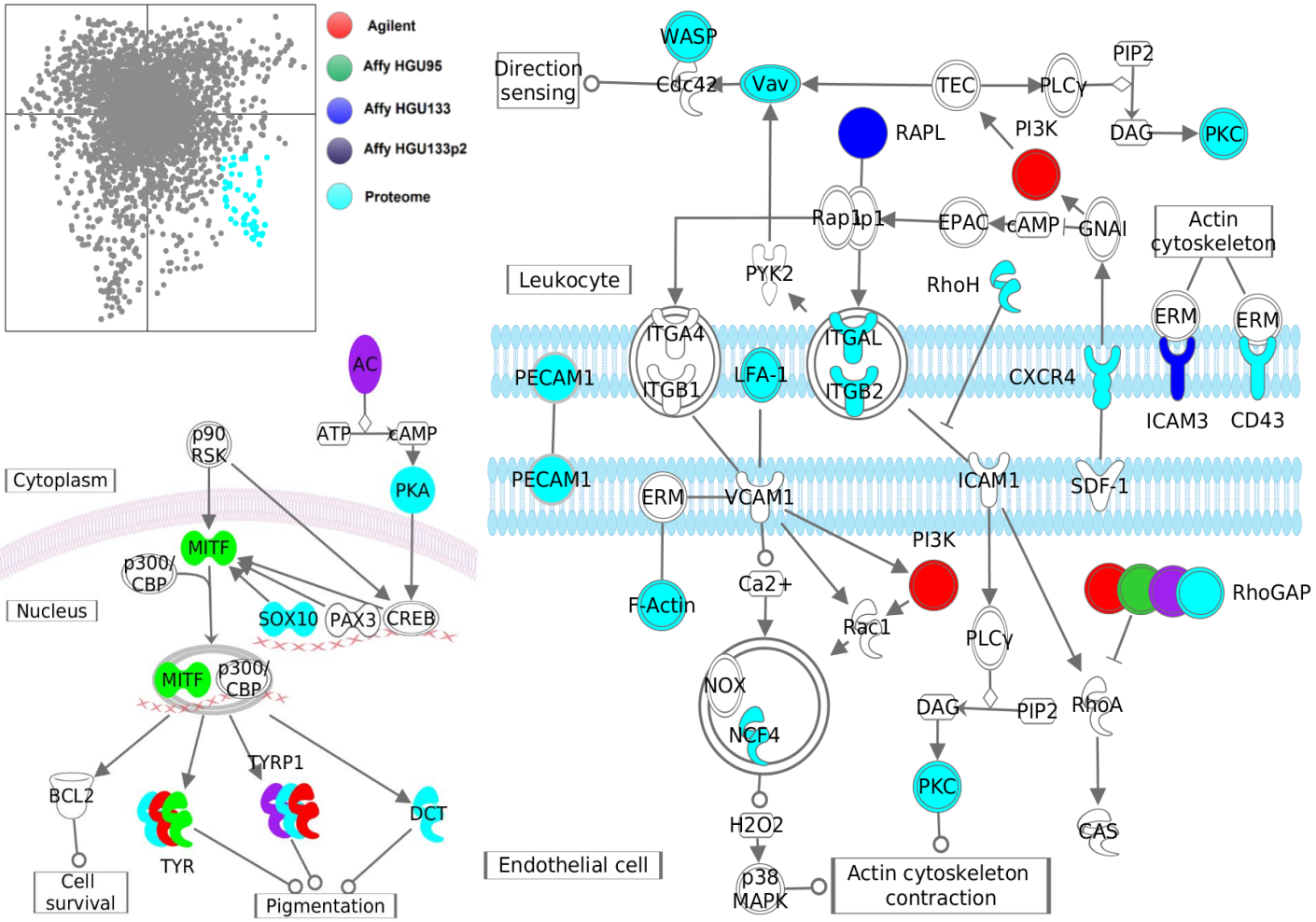
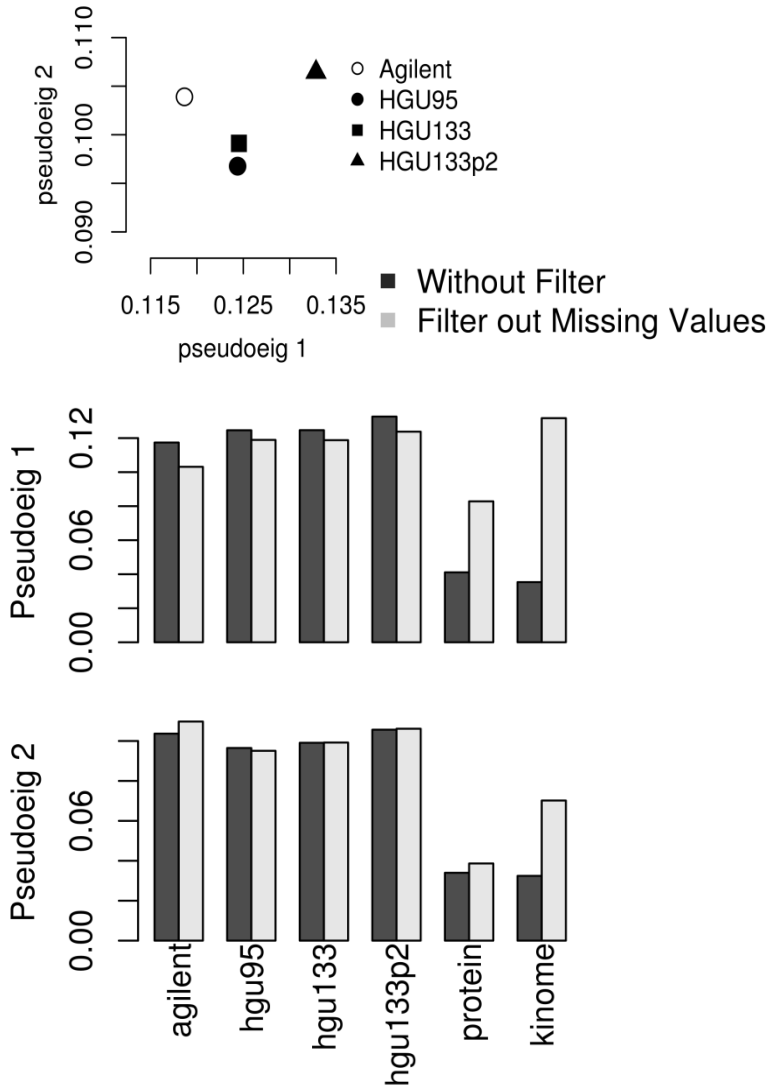


