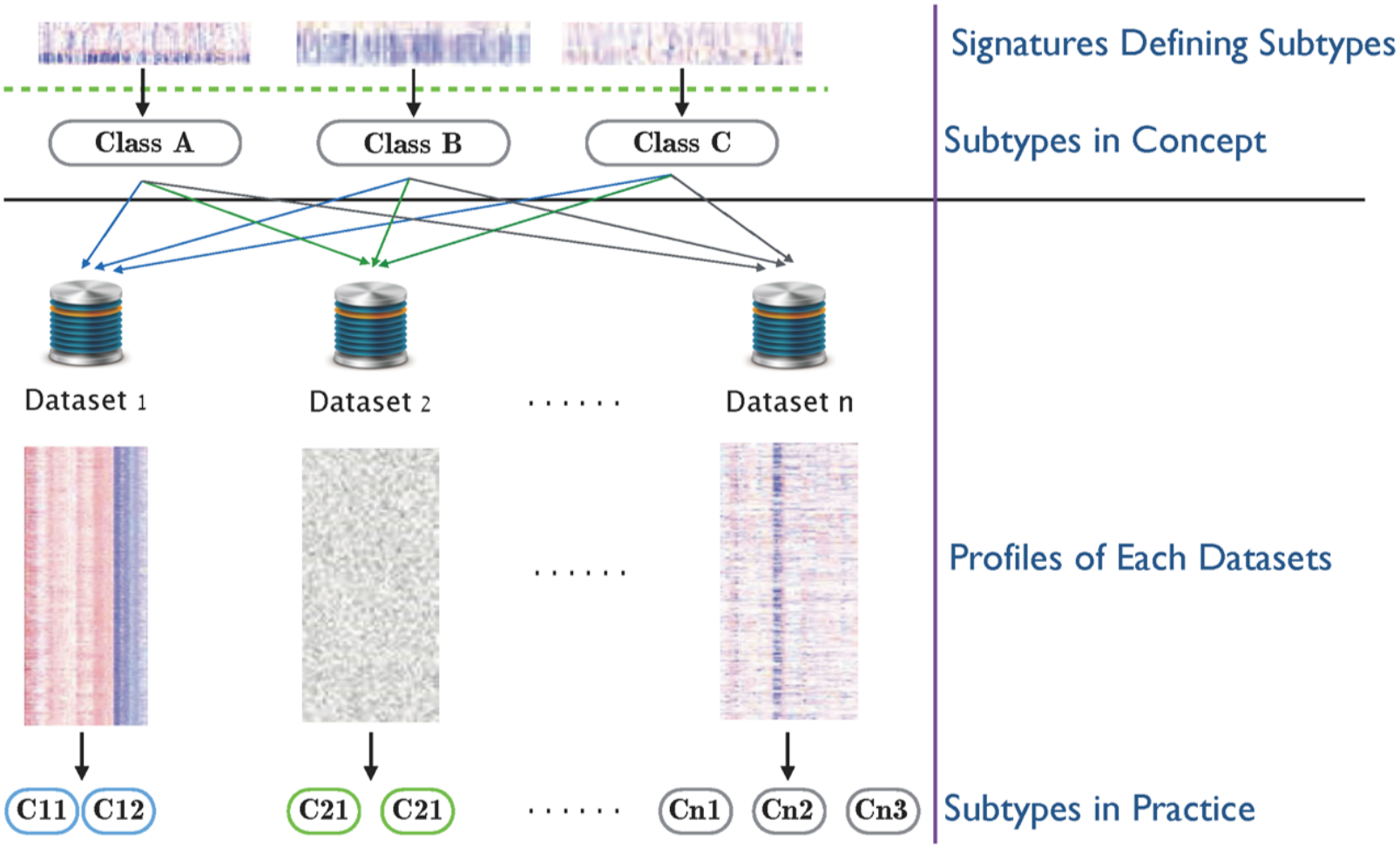
Cancer Subtype discovery by molecular signature network clustering: application to HCC

Dongfang Wang

Tsinghua University

2016-06-14


Signatures Defining Subtypes
Class A
Class B
Class C
Subtypes in Concept
Dataset 1
Dataset 2
Dataset n
Profiles of Each Datasets
C11
C12
C21
C21
Cn1
Cn2
Cn3
Subtypes in Practice

Content

- Review of current HCC-related gene signature sets
- Detailed description of our methods
- Possible improvement

Review of HCC-related gene signature sets

Some limited overlaps exist among different studies.

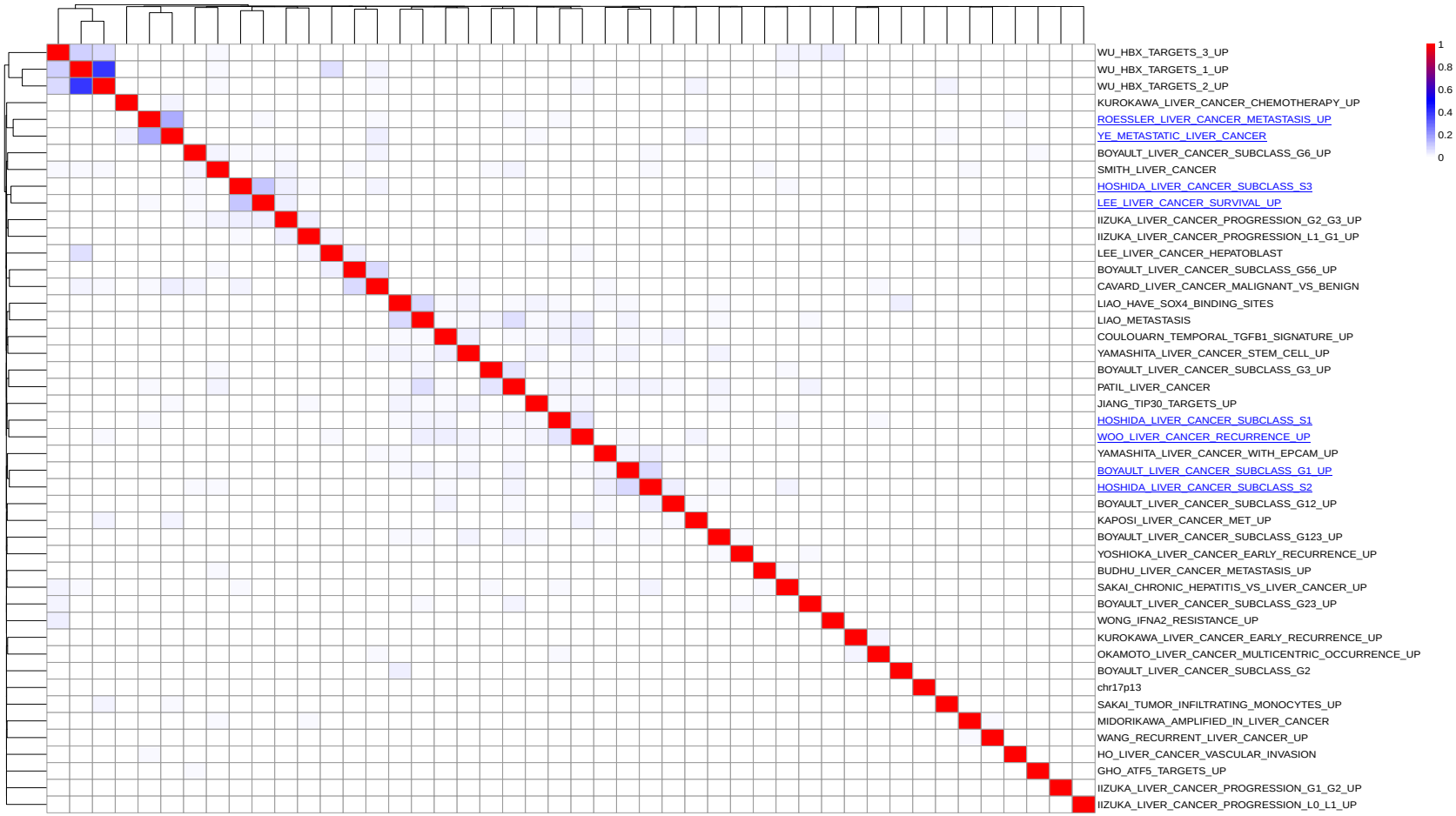


Figure: Similarity between HCC signature sets. Each point represents the Jaccard index of two corresponding sets. (Data from MSigDBv5.1)

Discovery of some gene signature sets

Example 1:

Published OnlineFirst September 1, 2009; DOI: 10.1158/0008-5472.CAN-09-1089

Molecular Biology, Pathobiology, and Genetics

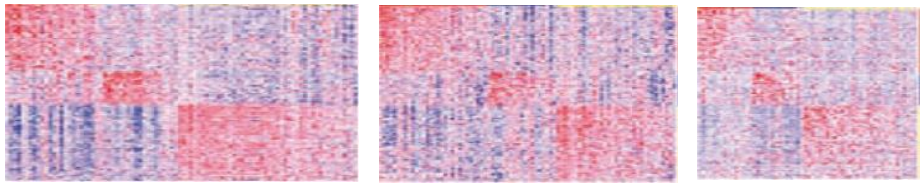
Integrative Transcriptome Analysis Reveals Common Molecular Subclasses of Human Hepatocellular Carcinoma

Yujin Hoshida,[1,2] Sebastian M.B. Nijman,[1,5] Masahiro Kobayashi,[6] Jennifer A. Chan,[1,7] Jean-Philippe Brunet,[1] Derek Y. Chiang,[1] Augusto Villanueva,[8] Philippa Newell,[10] Kenji Ikeda,[6] Masaji Hashimoto,[6] Goro Watanabe,[6] Stacey Gabriel,[1] Scott L. Friedman,[10] Hiromitsu Kumada,[6] Josep M. Llovet,[8,9,10] and Todd R. Golub[1,2,3,4]

[1]Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts; [2]Pediatric Oncology, Dana-Farber Cancer Institute; [3]Children's Hospital Boston, Harvard Medical School; [4]Howard Hughes Medical Institute, Boston, Massachusetts; [5]Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria; [6]Toranomon Hospital, Tokyo, Japan; [7]University of Calgary, Calgary, Alberta, Canada; [8]Barcelona-Clínic Liver Cancer Group, Liver Unit, CIBERehd, Hospital Clínic, IDIBAPS; [9]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain; and [10]Liver Cancer Program, Mount Sinai School of Medicine, New York, New York

HOSHIDA_LIVER_CANCER_SUBCLASS_S1	237	> Genes from 'subtype S1' signature of hepatocellular carcinoma (HCC): aberrant activation of the **WNT signaling pathway**.
HOSHIDA_LIVER_CANCER_SUBCLASS_S2	115	> Genes from 'subtype S2' signature of hepatocellular carcinoma (HCC): **proliferation, MYC and AKT1 [GeneID=4609;207] activation**.
HOSHIDA_LIVER_CANCER_SUBCLASS_S3	266	> Genes from 'subtype S3' signature of hepatocellular carcinoma (HCC): **hepatocyte differentiation**.

