TCGA数据分析

作者邮箱：2740881706@qq.com

主要内容

- TCGA数据库简介
- 下载整理TCGA数据
- 差异表达
- 共表达网络
- 生存分析

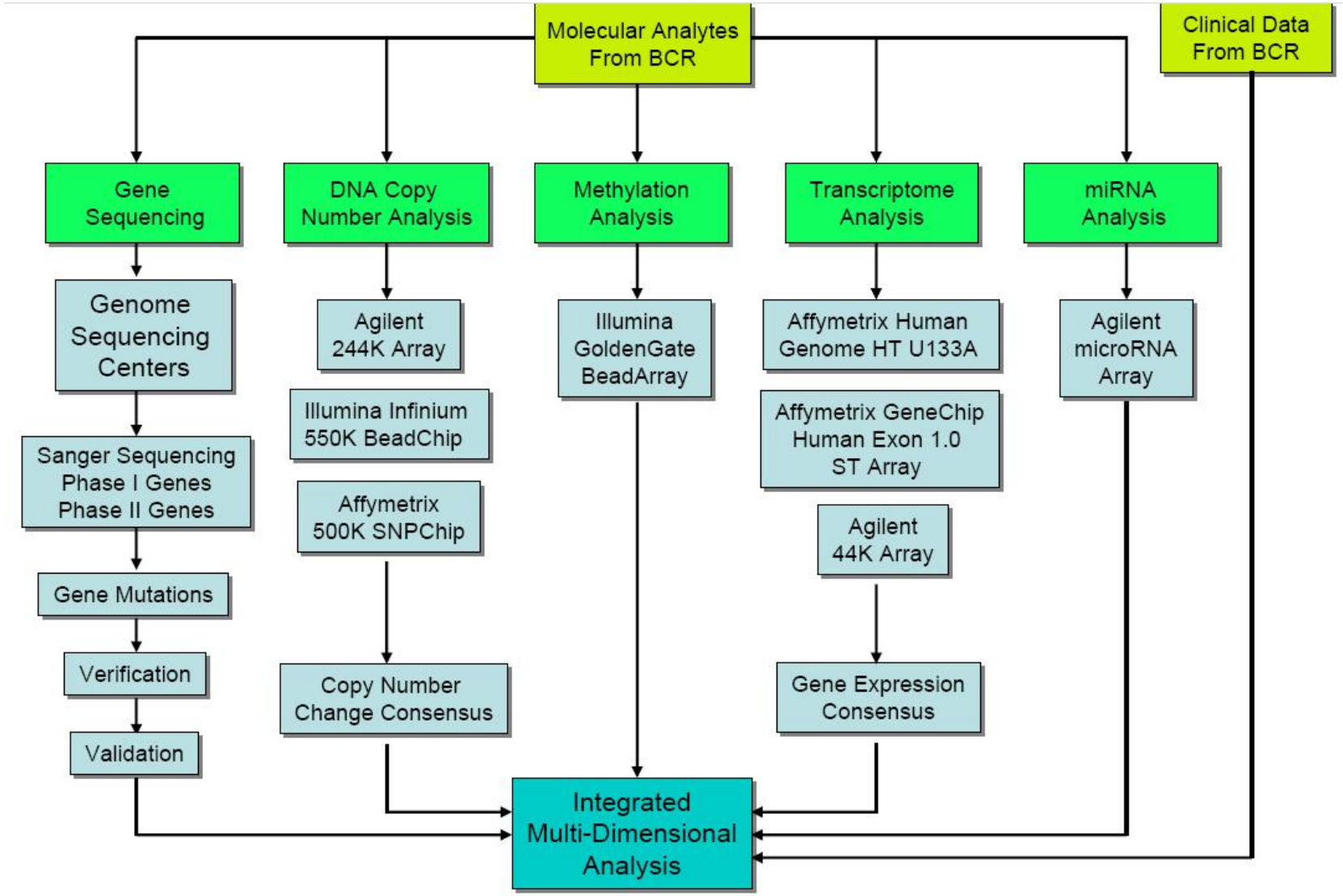
什么是TCGA?