Figure 3 The pseudo-eigenvalue space of NCI60 data. The pseudo-eigenvalues represent the contribution of each dataset to an axis. A. The pseudo-eigenvalue space corresponding to figure 1 and 2. B and C show the comparison of variance before filtering out and after filtering genes with missing values in the proteomic and kinomic data. B. The changes on the first principal component. C. The changes on the second principal component.



Summary

- MCIA builds from CIA. Projects molecular profiles of matched samples onto the same space and scale
- Identified most covariant/divergent trends across different data types
- Useful for pooling and extracting most variant features across data sets
  - genes, proteins transformed onto same scale
- Not suited for large scale studies of hundreds of studies.

Meng C, Kuster B, **Culhane AC*** and Moghaddas Gholami A * (2013) A multivariate approach to the integration of multi-omics datasets. To be submitted

TCGA data- Ovarian

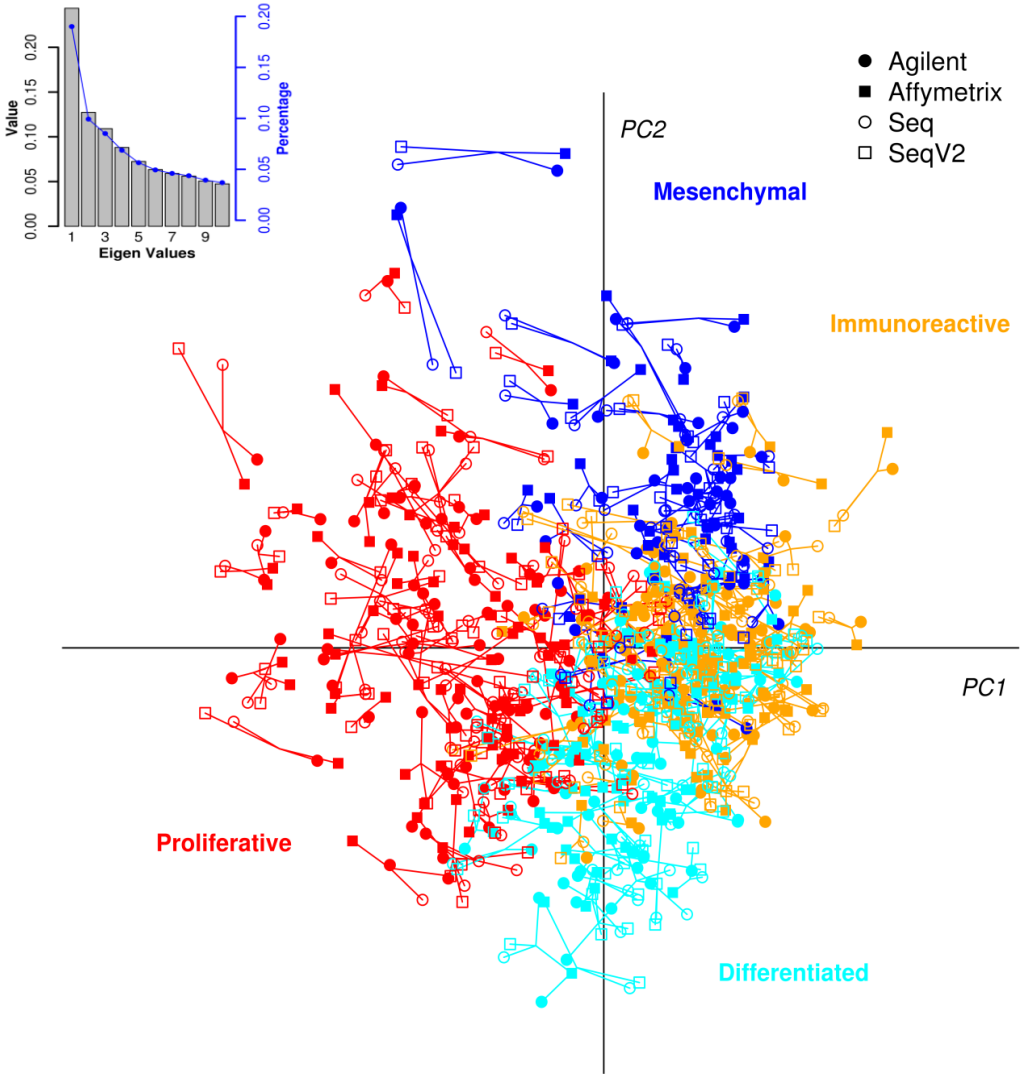


Figure 4 Visualization of the patient from multiple platforms. Each patient is projected on the first two axes and same patients are linked by edges to synthetic center. The scree plot on the top left corner shows the eigenvalue of each axis.

Figure 5 Visualization of molecule space of ovarian data. The first two principal component is plotted. The genes in CLOVAR subtype signature are coloured. The colour code of genes corresponds to the subtype defined in figure 4.

