A Molecular Features-Based Meta-classification System of Hepatocellular Carcinoma

Dongfang Wang

Tsinghua University

2016-03-25

Introduction

Currently, we've collected 13 datasets of hepatocellular carcinoma, from different platforms, different types ...

HCCDB: Hepatocellular Carcinoma Expression Atlas

Latest updated
January 5, 2016

Number of Datasets
mRNA: 13
miRNA: 4

Number of Samples
HCC: 2109
Adjacent: 1277
Healthy: 18
Cirrhotic: 40

SEARCH

DOWNLOAD

LINKS

Gu's Lab

OncomiRDB

LRAcluster

Hepatocellular carcinoma (HCC), the major primary liver cancer, is one of the most common cancer type world-wide. Current studies show that HCCs is highly genetic heterogeneous. Large-scale molecular profiles are crucial for identifying the altered molecular networks, diverse biomarkers and molecule-based subtypes of HCCs.

HCCDB aims at collecting large-scale gene expression profiles (phase I) and other molecular data of HCC clinical samples. The database search engine provides gene centered meta-data interface: it visualizes the expression details and prognostic values of each gene. The database also provides basic data analysis tools, including differential analysis and survival analysis. HCCDB should be a useful resource for HCC research.

DATASET	ANALYSIS
HCCDB-1 **Source**: GSE22058 **Sample**: Adjacent 97; HCC 100; Paired **Platform**: mRNA array, miRNA array	**Diff_Analysis**: HCC/Adjacent (mRNA) **Diff_Analysis**: HCC/Adjacent (miRNA)
HCCDB-3 **Source**: GSE25097 **Sample**: Healthy: 6; Cirrhotic: 40; Adjacent: 243; HCC: 268 **Platform**: mRNA array	**Diff_Analysis**: Adjacent/HCC **Diff_Analysis**: Healthy/Cirrhotic/Adjacent/HCC
HCCDB-4 **Source**: GSE36376	**Diff_Analysis**: Adjacent/HCC

Credit: Zhang Guchao

We want to build a meta-classification system, which could:

- find consistent and replicatable subtypes among different datasets;
- integrate biological prior knowledge effectively;
- and finally, discover new molecular features of highly heterogenous HCC samples.

Compared with previous methods, we want to focus on the specific immune-related or metabolic process of tumour cells ,...

Paper Introduction

LETTER

doi:10.1038/nature16982

Tumour–specific proline vulnerability uncovered by differential ribosome codon reading

Fabricio Loayza-Puch[1]*, Koos Rooijers[1]*, Levi C. M. Buil[2], Jelle Zijlstra[1], Joachim F. Oude Vrielink[1], Rui Lopes[1], Alejandro Pineiro Ugalde[1], Pieter van Breugel[1], Ingrid Hofland[3], Jelle Wesseling[4], Olaf van Tellingen[2], Axel Bex[5] & Reuven Agami[1,6]

[1]Division of Biological Stress Response, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. [2]Department of Bio-Pharmacology/ Mouse Cancer Clinic, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. [3]Core Facility Molecular Pathology and Biobanking, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. [4]Molecular Pathology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. [5]Division of Surgical Oncology, Department of Urology The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. [6]Department of Genetics, Erasmus University Medical Center, Wytemaweg 80, 3015 CN Rotterdam, The Netherlands.
*These authors contributed equally to this work.

490 | NATURE | VOL 530 | 25 FEBRUARY 2016

In this paper, they defined a new molecular feature which may be related with cancer proliferation.

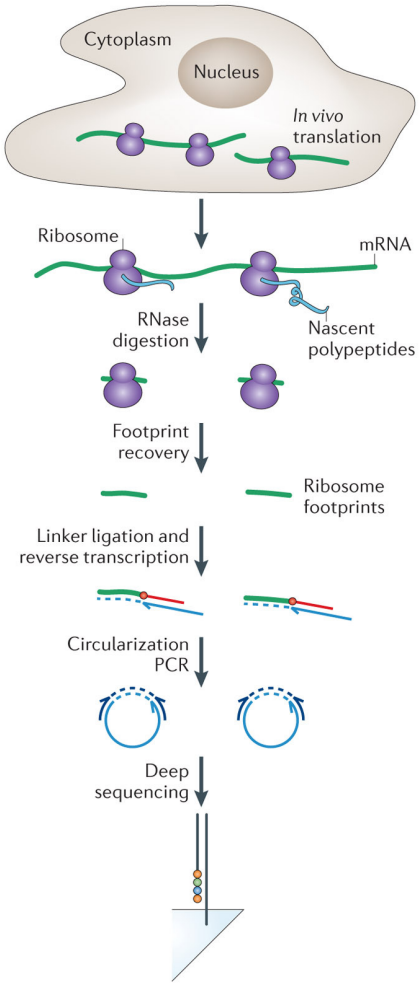
Background

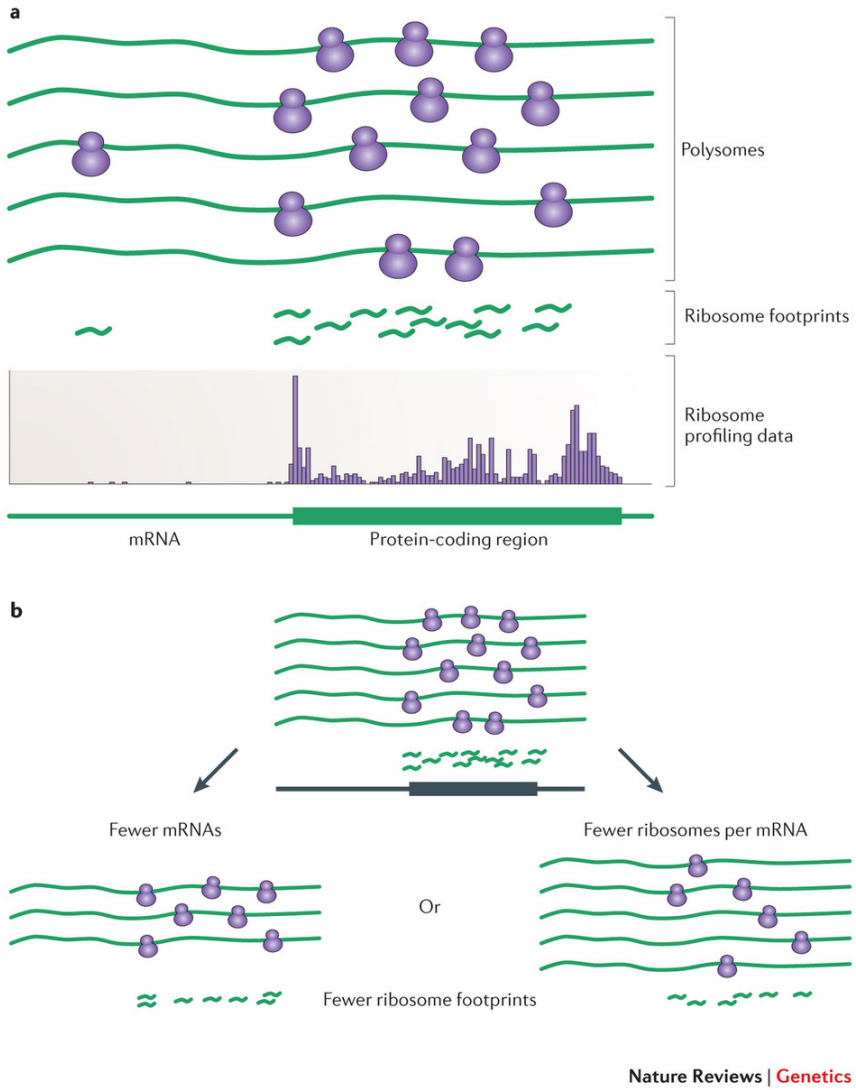
- *Amino Acids* are necessary for cancer proliferation.

- Different cancers, different samples need different amino acids to different extent. (Heterogeneity)

- If we know, for a certain patient, which kind of amino acid is the most important and a restrictive factor of its growth, we could deprive it to starve his tumour cells.

