Molecular features based cancer classification and stratification

Dongfang Wang

Tsinghua University

2016-09-02

Cancer heterogeneity and stratification

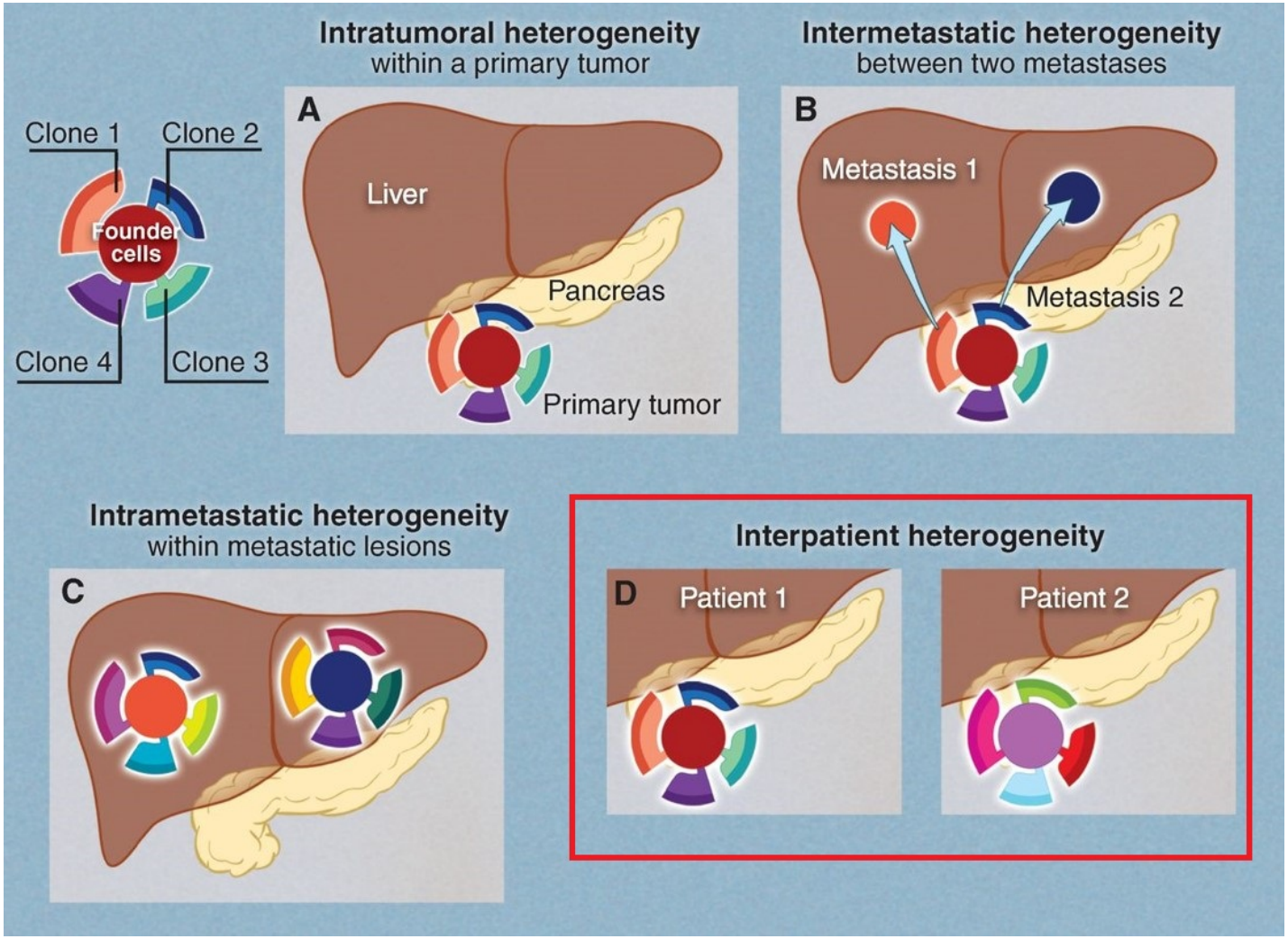
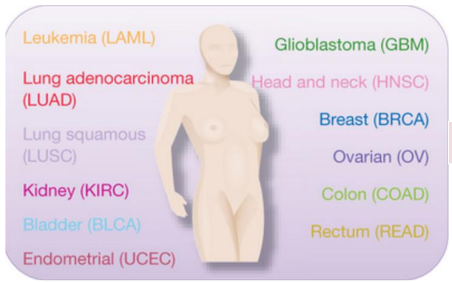
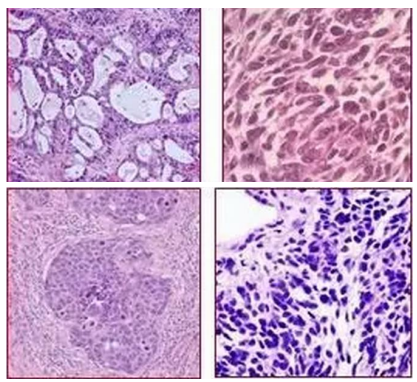


Figure: **Four types of genetic heterogeneity in tumors.** Vogelstein et al.,*Science*,2013

Cancer stratification:

- Better undersatnding of tumor biology
- Subtype specific drug and therapy
- Patient tailored treatment
- ...





1.0 Primary tumor apparatus

2.0 Clinical Pathology

Fig: four types of lung cancer

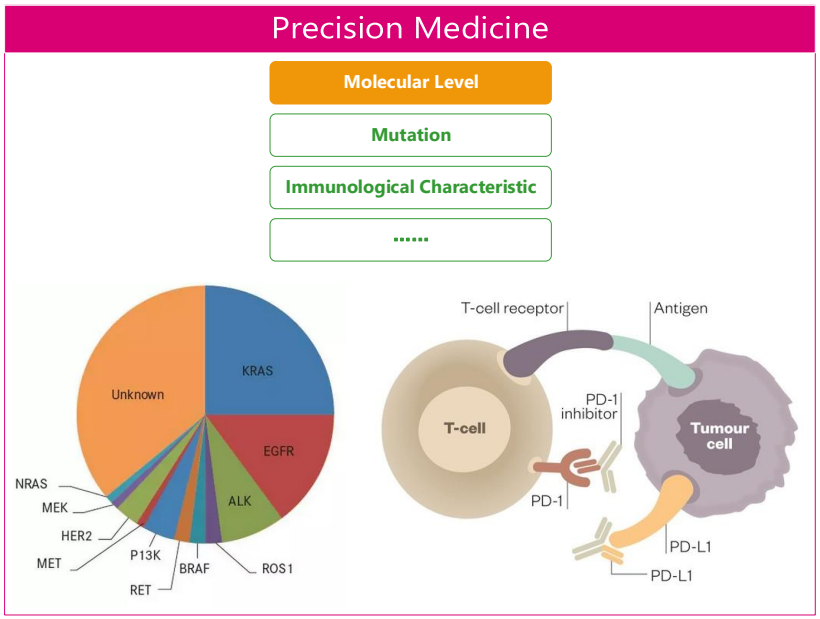


Figure: **Development of cancer stratification.** Figs from Internet.