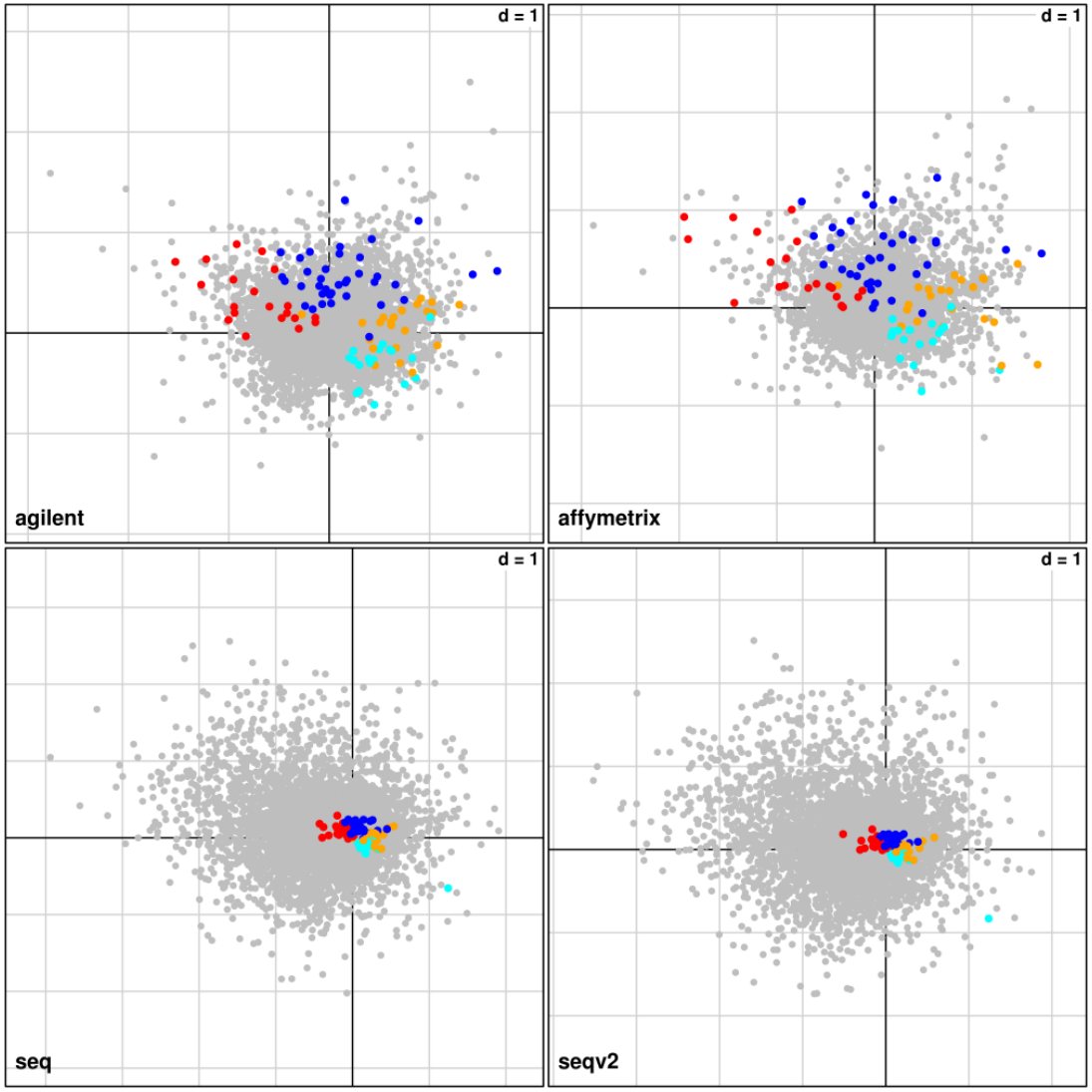
