大数据医学

Zhang et al. JTB 2013 ; Zhang et al. JMCB 2015
Liu, et al. NAR 2016 ; Zhao et al. PNAS 2016

分子标志物

揭示表达暗物质

动态网络标志物

中国科学院系统生物学重点实验室
陈洛南

Zhang et al. JTB 2013 ; Zhang et al. JMCB 2015; Liu et al. NAR 2016; Zhao et al. PNAS 2016

分子标志物 → occurred disease

网络标志物 → occurred complex disease

动态网络标志物 → un-occurred disease

Characterizing one sample

One sample →

- ✔ Sequence
- ✔ Gene expression
- ✔ Protein expression
- ✔ Metabolomics
- ✔ Methylation

Unstable feature
不稳定

Multiple samples → ✘ **Network**

Stable feature
稳定

Clinic data or Single-cell data

One sample or a few samples for an individual

Large samples for population

Small samples for individuals

Single-cell data

Question ?

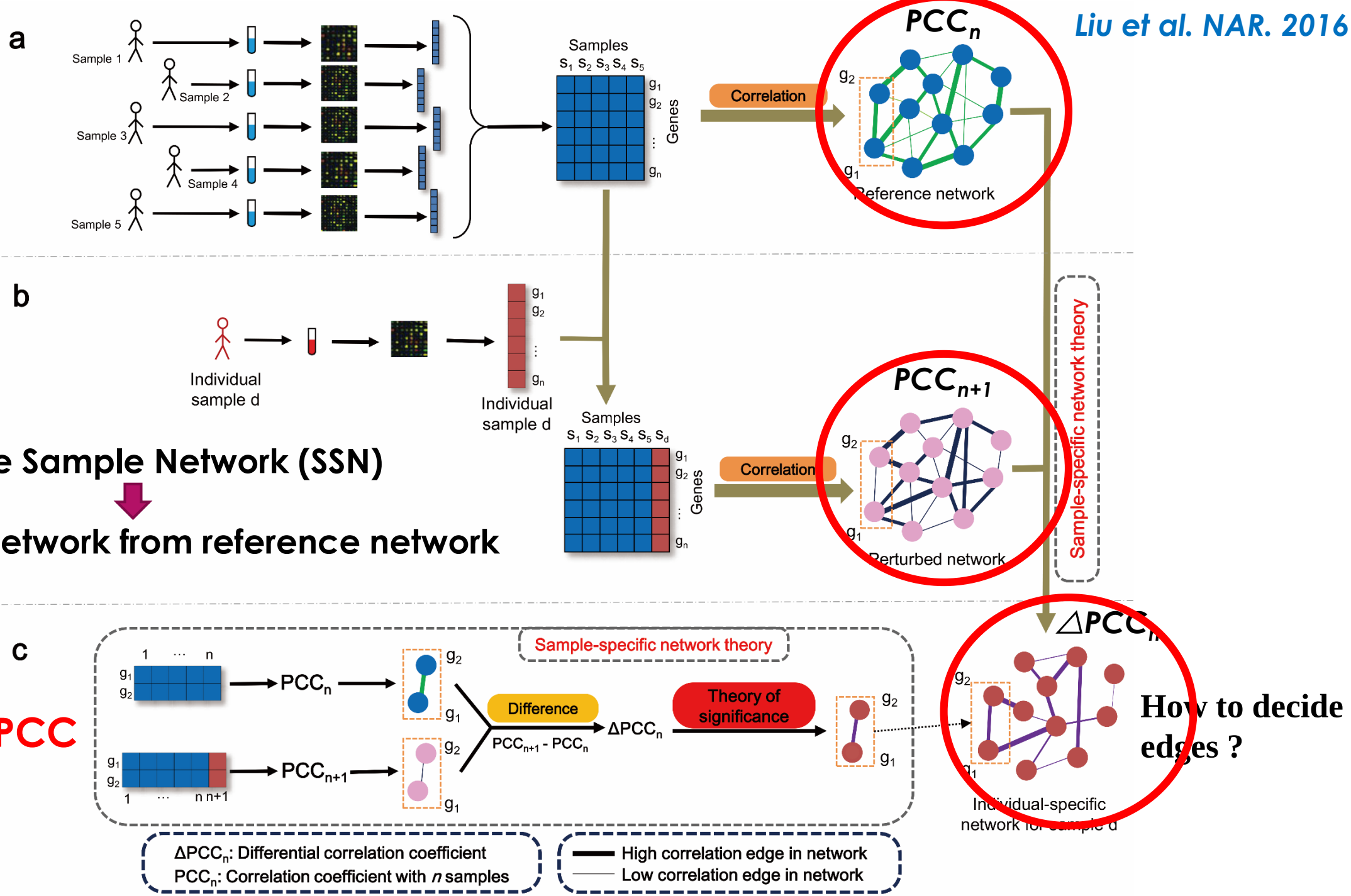
- Can we construct a network by a single sample data?

YES

Single Sample Network (SSN)

Liu et al. NAR. 2016

PCC



Single Sample Network (SSN)

Differential network from reference network

Differential PCC

How to decide edges ?

单样本网络理论

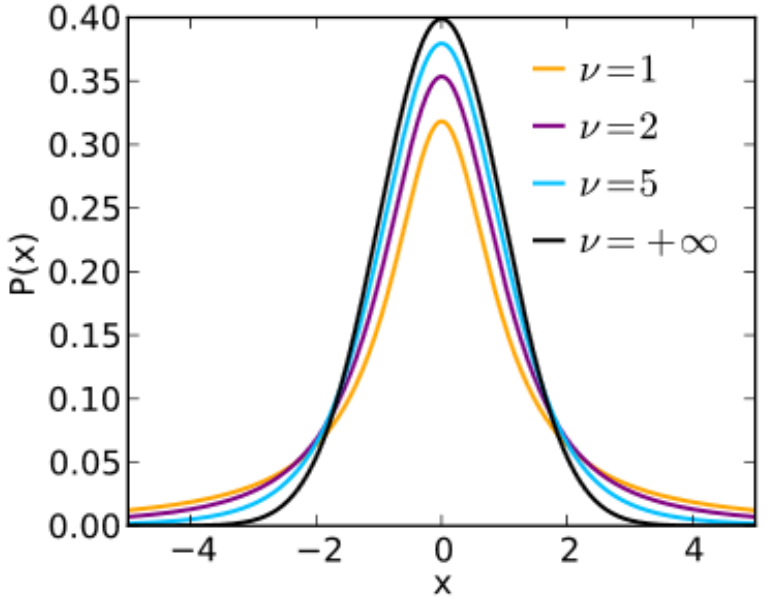
一个样本 → 一个网络

Statistic of PCC

$\boldsymbol{PCC_n =}$ $R_n = \frac{1}{n}\sum(\frac{x-\mu_x}{\sigma_x})(\frac{y-\mu_y}{\sigma_y})$

$$\mu_{PCC} = E(R_n) = 0$$

$$\sigma_{PCC} = \sqrt{E(R_n^2)} = \sqrt{\frac{1-R^2}{n-2}}$$



PCC distribution

PCC_n follows Student T_{n-2} distribution

As n → ∞, T_{n-2} → Gaussian Distribution

Differential PCC

- **Differential PCC distribution between n samples and n+1 samples** with n common samples.