Ribosome Profiling

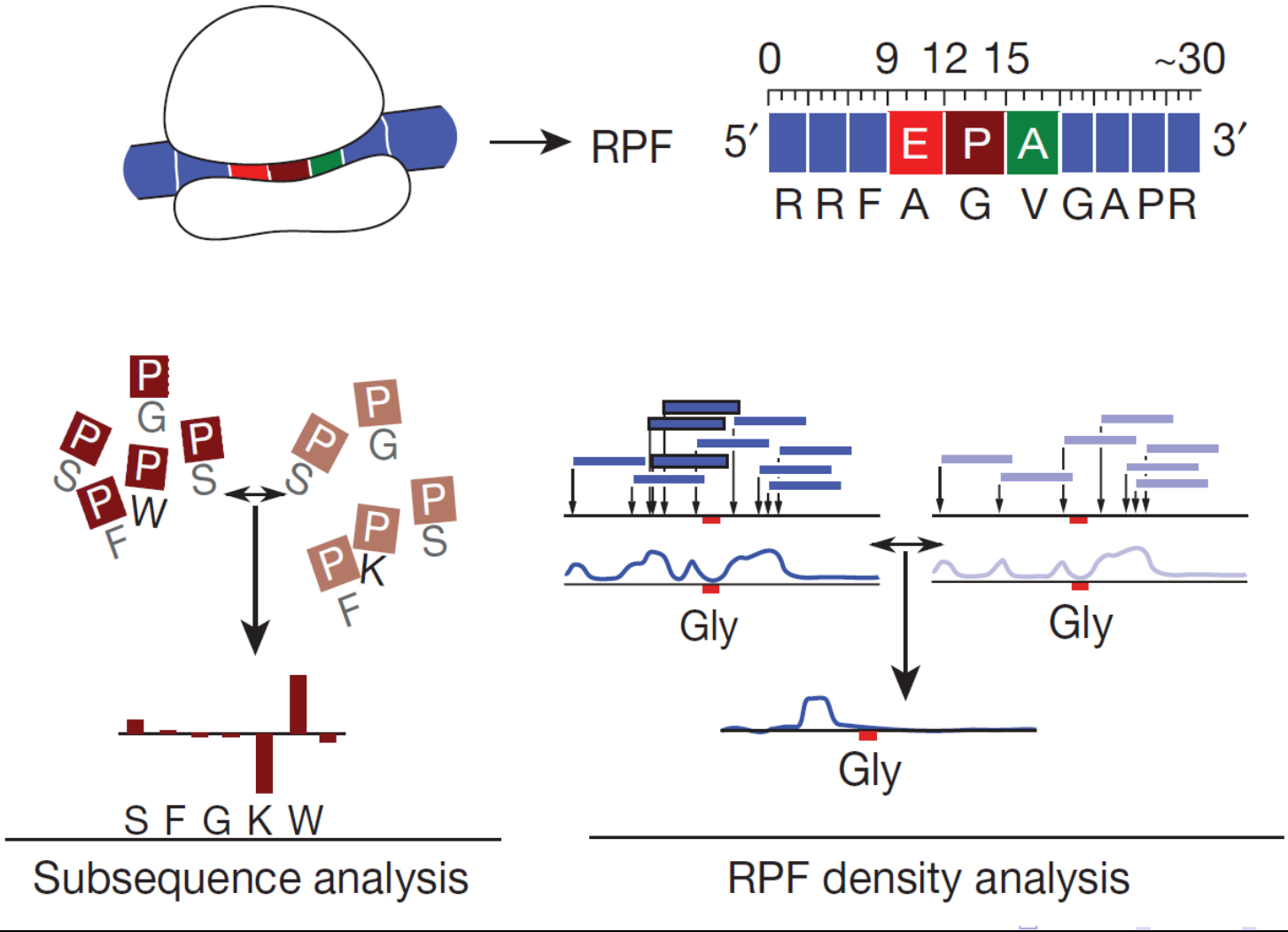




Credit: Nicholas T. Ingolia,et al,*Nature Reviews Genetics*, 2014

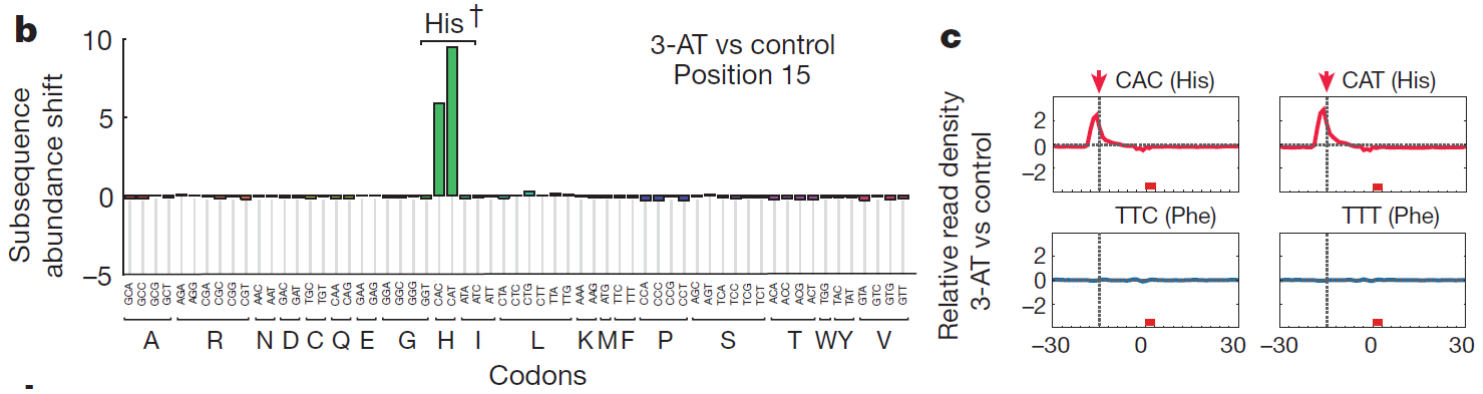
Differential ribosome measurements of codon reading

RPF: ribosome-protected fragments

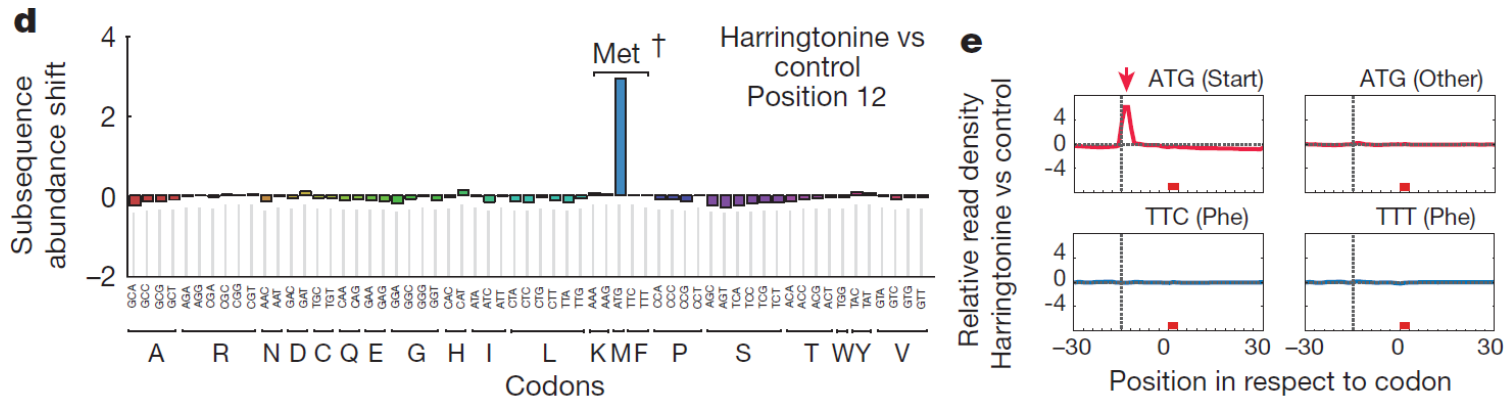


Functionality of diricore

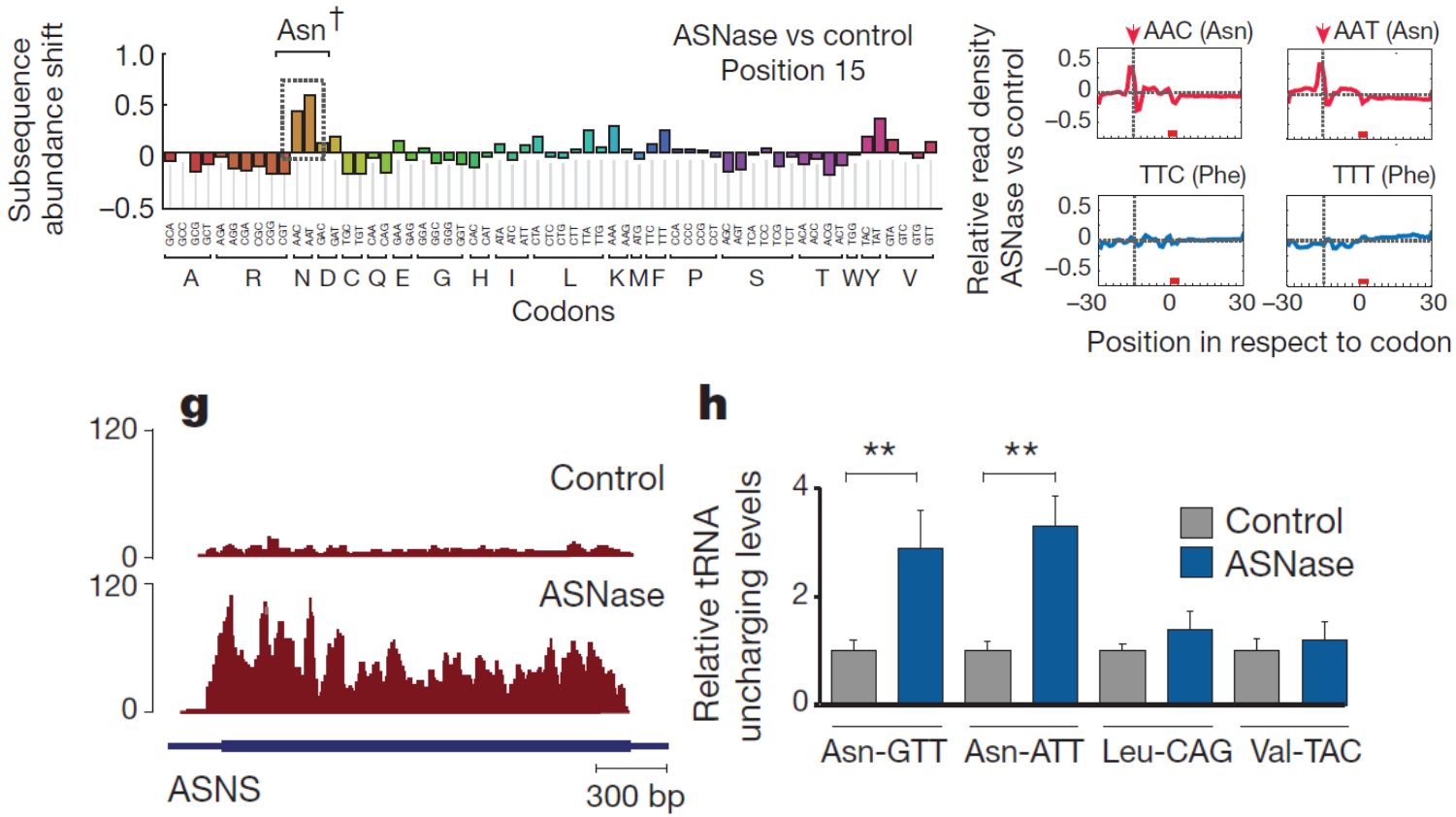
- 3-AT inhibitor of histidine synthesis



- Harringtonine: immobilize initiating ribosomes (ATG codon: initiator)