Method Overview:



Gene Expression Matrix

Hierarchical Clustering

K-means

NMF

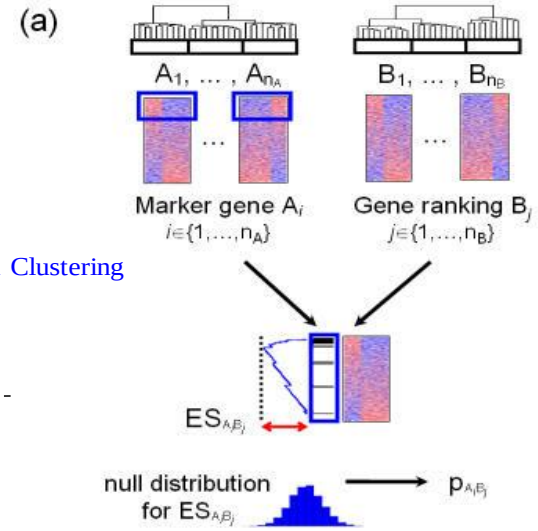
SubClass Mapping

SubClass HC | SubClass kmeans | SubClass NMF

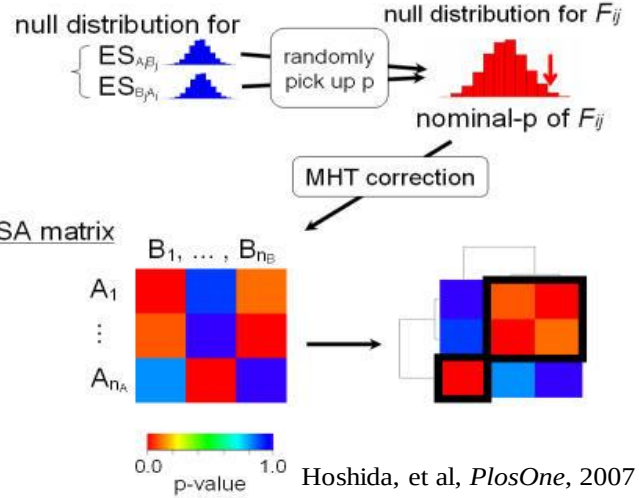
Sig HC | Sig kmeans | Sig NMF

HCC Subclass signature

Prediction



(b) $F_{ij} = -\log(p_{A_iB_j}) - \log(p_{B_jA_i})$



Hoshida, et al, *PlosOne*, 2007

NearestTemplatePrediction: cosine dist from 0-1 sig vector

Results:

1. Clinical significance:

Table 1. Clinical phenotypes associated with HCC subclasses

Variable	S1	S2	S3	*P*
Tumor size (cm)*	3.0 [2.0,4.5]	4.5 [2.5,7.0]	2.5 [1.8,4.3]	0.003
Tumor differentiation*				
Well	8 (16%)	4 (10%)	37 (44%)	
Moderate	27 (53%)	23 (59%)	45 (53%)	<0.001
Poor	16 (31%)	12 (31%)	3 (4%)	
AFP (ng/mL) †	50 [14,332]	171 [27,1,251]	13 [5,43]	<0.001
Hepatitis B virus infection ‡	39 (38%)	27 (36%)	39 (25%)	0.05
Hepatitis C virus infection ‡	55 (53%)	44 (58%)	109 (69%)	0.03

NOTE: Median [25%,75%]. Wilcoxon rank sum test for continuous data, and Fisher's exact test for categorical data.
*HCC-F, HCC-H, and HCC-I: S1, n = 55; S2, n = 46; S3, n = 96.
†HCC-H and HCC-I: S1, n = 48; S2, n = 39; S3, n = 83.
‡HCC-B, HCC-C, HCC-F, HCC-H, and HCC-I: S1, n = 103; S2, n = 76; S3, n = 158.

- S1: early recurrence ($p = 0.03$,See Figure@3), vascular invasion

- S1 and S2: known poor survival signatures; S3: known good survival signatures (See Figure@3)

2. Statistical significance of prediction:

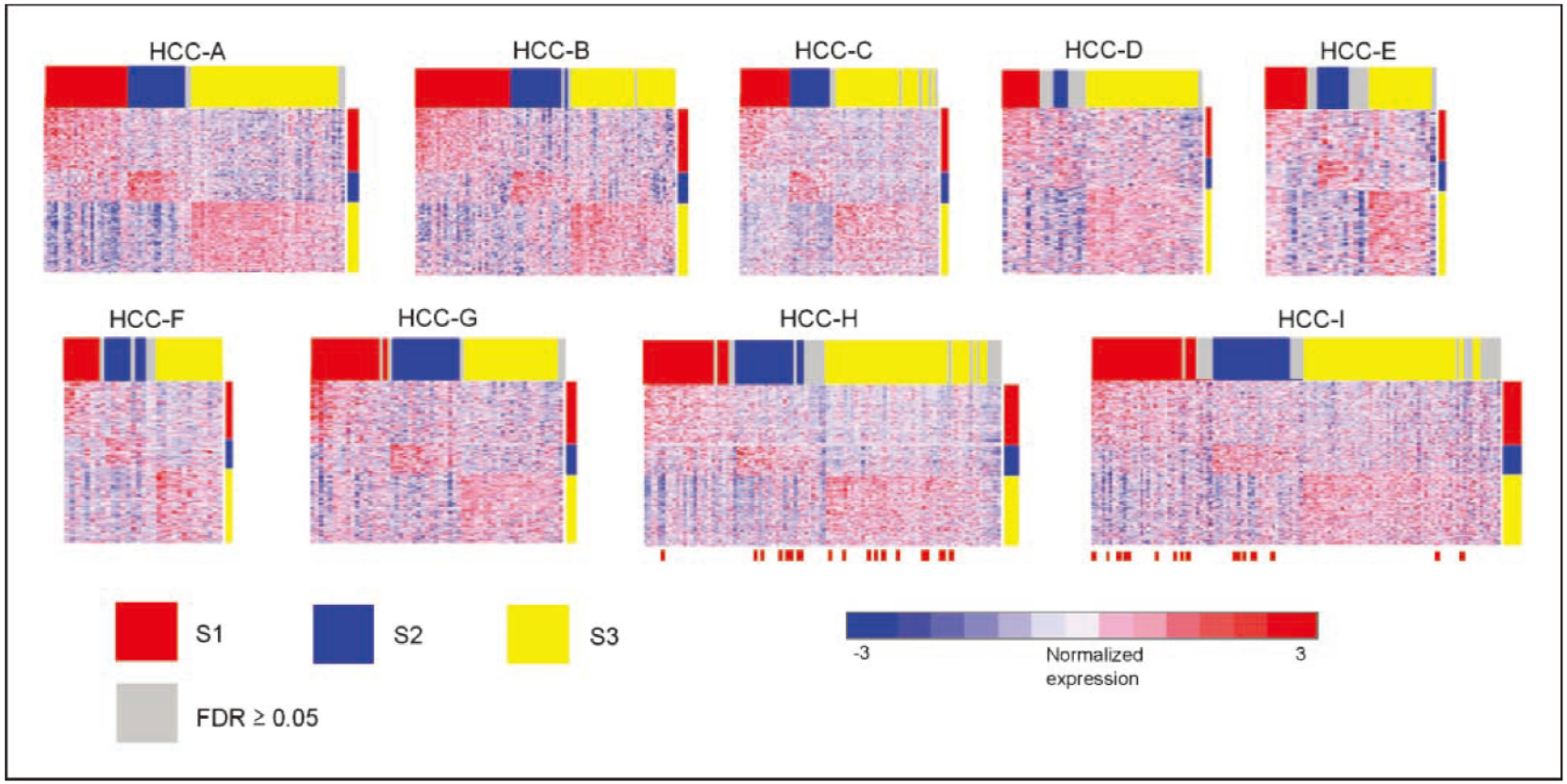


Figure 1. HCC subclasses predicted in nine independent data sets. Predicted subclasses are shown in red (S1), blue (S2), and yellow (S3) with expression pattern of the HCC subclass signature. The proportion of the cases with confident prediction (FDR < 0.05) in HCC-A, HCC-B, HCC-C, HCC-D, HCC-E, HCC-F, HCC-G, HCC-H, and HCC-I were 96%, 96%, 90%, 81%, 79%, 87%, 94%, 83%, and 83%, respectively. Red bars attached to HCC-H and HCC-I indicate positive *β-catenin* mutations and nuclear staining of p53, respectively.