Molecular features based stratification

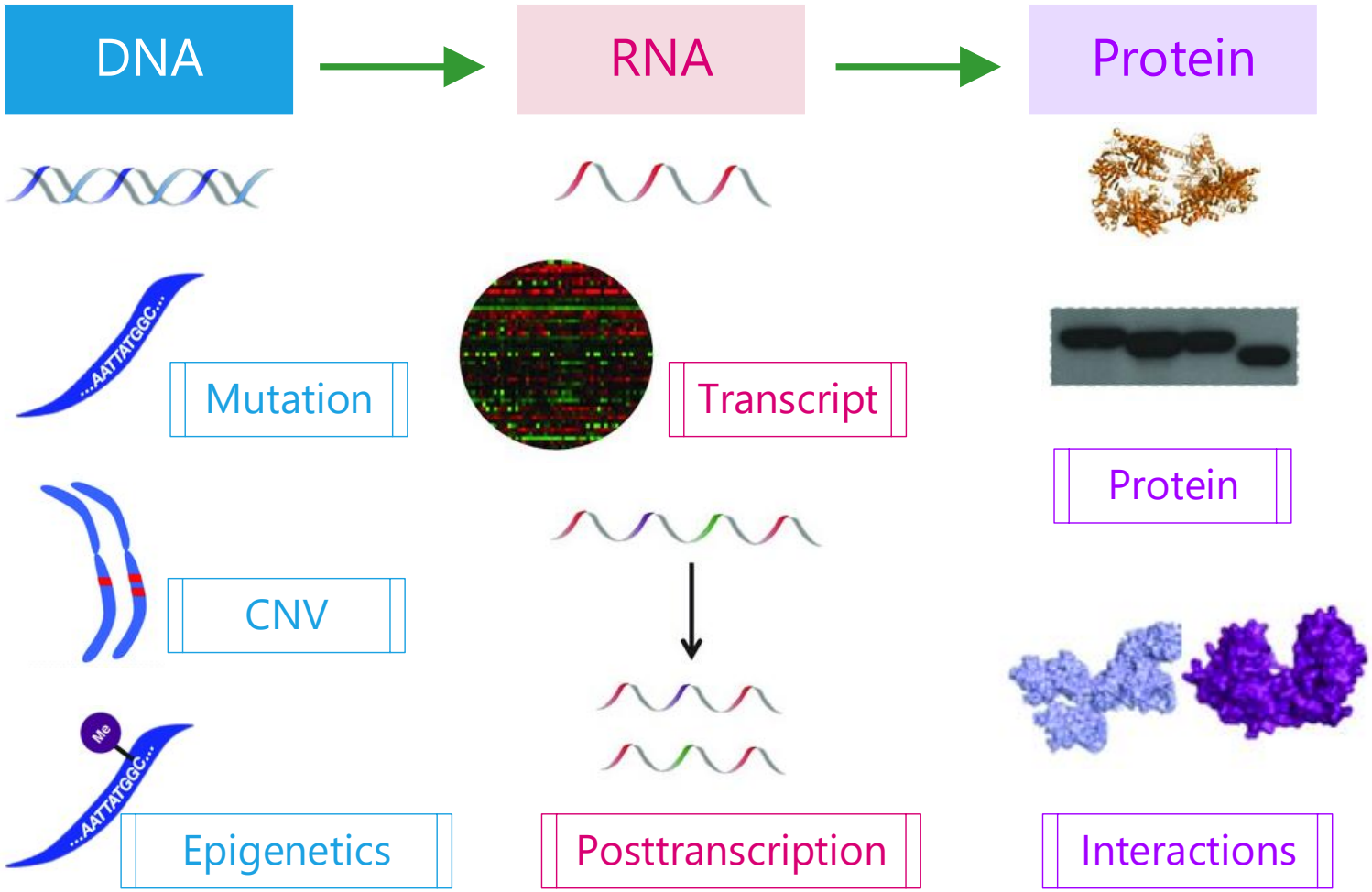


Figure: **Various molecular features available for tumor patients.** Figs from Internet.

Different molecular levels indicates various kinds of information for subtyping:

- Different driver mutations (Ciriello et al., *Nat.Genetics*, 2013, etc.)
- CpG island hypermethylation defines CIMP(CpG island methylator phenotype) subtypes (Issa et al.,*Nat.Rev.Cancer*,2004, etc.)
- ER/HER2/PR define breast cancer subtypes (Cancer Genome Atlas,*Nature*, 2012, etc.)
- ...

Computational Concerns

1. **Collections of available datasets:**
   1. **Public available cancer datasets**
      - Multi-omics cancer samples: TCGA, ICGC ($\sim 10^2 - 10^3$)
      - Gene Expression data: GEO, ArrayExpress ($\sim 10^3 - 10^5$)
      - More omics data in cancer cell lines: CCLE
      - Biological knowledge: COSMIC, MSigDB, DAVID
   2. **Own collected datasets**: specific and accurate information ($\sim 10^2$)
2. **Characteristics and difficulties:**
   - High-dimensional features with limited samples
   - Multi-source information
   - Heterogenous datasets

Content

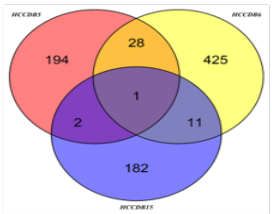
1. Cancer subtypes across multiple datasets
2. Cancer subtypes indicated by multi-omics
3. Gene mutation and Sorafenib drug response

Cancer subtypes across multiple datasets

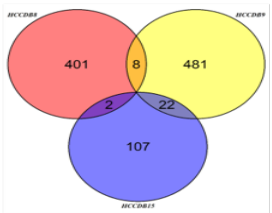
- More samples ($\sim$ 2000 HCCs) to support possible discovery.

Group	mRNA type	miRNA type	Total
HCC	2031	625	2109
Adjacent	1199	224	1277
Cirrhotic	40	0	40
Healthy	8	10	18
Total number of samples	3286	859	3444

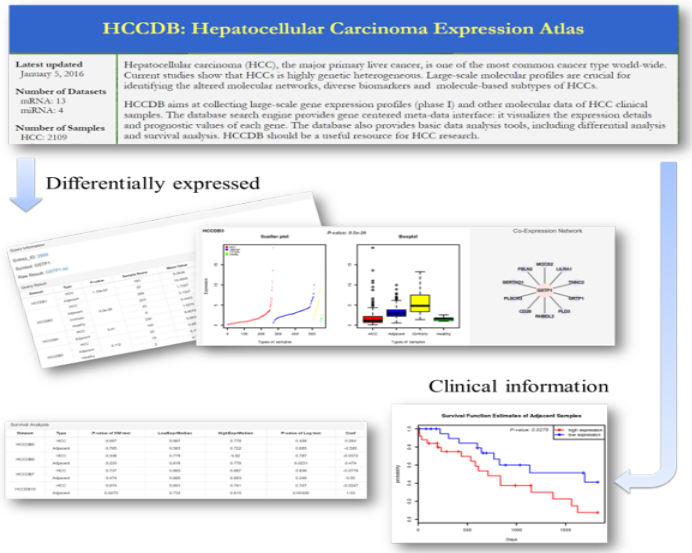
Table I: Number of samples in cruated HCC datasets



(a) Number of genes associated with overall survival



(b) Number of genes associated with vascular invasion



(c) HCCDB database

Figure: HCC Databse made by Zhang Guchao

Summary of current methods

- **Strategy 1:**

  Stack all datasets into one big matrix after "removing" batch effects.

  (J Cuinney et al., *Nature Medicine*, 2015; Johnson WE et al., *Biostatistics*, 2007; etc.)

- **Strategy 2:**

  Compare "similarity" of clusters from different datasets.

  (Hoshida Y, et al. *Cancer Research*,2009; CR Planey, et al., *Genome Medicine*, 2016.)

  Similarity computation is based on signatures of each dataset.

- **Key questions:** How to **represent shared information** across different datasets/studies/cohorts?

- **Hypothesis:** driver signature modules and their degrees of perturbation are shared among datasets.

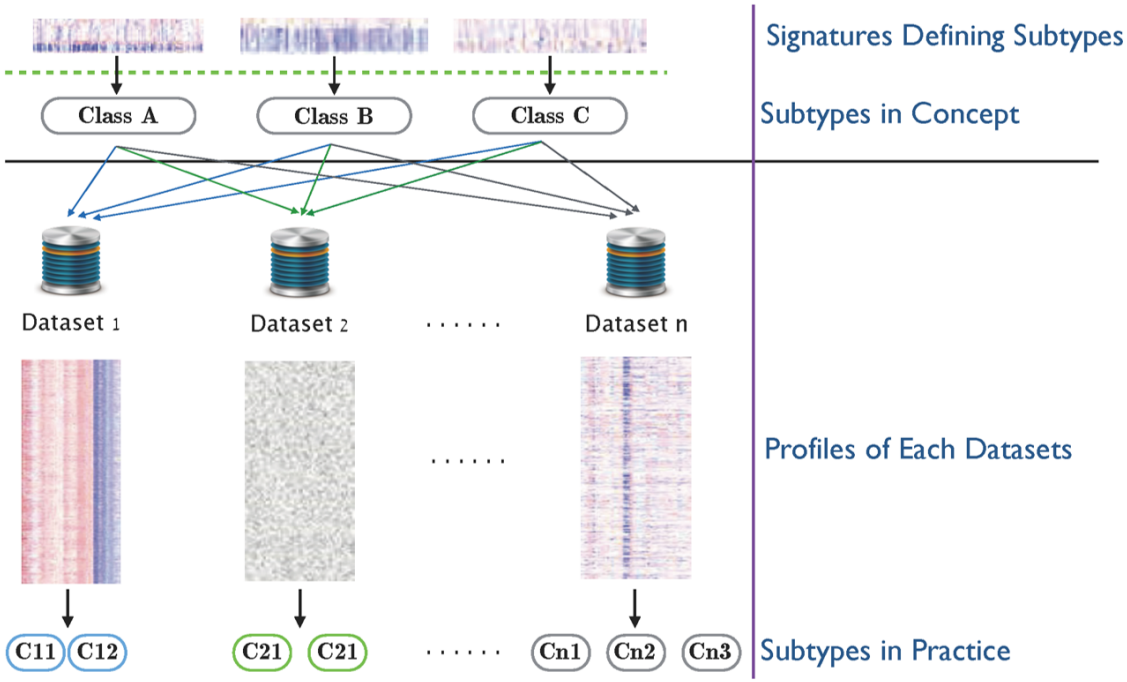


Figure: Basic framework of our model.

- **Definition of signature modules:**

1. Firstly, we divide every dataset into appropriate clusters c_{ij}, and obtain representative gene lists $\boldsymbol{g}_{ij}$, where i indexes dataset and j cluster.
2. Then integrating with literature genes and PPI network, we define signature modules M_k.