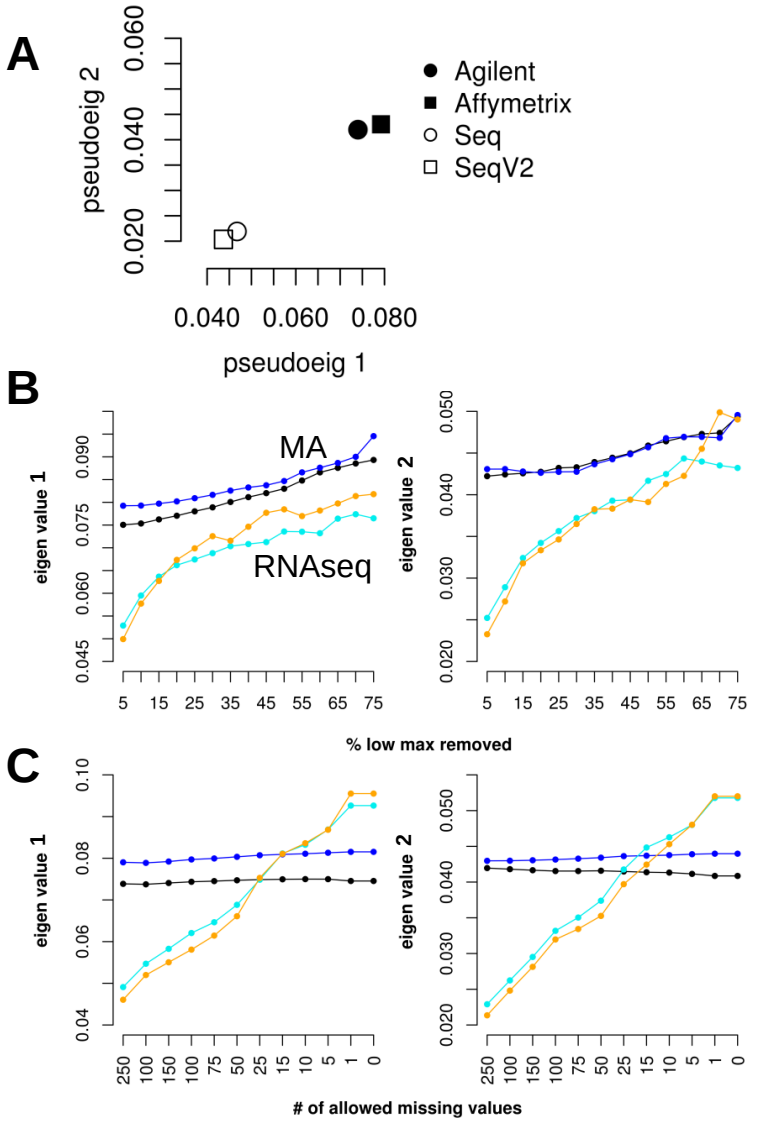