- 美国政府发起的癌症和肿瘤基因图谱（Cancer Genome Atlas，TCGA）计划，试图通过应用基因组分析技术，特别是采用大规模的基因组测序，将人类全部癌症（近期目标为50种包括亚型在内的肿瘤）的基因组变异图谱绘制出来，并进行系统分析，旨在找到所有致癌和抑癌基因的微小变异，了解癌细胞发生、发展的机制，在此基础上取得新的诊断和治疗方法，最后可以勾画出整个新型“预防癌症的策略”。

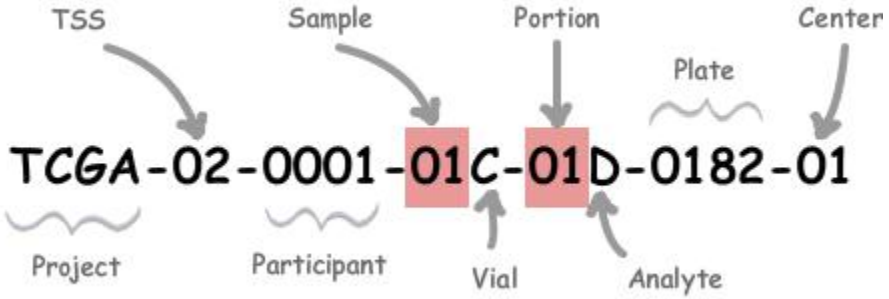
- TCGA 使命:提高人们对癌症发病分子基础的科学认识及提高我们诊断、治疗和预防癌症的能力

- TCGA 目标:完成一套完整的与所有癌症基因组改变相关的“图谱”

基因组图谱



The Cancer Genome Atlas Research Network, Nature, 2008



Label	Identifier for	Value	Value description	Possible values
Vial	Order of sample in a sequence of samples	C	The third vial	A to Z
TSS	Tissue source site	02	GBM (brain tumor) sample from MD Anderson	See Code Tables Report
Sample	Sample type	[illegible]	[illegible]	Tumor types range from 01 - 09, normal types from 10 - 19 and control samples from 20 - 29. See Code Tables Report for a complete list of sample codes
Project	Project name	TCGA	TCGA project	TCGA
Portion	Order of portion in a sequence of 100 - 120 mg sample portions	01	The first portion of the sample	01-99
Plate	Order of plate in a sequence of 96-well plates	0182	The 182nd plate	4-digit alphanumeric value
Participant	Study participant	0001	The first participant from MD Anderson for GBM study	Any alpha-numeric value
Center	Sequencing or characterization center that will receive the aliquot for analysis	01	The Broad Institute GCC	See Code Tables Report

主要内容

- TCGA数据库简介
- 下载整理TCGA数据
- 差异表达
- 共表达网络
- 生存分析

下载TCGA数据

- 网页下载
- 命令行下载

UQ--upper-quartile normalization (EDASeq)

整理TCGA数据

- 分散文件整理成矩阵
- id转换成gene symbol
- 生存数据

主要内容

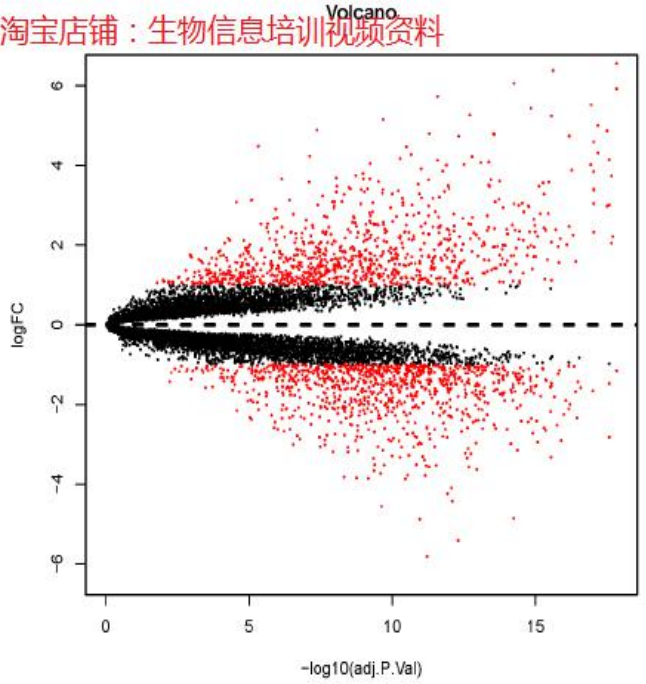
- TCGA数据库简介
- 下载整理TCGA数据
- 差异表达
- 共表达网络
- 生存分析