- L-asparaginase treatment to cancer cells (inhibit asparagine)



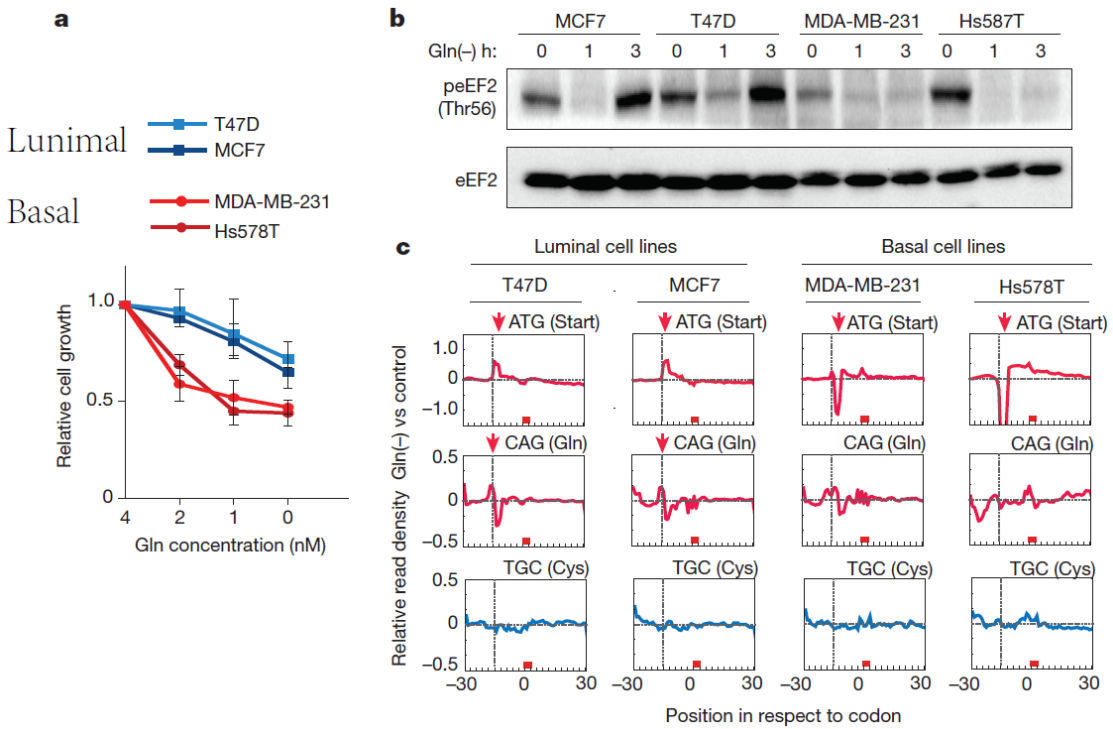


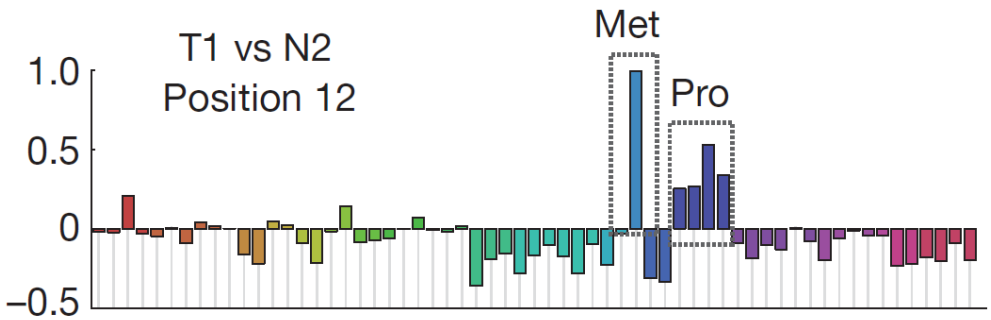
Figure 2 | Estimation of global changes in translation and specific codon signals by diricore. **a**, Normalized cell growth of the indicated breast cancer cell lines at different concentrations of glutamine. Measurements were taken 48 h after plating. peEF2, phosphorylated eEF2. **b**, Immunoblot analysis of cells grown in complete media or under glutamine starvation for the indicated times. **c**, Diricore analysis of cell lines deprived of glutamine for 48 h. For **a**, values represent the mean of three independent biological replicates; error bars represent s.d. Cutoff for RPF density shift plots in **c** was 25 reads per window.

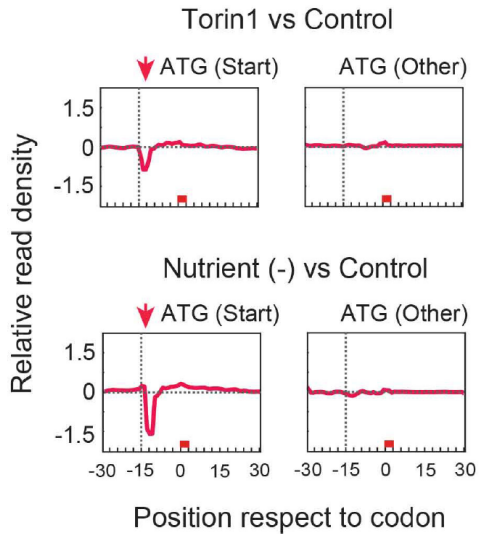
Note:

- Basal: sensitive to Glu; Global translation reduction.
- Luminal: specific signals of Glu codon.

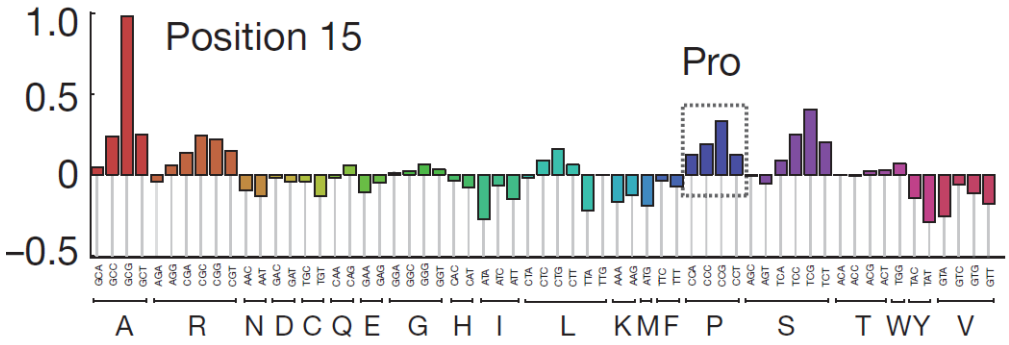
Detect restrictive amino acids in kidney cancer

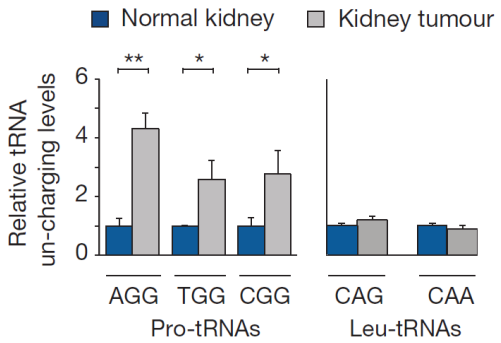
- Global increased translation initiation rate:



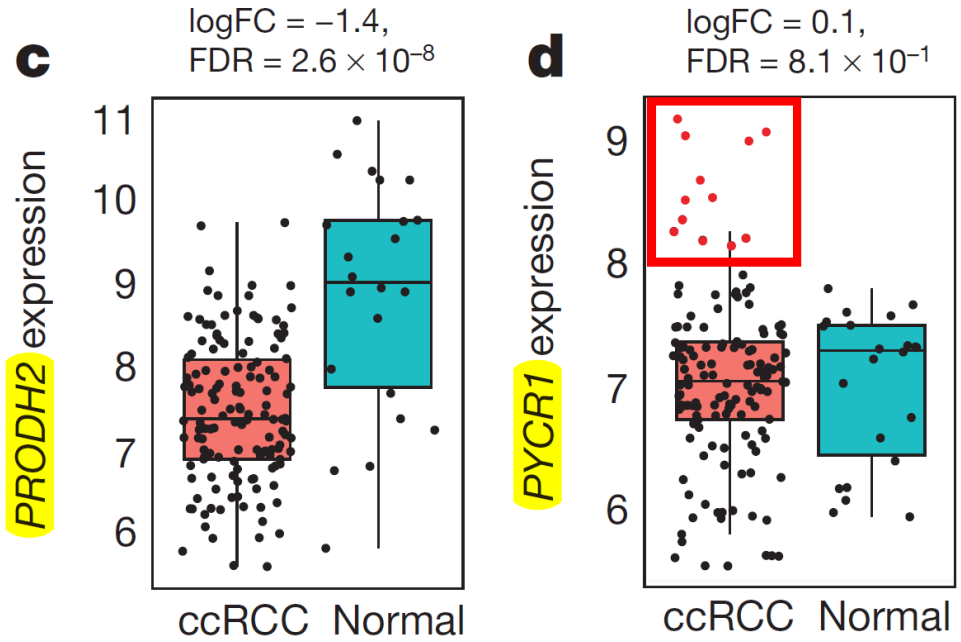


- Limiting availability of proline for protein synthesis.