Figure 6 The pseudo-eigenvalue space of ovarian data. A: The pseudo-eigenvalue space corresponding to figure 4 and 5. B: all datasets are filtering by the maximum expression of genes. A percentage of genes are excluded. Eigenvalue 1 and eigenvalue 2 are compared. C: The RNA sequencing data are filtered by the number of missing value of genes. Top two PCs are compared.



Projecting Annotation on MCIA

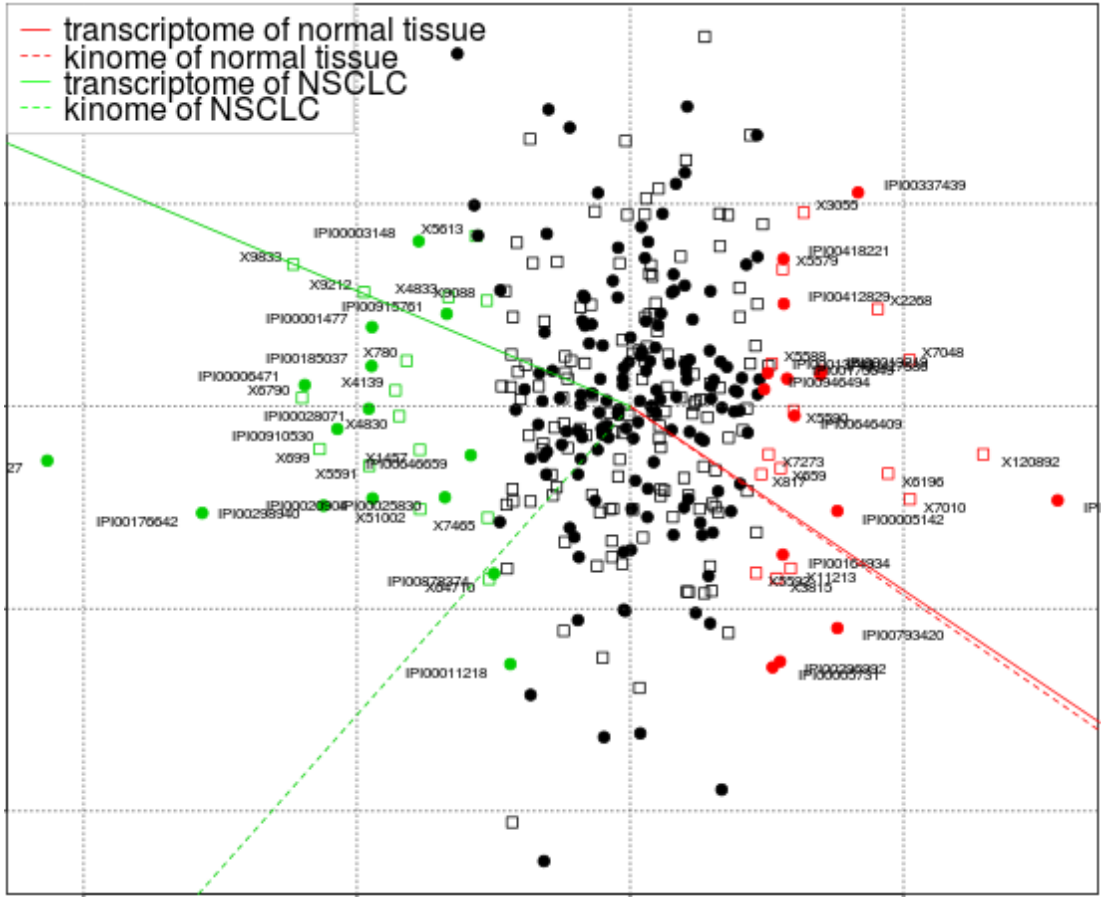
- There are four major entities to be represented (Normal, NSCLC, proteins and transcripts)
  - transcripts from normal tissue
  - transcripts from NSCLC
  - kinases from normal tissue
  - kinases from NSCLC
- Data are grouped into four clusters. Lines to each cluster centers (median of each cluster) are plotted to indicate the direction of clusters. Terms best associated with each cluster are colored
- Euclidean distance of the same term from different data is calculated, then use random permutations to select the most distanced terms according to the empirical p-value