差异分析

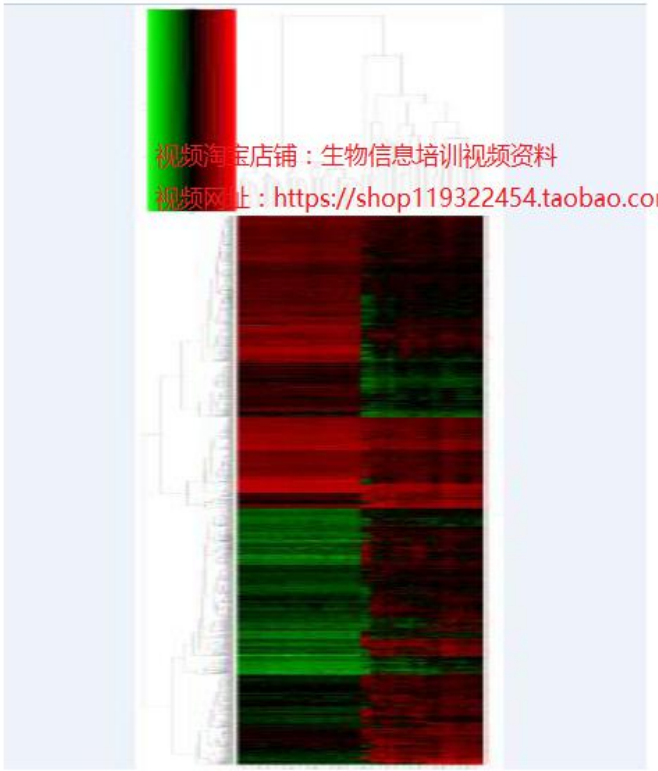
gene	logFC	logCPM	PValue	FDR
CD5L	-10.07233415	1.061156098	2.54E-91	8.13E-87
CTD-2561B21.5	-6.101725443	-1.810768595	1.71E-86	2.73E-82
RP11-414J4.2	-8.096655852	-2.15498115	5.25E-83	5.60E-79
LMAN1L	-7.542295697	-3.431747713	7.43E-54	5.94E-50
FAM9C	-5.558207575	-3.309251405	1.07E-35	6.88E-32
SPIC	-6.054372154	-3.102244201	1.93E-27	1.03E-23
KCNT1	-3.427448565	-0.640116704	3.52E-25	1.61E-21
RP11-752D24.2	-4.258163522	-2.66550925	1.54E-23	6.17E-20
STAB2	-4.488377941	0.889271681	1.98E-22	7.06E-19
CD160	-2.94084641	-0.984586321	8.56E-21	2.74E-17