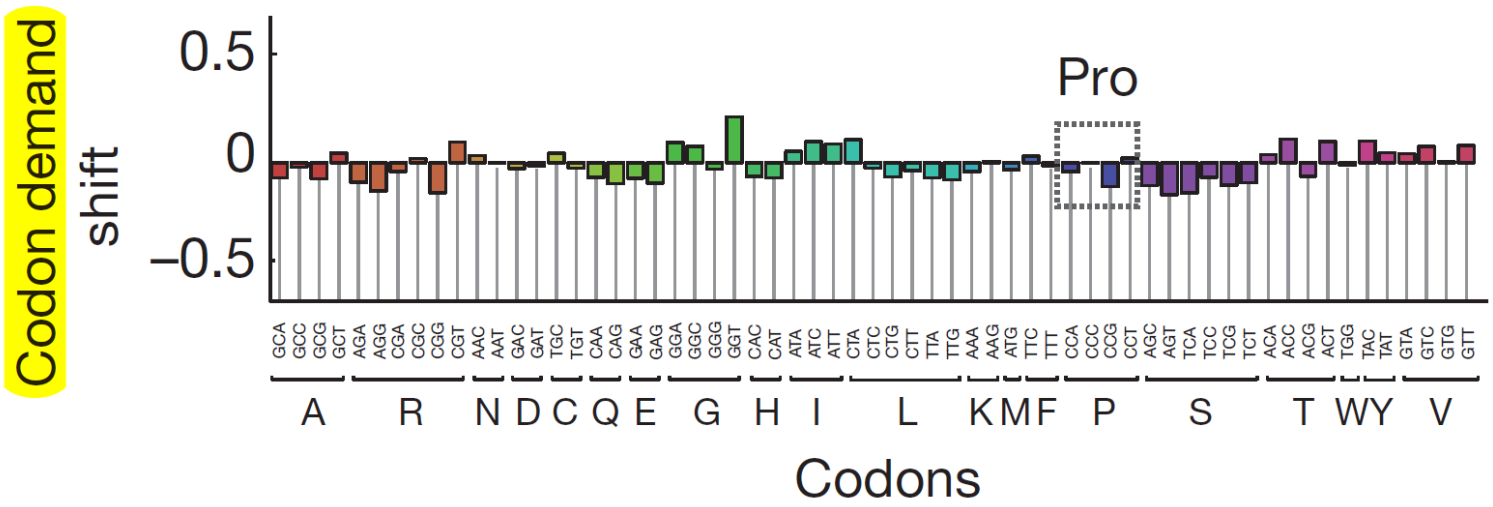
A compensatory feedback mechanism in the kidney cells to enable proliferation in conditions of relative proline shortage:



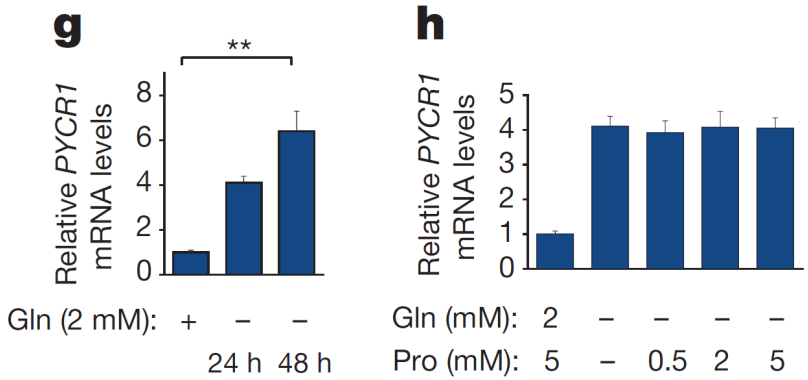
- PYCR1: a key enzyme that catalyses the last step in proline synthesis.
- PRODH2: proline catabolic enzymes.

The reason of proline deficiency

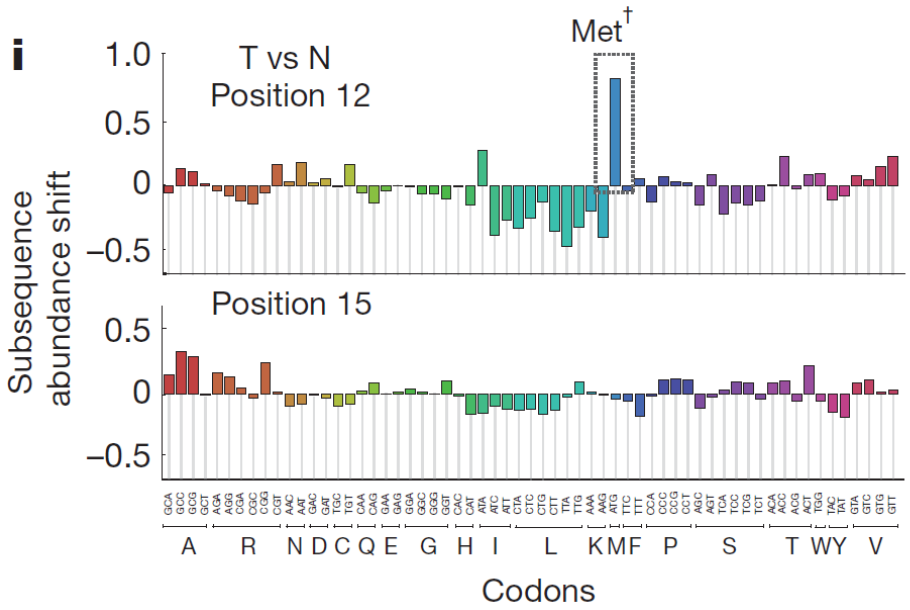
- No increased codon demands

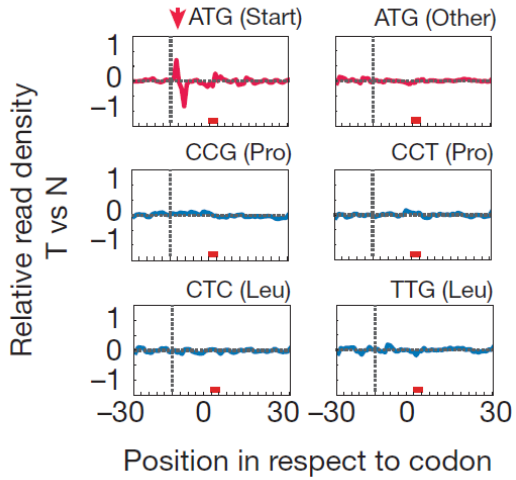


- Lack of proline precursor, decreased proline production.