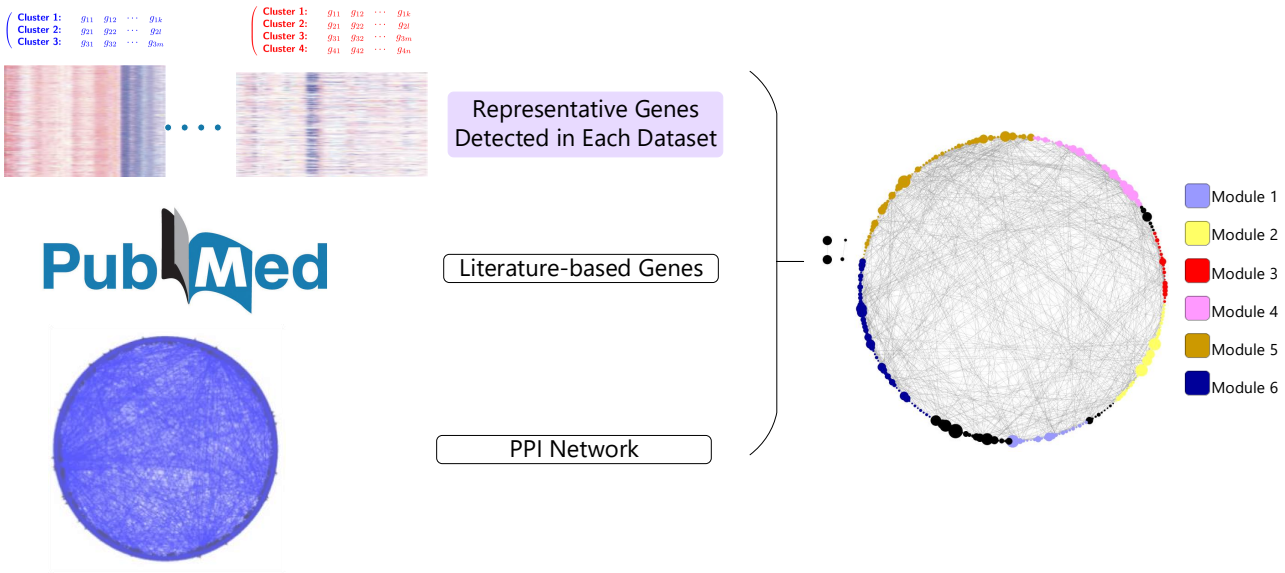


Figure: Discovery of signature modules across multiple datasets.

- Functional annotation of signature modules:

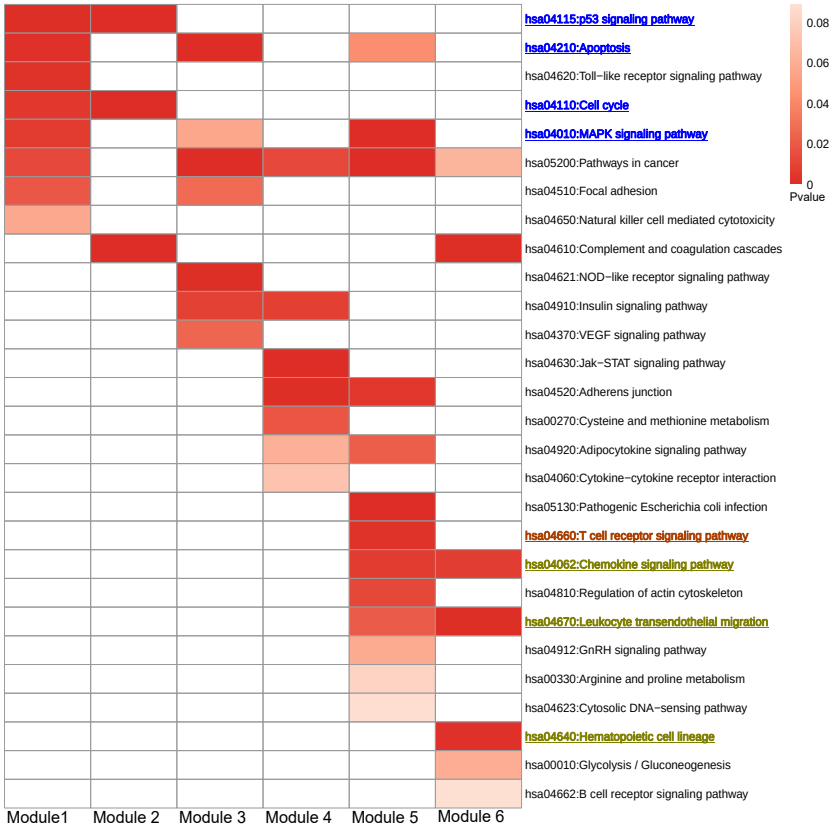


Figure: Function annonation of signature modules by DAVID.(Huang DW, et al., *Nature Protoc.*, 2009)

- **Perturbation of signature modules in each cluster c_{ij}:**
  - Compute median expression as the template of every cluster in each dataset.
  - Compute z-score of each template's enrichment in all 6 modules.

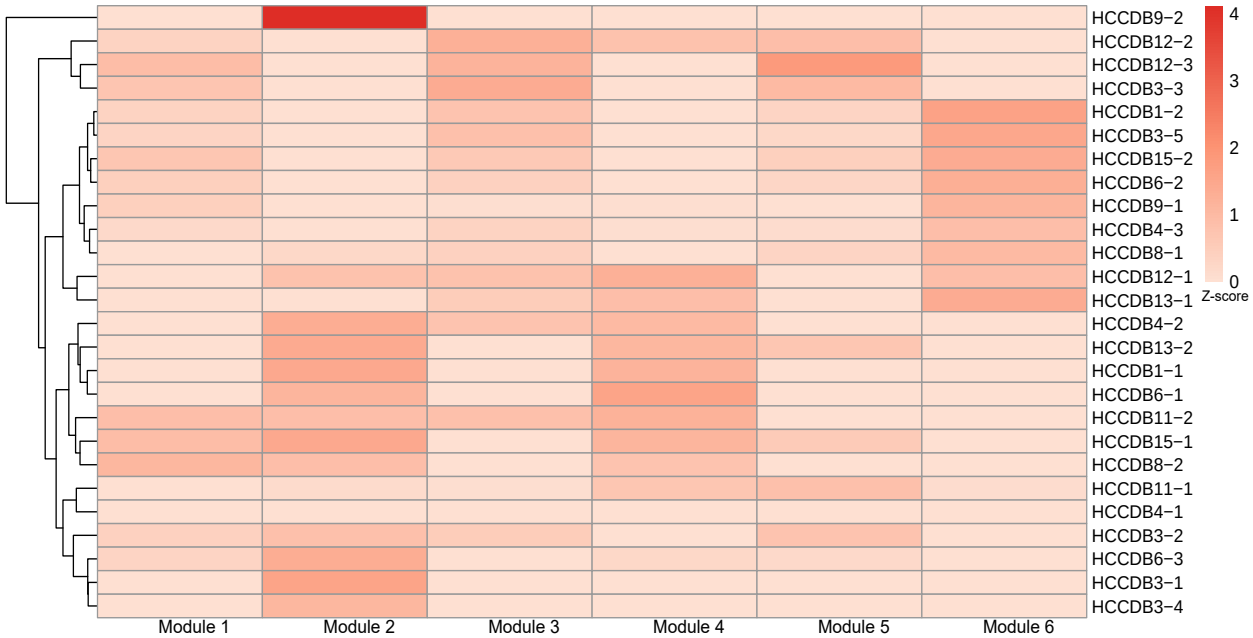


Figure: Perturbation vectors of every cluster in each dataset.

- **Consensus perturbation vectors across datasets:**

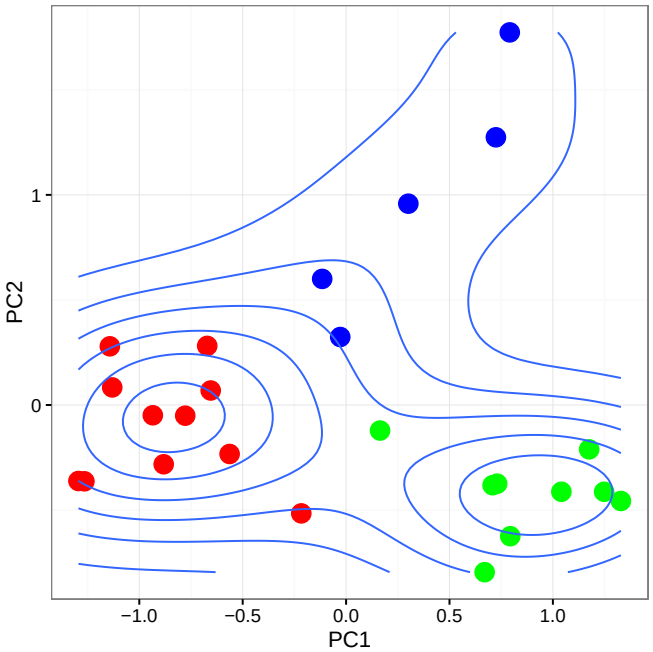


Figure: PCs of perturbation vectors for each cluster.

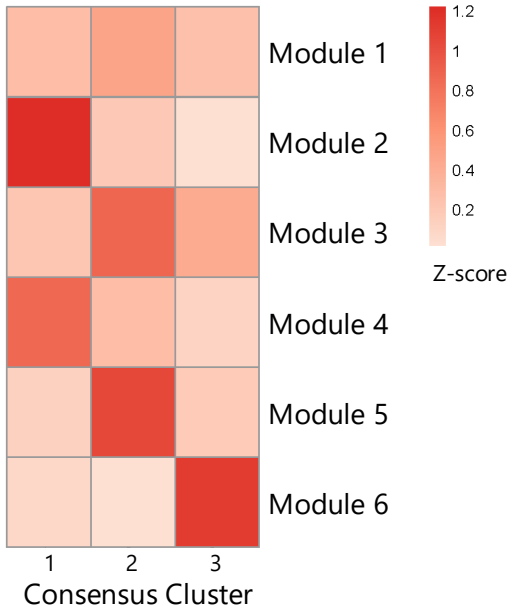


Figure: Heatmap for three consensus perturbation centers.

