Paper Introduction:

Seifert *et al. Genome Biology* (2016) 17:204
DOI 10.1186/s13059-016-1058-1

Genome Biology

METHOD Open Access

Importance of rare gene copy number alterations for personalized tumor characterization and survival analysis

Michael Seifert[1,2,4*], Betty Friedrich[3] and Andreas Beyer[4]

Introduction

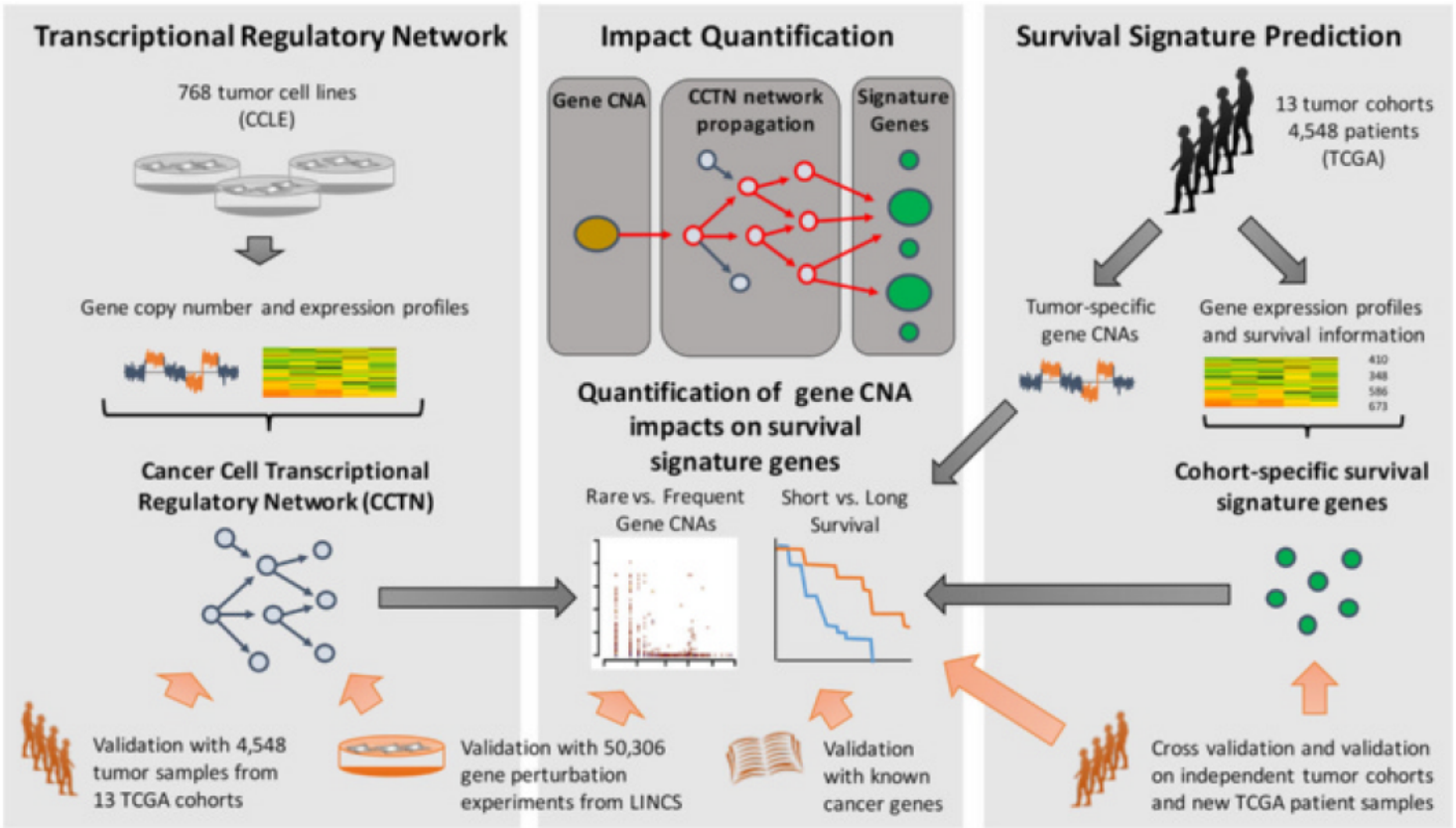
- Individual cancer risks are most likely not fully explained by frequent mutations alone.
- Quantifying the risks associated with rare mutations are difficult:
  1. Low observation probability;
  2. Insufficient statistical power;
  3. Hidden effects by complex interactions;
  4. Usual co-occurence of weak effects.

This study:

- Focus on the rare DNA copy number alterations;
- By exploiting the additional information contained in gene expression data;

- **Hypothesis**: Regulatory relationships between genes are fairly robust across tumours, whereas the specific mutational pattern of a given tumor is virtually private.