$$\triangle PCC_n = PCC_{n+1} - PCC_n \text{ follows } ? \text{ distribution}$$

T_{n-1} distribution

T_{n-2} distribution

Distribution of differential PCC

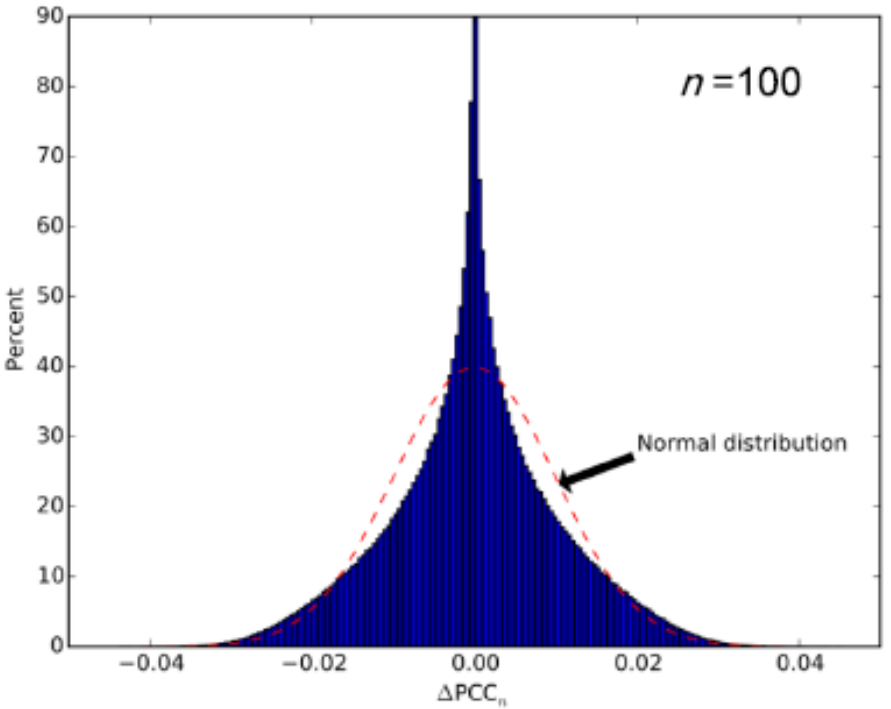
Thus, the mean and standard deviation of the differential *PCC*

$$\mu_{\Delta PCC} = E(\Delta R_n) = 0$$

$$\sigma_{\Delta PCC} = \sqrt{E(\Delta R_n^2)} = \frac{1}{n-1}(1-R_n^2)$$

We can use Z statistic to quantify differential PCC

$$v = \frac{\Delta R_n - \mu_{\Delta PCC}}{\sigma_{\Delta PCC} / m}$$



Volcano distribution

Eiffel Tower Distribution
埃菲尔铁塔 分布

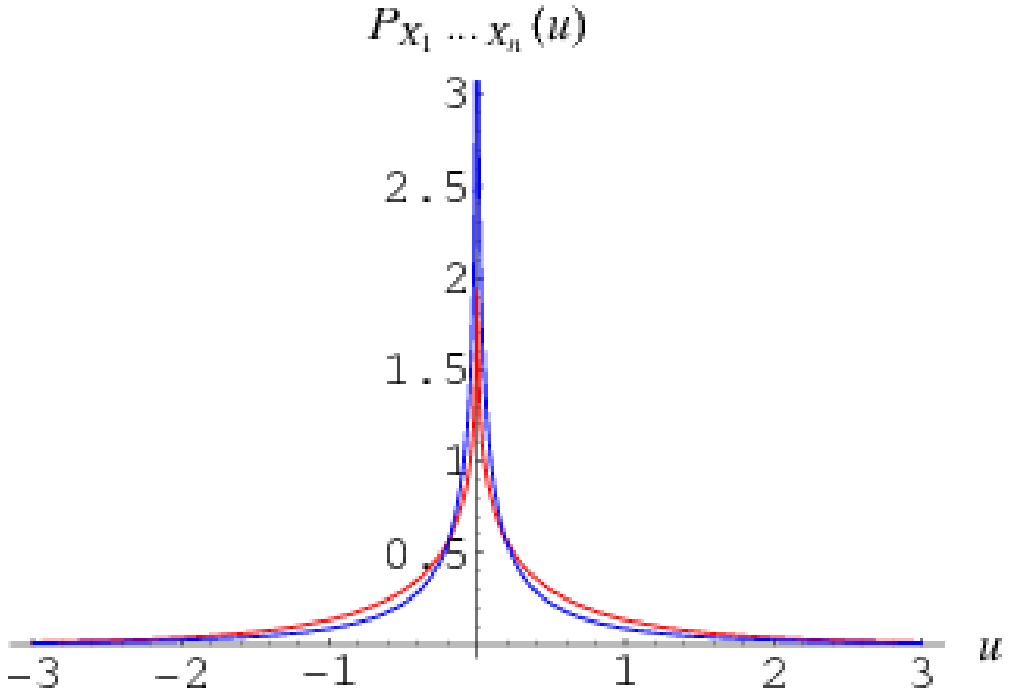
Theoretically, when n → infinite (Gaussian assumption),
Volcano distribution approaches normal product distribution with the correlation *PCCn*