3. Molecular pathway: see the Table@4; GSEA analysis.

Check S1,S2,S3 in TCGA:

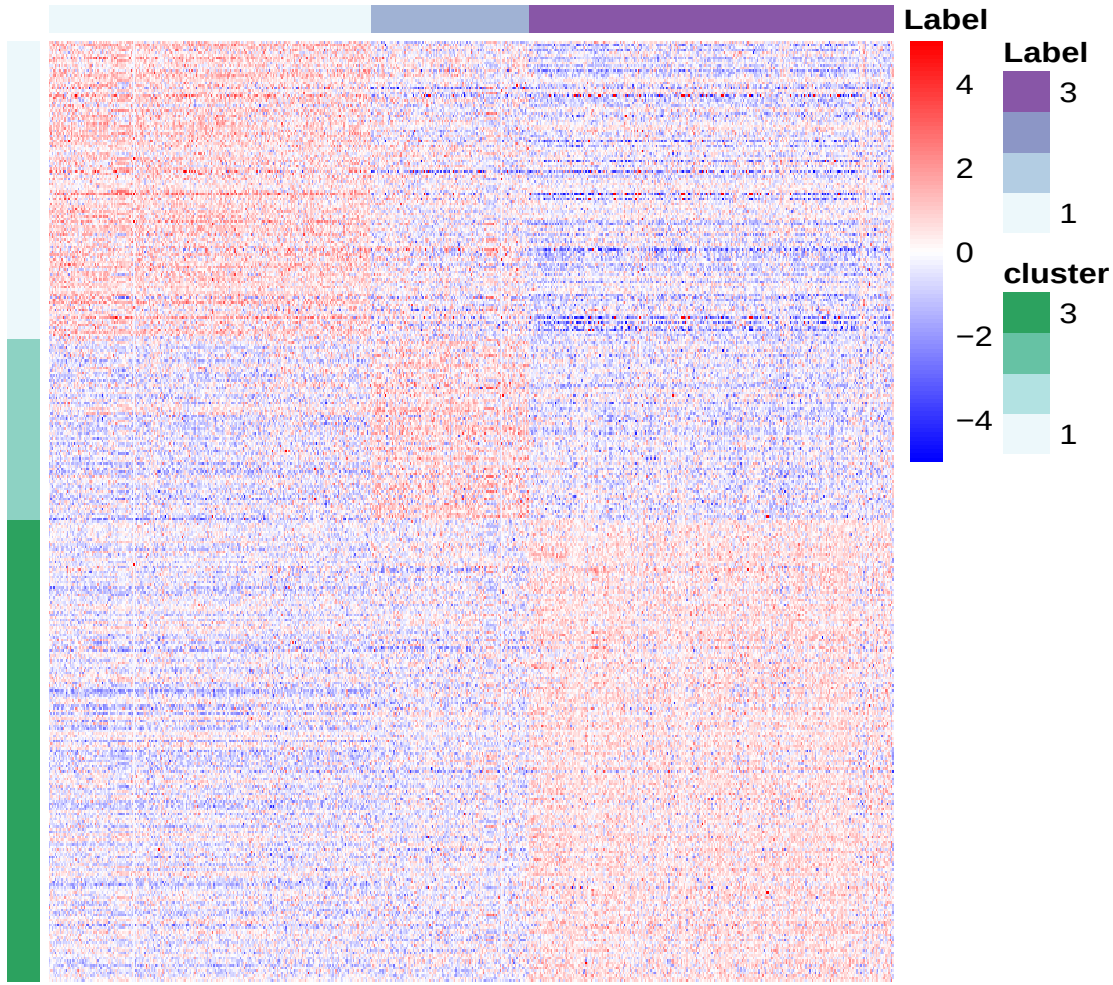


Figure: Normalized gene expression matrix of HCC samples in TCGA. Row: genes in each signature sets; Col: patients classified by NTC

S1,S2,S3 association with clinical variables:

(1). Survival analysis

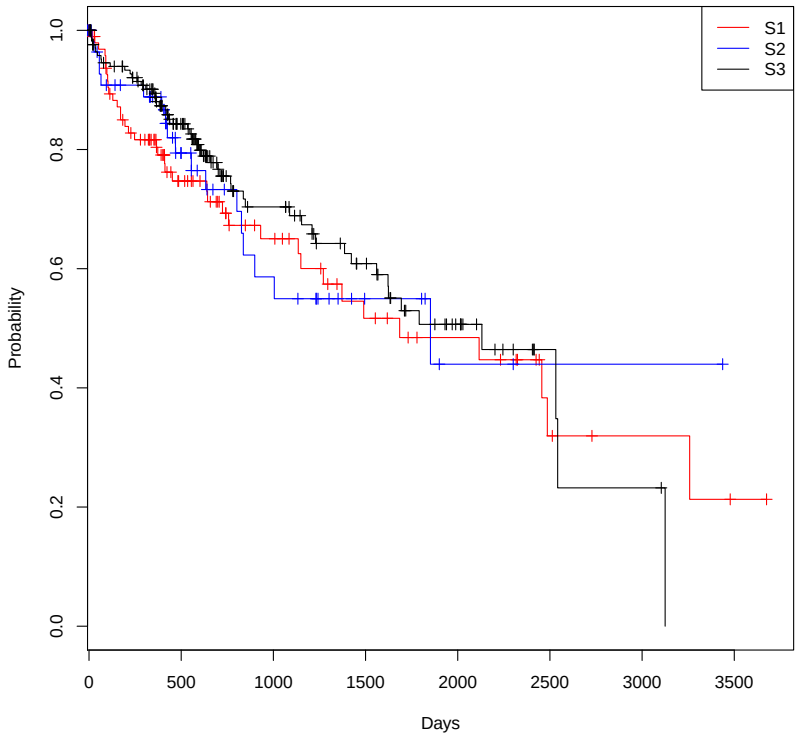


Figure: KM plot of samples in TCGA based on S1,S2,S3 . Cox: 0.376; KM: 0.673

(2). Vascular invasion

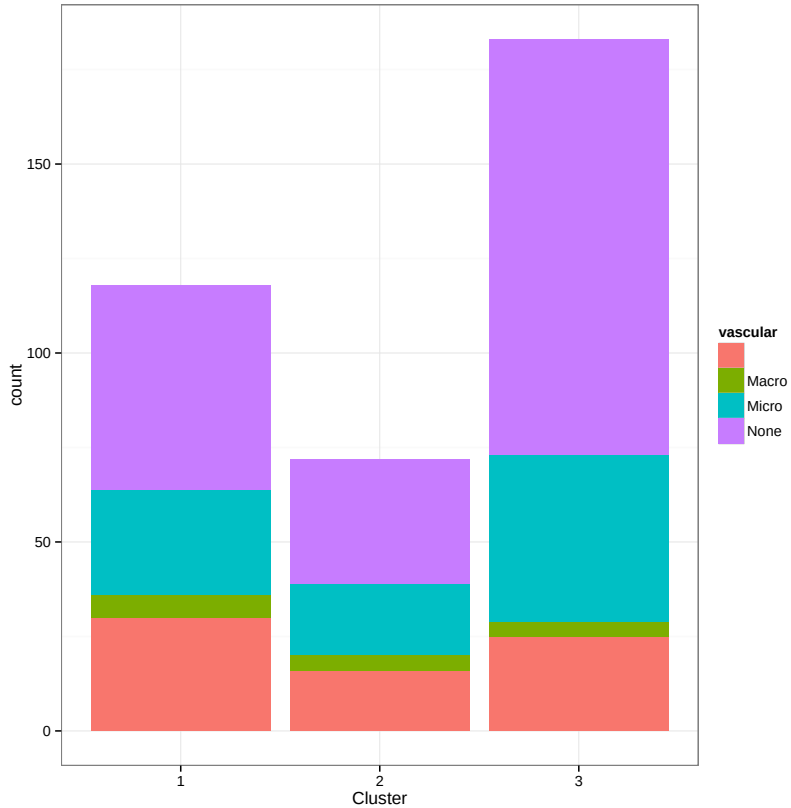


Figure: Association between vascular invasion and S1,S2,S3 clusters in TCGA.

Example 2:

Research Article

EpCAM and α-Fetoprotein Expression Defines Novel Prognostic Subtypes of Hepatocellular Carcinoma

Taro Yamashita,[1] Marshonna Forgues,[1] Wei Wang,[1] Jin Woo Kim,[1] Qinghai Ye,[4] Huliang Jia,[4] Anuradha Budhu,[1] Krista A. Zanetti,[1,3] Yidong Chen,[2] Lun-Xiu Qin,[4] Zhao-You Tang,[4] and Xin Wei Wang[1]

[1]Liver Carcinogenesis Section, Laboratory of Human Carcinogenesis, [2]Genetics Branch, Center for Cancer Research, and [3]Cancer Prevention Fellowship Program, Division of Cancer Prevention, National Cancer Institute, Bethesda, Maryland; and [4]Liver Cancer Institute and Zhongshan Hospital, Fudan University, Shanghai, China

YAMASHITA_LIVER_CANCER_WITH_EPCAM_DN	16	Down-regulated genes distinguishing hepatocellular carcinoma (HCC) samples positive for EPCAM [GeneID=4072] from the negative ones.
YAMASHITA_LIVER_CANCER_WITH_EPCAM_UP	53	Up-regulated genes distinguishing hepatocellular carcinoma (HCC) samples positive for EPCAM [GeneID=4072] from the negative ones.

Check EpCAM and AFP in TCGA:

(1). EpCAM$^+$ subtypes: ≥ 2 fold increase in the level of EPCAM compared with (paired) non-tumor samples (50 pairs in TCGA, of which 12 were identified EPCAM$^+$).

(2). Representative genes: differentially expressed (wilcox's test, $p < 0.005$) and ≥ 2 fold difference. (We found no overlap between our results with original papers.)

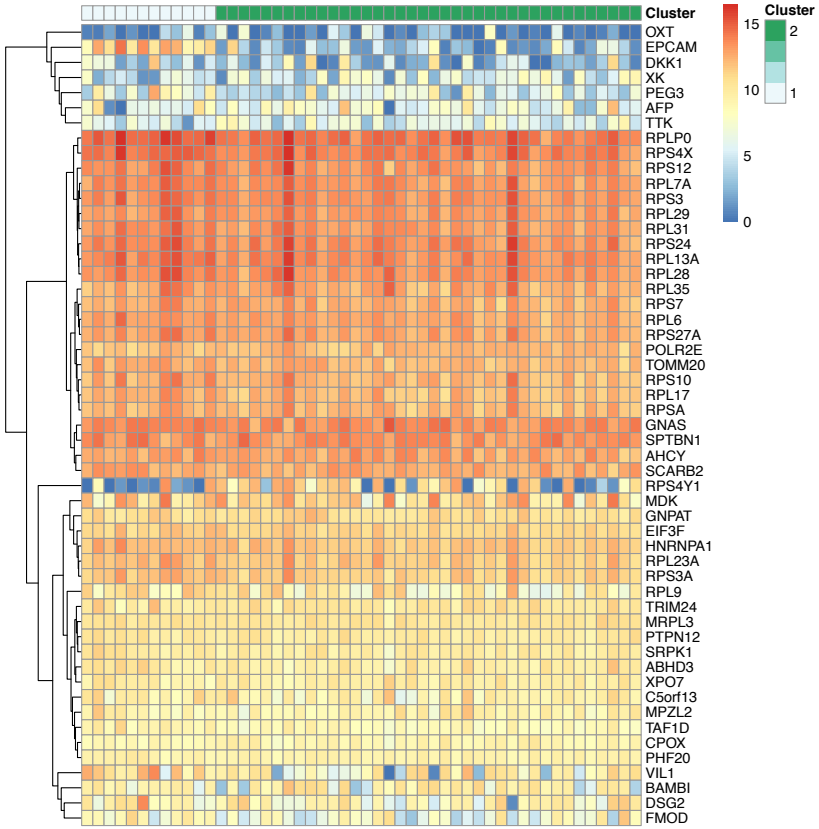


Figure: Heatmap of TCGA data. Rows: genes from original Paper.