Methodological Overview



Cancer Cell Transcriptional Network

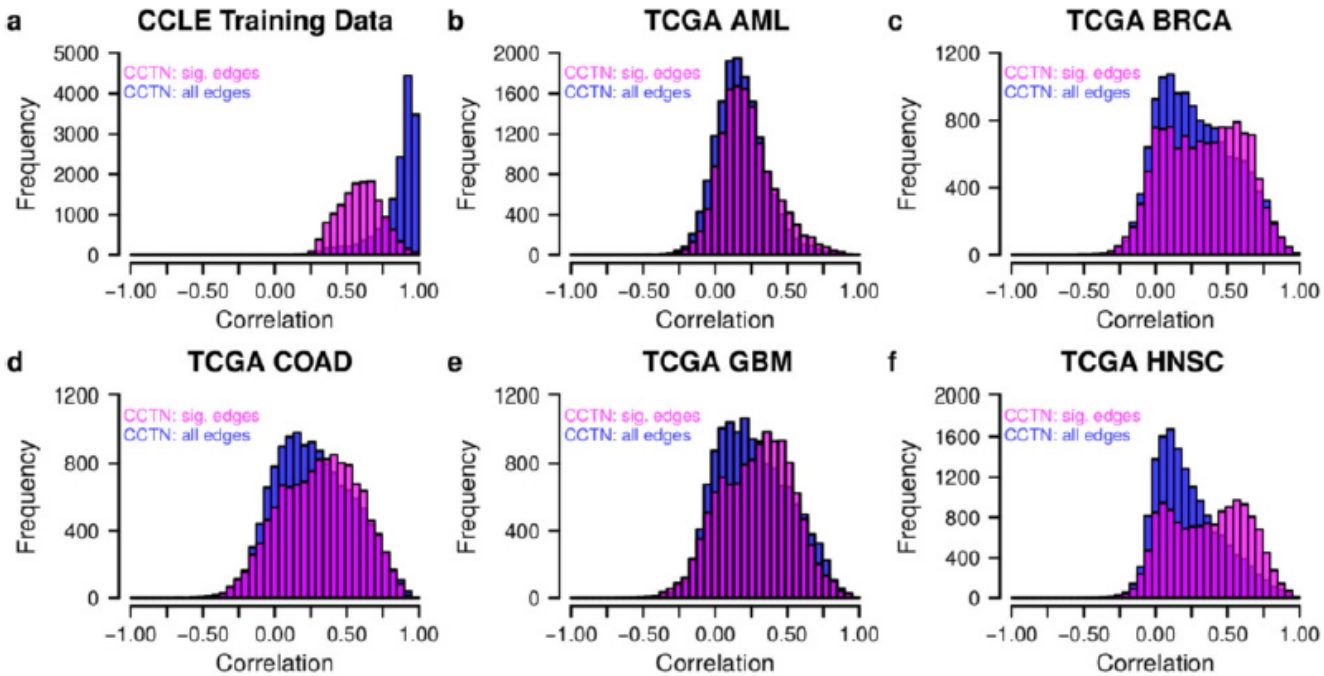
- **Data:** 768 cell lines, 15,942 genes (gene level copy number and expression data)
- For each gene, use LASSO to select regulator genes:

$$\boldsymbol{a}_i^* = \arg\ \min_{\boldsymbol{a}} \sum_{d=1}^{D} \left(e_{id} - \left(a_{ii} c_{id} + \sum_{j \neq i} a_{ji} e_{jd} \right) \right)^2 + \lambda_i \sum_{j=1}^{N} |a_{ji}|$$

  where D is the number of cell lines and N is the number of genes.
- Those with non-zero and significant coefficients are regarded as regulators.

Summary and validation of CCTN

- The use of cell lines circumvented the variation in tumour cell purity between tumor samples.
- Validated in tumor samples:



Impact of CNAs on survival signatures

By far,

- Define Cancer Cell Transcriptional Network;
- Obtain an $\boldsymbol{a}_i^*$ for each gene.

To test the impact of individual gene CNAs on survival signatures:

- Identification of tumour-specific survival signature genes;
- Quantify the contribution of each gene's **copy number state** on the expression of all genes.
- Impacts of individual gene CNAs on these signature genes.

Identification of survival signature genes

Determine genes whose expression levels are predictive of patient survival:

- **Random forest**: one of the best performing methods for the prediction of patient survival based on gene expression data.
- **Top-ranking genes** in two groups: positively and negatively correlated with survival. (8 to 199 for different tumours)
- The correlation of individual gene expression levels with survival was weak, thus underlining the need to consider multiple marker genes in combination to obtain significant predictions of patient survival.