差异分析

火山图



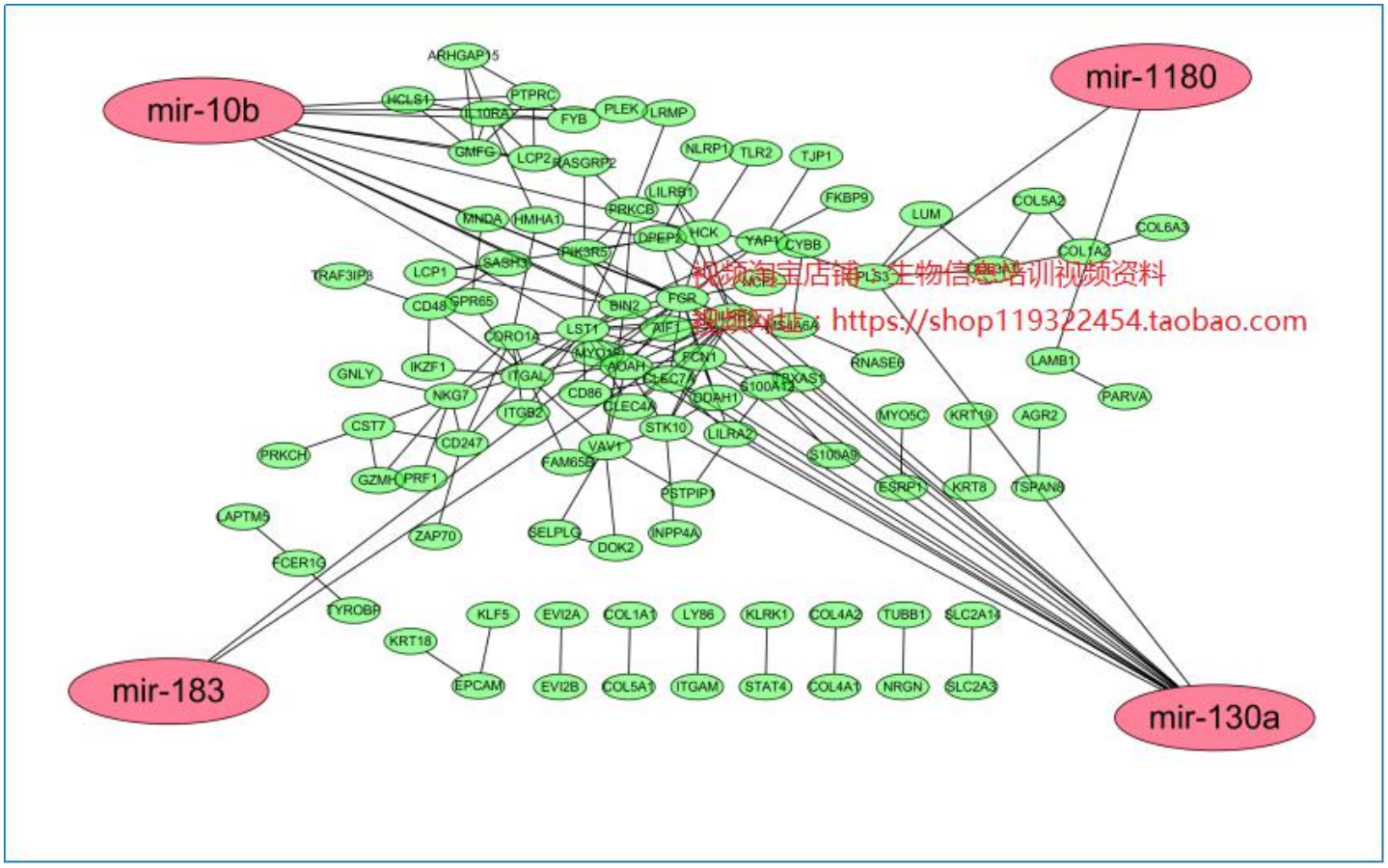
热图



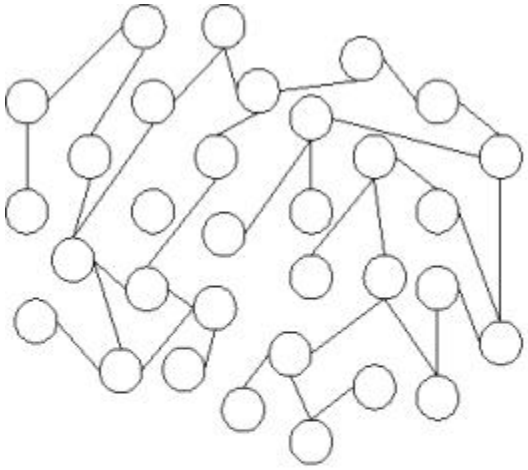
主要内容

- TCGA数据库简介
- 下载整理TCGA数据
- 差异表达
- 共表达网络
- 生存分析

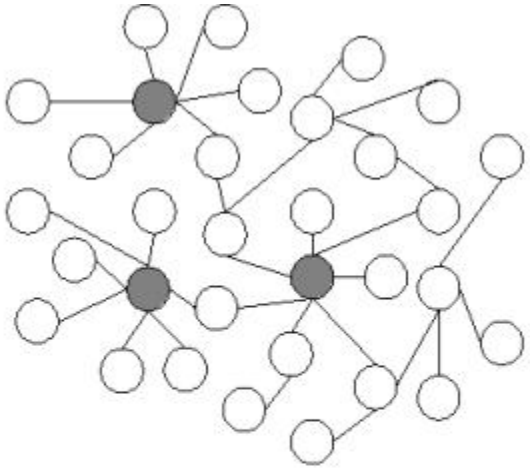
共表达网络


mir-10b
mir-1180
mir-183
mir-130a
ARHGAP15
HCLS1
PTPRC
IL10RA
FYB
PLEK
LRMP
GMFG
LCP2
RASGRP2
NLRP1
TLR2
TJP1
LILRB1
FKBP9
PRKCB
MNDA
HMHA1
DPEP2
HCK
YAP1
CYBB
LUM
COL5A2
COL6A3
COL1A2
PIK3R5
SASH3
LCP1
TRAF3IP3
CD48
GPR65
BIN2
FGR
AIF1
LST1
CORO1A
MYO1F
FCN1
AOAH
CLEC7A
GNLY
IKZF1
ITGAL
NKG7
CD86
CLEC4A
ITGB2
CST7
PRKCH
CD247
VAV1
FAM65B
GZMH
PRF1
STK10
LILRA2
PSTPIP1
SELPLG
DOK2
NPP4A
ZAP70
LAPTM5
FCER1G
TYROBP
KLF5
EVI2A
COL1A1
LY86
KLRK1
COL4A2
TUBB1
SLC2A14
KRT18
EPCAM
EVI2B
COL5A1
ITGAM
STAT4
COL4A1
NRGN
SLC2A3
RNASE6
LAMB1
PARVA
MYO5C
KRT19
AGR2
S100A9
ESRP1
KRT8
TSPAN8
视频淘宝店铺：生物信息培训视频资料
视频网址：https://shop119322454.taobao.com

一种random network，即每一个节点的度相对平均。然而第二种图，即scale-free network才是一种更稳定的选择。Scale-free network具有这样的特点，即存在少数节点具有明显高于一般点的度，这些点被称为hub。