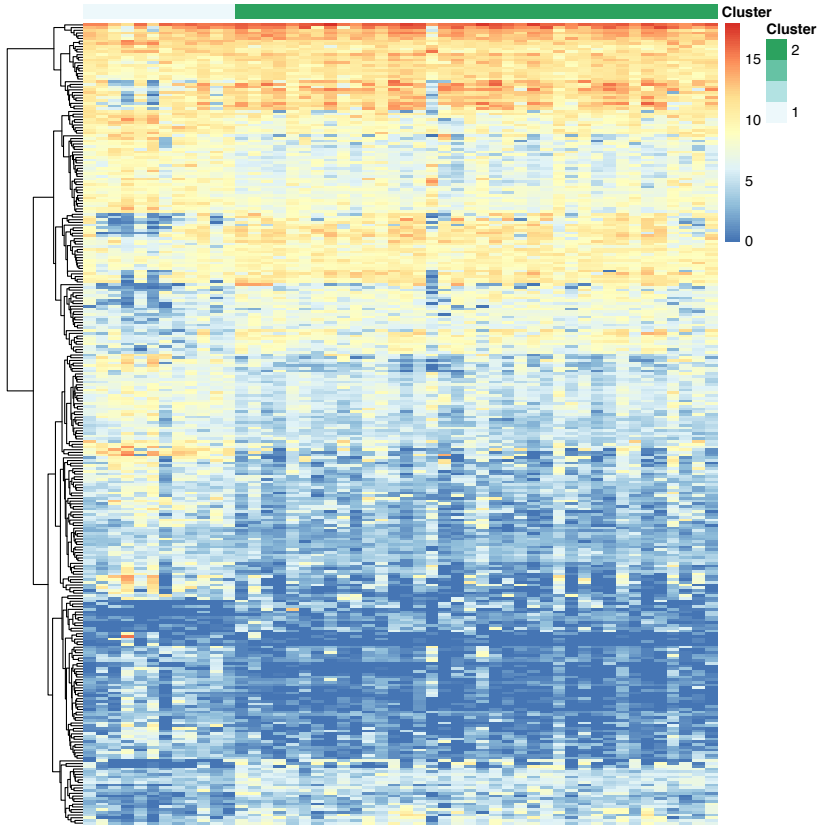


Figure: Heatmap of TCGA data. Rows: genes from ourselves analysis. (296 genes)

Survival analysis:

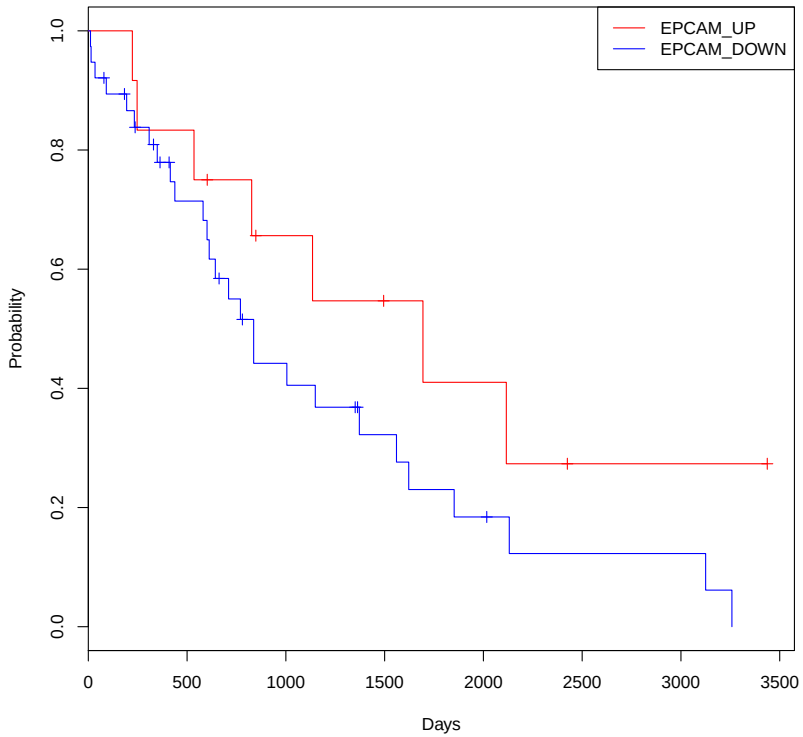


Figure: KM plot of TCGA samples based on EPCAM. Cox:0.115; KM:0.131

Another strategy: Cause the limited pairs of samples, we used median level of non-tumor samples as a reference, and tumors with ≥ 2 fold EPCAM level than it were regarded as EPCAM$^+$ samples.

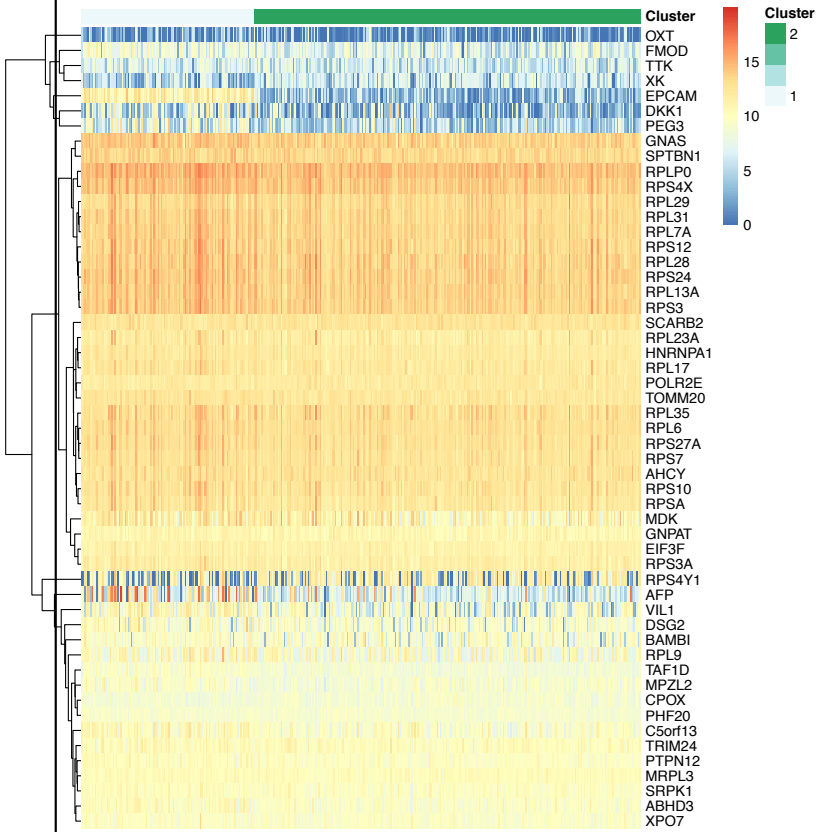


Figure: Heatmap of TCGA data. Rows: genes from original Paper.

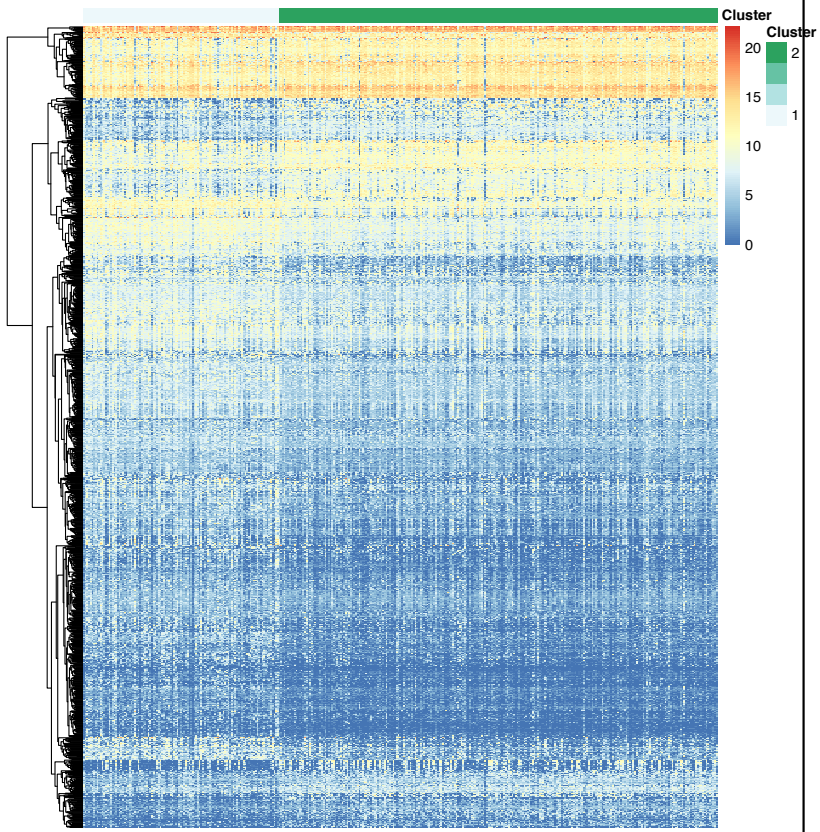


Figure: Heatmap of TCGA data. Rows: genes from ourselves analysis.(1431 genes)