Impacts of individual regulators

For each gene $i \in \{1, \cdots, N\}$ in a cohort with D patients,

- $\boldsymbol{a}_i$ from CCTN;
- r_i: correlation between predicted value and observed expression value;
- $R^2 = r_i \cdot r_i$: goodness of fit.
- Regulator gene j's direct contribution:

$$p_{ji} = \frac{1}{D} \sum_{d=1}^{D} \frac{|a_{ji} e_{jd}|}{|a_{ii} c_{id}| + \sum_{v \neq i} |a_{vi} e_{vd}|}$$

- CNA contribution:

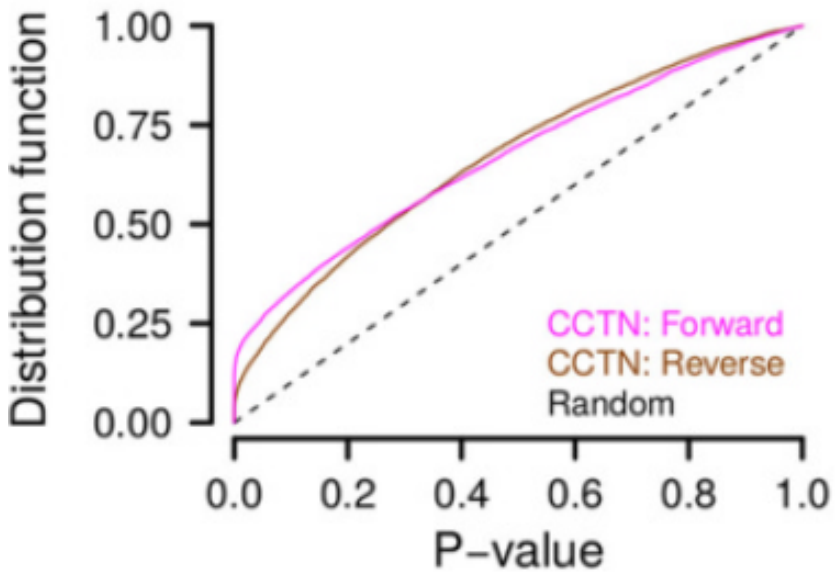
$$p_{ii} = \frac{1}{D} \sum_{d=1}^{D} \frac{|a_{ii} c_{id}|}{|a_{ii} c_{id}| + \sum_{v \neq i} |a_{vi} e_{vd}|}$$

- Network propagation: (cis and trans acting?)

$$F = (f_{ji})_{1 \le i,j \le N} := p_{ji} \cdot R_i^2$$

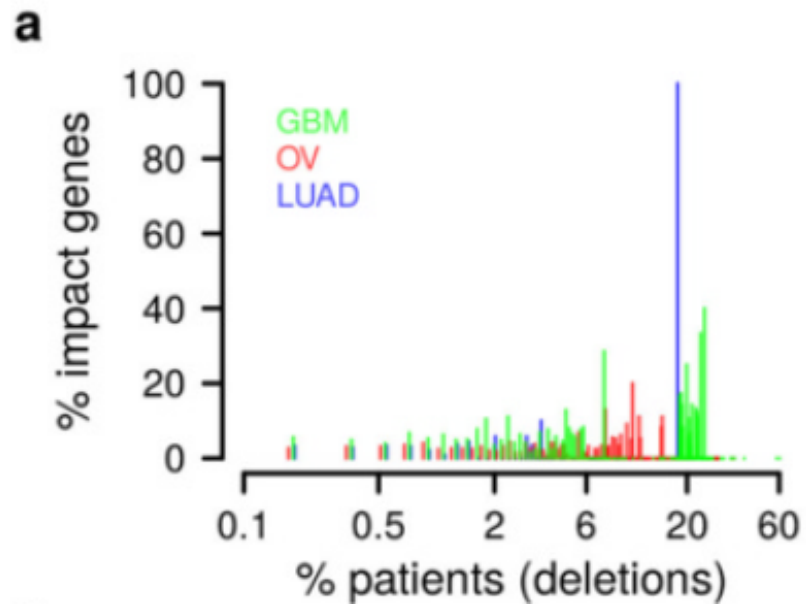
Split explained variance into average proportions according to the direct contribution.