(a) Random network



(b) Scale-free network

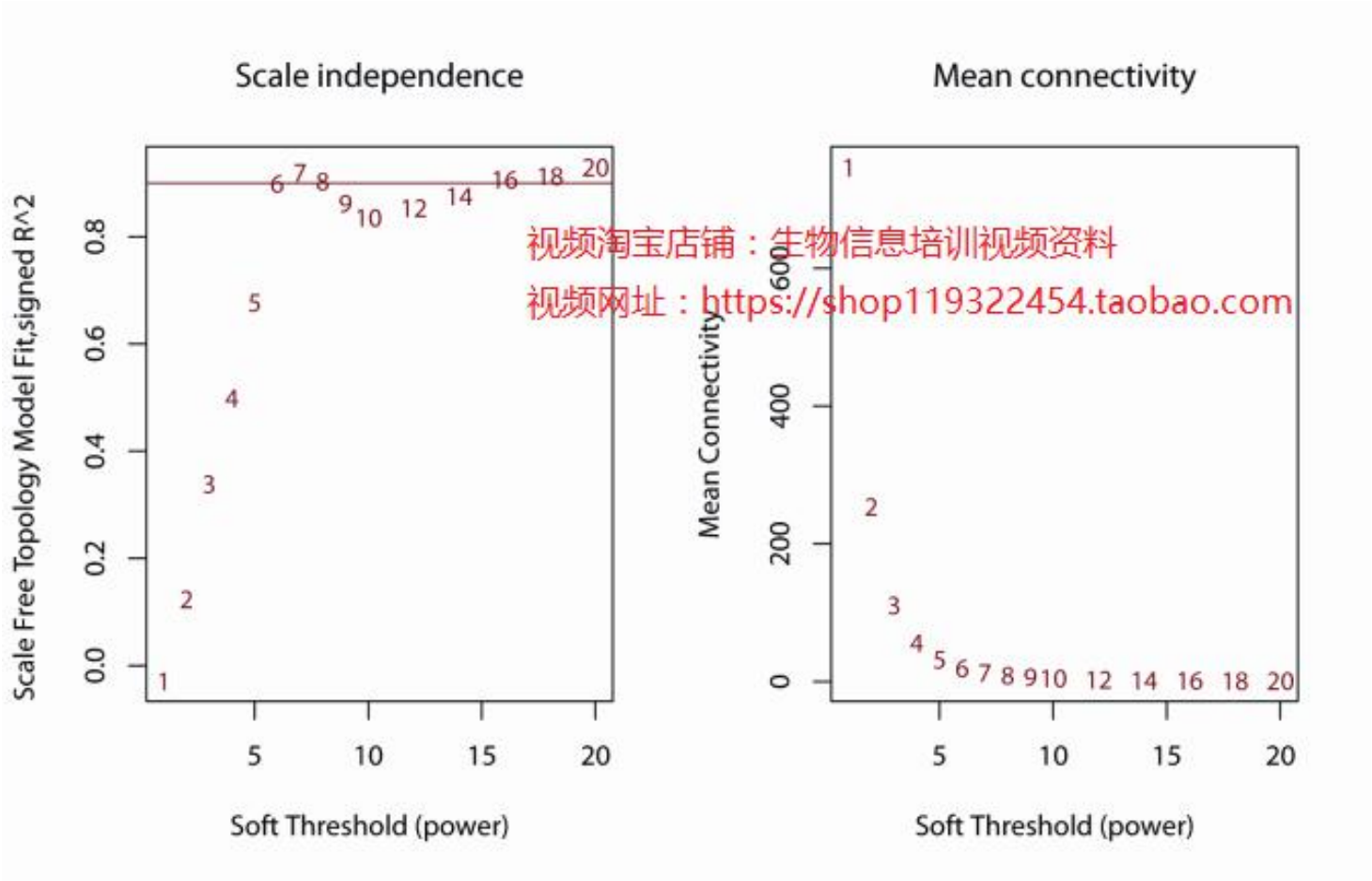
Correlation and distance are transformed as follows:
for type = "unsigned", adjacency = |cor|^power;

for type = "signed", adjacency = (0.5 * (1+cor))^power;

for type = "signed hybrid",adjacency = cor^power if cor>0 and 0 otherwise;

and for type = "distance", adjacency = (1-(dist/max(dist))^2)^power.

power选择



主要内容

- TCGA数据库简介
- 下载整理TCGA数据
- 差异表达
- 共表达网络
- 生存分析

生存分析

- 所谓生存期（survival time）是指从某个标准时刻(如发病,确诊,开始治疗或进行手术的时间)算起至死亡或复发为止的时间。

生存函数

- 一. 生存率(Survival Rate)
- 又称为生存概率或生存函数,它表示一个病人的生存时间长于时间t的概率,用S(t) 表示： s(t)=P（T≥t）
- 如5年生存率： s(5)=P（T≥5）
- 以时间t为横坐标,S(t)为纵坐标所作的曲线称为生存率曲线,它是一条下降的曲线,下降的坡度越陡,表示生存率越低或生存时间越短,其斜率表示死亡速率。