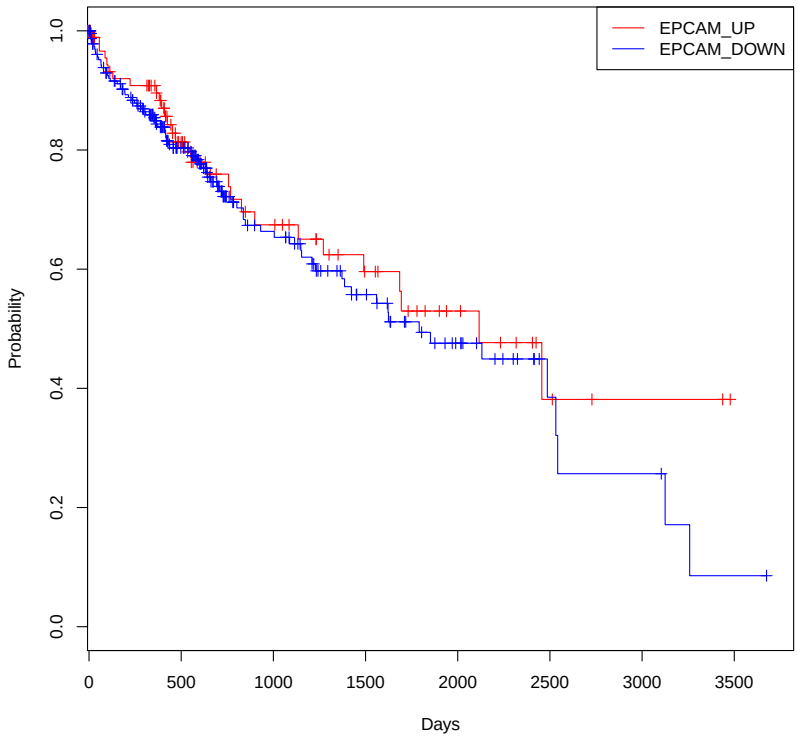


Figure: KM plot of TCGA samples based on EPCAM. Cox:0.413; KM:0.419

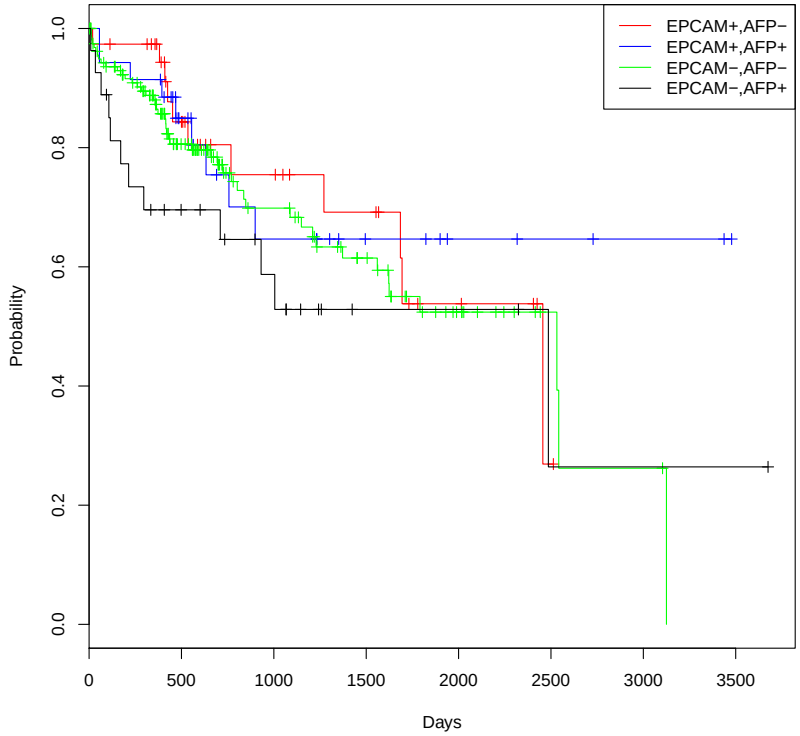
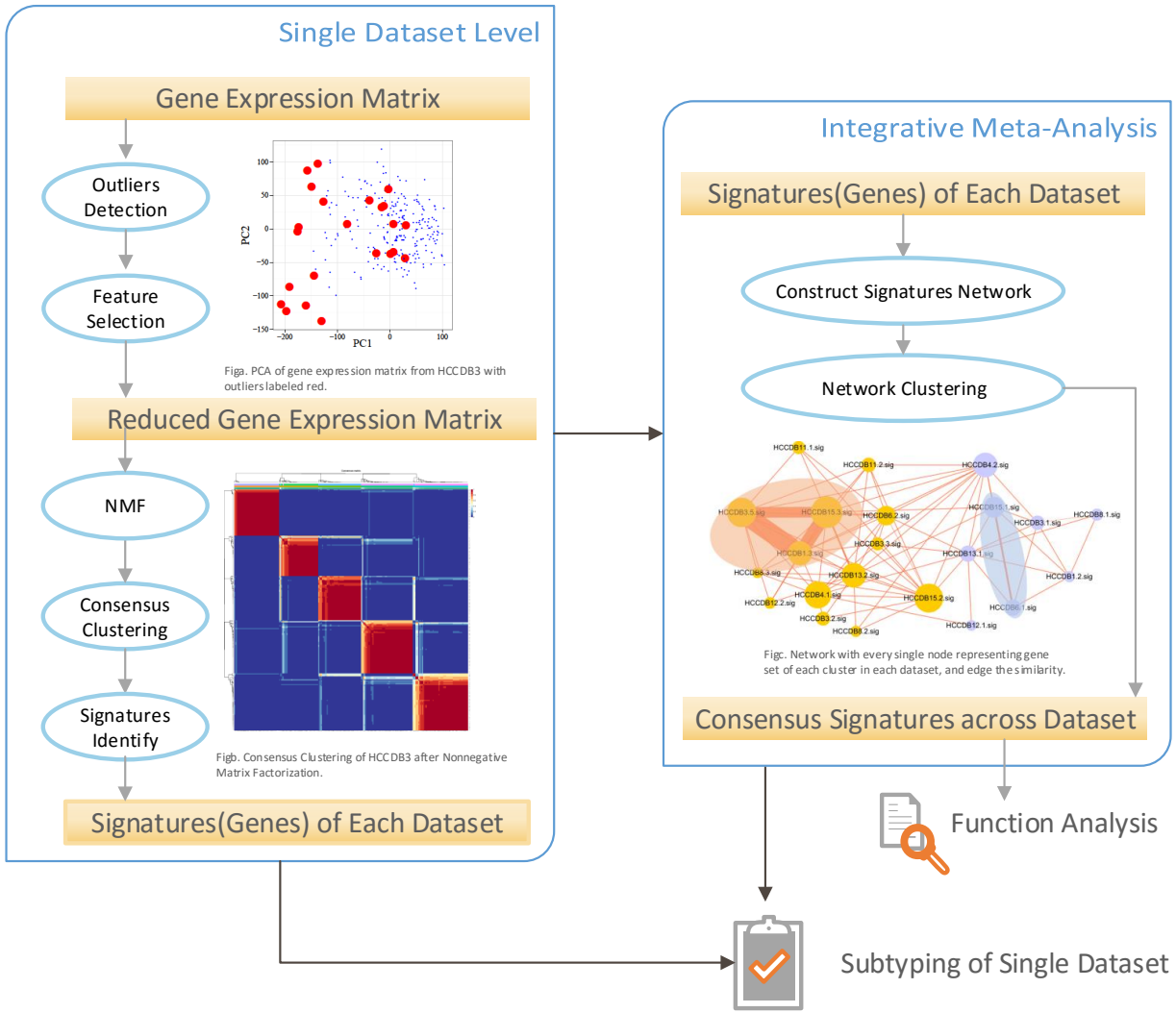


Figure: KM plot of TCGA samples based on EPCAM and AFP (300). Cox:0.167; KM:0.367

Signature network-based subtyping



1. Single dataset analysis

(1) 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:

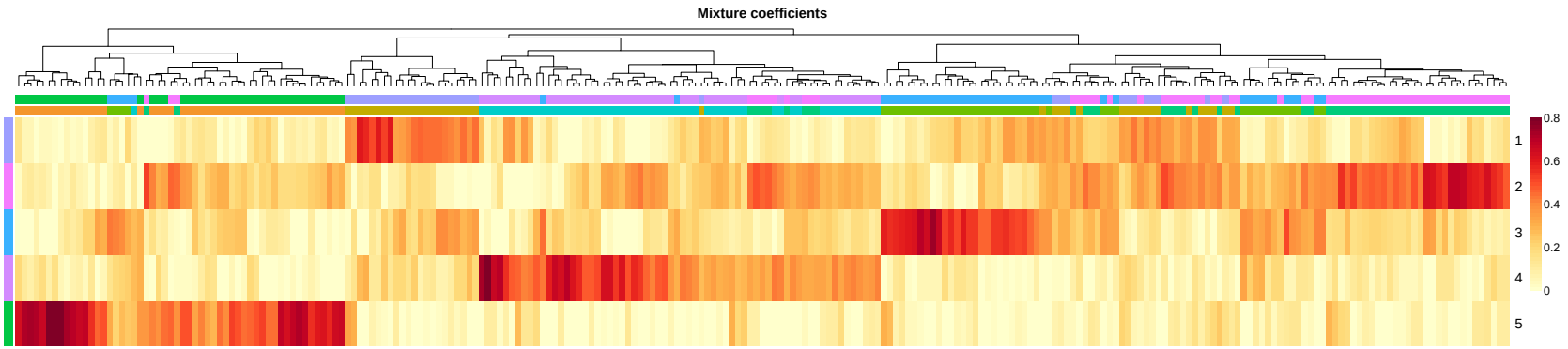


Figure: An example of NMF coef matrix. (HCCDB3)

- **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)