- **Survival analysis results compared with clustering by single dataset:**

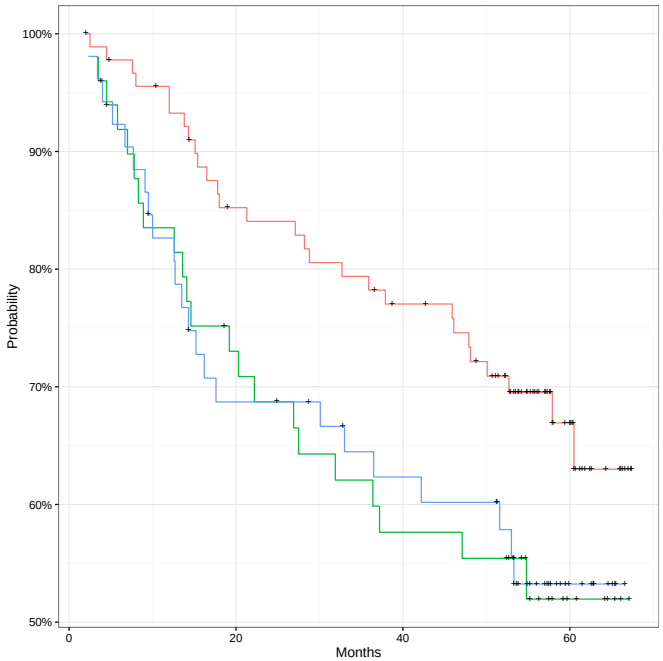


Figure: Survival analysis by single clustering of HCCDB6. cox: 0.05

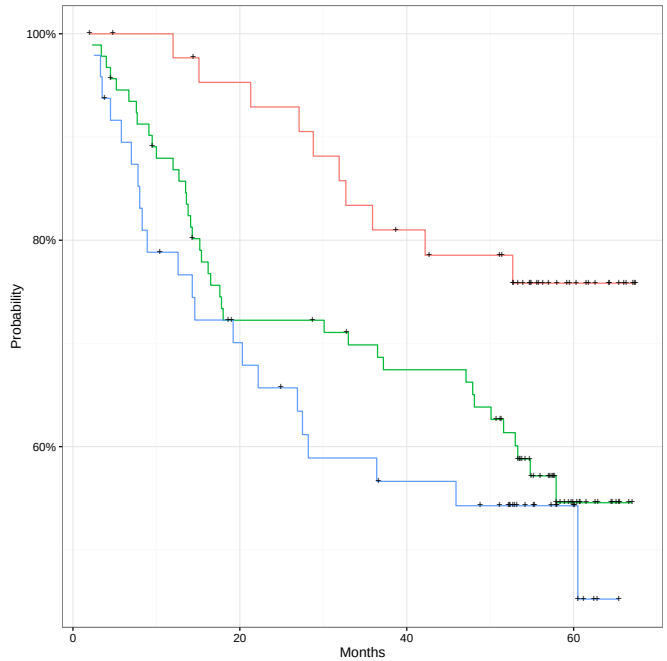


Figure: Survival analysis by consensus clustering of HCCDB6. cox: 0.008

Summary

- Try to perform collective clustering across different datasets from one type of cancer,
- Based on their shared driver signature modules.
- Preliminary results support our hypothesis.

- The current model is still very simple.
- Further improvements and some details need to be re-considered.

Cancer subtypes indicated by multi-omics

An example:

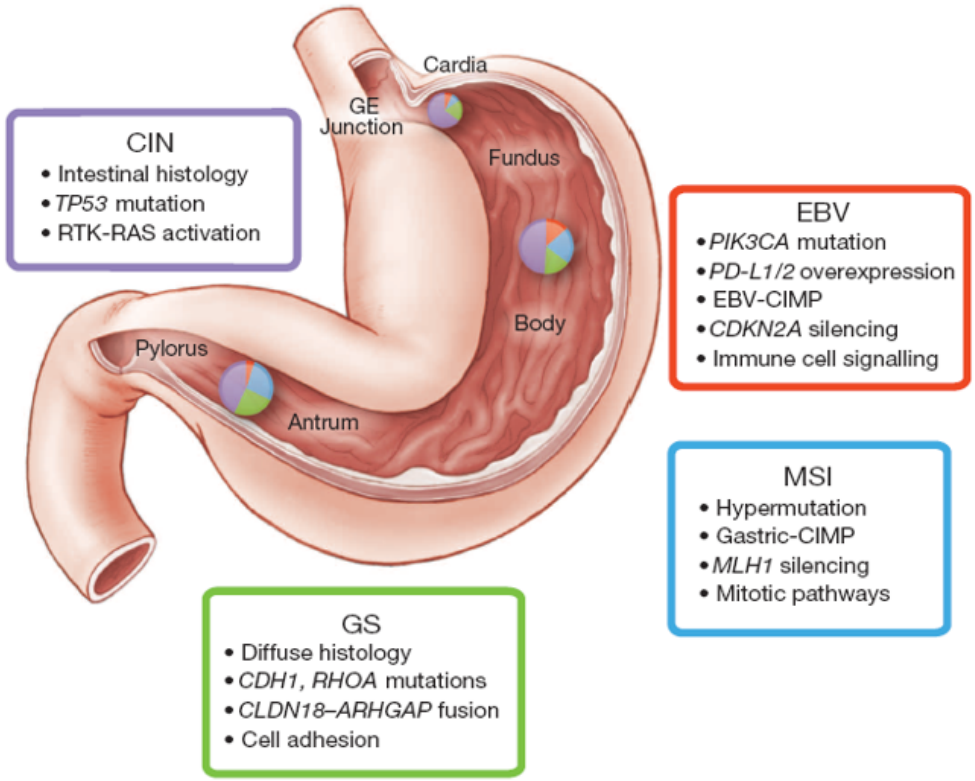


Figure: Subtypes of gastric adenocarcinoma, (TCGA, *Nature*, 2014)

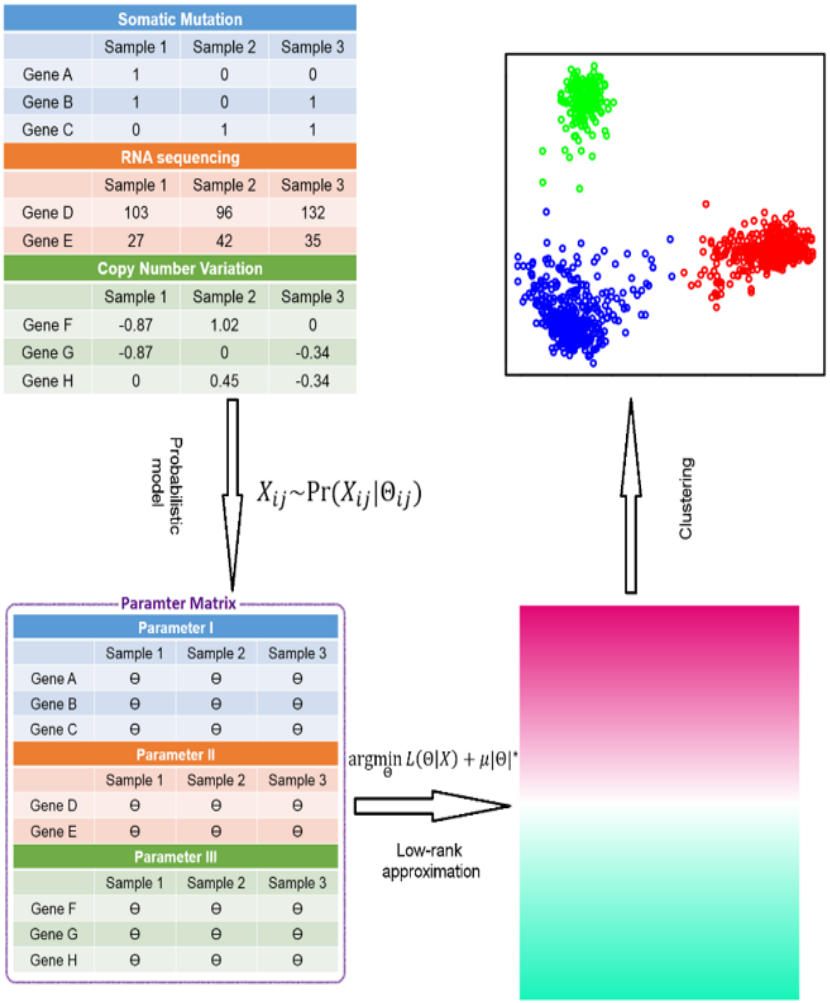
Summary of current methods

Table 1. Selected methods for unsupervised clustering of multi-omics data.

Strategy	Description	Methods	Basic model	Refs.
Direct integrative clustering	Put multi-omics datasets into a stacked matrix, and take this stacked matrix as input for the following clustering analysis	Super *k*-means	Direct clustering with BIC	[12]
		iCluster+	Latent factor analysis	[13]
		JIVE	Low rank-based approximation	[14]
		LRAcluster	Low rank-based approximation	[15]
		jNMF	Non-negative matrix factorization	[16]
		Pathifier	Pathway-based integration	[17]
Clustering of clusters	Perform clustering analysis on every single omics dataset and then integrate the primary clustering results into final cluster assignments	COCA	Clustering of intermediated clusters	[3]
		MDI	Latent *Dirichlet* allocation	[18]
		BCC	Latent *Dirichlet* allocation	[19]
		SNF	Similarity network fusion	[20]
Regulatory integrative clustering	Focus on driver variations by considering the regulatory structures between different molecular layers	PARADIGM	Pathway-based integration	[21]

Figure: Summary of current integrative clustering methods. Wang,et al., *Quantitative Biology*, 2016.