Normal Product Distribution

The distribution of a product u of two normally distributed variates x and y with zero means and variances σ_x^2 and σ_y^2 without the correlation is given by

$$P_{X\,Y}(u) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{e^{-x^2/(2\sigma_x^2)}}{\sigma_x\sqrt{2\pi}} \frac{e^{-y^2/(2\sigma_y^2)}}{\sigma_y\sqrt{2\pi}} \delta(x\,y-u)\,dx\,dy$$

where $\delta(x)$ is a delta function



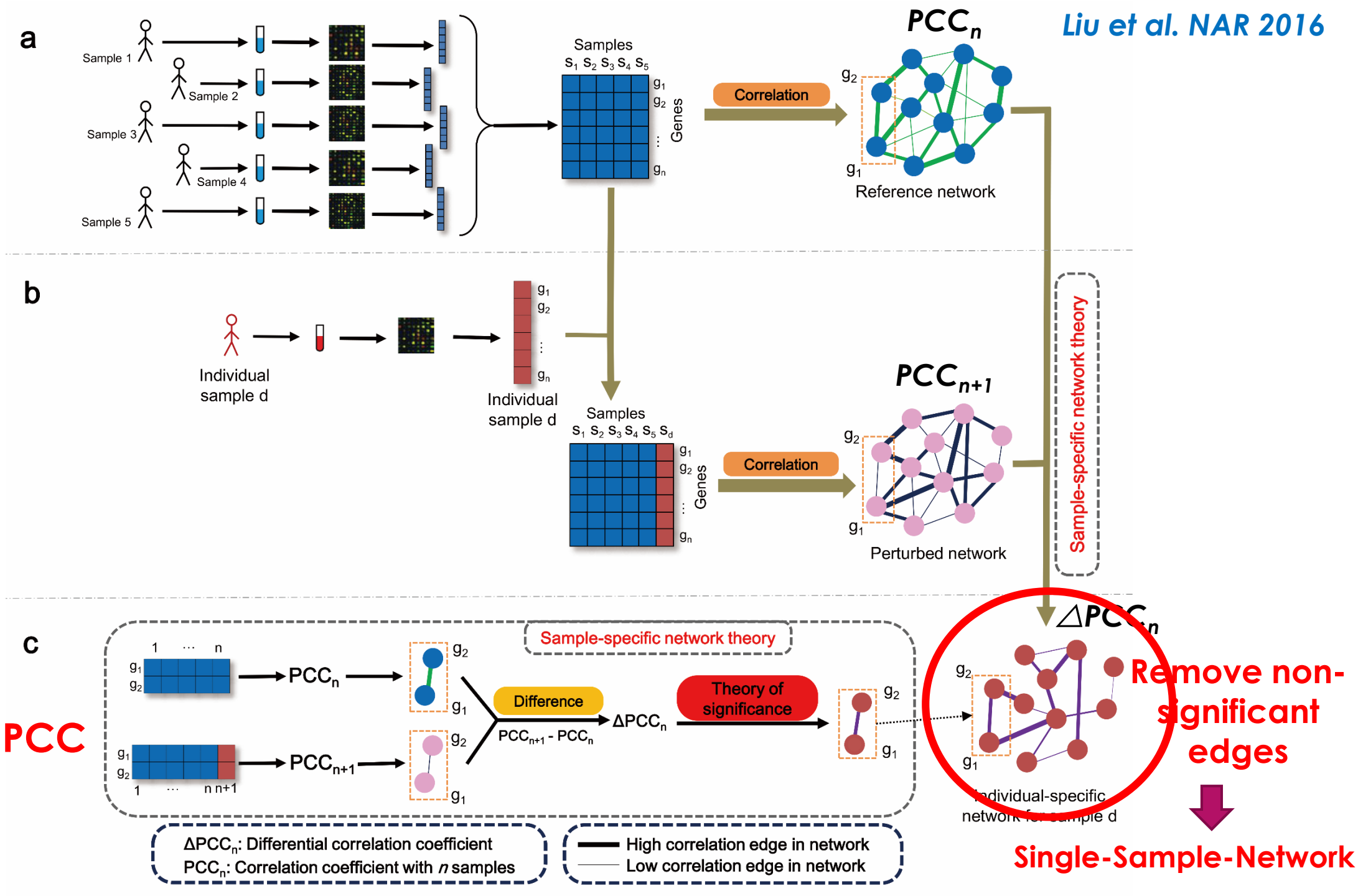
Theoretically, when n → infinite, $\frac{x-\mu_x}{\sigma_x}\frac{y-\mu_y}{\sigma_y} \rightarrow$

1. With Gaussian assumption, Volcano distribution approaches normal product distribution with the correlation.
2. Without Gaussian assumption, Volcano distribution approaches the product distribution of variables x and y.