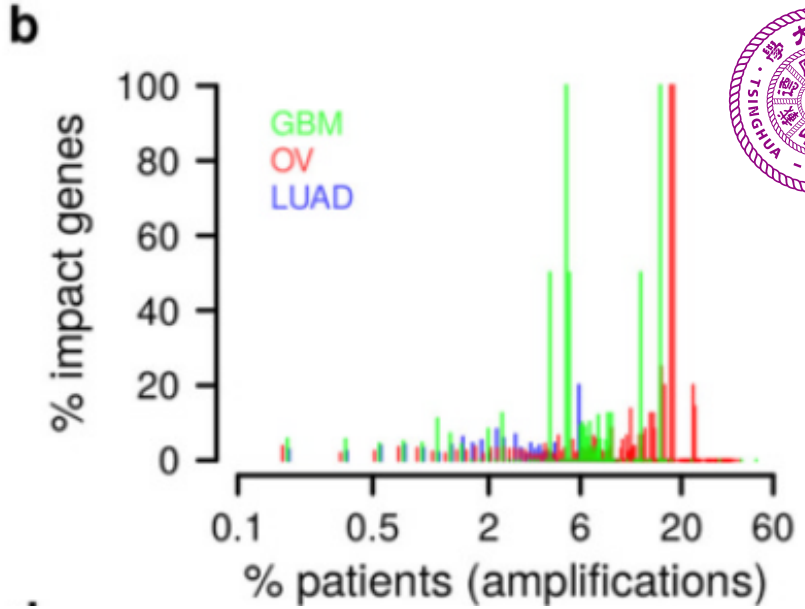
$$F^* = \sum_{k=1}^{\infty} F^k$$

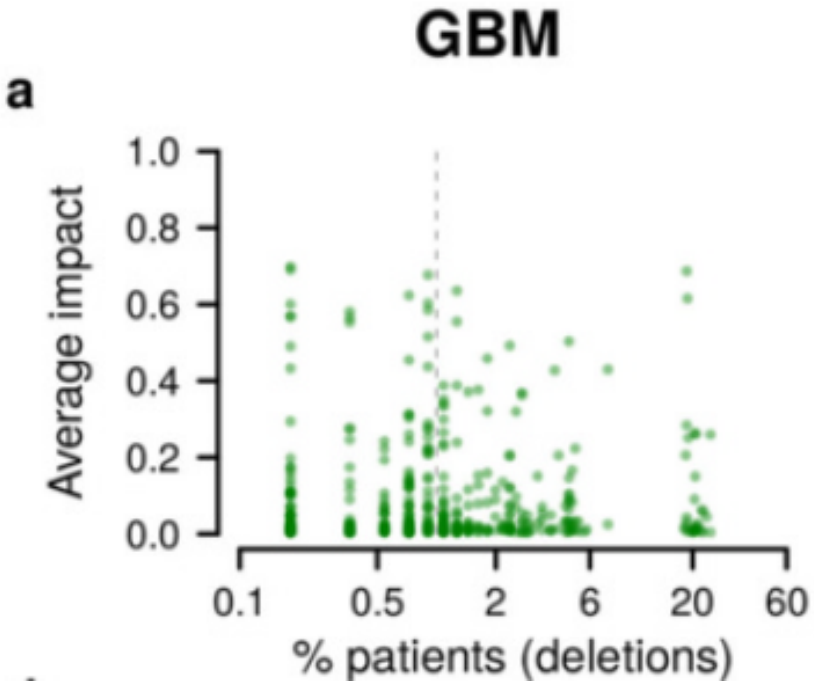


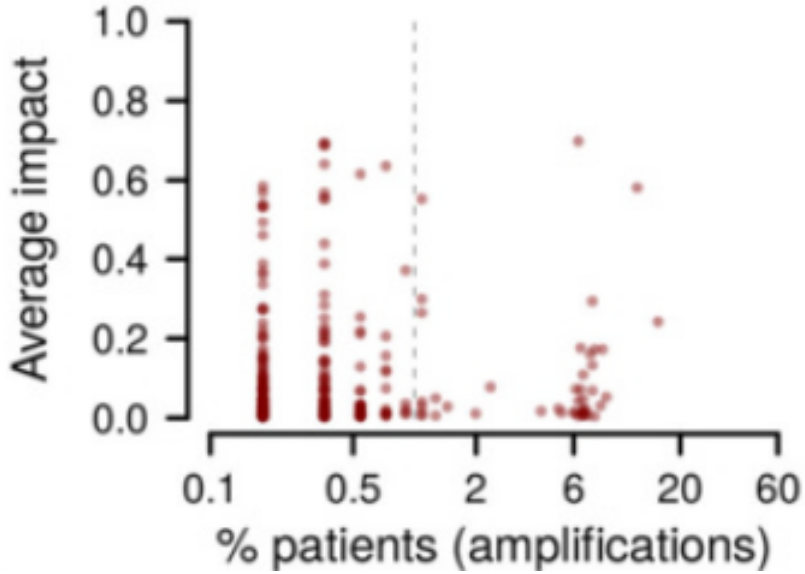
Impact of CNAs on survival signatures

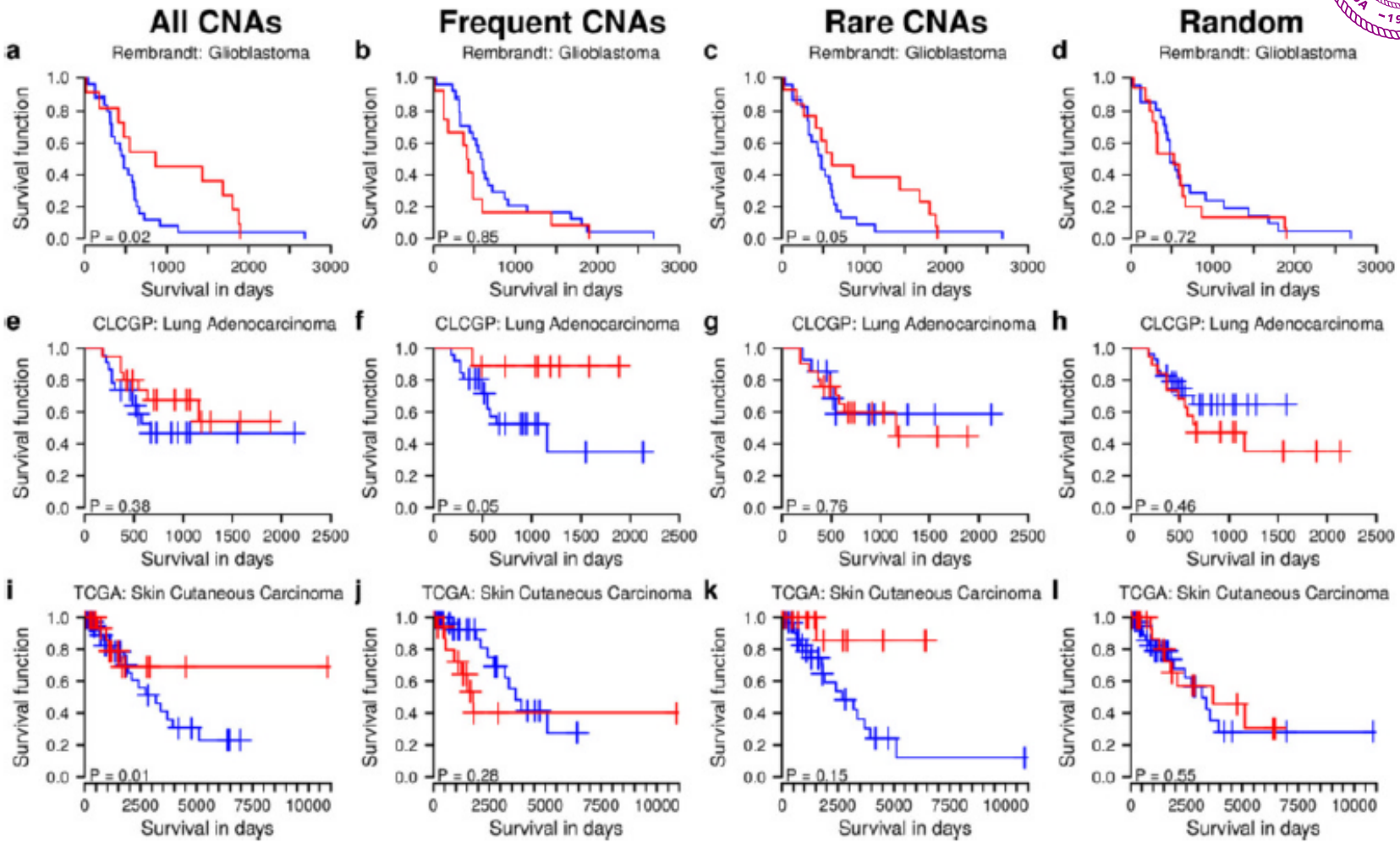
For each gene CNA, average its correponding impact scores across all signature genes, yielding a single impact score that quantifies the contribution of this specifci gene CNA on the expression of all survival signature gens.


a
% impact genes
100
80
60
40
20
0
GBM
OV
LUAD
0.1
0.5
2
6
20
60
% patients (deletions)


b
% impact genes
100
80
60
40
20
0
GBM
OV
LUAD
0.1
0.5
2
6
20
60
% patients (amplifications)


GBM
a
Average impact
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.5
2
6
20
60
% patients (deletions)


Average impact
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.5
2
6
20
60
% patients (amplifications)


All CNAs
Frequent CNAs
Rare CNAs
Random
a
Rembrandt: Glioblastoma
P = 0.02
b
Rembrandt: Glioblastoma
P = 0.85
c
Rembrandt: Glioblastoma
P = 0.05
d
Rembrandt: Glioblastoma
P = 0.72
e
CLCGP: Lung Adenocarcinoma
P = 0.38
f
CLCGP: Lung Adenocarcinoma
P = 0.05
g
CLCGP: Lung Adenocarcinoma
P = 0.76
h
CLCGP: Lung Adenocarcinoma
P = 0.46
i
TCGA: Skin Cutaneous Carcinoma
P = 0.01
j
TCGA: Skin Cutaneous Carcinoma
P = 0.28
k
TCGA: Skin Cutaneous Carcinoma
P = 0.15
l
TCGA: Skin Cutaneous Carcinoma
P = 0.55
Survival function
Survival in days

Summary

- The space of possible mutational patterns affecting the aggressiveness of a tumour is practically unlimited.
- Expression variation of individual regulators changes the activity of molecular subnetworks, while the topology of regulatory relationship turns out to be remarkably robust across cell types.
- Rare variants can be as important as frequent variants.