LRAcluster: overview



$$\min_{\boldsymbol{\Theta}} \; -\log \mathcal{P}(\boldsymbol{X}|\boldsymbol{\Theta}) + \mu|\boldsymbol{\Theta}|_*$$

- $\boldsymbol{X}$: omics data matrix;
- $\boldsymbol{\Theta}$: parameter matrix;
- $\mathcal{P}$: probability distribution, Gaussian for real number, Binomial for binary, etc.
- $||_*$: nuclear norm.

Note: cooperate with Wu Dingming (Wu et al., *BMC Genomics*, 2016). I helped design this algorithm and did experiments.

LRAcluster: results

Simulation: mixture of samples from three kinds of cancers

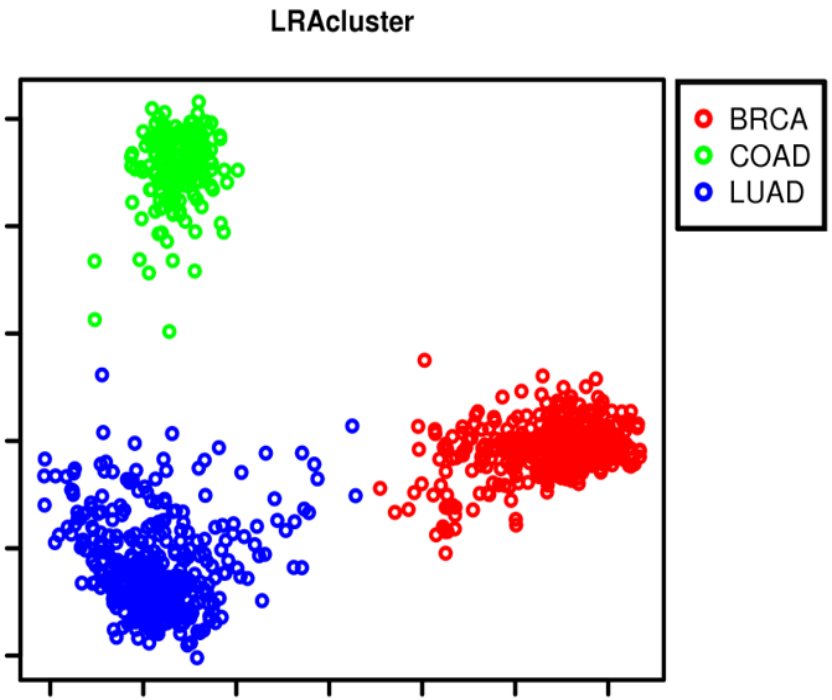


Figure: Clustering results of LRAcluster in simulation datasets.

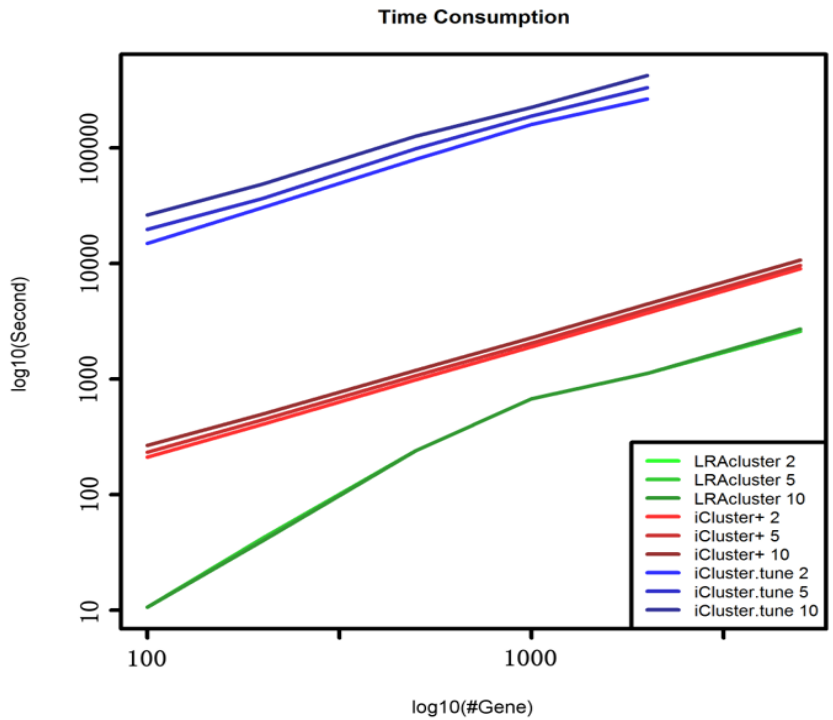


Figure: Time consumption of LRAcluster compared with similar methods. x-axis: number of dimensions.

LRAcluster: summary

- A unified probabilistic framework for analyzing high dimensional multi-omics data.
- A fast tool to analyze multi-omics data and visualize them in a low dimensional space.
- Applied to pan-cancer and single cancer datasets.

Table 1. The results of pan-cancer analysis.

	BRCA	COAD	GBM	HNSC	KIRC	LGG	LUAD	LUSC	PRAD	STAD	THCA	Total
C1	1	0	0	286	0	0	0	6	0	0	0	293
C2	0	0	0	0	0	1	0	0	0	0	411	412
C3	0	0	41	0	0	451	0	0	0	0	0	492
C4	0	0	0	0	0	0	0	0	0	231	0	231
C5	0	0	0	0	0	0	0	0	293	0	0	293
C6	0	190	0	1	0	0	2	0	1	0	0	194
C7	3	17	0	0	1	0	406	7	0	0	3	437
C8	0	0	0	0	240	0	0	0	0	0	0	240
C9	448	0	1	2	1	0	4	1	0	0	0	457
C10	8	1	0	195	0	0	6	60	0	0	0	270
Total	460	208	42	484	242	452	418	74	294	231	414	3319

Figure: Apply LRAcluster on pan-cancer analysis.