Liu et al. NAR 2016

PCC

Differential PCC

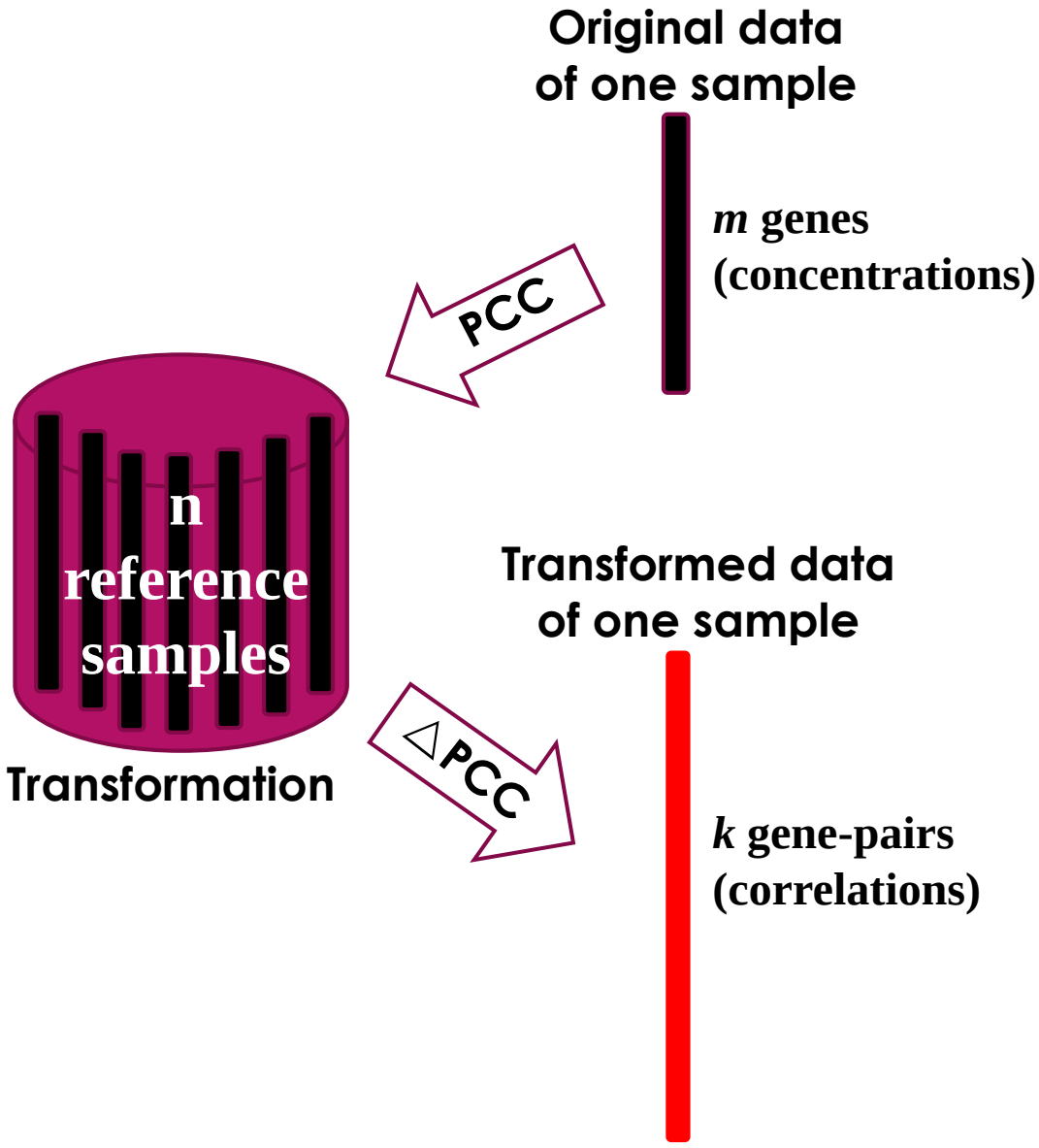


Remove non-significant edges

Single-Sample-Network

Intuitive Illustration of SSN

SSN → Data Transformation

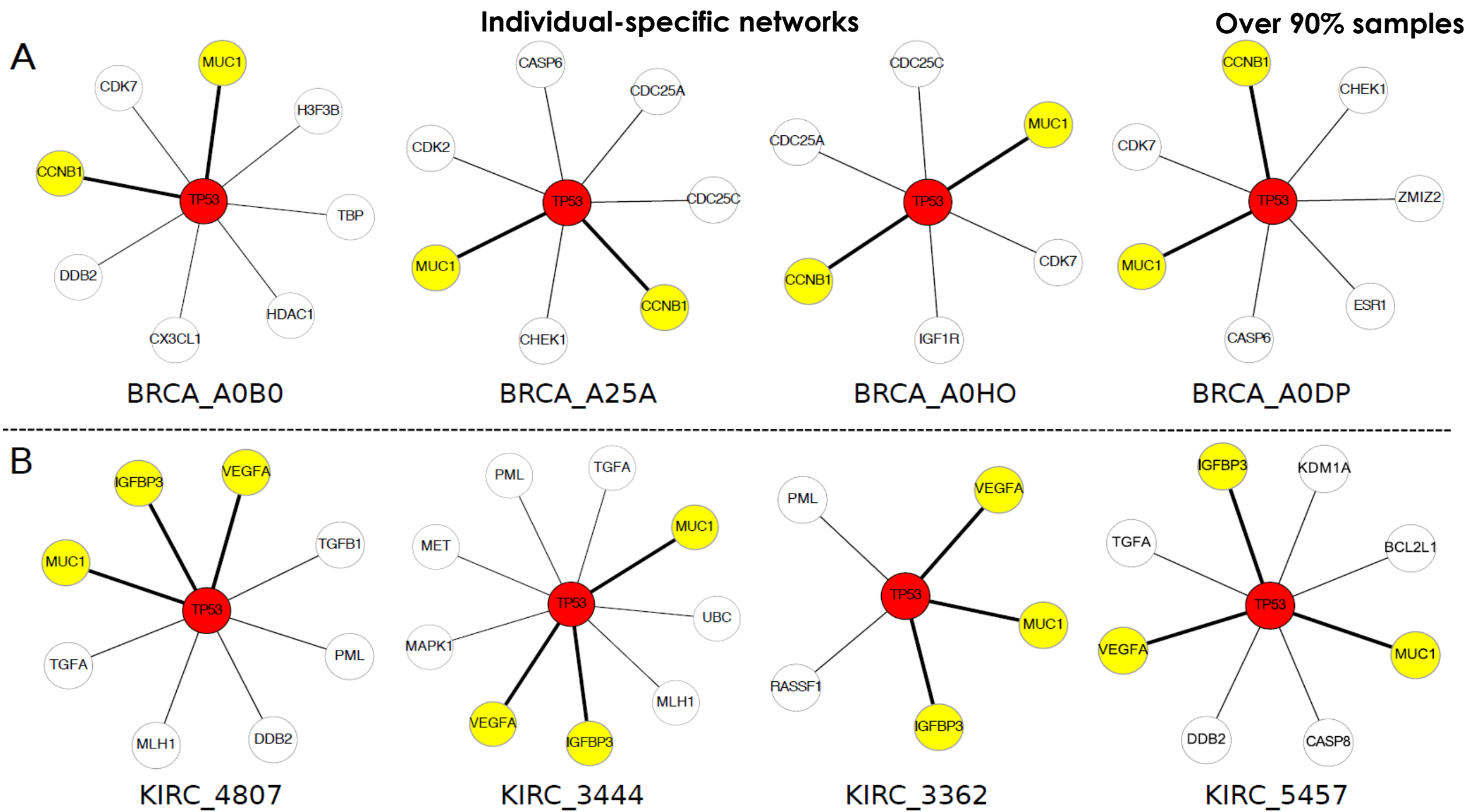


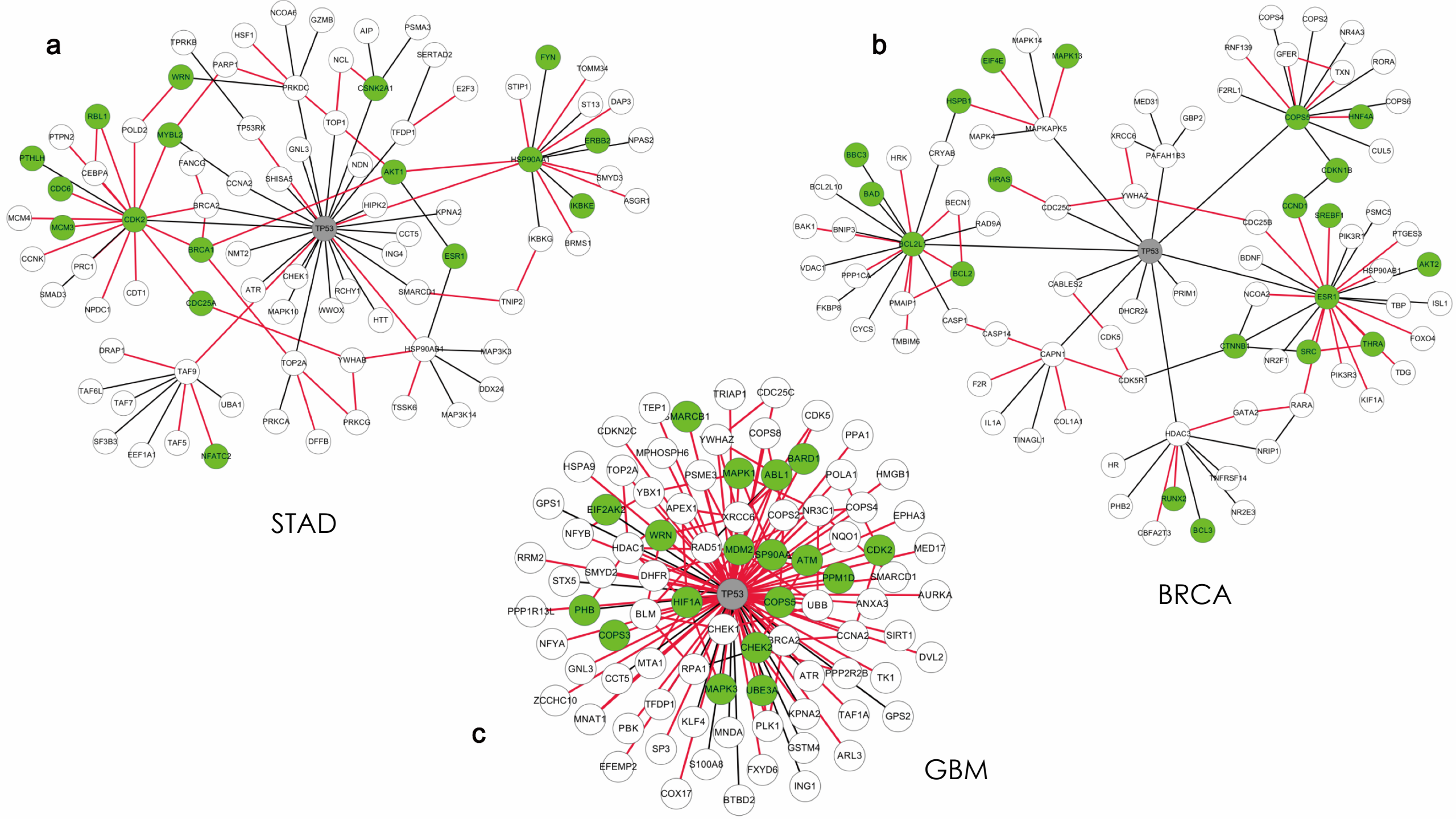
Similar to
Fourier or Wavelet transformation,
changing one type data to another domain

Validation