整合多批次、多中心的组学数据的癌症分类方法

背景

- 基于组学的癌症研究在临床诊疗中取得进展有限；
- 很多方法与发现的可重复性较差。

- 考虑多批次、多中心的组学数据的异质性与一致性；
- 组学数据与机理模型、先验信息的整合。

现有方法综述

- **策略一:**

  将所有数据集组成一个组合矩阵

  关键：“消除”批次影响 (batch effect)

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

- **策略二:**

  比较不同数据集得到的“类”的“相似性”

  关键：相似性计算是基于数据集得到的信号特征 **(signatures)**

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

关键问题：如何表示不同批次数据集共享的信息?

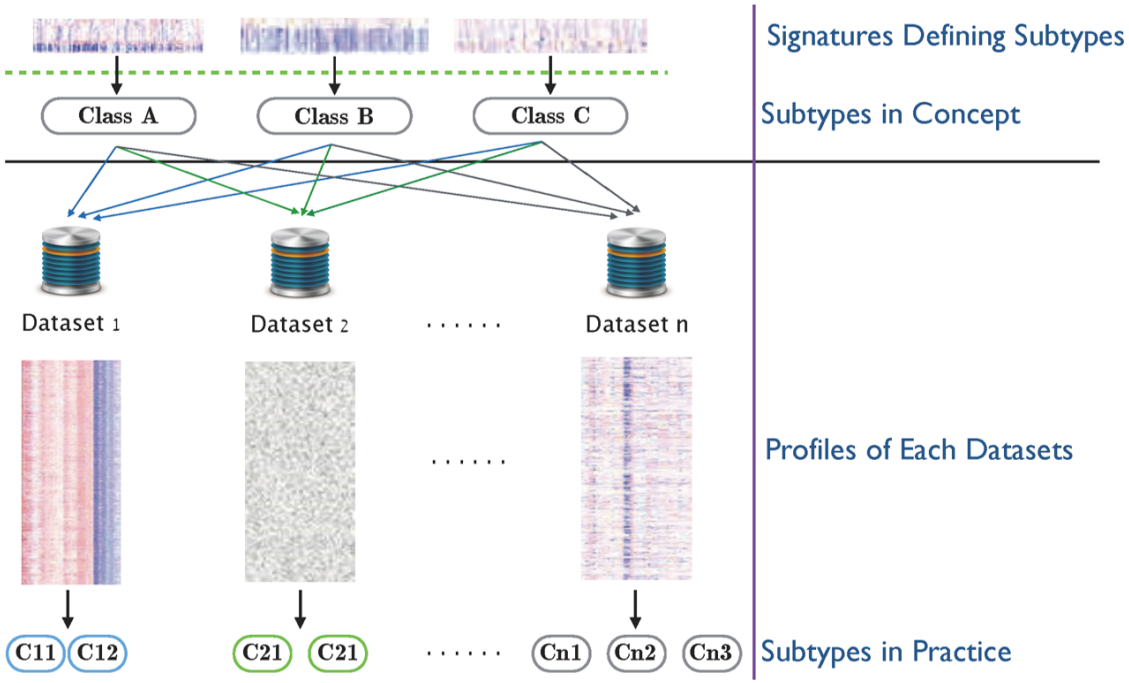


Figure: 模型基本构想

假设： 信号模块以及其扰动程度决定了癌症的子类型。

初步方法与结果 (1)

1. **单数据集信号模块：** 对单数据集病人进行聚类，通过差异表达分析寻找每类的代表性基因;
2. **跨数据集信号模块：** 结合文献挖掘基因和各个数据集各自找到的代表性基因，在 *PPI* 网络上进行网络扩散与聚类。