Gene mutation and Sorafenib drug response

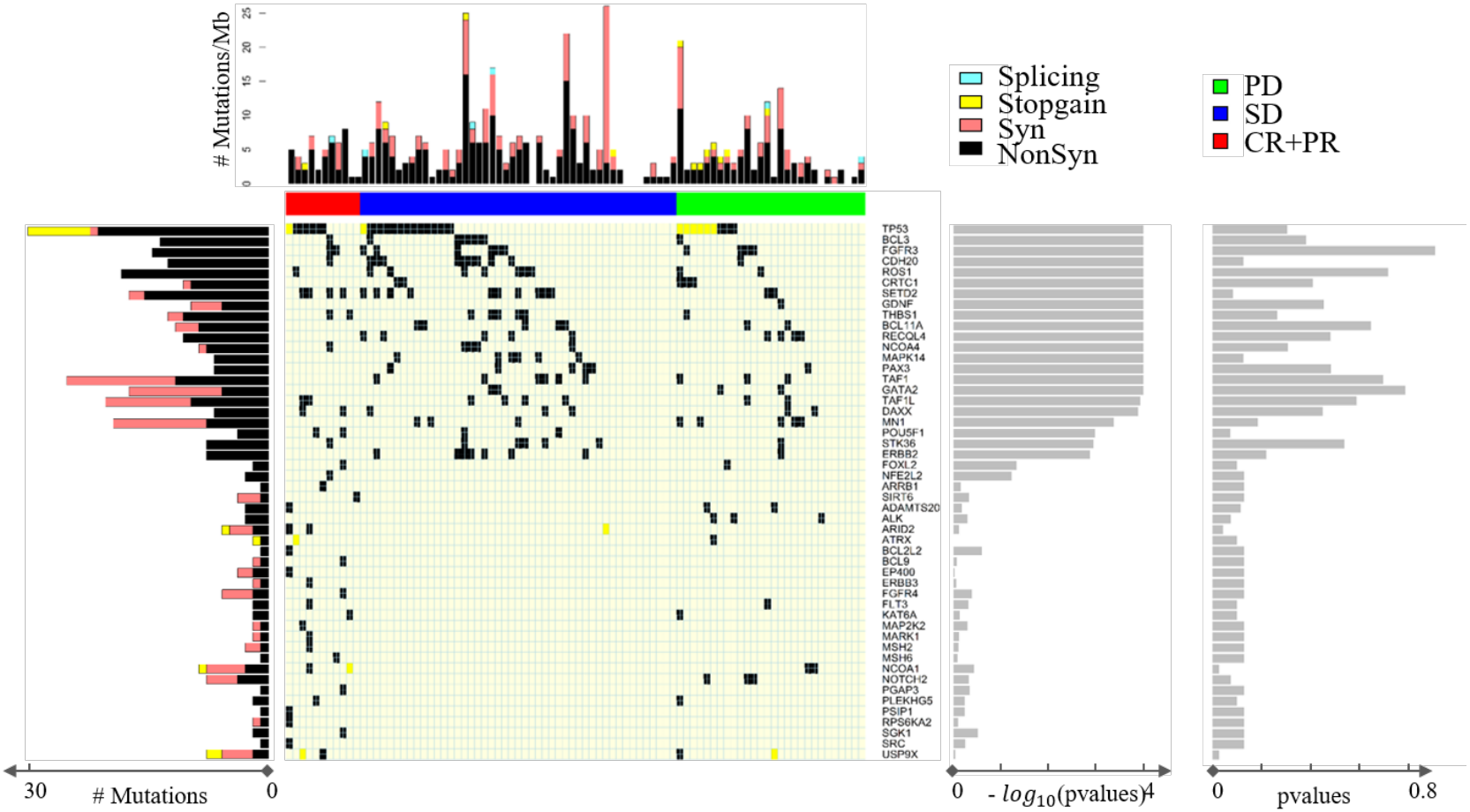
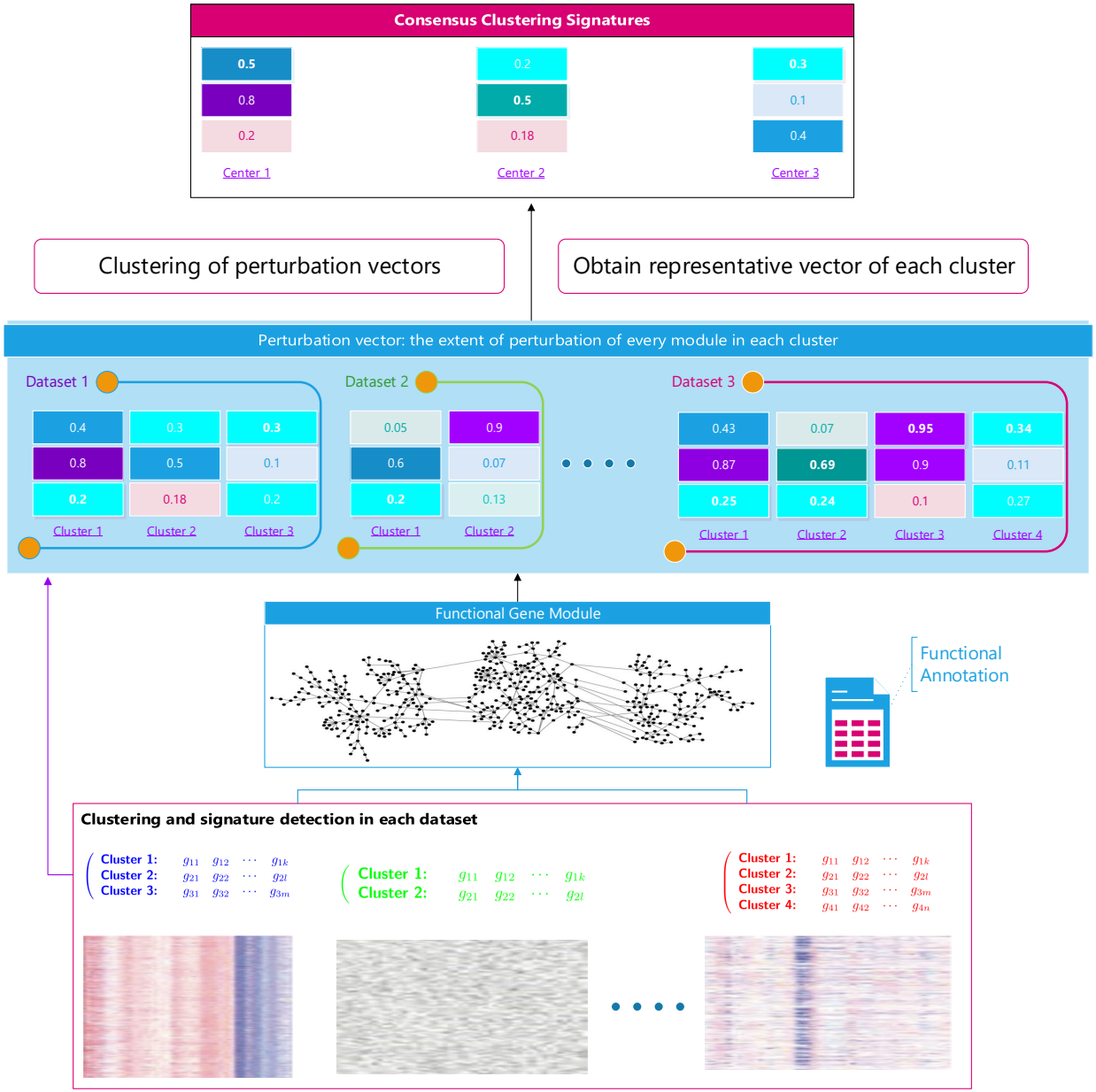


Figure: Overview of sorafenib reponse associated mutations.

Conclusion

- Molecular features-based cancer stratification and classification have profound impacts on understanding cancer biology and cancer therapy.
- To overcome the computational difficulties, such as high-dimension, limited sampels, multi-source information, we've builted some models, including:
  - Cancer subtypes across multiple datasets
  - Cancer subtypes indicated by multi-omics
- It's still arguable about how to validate cancer stratification results.

Supplement


Consensus Clustering Signatures
0.5
0.8
0.2
0.2
0.5
0.18
0.3
0.1
0.4
Center 1
Center 2
Center 3
Clustering of perturbation vectors
Obtain representative vector of each cluster
Perturbation vector: the extent of perturbation of every module in each cluster
Dataset 1
0.4
0.3
0.3
0.8
0.5
0.1
0.2
0.18
0.2
Cluster 1
Cluster 2
Cluster 3
Dataset 2
0.05
0.9
0.6
0.07
0.2
0.13
Cluster 1
Cluster 2
Dataset 3
0.43
0.07
0.95
0.34
0.87
0.69
0.9
0.11
0.25
0.24
0.1
0.27
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Functional Gene Module
Functional Annotation
Clustering and signature detection in each dataset
Cluster 1: g_{11} g_{12} ⋯ g_{1k}
Cluster 2: g_{21} g_{22} ⋯ g_{2l}
Cluster 3: g_{31} g_{32} ⋯ g_{3m}
Cluster 1: g_{11} g_{12} ⋯ g_{1k}
Cluster 2: g_{21} g_{22} ⋯ g_{2l}
Cluster 1: g_{11} g_{12} ⋯ g_{1k}
Cluster 2: g_{21} g_{22} ⋯ g_{2l}
Cluster 3: g_{31} g_{32} ⋯ g_{3m}
Cluster 4: g_{41} g_{42} ⋯ g_{4n}

Clustering of single dataset

Outliers detection

- PCA: use the first two PCs, and compute the distance from the origin.
- KS test: check whether the distribution of one sample's gene expression profile was significantly different from the overall distribution.

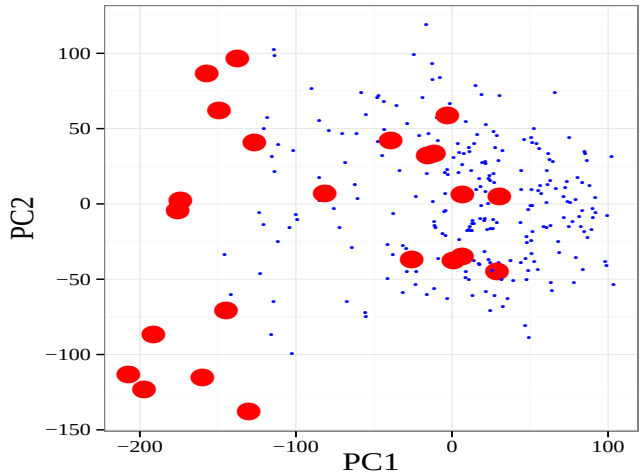


Figure: PCA of HCCDB3 expression profile with outliers labeled.

Review of NMF methods

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H} \text{ s.t. } \boldsymbol{W}, \boldsymbol{H} \geq \boldsymbol{0}$$

- Learn parts of an object.
- Applied widely in biological subtyping since (Brunet et al, *PNAS*, 2004).
- The obtained coefficients matrix could be used for clustering:

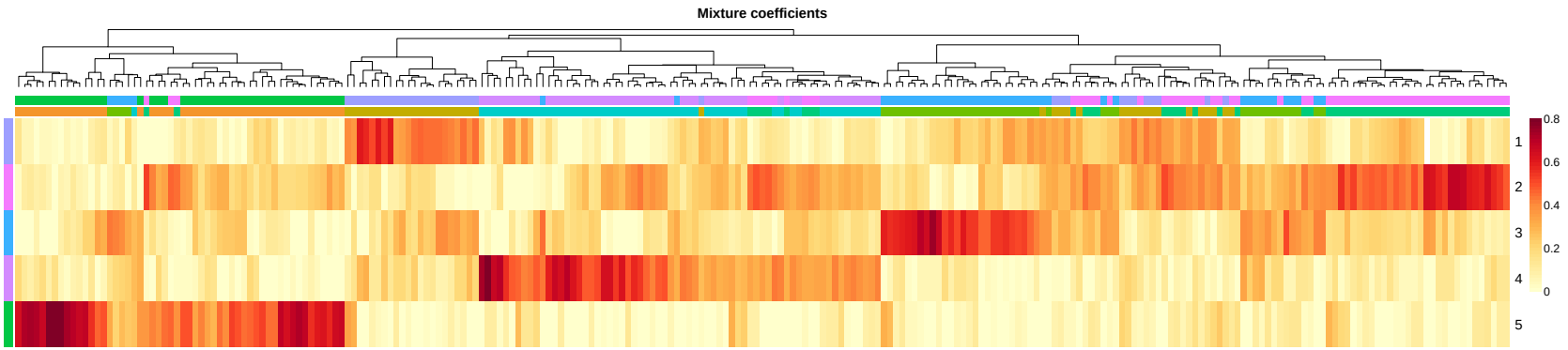


Figure: An example of NMF coef matrix. (HCCDB3)

- **NMF consensus clustering**: bootstrap samples, perform clustering seperately and construct a consensus matrix. (See nect page for an example)
- Use **cophenetic correlation** to determine the factorization rank. (how faithfully a dendrogram preserves the pairwise distances between the original unmodeled data points)