- TCGA (expression data and sequence information)
  - Breast cancer
  - Glioblastoma multiforme
  - Kidney renal clear cell carcinoma
  - Lung adenocarcinoma
  - Lung squamous cell carcinoma
  - Ovarian serous cystadenocarcinoma
  - Stomach cancer
  - Thyroid carcinoma
  - Uterine Corpus Endometrial Carcinoma

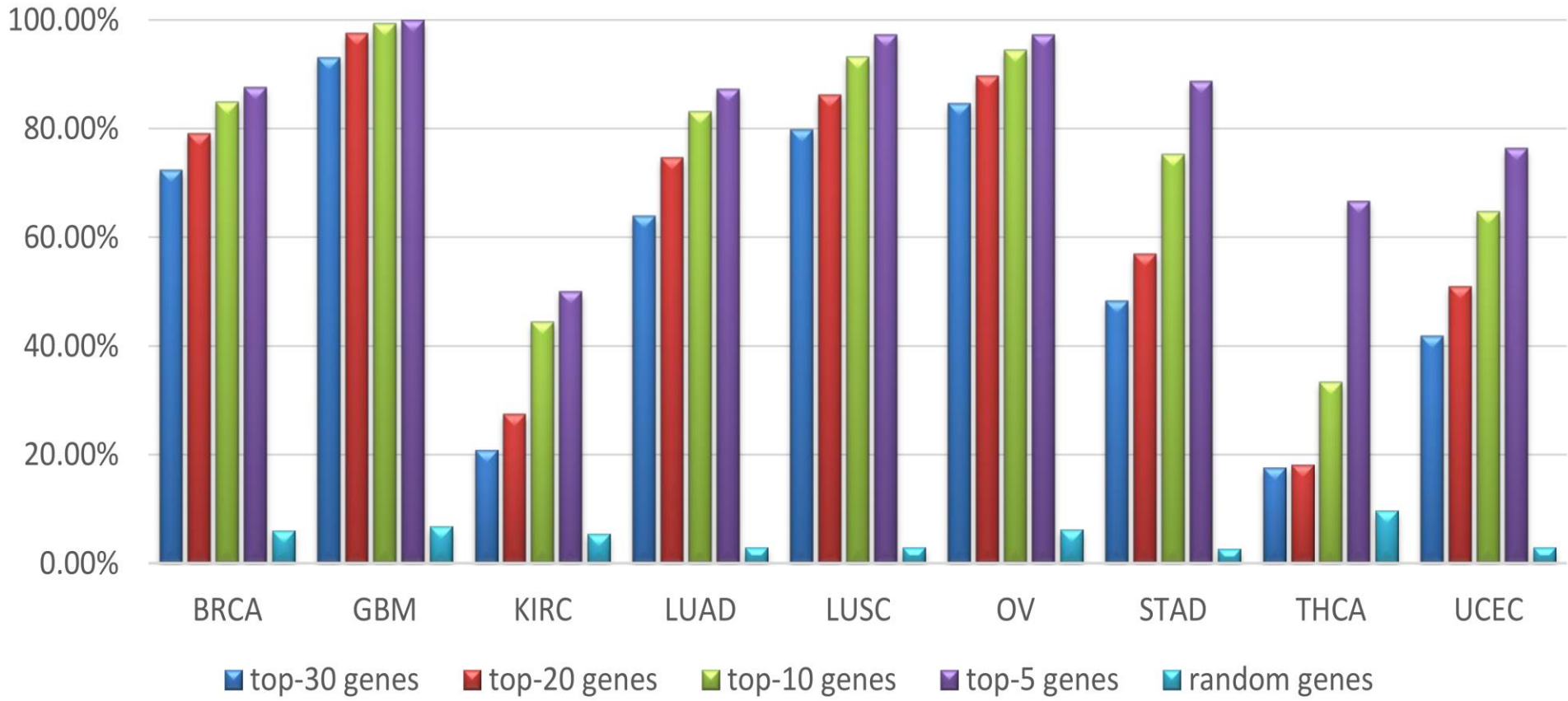
- Constructing individual-network for every disease sample