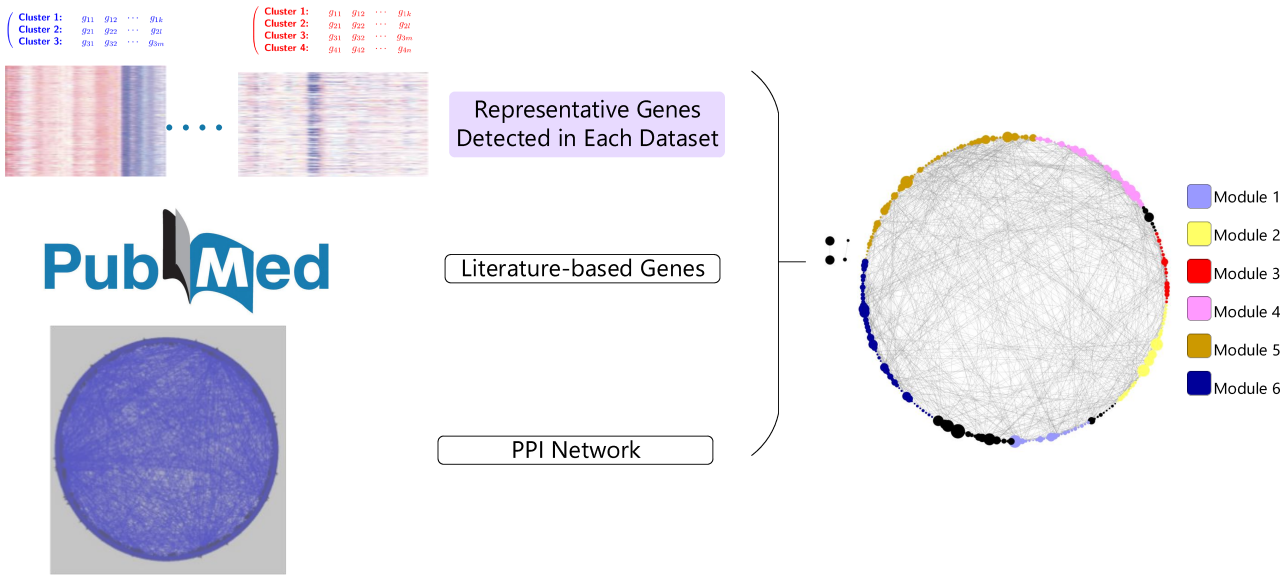


Figure: 整合多数聚集定义信号模块

Figure: 使用 DAVID(Huang DW, et al., *Nature Protoc.*, 2009)对信号模块进行功能注释

初步方法与结果 (2)

1. **定义单个数据集各类在功能模块上的扰动程度：** 计算每类的所有样本基因表达量的中位数组成该类的表示向量，计算该向量在功能模块上富集的 z—score；
2. **定义跨数据集的聚类中心：** 对每个数据集每类在各功能模块富集的 z—score 向量进行聚类。