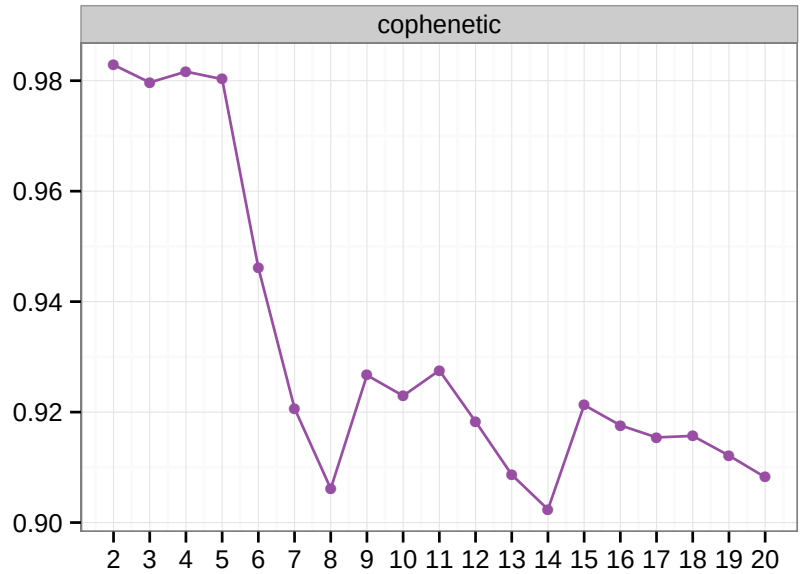


Figure: Cophenetic correlation for different K. (HCCDB3)

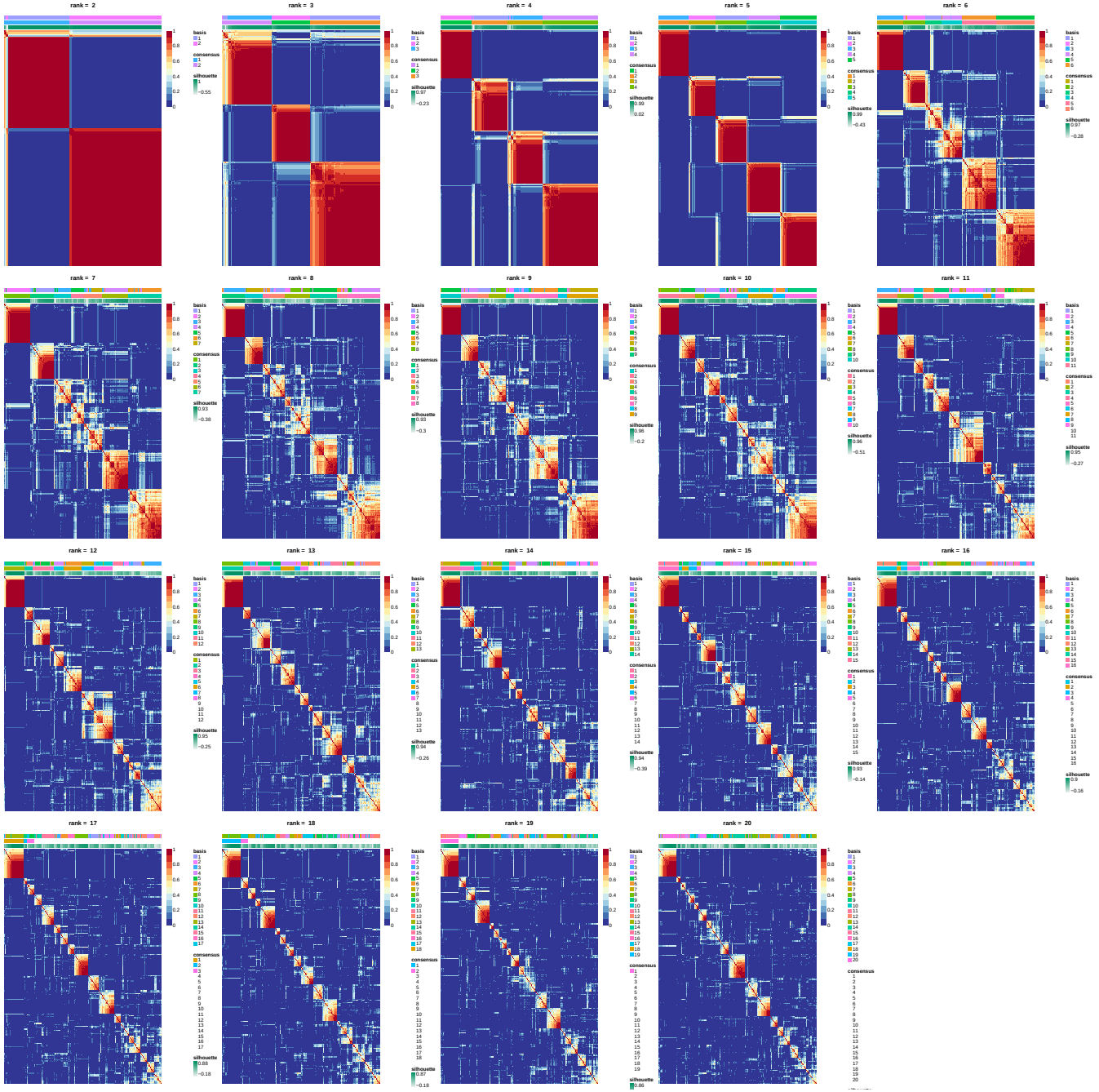


Figure: An example of NMF consensus matrix for different rank K. (HCCDB3)

Discovery of signature modules

1. For each cluster in every single dataset, define its signature gene list by:
   - Differentially expressed (t.test and wilcoxon test, fdr $< 1e-5$)
   - Higher mean expression in this cluster than others.
2. Map literature genes on PPI to extract PPI sub-network.
3. For every gene at the PPI sub-network, count its number of appearance in signature gene lists of step 1 as its initial value.
4. Peform random walk with restart on sub-network of step 2 with initial values defined at step 2.

Definition of perturbation vectors

- Now we have 6 signature modules (6 signature gene lists), and for each cluster in every single dataset, we could define its representative "sample" vector $\boldsymbol{g}$ which is median of all samples in this cluster.
- Compute enrichment score s of $\boldsymbol{g}$ in every signature module.
- Then permute the gene lists, and generate a null distribution for s.
- Compute the $z-$score z for s, denoted by $z_i, i = 1, \cdots, 6$
- The dim 6 vector $\boldsymbol{z}$ is the perturbation vector for every cluster.

Re-classification of single samples

- For every sample, we could also define a perturbation vector z.
- Then we compute the distance from it to all perturbation vectors defined for this dataset.

Survival analysis of HCCDB15

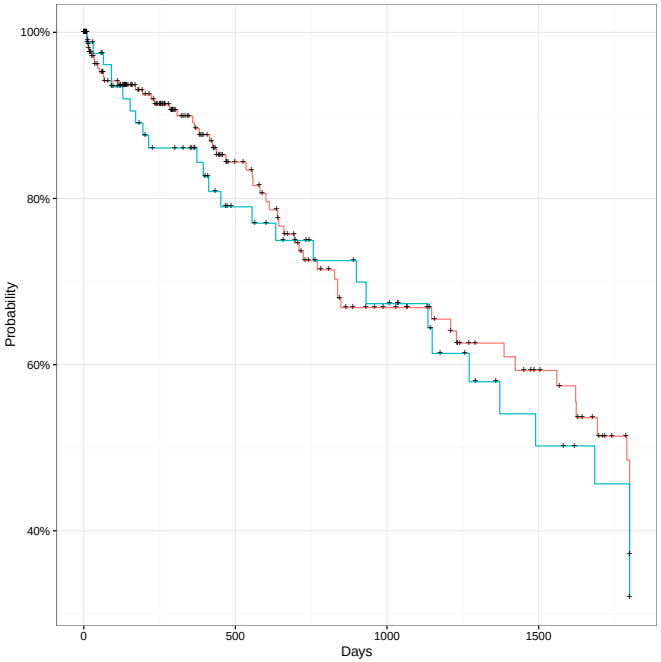


Figure: Survival analysis by single clustering of HCCDB15. cox: 0.56

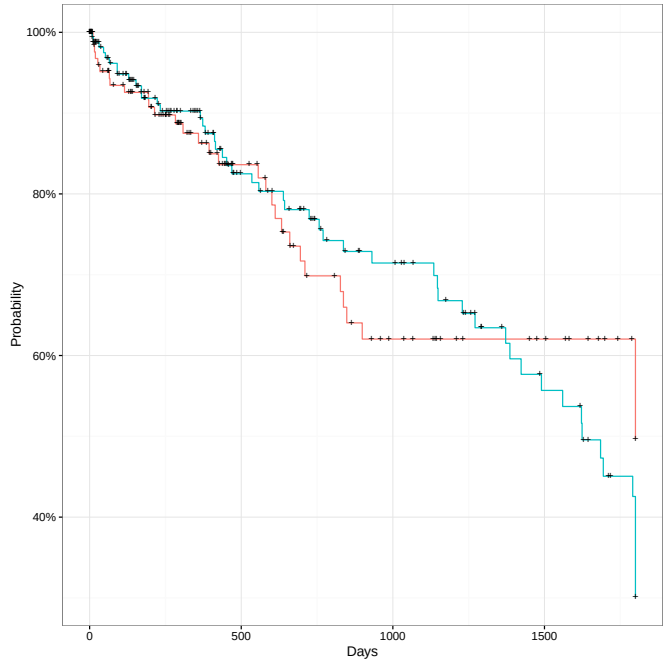


Figure: Survival analysis by consensus clustering of HCCDB15. cox: 0.62

Optimization methods of LRAcluster

基于低秩逼近的方法

- 迭代求解框架

$$(1)\Theta^{(2n+1)} = \Theta^{(2n)} + \delta_n \partial \ln(\Pr(X; \Theta^{(2n)}))$$