Individual-specific networks
Over 90% samples
A
MUC1
CDK7
H3F3B
CCNB1
TP53
TBP
DDB2
HDAC1
CX3CL1
BRCA_A0B0
CASP6
CDC25A
CDK2
TP53
CDC25C
MUC1
CCNB1
CHEK1
BRCA_A25A
CDC25C
CDC25A
MUC1
TP53
CDK7
CCNB1
IGF1R
BRCA_A0HO
CCNB1
CHEK1
CDK7
TP53
ZMIZ2
MUC1
ESR1
CASP6
BRCA_A0DP
B
IGFBP3
VEGFA
MUC1
TGFB1
TP53
TGFA
PML
MLH1
DDB2
KIRC_4807
PML
TGFA
MET
MUC1
TP53
UBC
MAPK1
MLH1
VEGFA
IGFBP3
KIRC_3444
PML
VEGFA
TP53
MUC1
RASSF1
IGFBP3
KIRC_3362
IGFBP3
KDM1A
TGFA
BCL2L1
TP53
VEGFA
MUC1
DDB2
CASP8
KIRC_5457

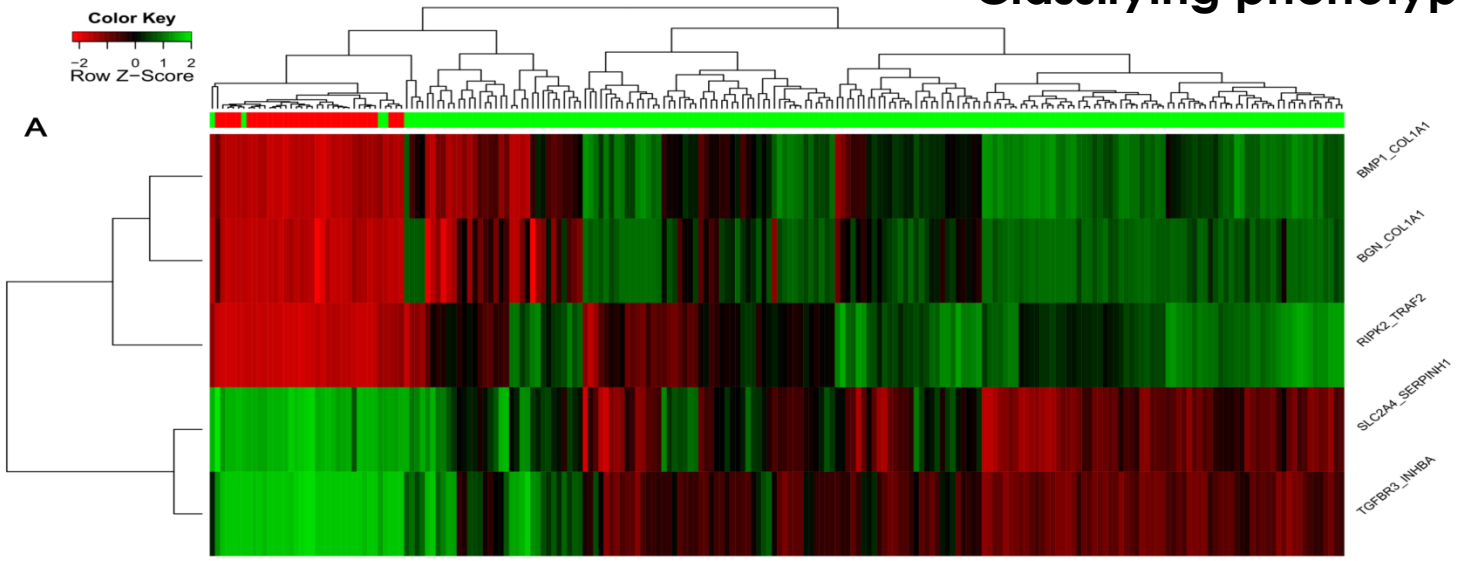

a
NCOA6
GZMB
AIP
PSMA3
HSF1
TPRKB
SERTAD2
PARP1
NCL
WRN
PRKDC
CSNK2A1
E2F3
FYN
TOMM34
STIP1
ST13
DAP3
RBL1
POLD2
MYBL2
TP53RK
TOP1
TFDP1
ERBB2
NPAS2
PTPN2
GNL3
HSP90AA1
PTHLH
FANCG
NDN
AKT1
SMYD3
CEBPA
CCNA2
SHISA5
CDC6
BRCA2
HIPK2
IKBKE
ASGR1
KPNA2
MCM4
CDK2
MCM3
TP53
CCT5
IKBKG
BRMS1
CCNK
BRCA1
NMT2
ING4
ESR1
PRC1
CHEK1
SMAD3
ATR
RCHY1
SMARCD1
CDT1
MAPK10
WWOX
TNIP2
NPDC1
CDC25A
HTT
DRAP1
HSP90AB1
MAP3K3
TOP2A
YWHAB
TAF9
DDX24
TAF6L
TSSK6
TAF7
UBA1
MAP3K14
PRKCA
PRKCG
SF3B3
DFFB
TAF5
EEF1A1
NFATC2
STAD
b
MAPK14
EIF4E
MAPK13
HSPB1
MED31
GBP2
MAPK4
MAPKAPK5
XRCC6
PAFAH1B3
BBC3
HRK
CRYAB
BCL2L10
BAD
BECN1
HRAS
CDC25C
YWHAZ
BAK1
BNIP3
RAD9A
BCL2L1
TP53
VDAC1
PPP1CA
BCL2
FKBP8
PMAIP1
CASP1
CYCS
CASP14
TMBIM6
CAPN1
CDK5
CDK5R1
F2R
IL1A
COL1A1
TINAGL1
COPS4
COPS2
NR4A3
RNF139
GFER
RORA
TXN
F2RL1
COPS5
HNF4A
COPS6
CUL5
CDKN1B
CCND1
SREBF1
PSMC5
PIK3R1
PTGES3
BDNF
AKT2
HSP90AB1
CABLES2
PRIM1
DHCR24
NCOA2
ESR1
TBP
ISL1
FOXO4
CTNNB1
SRC
THRA
NR2F1
PIK3R3
TDG
KIF1A
GATA2
RARA
HDAC3
NRIP1
HR
TNFRSF14
PHB2
RUNX2
NR2E3
CBFA2T3
BCL3
BRCA
c
TRIAP1
CDC25C
TEP1
SMARCB1
CDK5
CDKN2C
YWHAZ
COPS8
PPA1
MPHOSPH6
BARD1
HSPA9
TOP2A
PSME3
MAPK1
ABL1
POLA1
HMGB1
YBX1
GPS1
APEX1
EIF2AK2
XRCC6
COPS2
NR3C1
COPS4
EPHA3
NFYB
WRN
NQO1
HDAC1
RAD51
MDM2
HSP90AA1
ATM
CDK2
MED17
RRM2
SMYD2
DHFR
PPM1D
SMARCD1
STX5
TP53
HIF1A
COPS5
UBB
ANXA3
AURKA
PPP1R13L
PHB
BLM
CHEK1
COPS3
CCNA2
SIRT1
NFYA
BRCA2
CHEK2
DVL2
GNL3
MTA1
RPA1
CCT5
ATR
PPP2R2B
TK1
ZCCHC10
MAPK3
UBE3A
TFDP1
KPNA2
TAF1A
GPS2
MNAT1
PBK
KLF4
MNDA
PLK1
GSTM4
ARL3
SP3
S100A8
FXYD6
EFEMP2
ING1
COX17
BTBD2
GBM