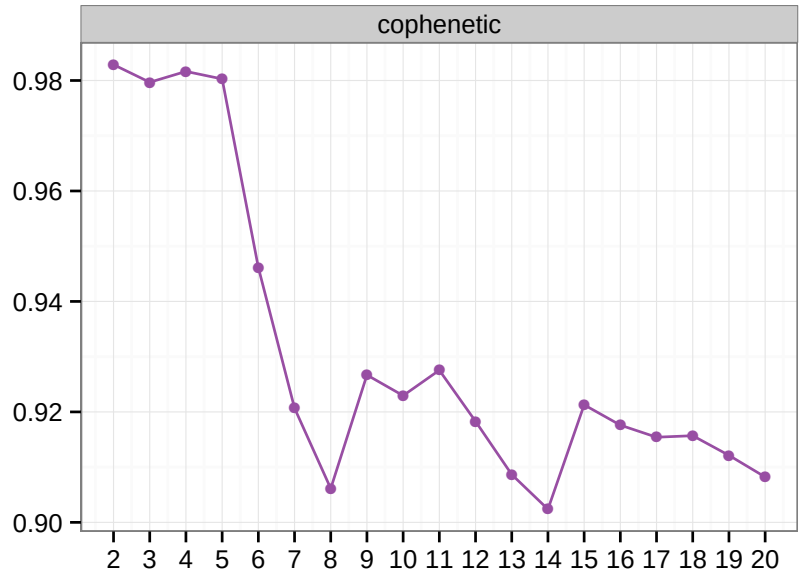


Figure: Cophenetic correlation for different K. (HCCDB3)

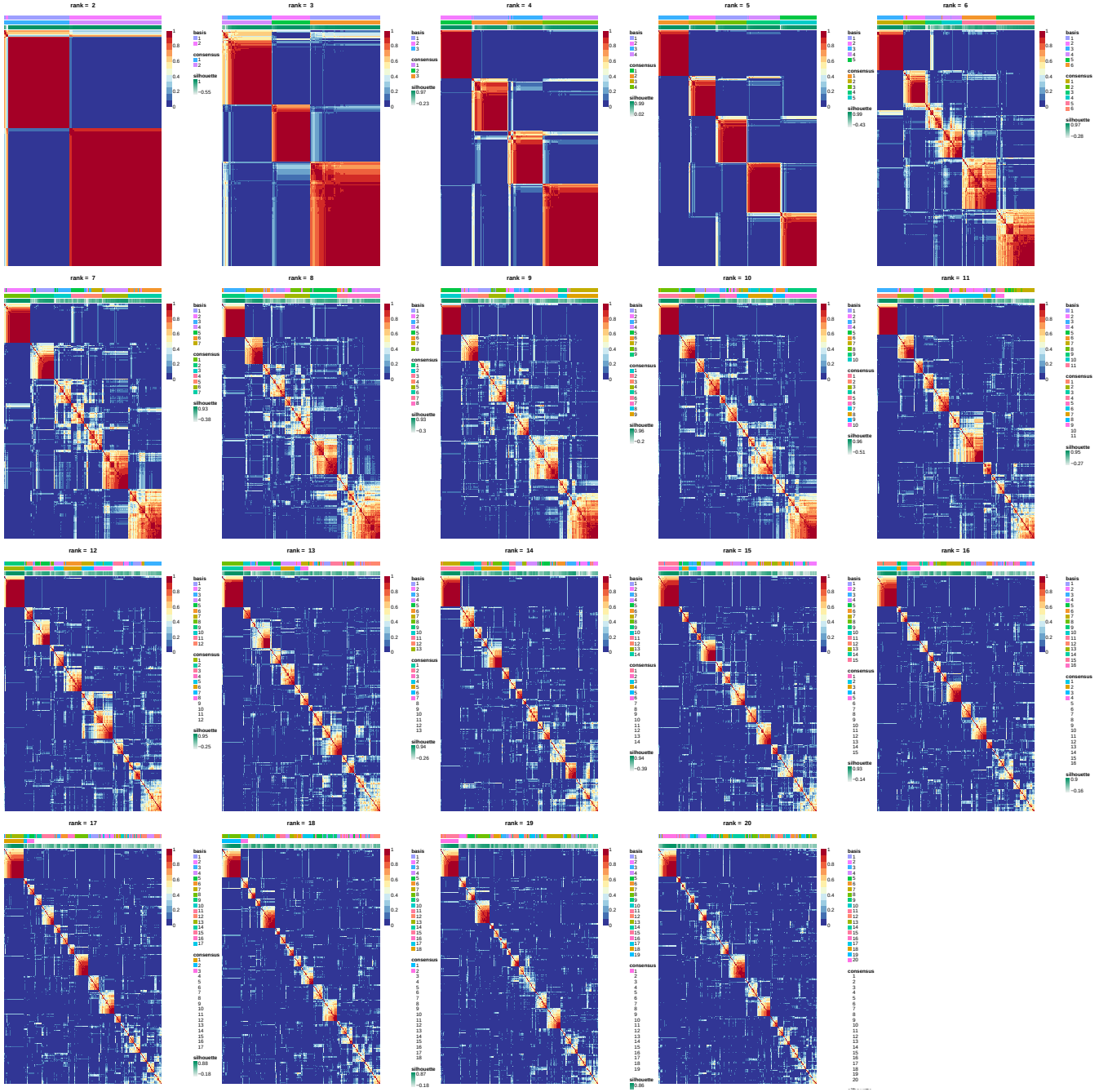


Figure: An example of NMF consensus matrix for different rank K. (HCCDB3)

(2)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.

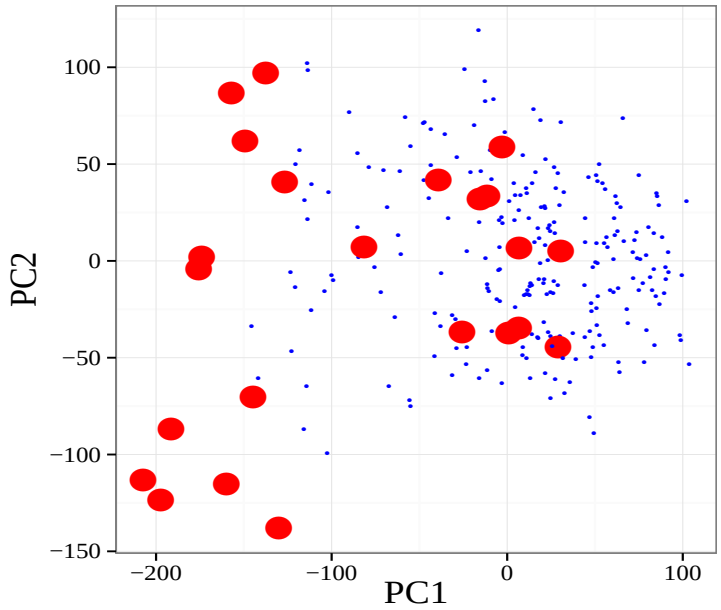


Figure: PCA of HCCDB3 expression profile with outliers labeled.

(3)Gene selection

Gene selection is an important procedure if we want to use small number of genes to detect biologically meaningful subtypes.

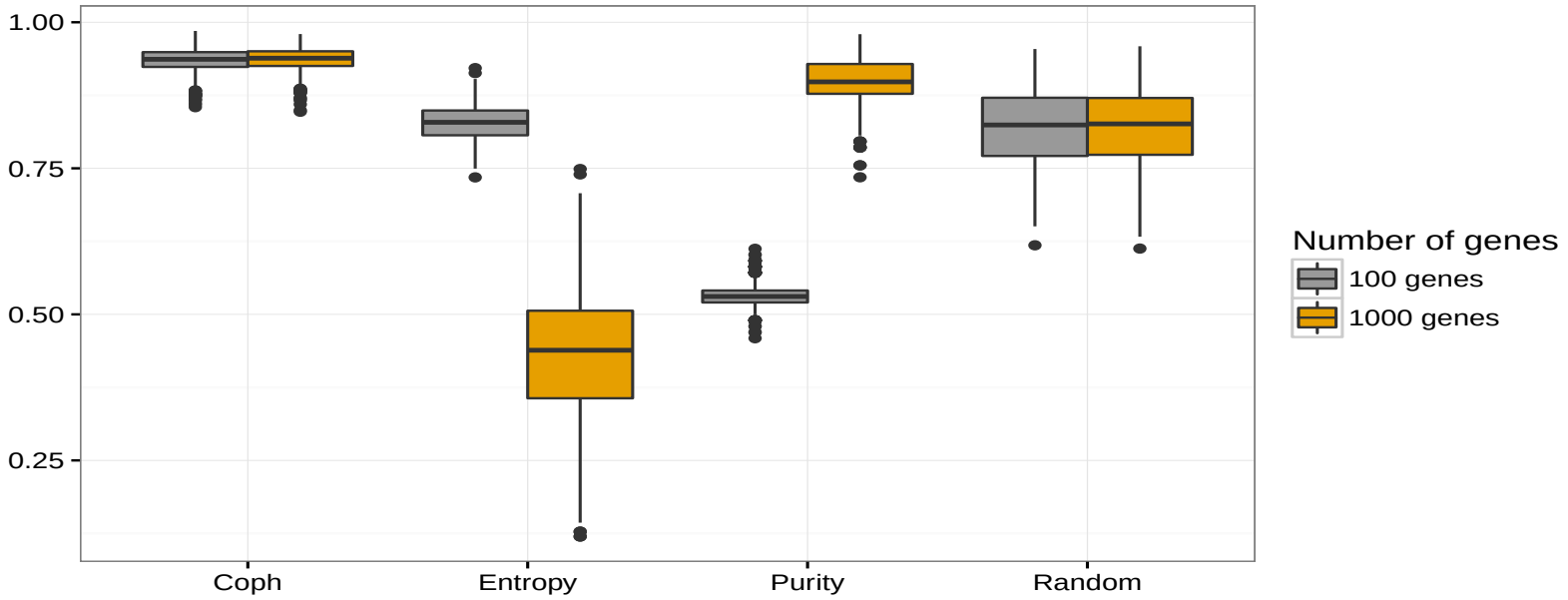


Figure: Randomly select 100 and 1,000 genes 1,000 times and perform NMF clustering. Coph: the cophnetic correlation distribution; Entropy and Purity: compare the clustering results with that of 1,000 most variated genes. Larger purity and smaller entropy mean more similar. Random: Shuffle the datasets, perform NMF clustering and compute the cophenetic correlation.

(4)Clustering and signature detection

- Use genes in KEGG hsa05200: pathway in cancer.
- Perform NMF consensus clustering
- Select representative genes for every cluster of each dataset.
  - Differentially expressed (t.test and wilcoxon test, fdr $< 1e-5$)
  - Higher mean expression in this cluster than others.