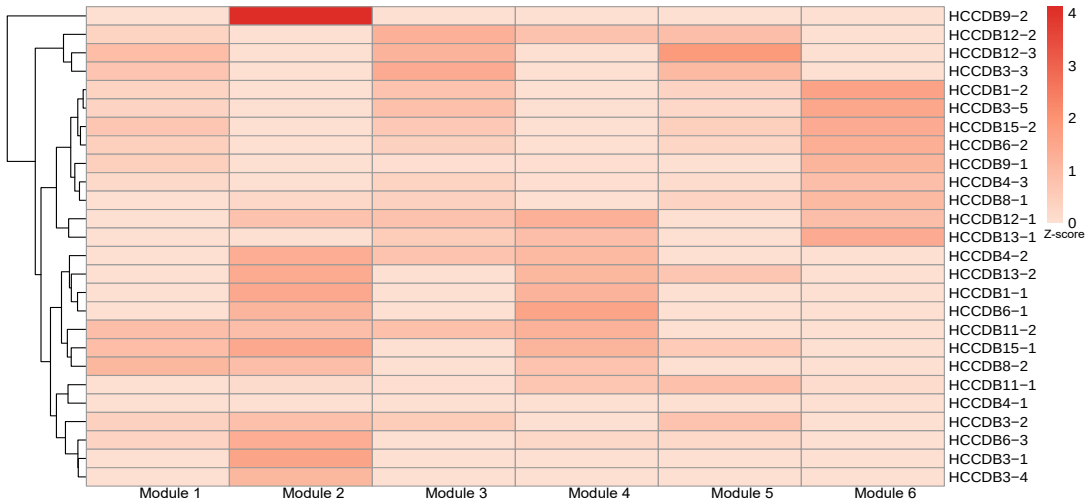


Figure: 原始分类在各功能模块上的扰动程度。

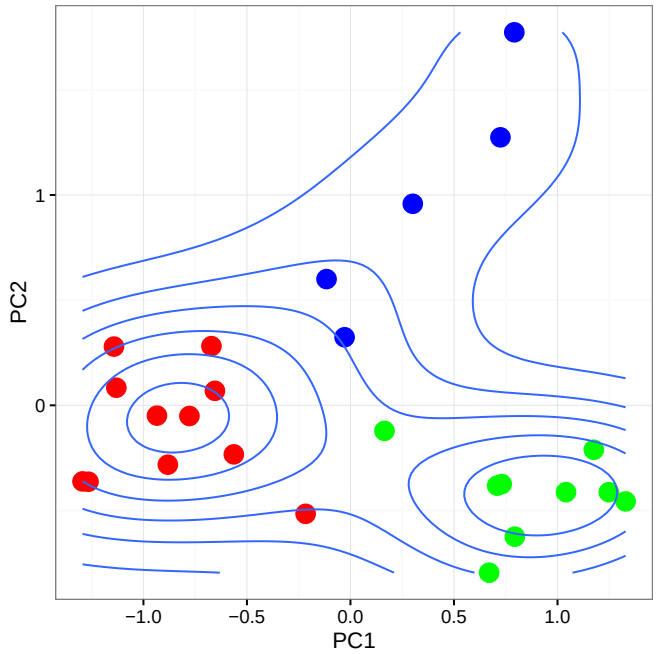


Figure: 各个数据集各类扰动向量的主成分分析结果。

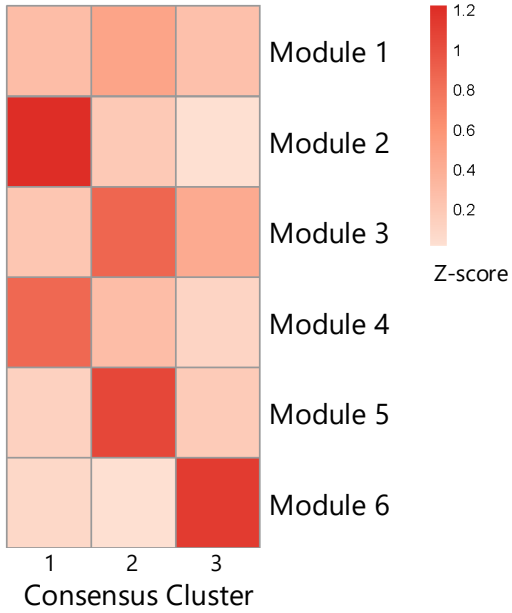


Figure: 对扰动向量进行聚类，得到的类中心。

初步方法与结果 (3)

- **单样本的重分类：**计算每个样本的扰动向量，将其划归到每个数据集最近的一类中。

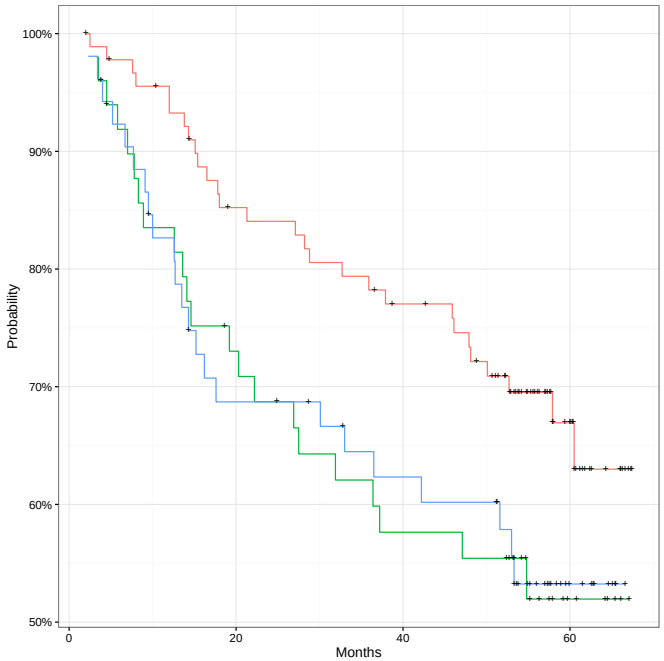


Figure: 单数据集聚类的生存期分析结果 (HCCDB6). cox: 0.05

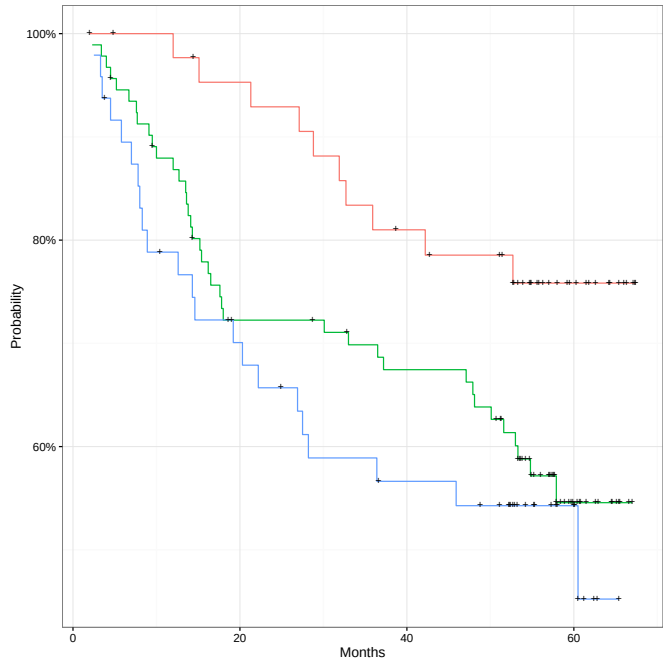


Figure: 重新分类以后的生存期分析结果 (HCCDB6). cox: 0.008

总结

- 提出一种基于多批次多中心的数据集对单一肿瘤进行联合聚类的方法。
- 我们的方法试图:
  - 获得多个数据集稳定的聚类结果;
  - 尽可能的整合先验信息.
- 核心是基于各类病人的特征信息在多数据集中是共享的。

下一步:

- 目前模型仅是初步结果，需要完善;
- 在多种癌症数据集上验证该模型，并与现有方法进行比较;
- 解决小样本新数据集的可靠聚类问题;
- 多组学数据支撑聚类结果。

基于“生物网络”的概念 -观测双层结构