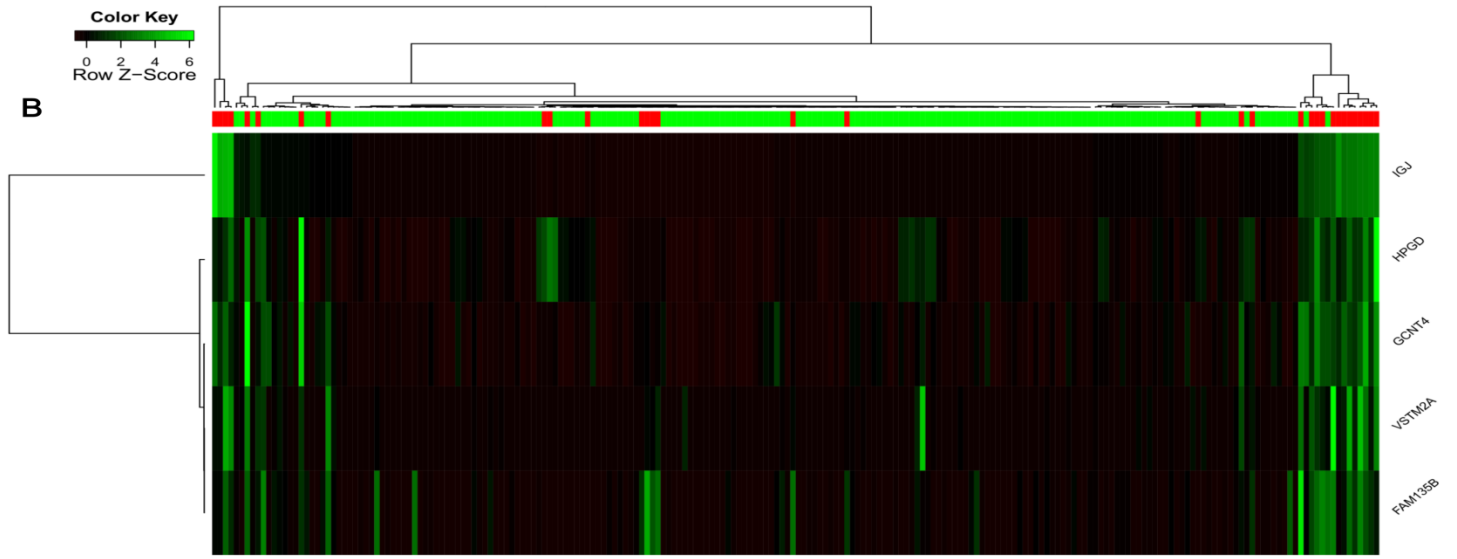
Driver mutation prediction by expression data

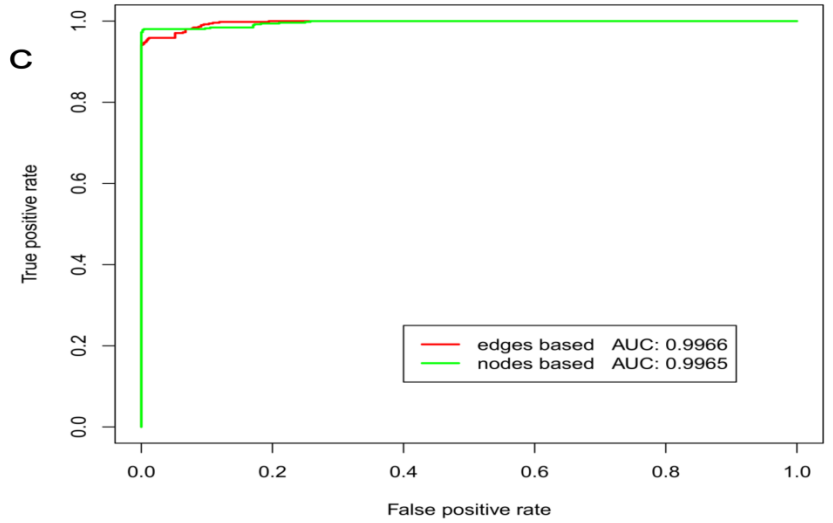
Potential driver genes: Top *n* highly connected nodes in the individual network for each sample

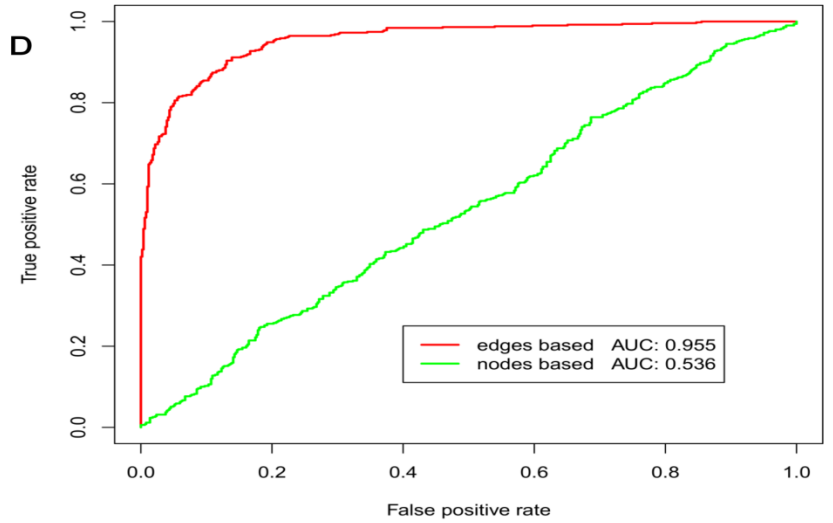


Classifying phenotypes and cancer subtypes









E

	BRCA	UCEC	KIRC	LUAD	OV	GBM	LUSC	THCA	STAD
Node markers	0.141	0.319	0.0026	0.0085	0.432	0.502	0.173	0.4	0.829
Edge markers	0.026	0.0037	<1E-6	0.0033	0.056	0.265	0.136	0.37	0.81

Experiment-1 as network biomarkers

Cholangiocarcinoma (胆管癌)