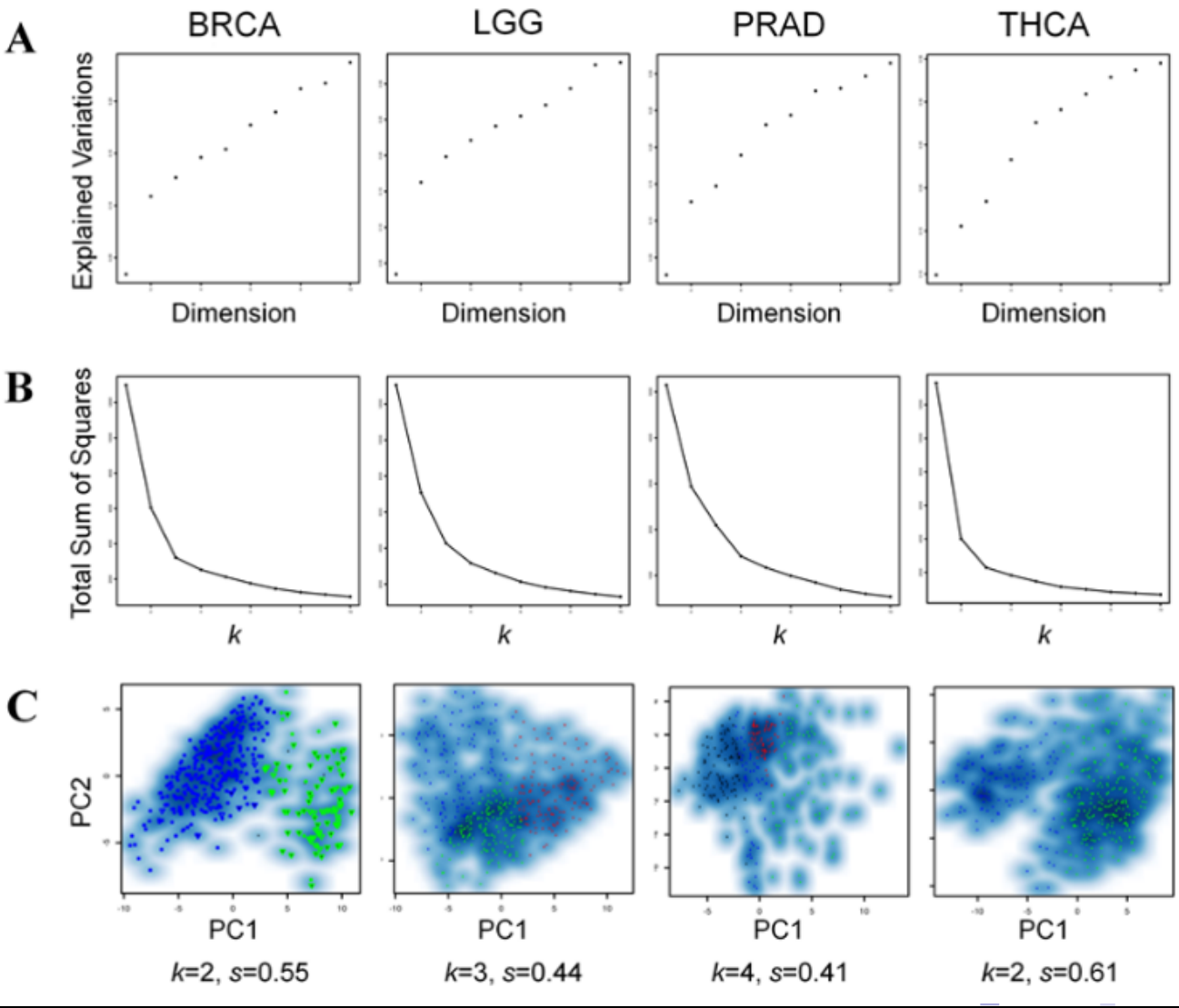
$$(2)\Theta^{(2n+2)} = \mathrm{D}_\mu(\Theta^{(2n+1)})$$

$$\mathrm{D}_\mu(X) = U \begin{bmatrix} \{s_1 - \mu\}^+ & 0 & \cdots & 0 \\ 0 & \{s_2 - \mu\}^+ & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \{s_R - \mu\}^+ \end{bmatrix} V^T$$

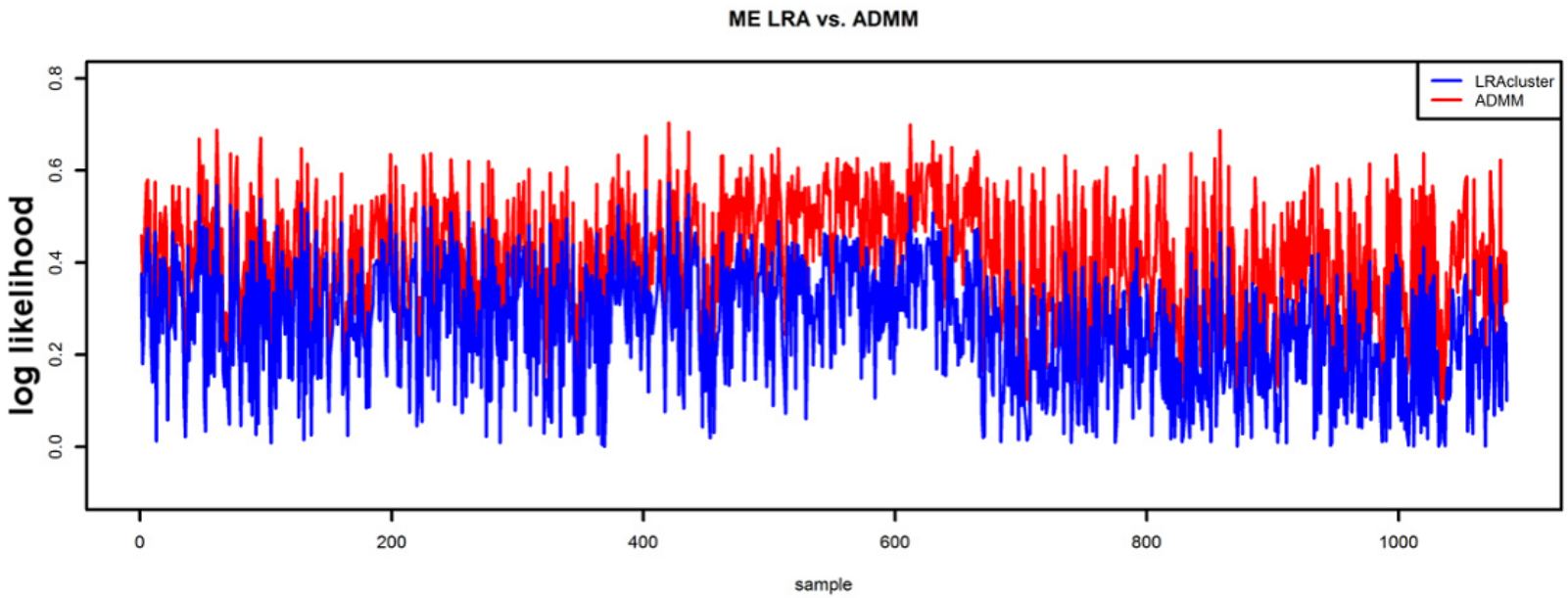
多组学数据整合降维聚类方法研究 31

Figure: Slides from Wu Dingming.

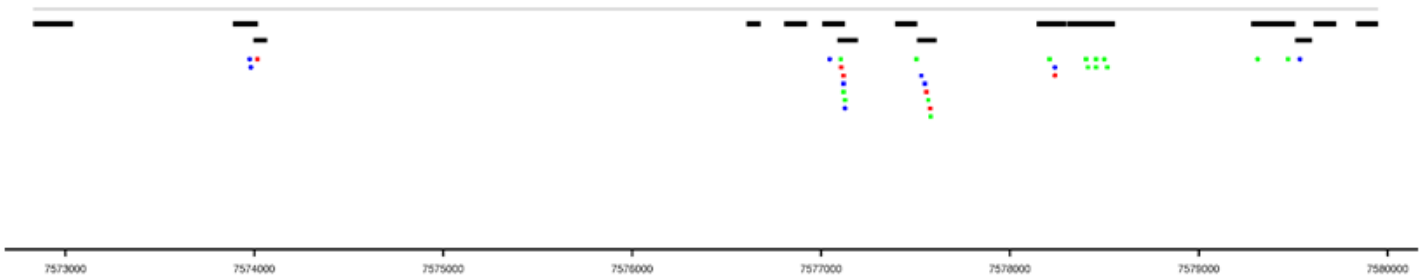
More LRAcluster results



More LRAcluster results



Introduction



// TP53: Probe, mutation sites

- Targeted exome sequencing 440 designed genes of advanced HCC patients (recurrent in 2 years) with Sorafenib treatment.
- 101 patients with PR+CR, SD, PD respectively
- We want to detect significant signatures that can predict the drug response and then validate it.