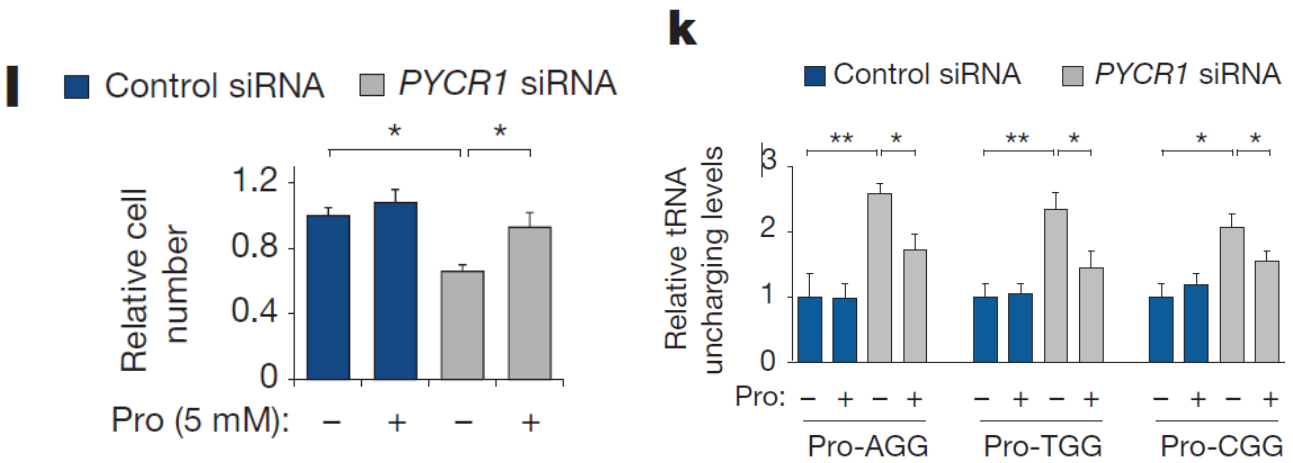
No presence in another tumour





Interference with the proline/PYCR1 regulatory pathway

- Used 786-O cell line, high PYCR1

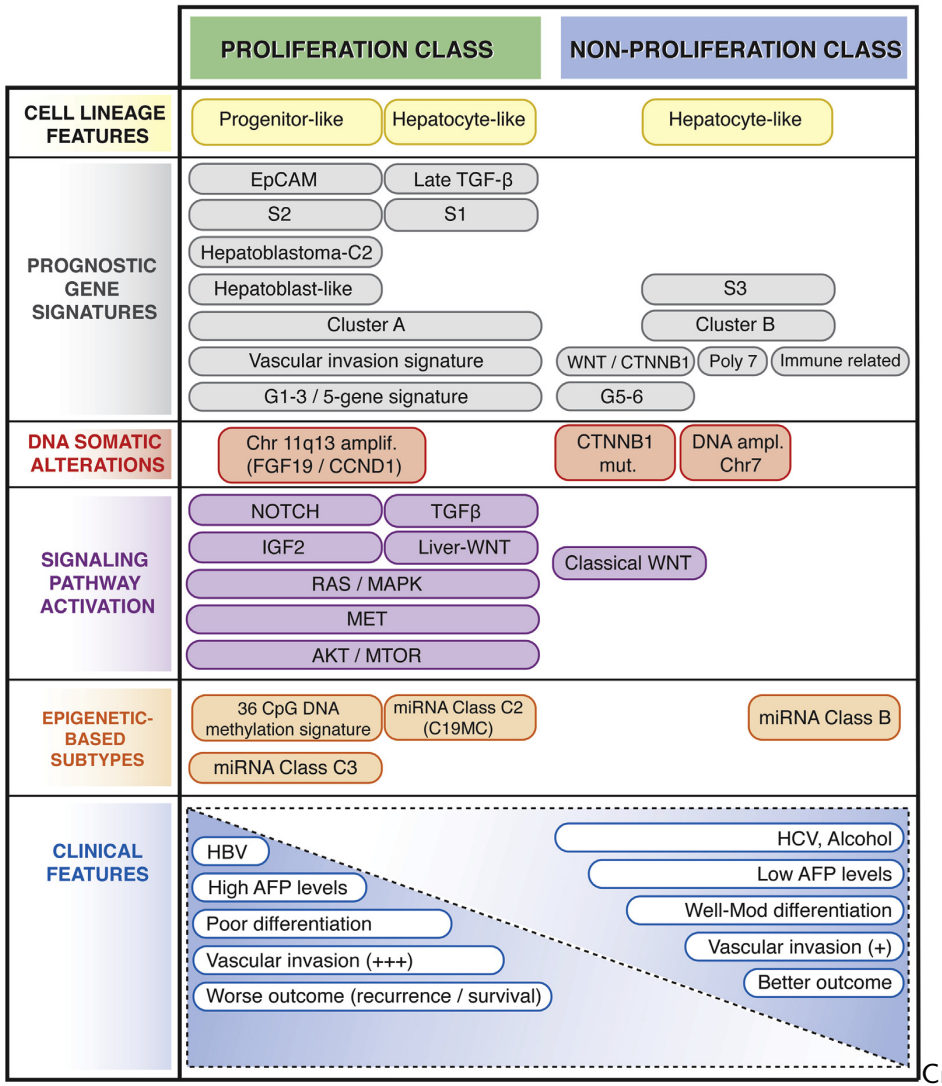


Interferring with intracelular proline production attenuates proliferation when proline availability is limiting.

Summary

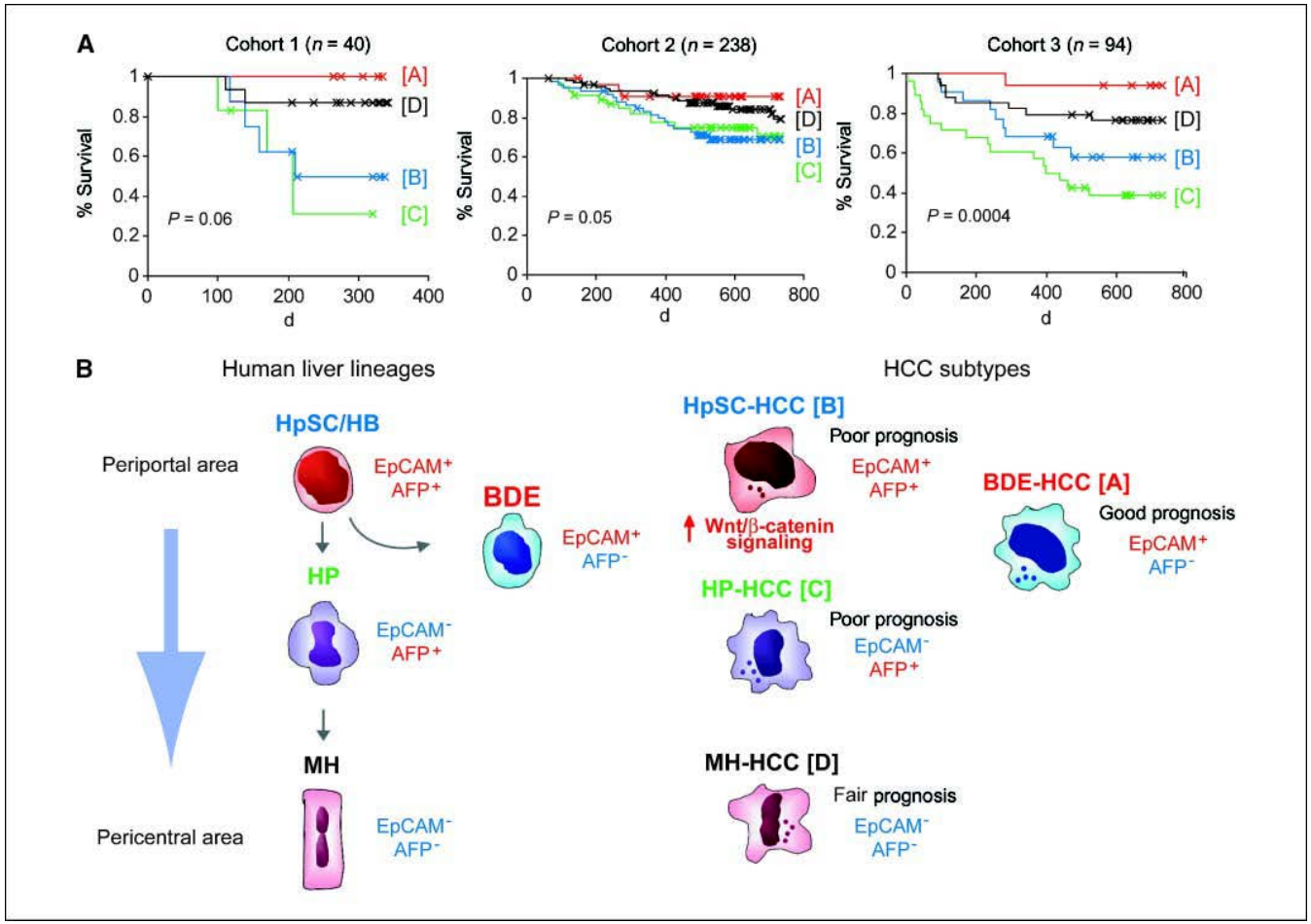
- **diricore**: reveal unknown amino acid deficiencies, vulnerabilities that could be used to target key metabolic pathways for cancer.
- Discover proline deficiencies in cancers and show the key pathway (PYCR1) allowing tumour expansion.
- Show a great example of using currently available technique to discover new molecular features and cancer subtypes.
- Is it possible to use RNAseq or Exome data to show metabolic heterogeneity of tumours?

Current molecular subtypes of HCC



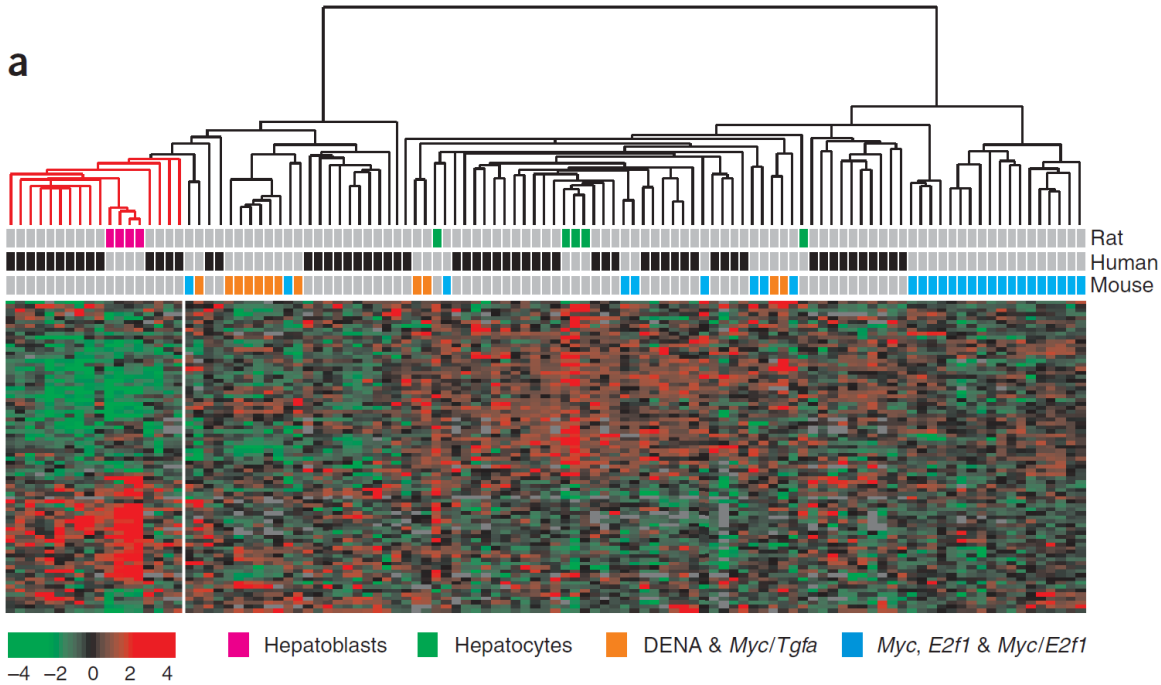
Credit: Jessica Zucman-Rossi,et al,*Gastroenterology*, 2015