Dataset	Cluster	Number of genes
HCCDB1	1	3
	2	2
	3	20
HCCDB3	1	0
	2	6
	3	48
	4	11
	5	9
HCCDB4	1	37
	2	41
HCCDB6	1	21
	2	14
HCCDB8	1	5
	2	2
	3	2

Dataset	Cluster	Number of genes
HCCDB11	1	8
	2	4
HCCDB12	1	0
	2	4
	3	0
HCCDB13	1	12
	2	33
HCCDB15	1	37
	2	19
	3	31
	4	48
HCCDB9	1	35
	2	55

Figure: Number of clusters and each cluster's representative genes in evry dataset. Only those with over 10 genes were used for further analysis.

2. Construction of signature network

Network construction:

- Node: each gene set in the previous page.
- Edge: negative log of statistical significance of the overlap between paired gene sets. (hypergeometric test)

Netwrok clustering:

- A random walk-based network clustering method was used.

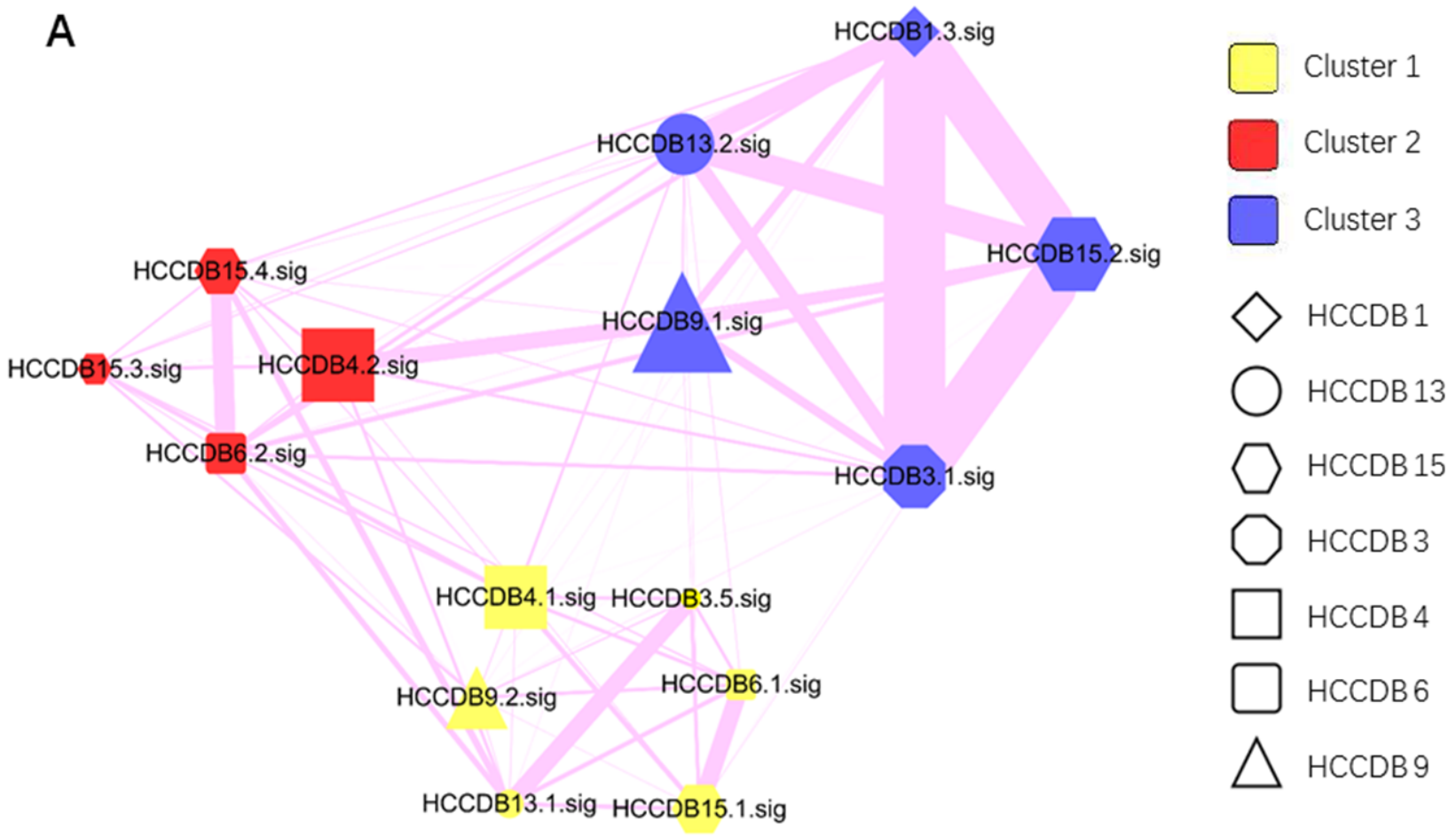
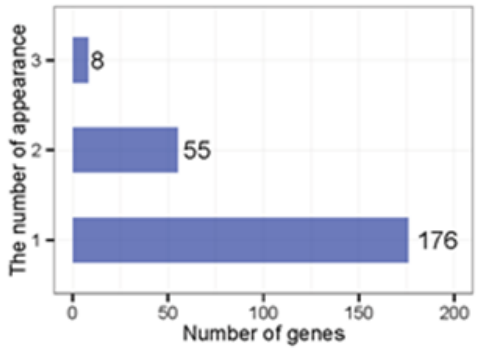


Figure: Signature Network clustering. Three signature cluster (Cluster1,Cluster2,Cluster3) was identified, which means there exists similar signature sets across different datasets.



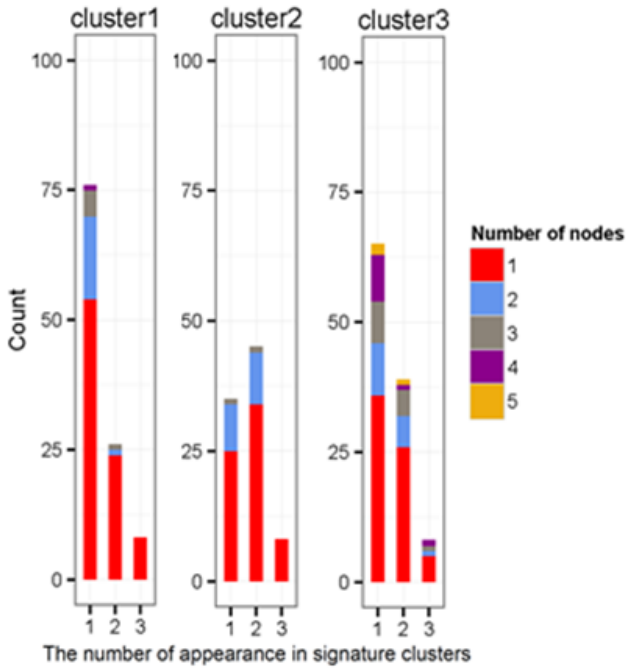


Figure: Count the appearance of genes in each signature cluster. (**Up**:) only 8 genes appeared in all 3 signature clusters, while 55 genes appeared in 2 of 3 clusters. (**Down**:) In Cluster1, those genes appeared in 3 clusters only were observed in one node. Many genes in cluster2 were also observed in the other two clusters.

3. Functional Annotation

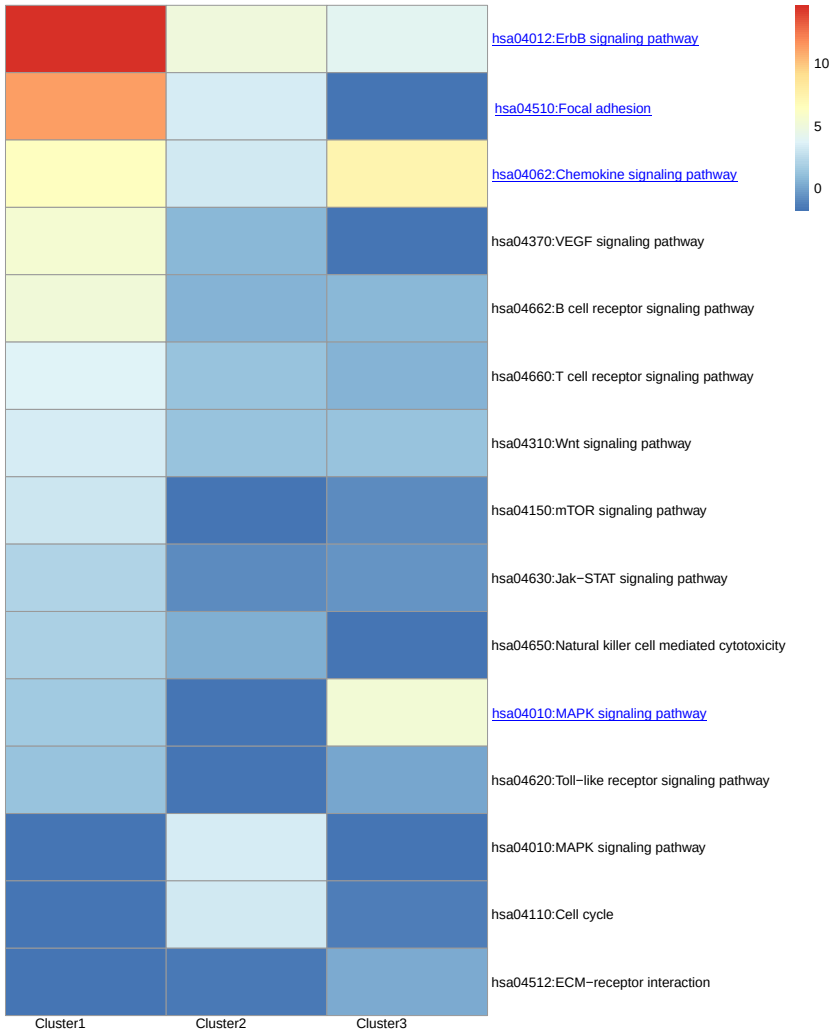


Figure: KEGG pathway annotation of signature clusters.

4. Recurrent modules analysis

Cluster	Number of Unique Genes	Most frequent genes
1	110	ROCK2 (4) AR (3) LAMA3 (3) BMP4 (3) RXRA (3) EGFR (3) MSH3 (3)
2	88	BIRC5 (3) CKS2 (3)
3	112	CXCR4(5) GSTP1(5) COL4A2(5) TGFB3(4) F2R(4) PDGFRA(4) CXCL12(4) GNG2(4) GNG11(4) ETS1(4) CBLB(4) GNB4(4) LAMA2(4) COL4A1(4)

Figure: Recurrent genes (appeared in more than 3/4 nodes) in each signature cluster.