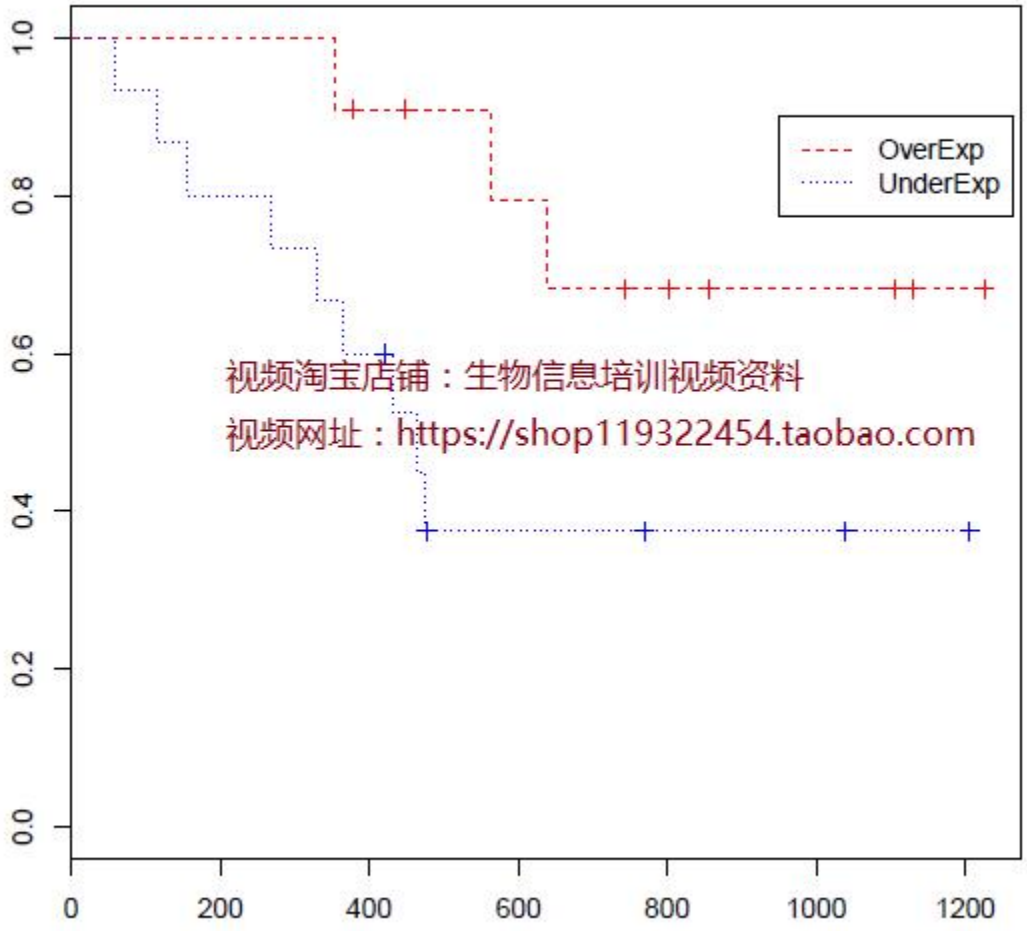
"survival" command

```
>library(survival)
>rt=read.table("ovarian.txt",header=T,sep="\t")
>fit <- survfit(Surv(futime, fustat) ~ expression, data = rt)
>plot(fit, lty = 2:3,col=c("red","blue"))
>legend(950, .9, c("OverExp", "UnderExp"), lty = 2:3, col=c("red","blue"))
>survdiff(Surv(futime, fustat) ~ expression,data=rt)
```

•"survival" input

sample	futime	fustat	expression
TCGA-DY-A1DE-01A	3932	0	1620.731512
TCGA-DY-A1DF-01A	734	1	620.7519135
TCGA-DY-A0XA-01A	3846	0	1797.340423
TCGA-DY-A1H8-01A	992	1	1028.026424
TCGA-F5-6864-01A	379	0	871.0717837
TCGA-F5-6863-01A	361	1	510.8608189
TCGA-F5-6861-01A	1160	0	1681.165085
TCGA-F5-6814-01A	1131	0	903.6017817
TCGA-EI-7004-01A	257	0	1064.617259
TCGA-EI-7002-01A	364	0	1598.569694
TCGA-EI-6917-01A	531	0	2163.575894

"survival" result



Thanks!!!

作者邮箱：2740881706@qq.com