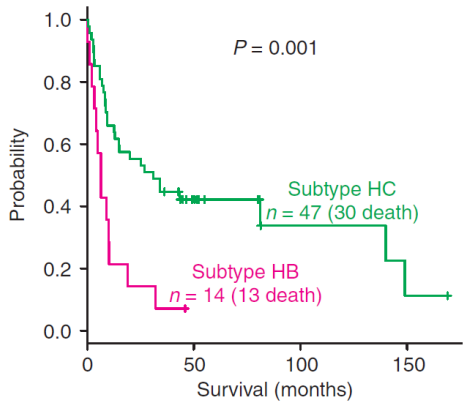
EpCAM and AFP subtypes



Credit: Taro Yamashita,et al,*Cancer research*, 2015

Progenitor-like Subtypes





Credit: Ju-Seog Lee,et al,*Nature Medicine*, 2006

Work Plan

Current Framework

A similar work

Planey and Gevaert *Genome Medicine* (2016) 8:27
DOI 10.1186/s13073-016-0281-4

Genome Medicine

METHOD Open Access

CoINcIDE: A framework for discovery of patient subtypes across multiple datasets

Catherine R. Planey and Olivier Gevaert*

Abstract

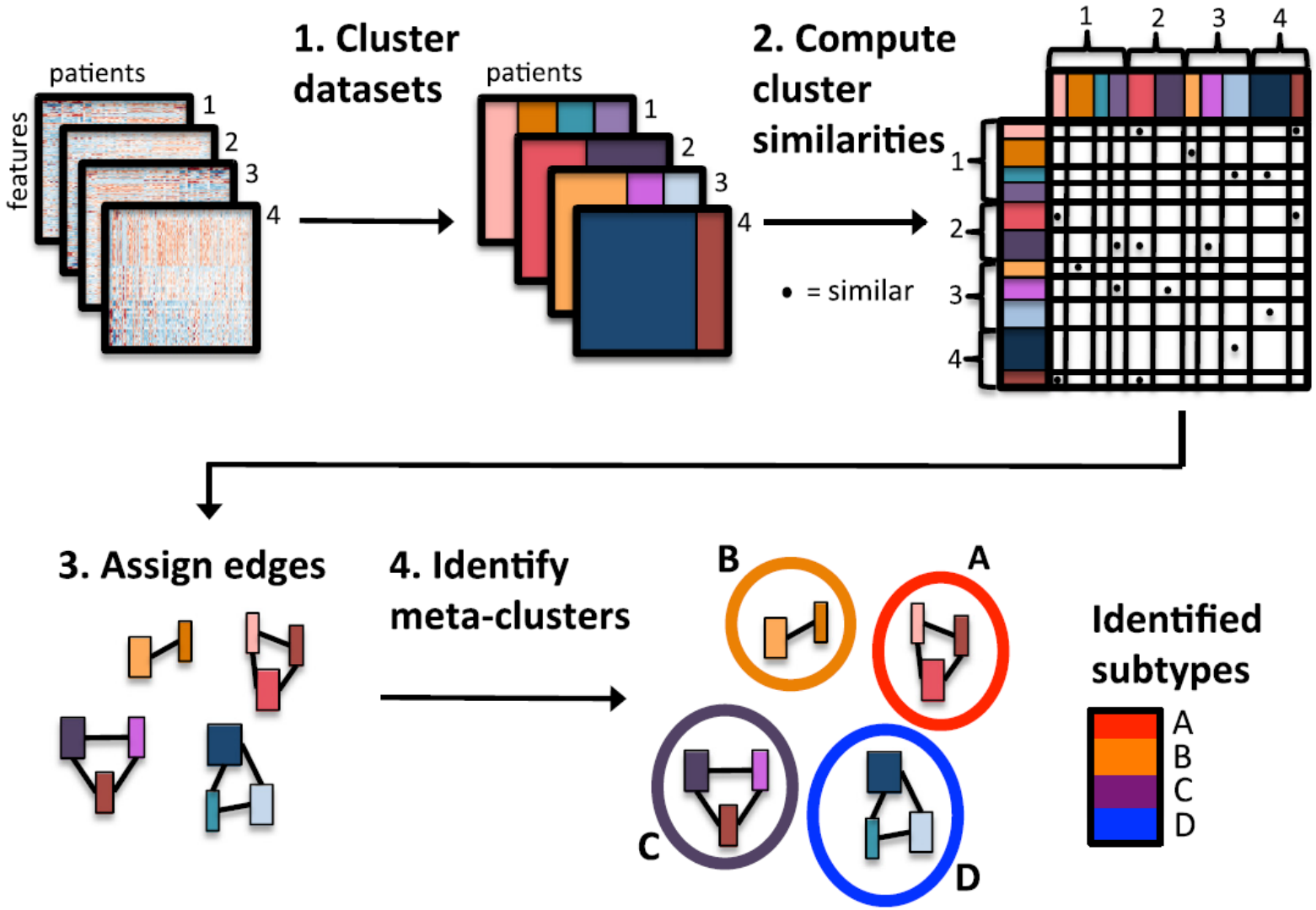
Patient disease subtypes have the potential to transform personalized medicine. However, many patient subtypes derived from unsupervised clustering analyses on high-dimensional datasets are not replicable across multiple datasets, limiting their clinical utility. We present CoINcIDE, a novel methodological framework for the discovery of patient subtypes across multiple datasets that requires no between-dataset transformations. We also present a high-quality database collection, curatedBreastData, with over 2,500 breast cancer gene expression samples. We use CoINcIDE to discover novel breast and ovarian cancer subtypes with prognostic significance and novel hypothesized ovarian therapeutic targets across multiple datasets. CoINcIDE and curatedBreastData are available as R packages.

Replicability: 'The chance that an independent experiment targeting the same scientific question will produce a consistent result'. (Subtypes with similar signaling of patient subtypes can be found across multiple datasets.)

There are two large hurdles to producing replicable patient subtypes: a lack of curated disease-specific data-set collections and a lack of methods that discover con-sensus across clusterings from multiple datasets. When

- **CoIncIDE**: Clustering Intra and Inter DatasEts.
- **cruatedBreastData**: 24 high-quality cruated breast cancer microarray datasets.

CoINcIDE method: Overview



Key steps of CoINcIDE

Step 1 **Clustering of single dataset**: gene selection + bootstrap based consensus clustering

Step 2 **Similarity between clusters from different datasets**:

- For patients from dataset 2, assign them to clusters in dataset 1 based on **Peasrson's correlation with the cluster centroid**.
- '2A' is matched to '1A' if most patients in '2A' are assigned to cluster 'A' in dataset 1. The **smiliarity metric** is the proportion of '2A"s patients assigned to '1A'.
- Switch the role of 1 and 2. Only those clusters **matched both times** are retained.

Step 3 **Assign edges**: select a similarity threshold.

Step 4 **Network community detection**:

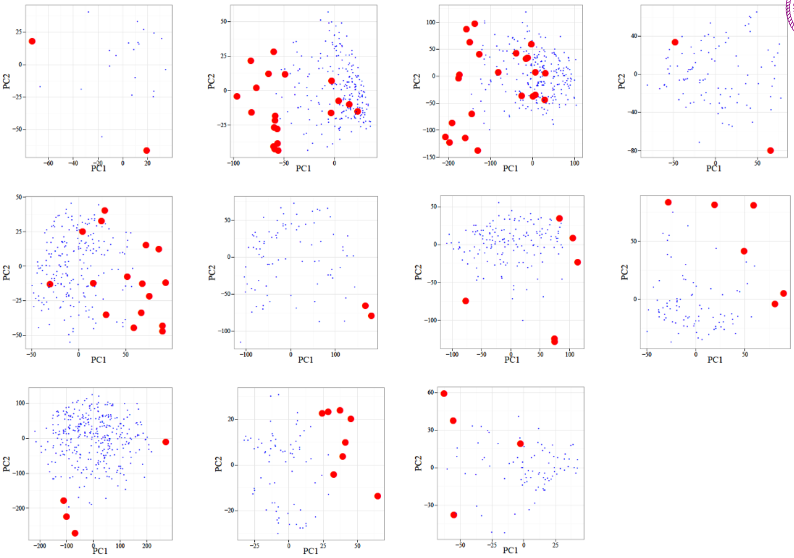
Preprocess

Gene Selection:

- Select highly expressed genes in each datasets;
- Select genes in cancer pathways.

Outlier detection:

- PCA: 1st, 2ed PCs;
- R package: library("arrayQualityMetrics") – *K-S test*

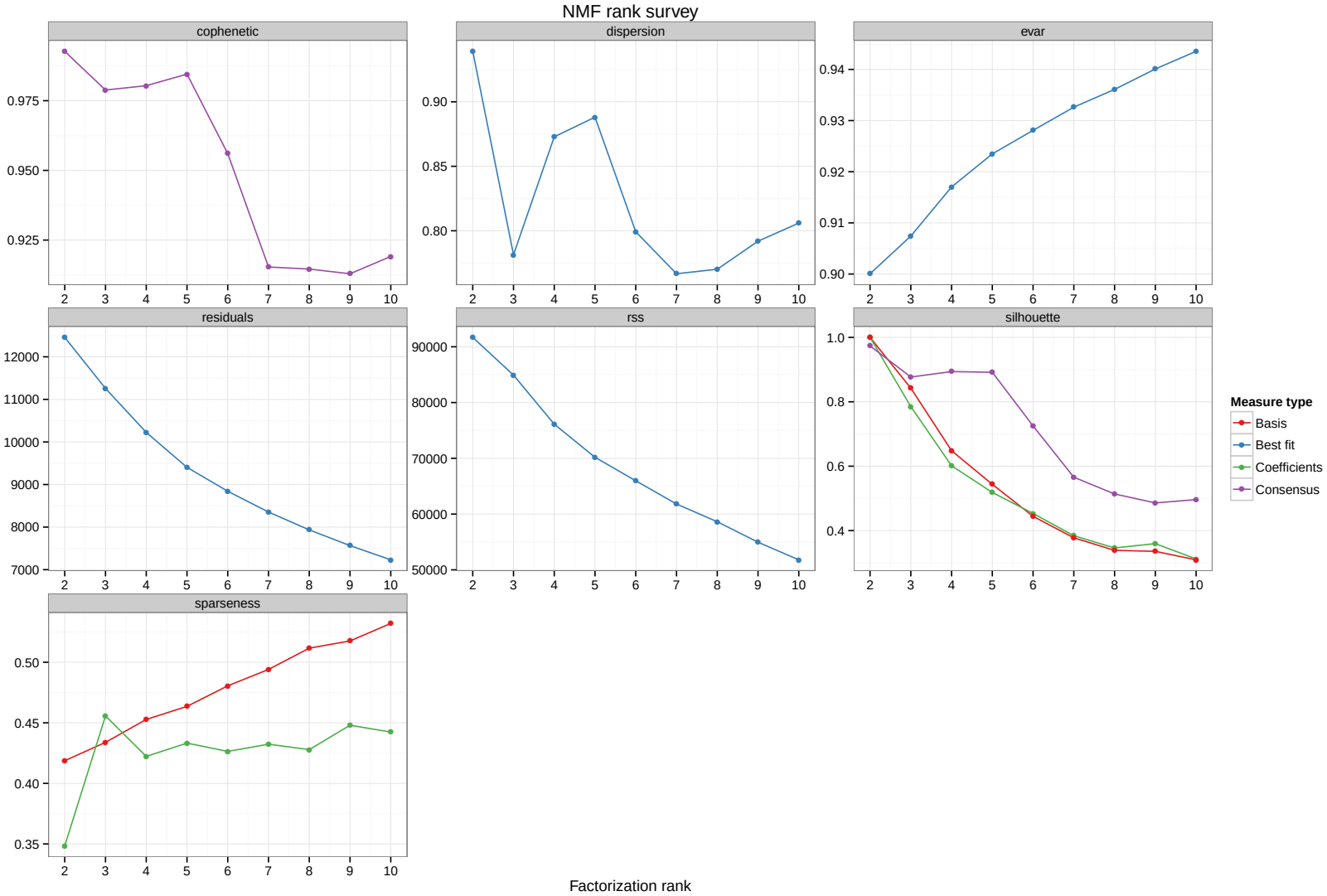

PC2
PC1

Clustering of Every Single Dataset

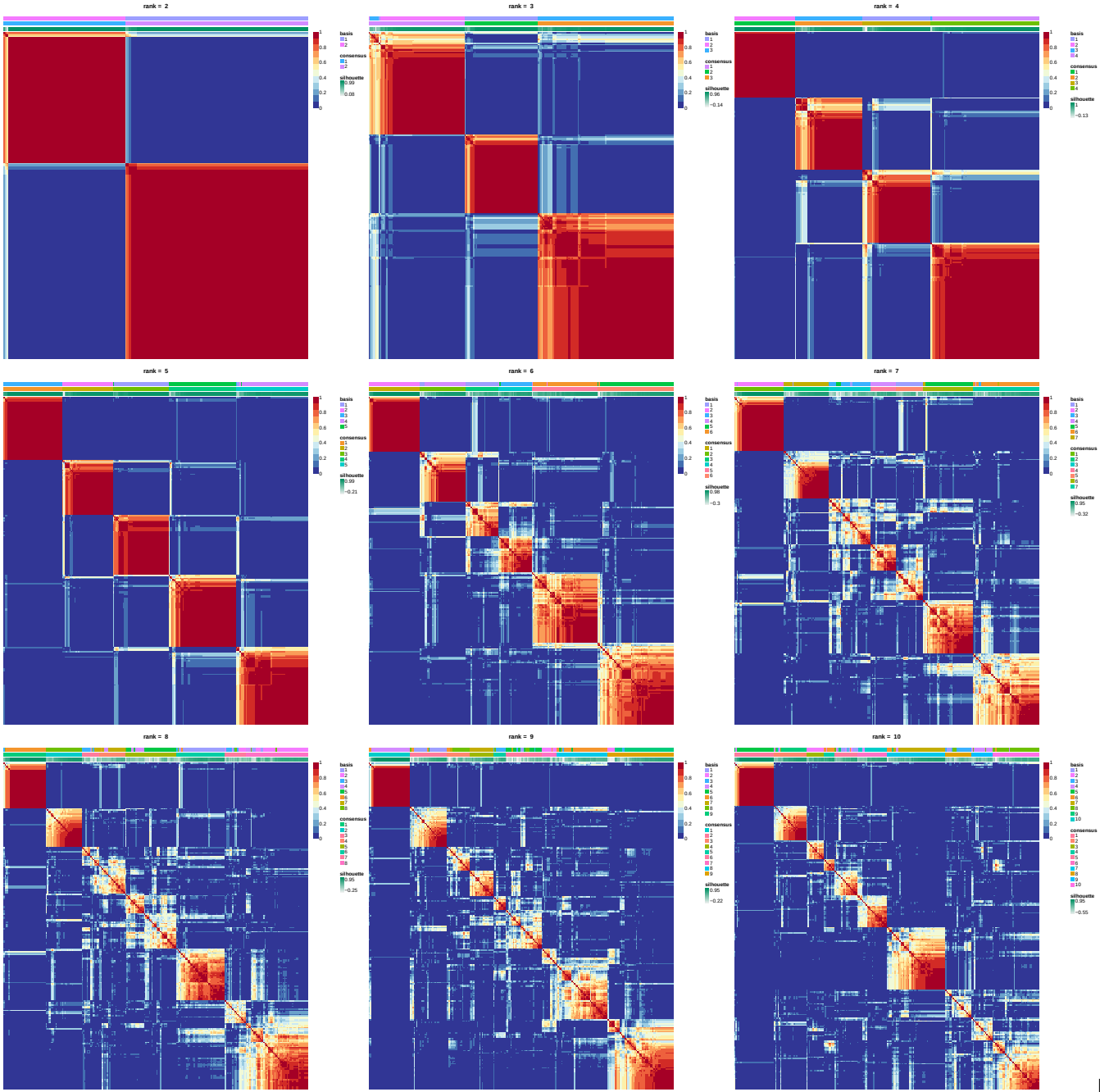
- Dimension reduction: NMF
- Consensus clustering

Questions:

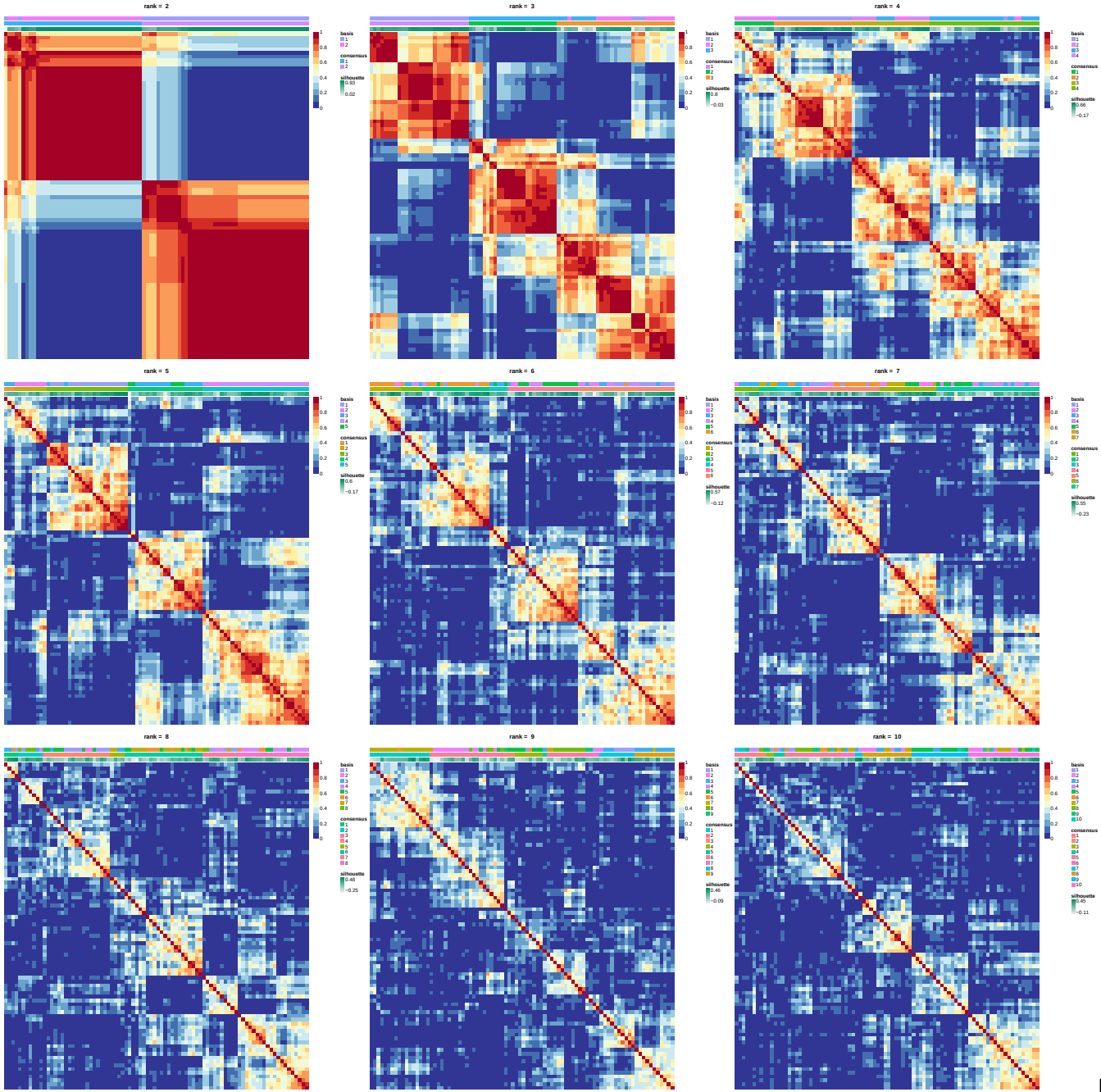
- Select factorization rank and number of clusters:
  - Let k change from 2 to 10, we could compute some measures, but ...
  - We could judge it from the consensus matrix , but ...
- Select algorithms to cluster


NMF rank survey
cophenetic
dispersion
evar
residuals
rss
silhouette
sparseness
Measure type
Basis
Best fit
Coefficients
Consensus
Factorization rank