Projecting kinomics and transcriptomics data

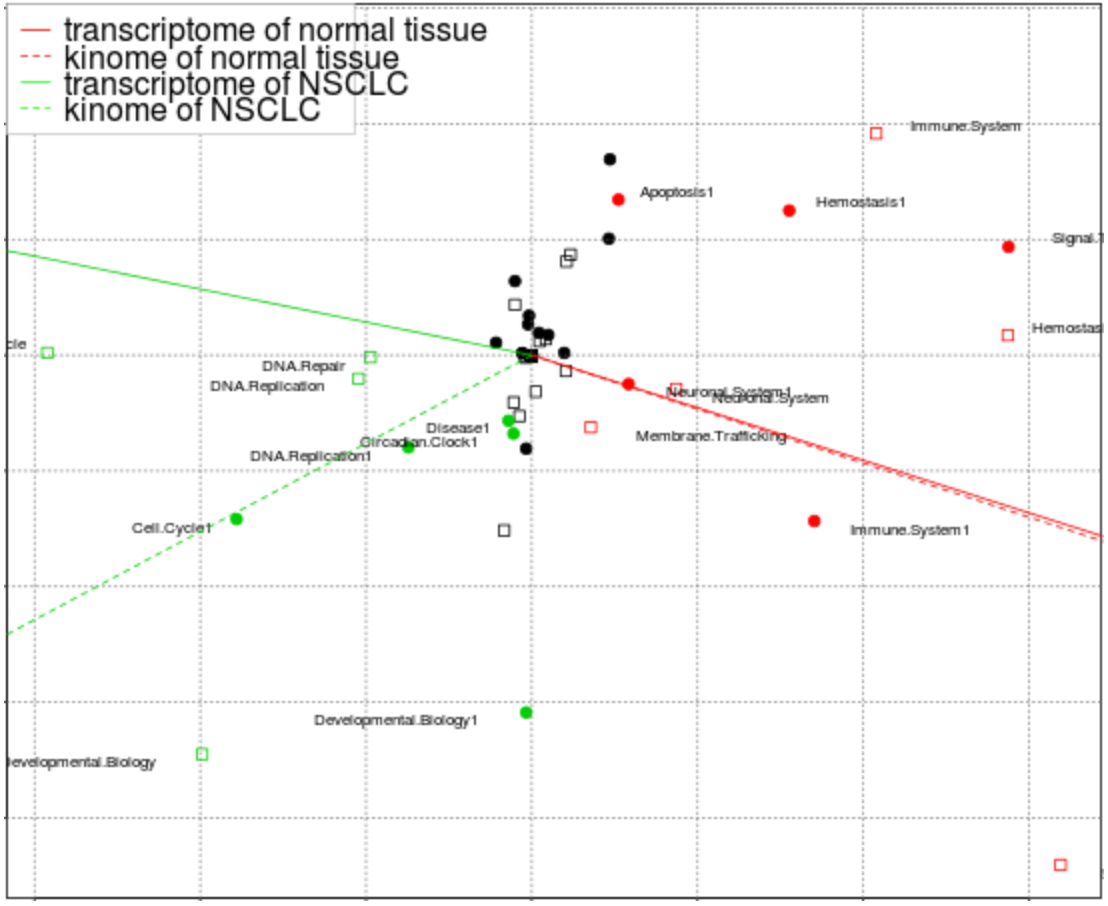
- 175 genes/proteins pairs.
- Solid circles are kinases, rectangles are transcripts.