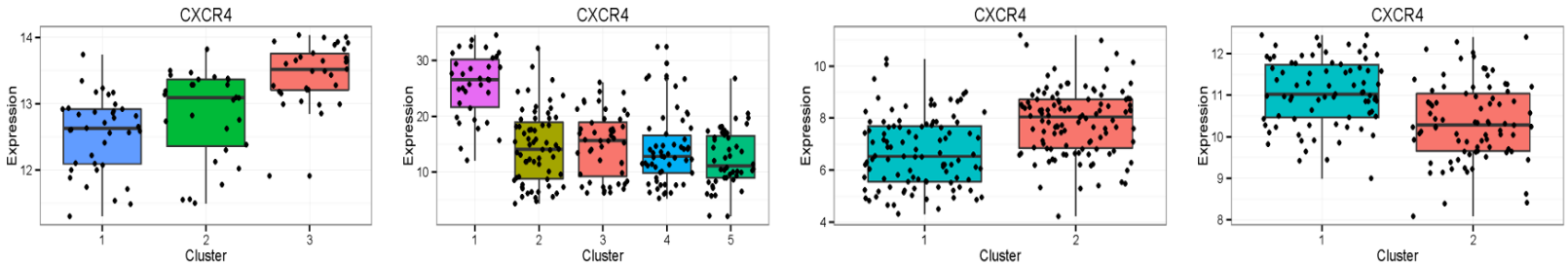
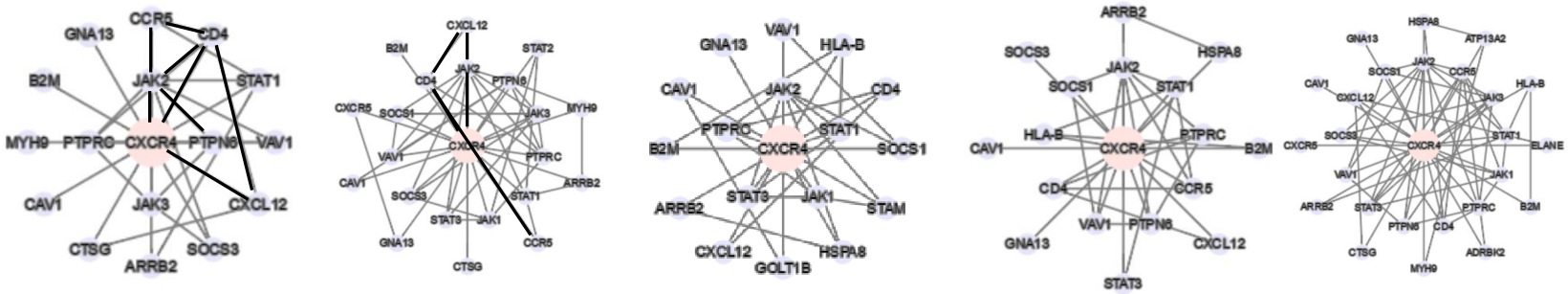


Figure: Expression of CXCR4 grouped by clusters of each dataset.

Next, we noticed some genes don't appear in hsa05200, but show strong co-expression with CXCR4/CXCL12 and are connected in PPI.



HCCDB1 HCCDB3 HCCDB4 HCCDB8 HCCDB15

Figure: Coexpression Network of CXCR4 in different datasets

	CD4	CXCL12	CCR5	JAK2	STAT1	PTPN6	PTPRC
One step	7	10	8	10	9	6	11

Figure: Recurrent genes appear in co-expression network of CXCR4. (11 datasets overall)

Some surival analysis results:

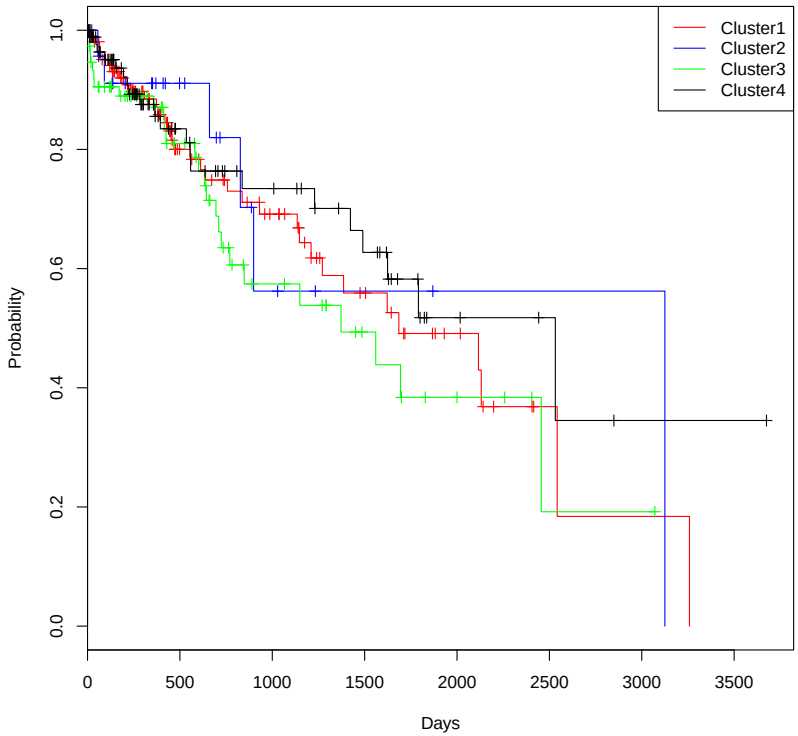


Figure: KM plot of HCCDB15 based on clustering results.

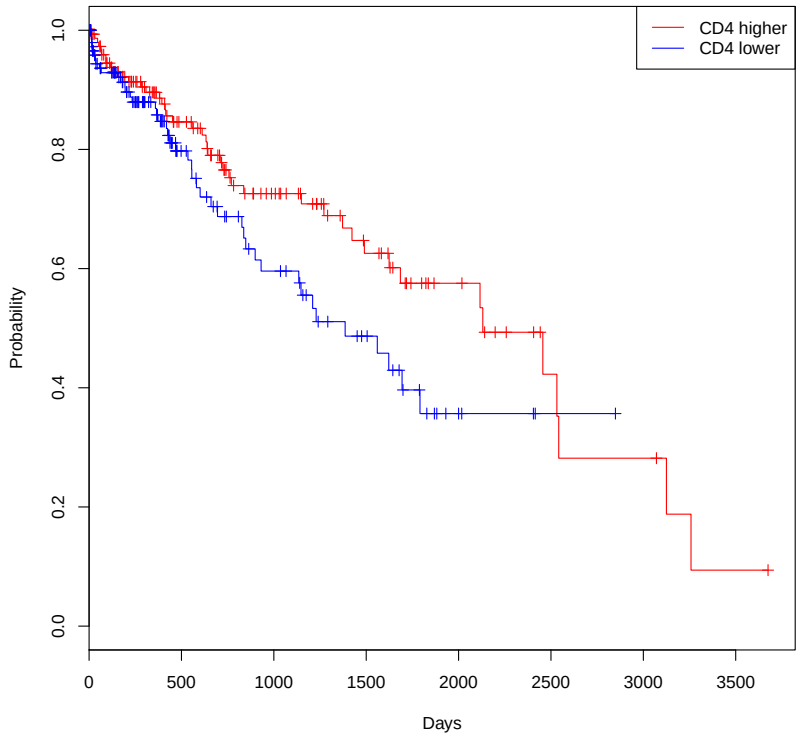


Figure: KM plot of HCCDB15 based on CD4 expression. KM:0.06; Cox:0.06

Others: CXCR4+/-;CXCL12+CXCR4;CD4+AFP show no association with survival time.

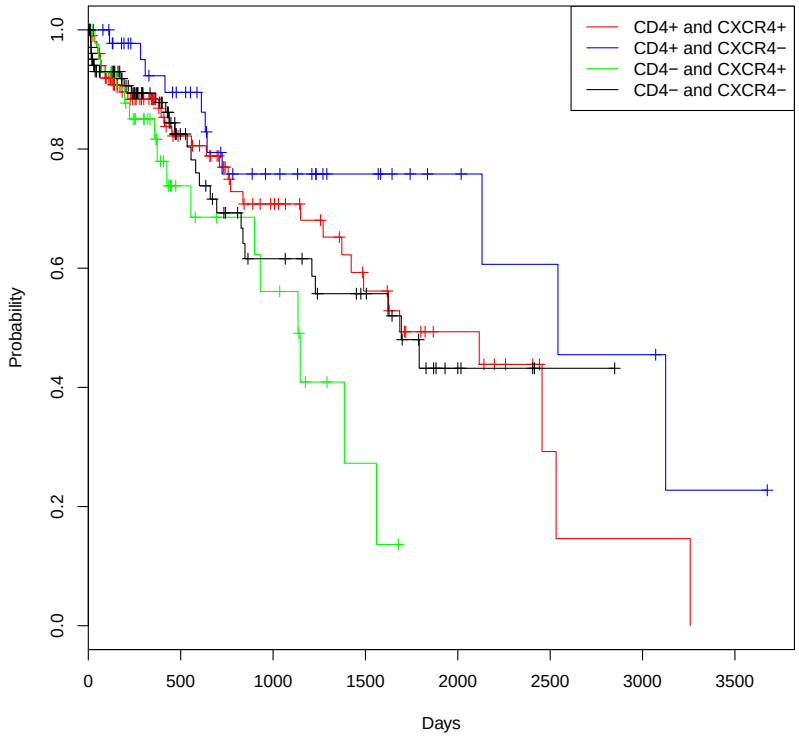


Figure: KM plot of HCCDB15 based on CD4 and CXCR4 expression. KM: 0.02; Cox: 0.37

5. Re-classify single dataset

- The main difficulty: The property of the gene expression signal widely varies across the datasets due to **Batch Effect** (platform difference, lab-to-lab/day-to-day variation, etc). This makes it practically infeasible to build a model of gene expression signal trained on one particular dataset and apply it to the rest as seen in standard machine learning prediction procedures.
- NearestTemplatePrediction? (A standard reference index)
- To be done ..

Further improvements