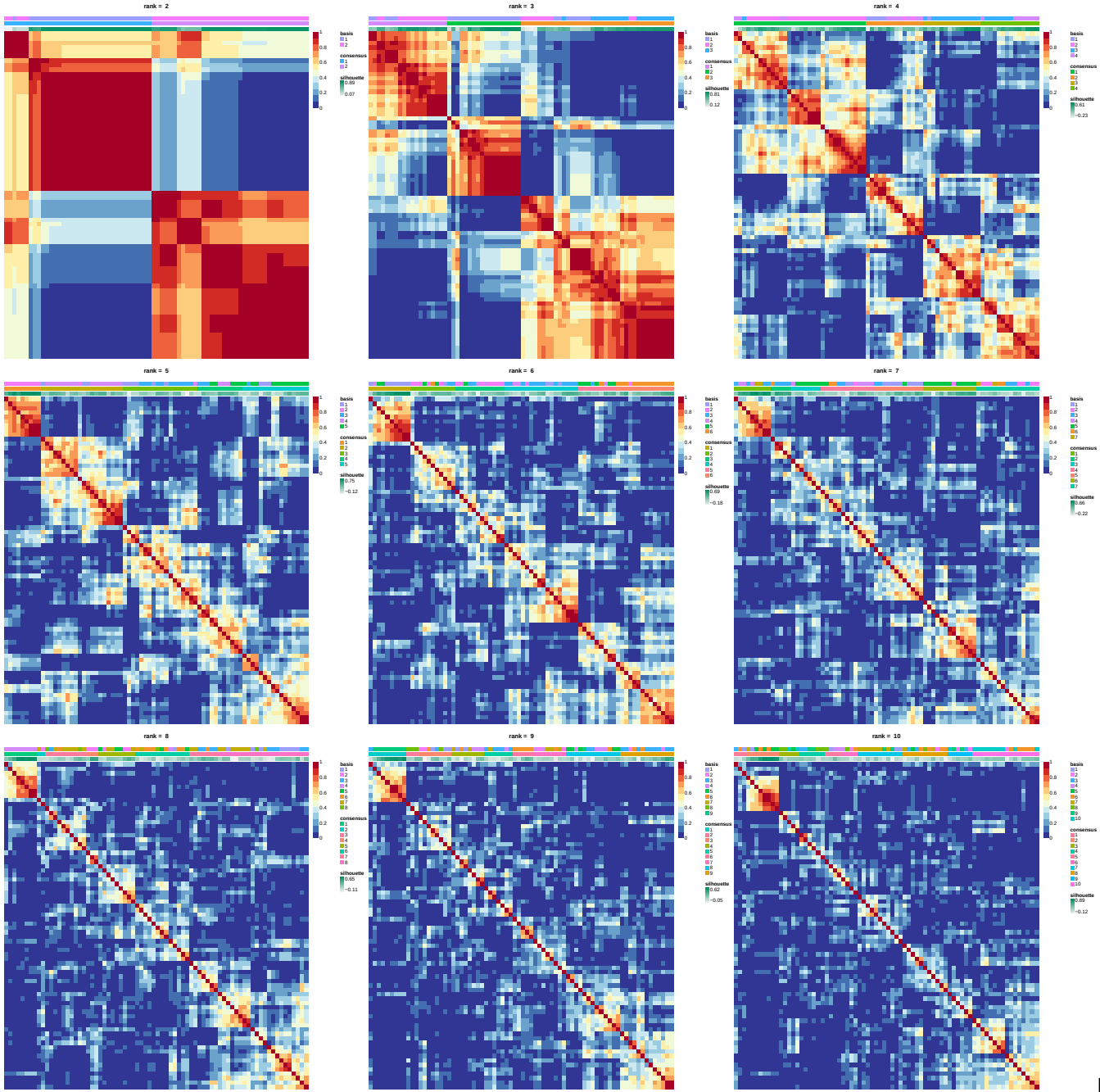
HCCDB3



HCCDB3



HCCDB11



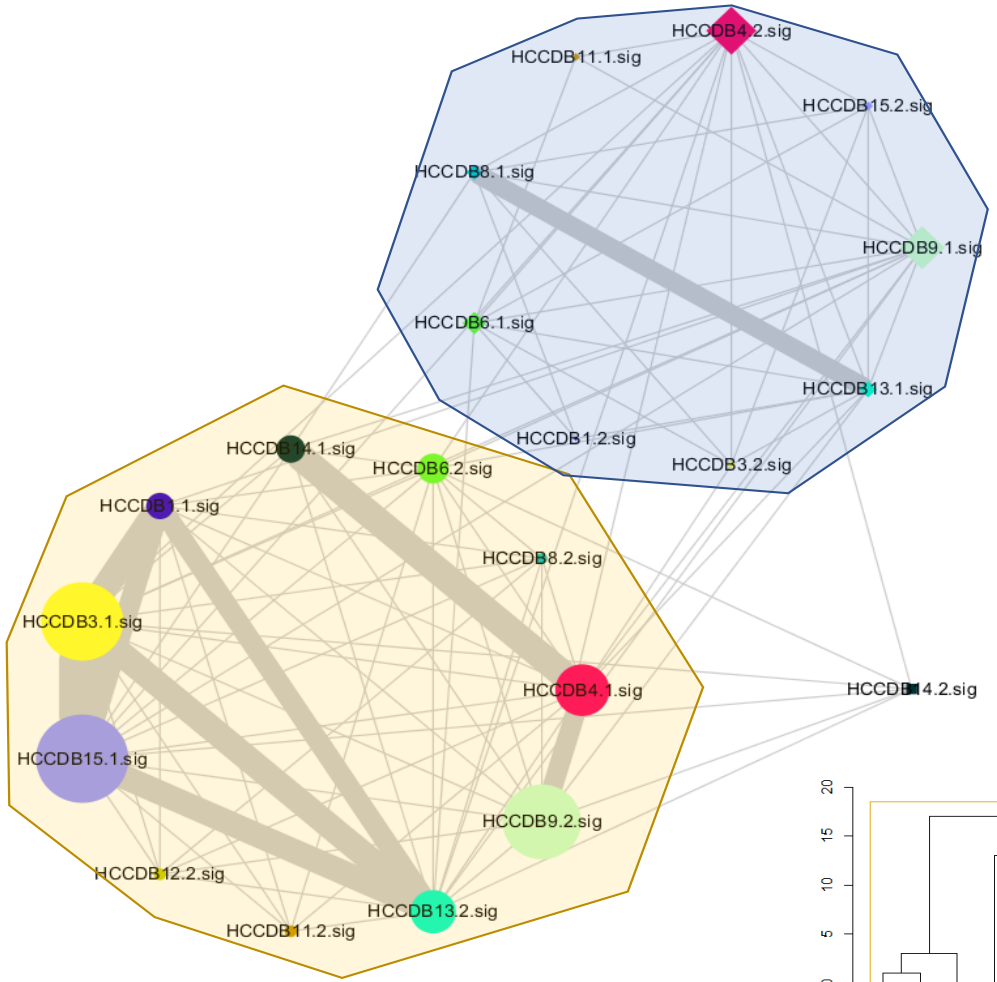
HCCDB12

Signals of every cluster in each dataset

Currently, cluster each dataset into 2 clusters.

- Use $t.test$ and $wilcoxon.test$ to obtain differentially expressed genes. (BH adjustment)
- Threshold: $t.pvalue < 1e-5$ and $wilcox.pvalue < 1e-5$.
- Mean expression: higher in cluster 1 as the signals of this cluster.

Signal Netwok and communities

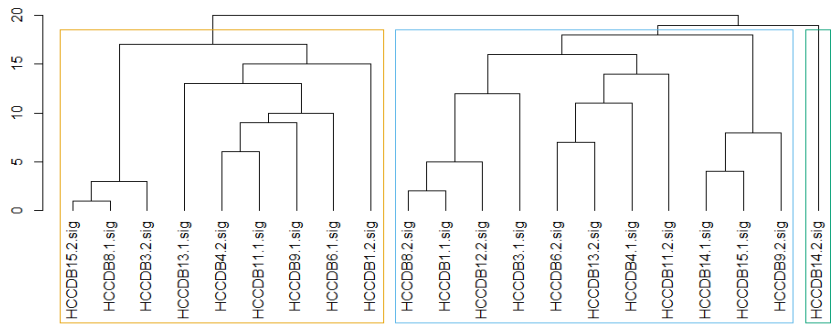


Cluster ○ ◇

Node size: Number of genes

Node color: Different datasets

Edge width: Jaccard similarity



GO Analysis of Cluster1

- Select the top 20 genes which appear the most times in cluster 1.

Biological Process	Adj.pvalue
response to chemical stimulus GO:0042221	6.43E-09
regulation of cell migration GO:0030334	2.87E-07
regulation of cell motility GO:2000145	3.08E-07
regulation of locomotion GO:0040012	4.12E-07
regulation of cellular component movement GO:0051270	4.12E-07
cell morphogenesis involved in differentiation GO:0000904	7.09E-07
angiogenesis GO:0001525	7.09E-07
axonogenesis GO:0007409	7.09E-07
vasculature development GO:0001944	7.09E-07
positive regulation of cell migration GO:0030335	7.09E-07

GO Analysis of Cluster2

- Select the top 20 genes which appear the most times in cluster 2.

Biological Process	Adj.pvalue
regulation of epithelial cell proliferation GO:0050678	4.19E-06
epithelial cell proliferation GO:0050673	6.01E-06
positive regulation of metabolic process GO:0009893	8.98E-05
regulation of apoptotic process GO:0042981	8.98E-05
regulation of programmed cell death GO:0043067	8.98E-05
type B pancreatic cell development GO:0003323	8.98E-05
regulation of locomotion GO:0040012	8.98E-05
type B pancreatic cell differentiation GO:0003309	8.98E-05
cell-substrate junction assembly GO:0007044	8.98E-05
regulation of cellular component movement GO:0051270	8.98E-05