- All samples are divided into 4 classes (their directions are indicated by lines).
- Transcriptomics represented by solid lines, kinomics by dashed lines. Red indicates normal tissue, green shows NSCLC)

- The molecules best aligned with each direction are highlighted by colors.



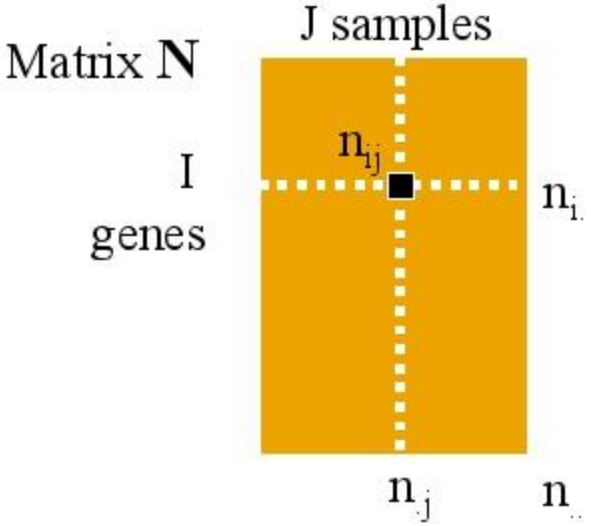
Projecting pathways (Reactome)



Terms from Transcriptome	Terms from Kinome
Developmental Biology	Developmental Biology
Cell Cycle	Cell Cycle
Metabolism	Disease
DNA Replication	DNA Replication
DNA Repair	Circadian Clock
Signaling Transduction	Signal Transduction
Hemostasis	Hemostasis
Immune System	Immune System
Neuronal System	Neuronal System
Membrane Trafficking	Apoptosis

COA: Initial Transformation

Matrix **N**

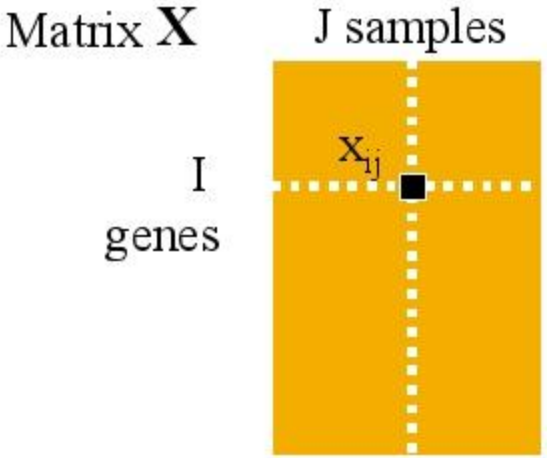


$$c_j = n_{.j}/n_{..}$$

$$r_i = n_{i.}/n_{..}$$

$$p_{ij} = n_{ij}/n_{..}$$

Matrix **X**



$$x_{ij} = (p_{ij} - r_i c_j)/\sqrt{r_i c_j}$$

Pearson chi-square statistic O_{ij}-E_{ij} / $\sqrt{E_{ij}}$