- GSE26566
- 163 samples
  - 104 cancerous
  - 59 non-cancerous

从不变 找到 变化

Find variant factors from invariant data

Choose only non-differential genes

1312 non-differential genes were selected with the criterion of p-value > 0.6

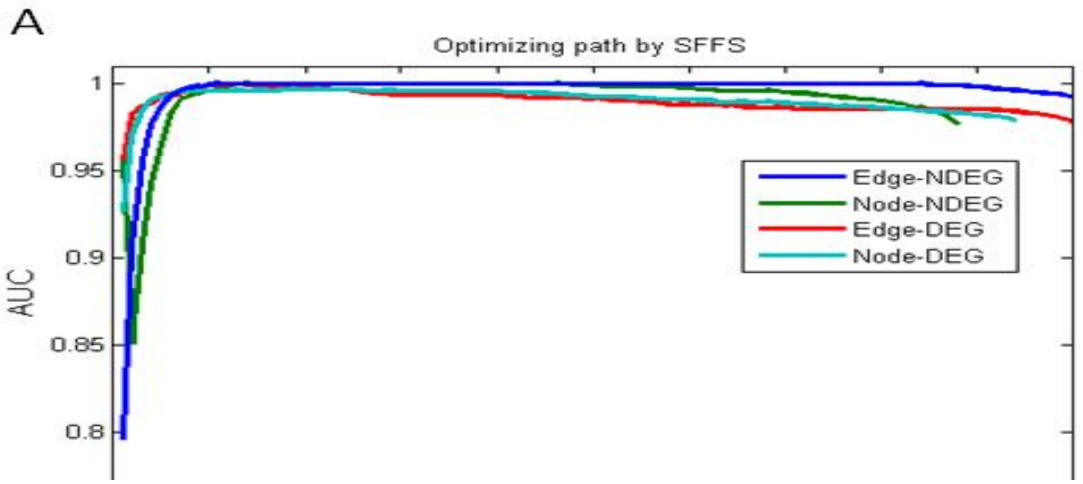
538 gene pairs are differentially correlated under the criterion of $|\Delta PCC| > 0.9$

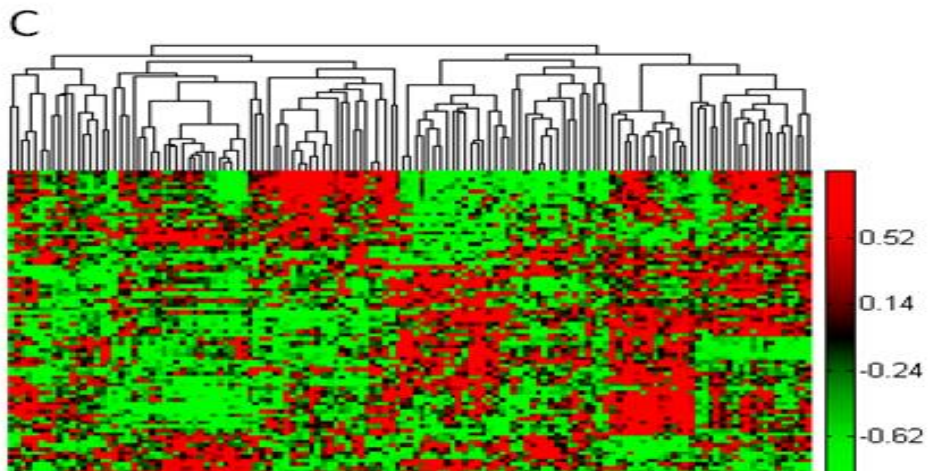
For comparison purpose,

delete all differential genes
or
choose only non-differential genes

Dark matter in gene expression 90%
Dark matter in gene sequence 90%
Dark matter in Universe 90%

Cholangiocarcinoma

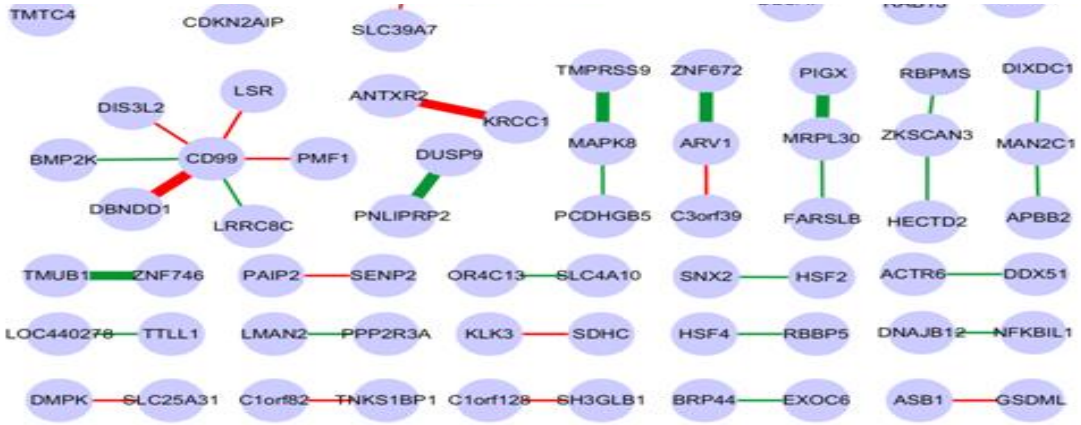


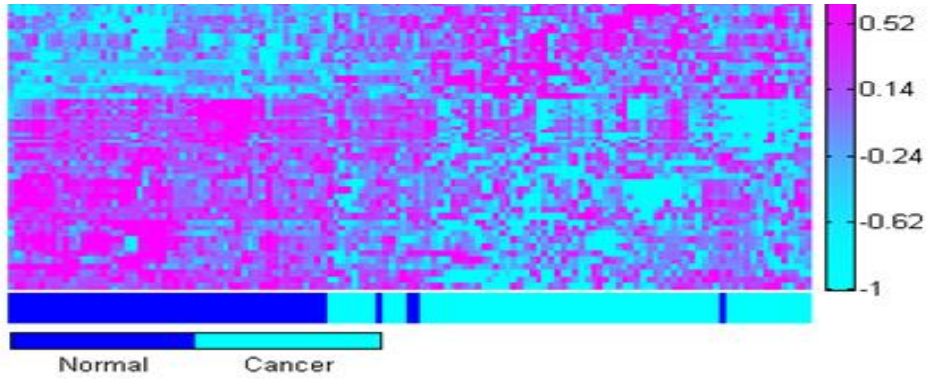


(A) For NDEG and DEG, the curve by SFFS methods on the top 100 edge features and the involved genes are plotted. (B) The discriminative edge features are presented, where bold lines are

Non-differential genes achieve the same or better effect than differential genes ! ! !

not in gene level but in network level





Heatmap of the edge data by (2).

Experiment-2 as network biomarkers

Type I diabetes

- GSE9006
- 67 samples
  - 43 newly T1D samples
  - 24 normal samples

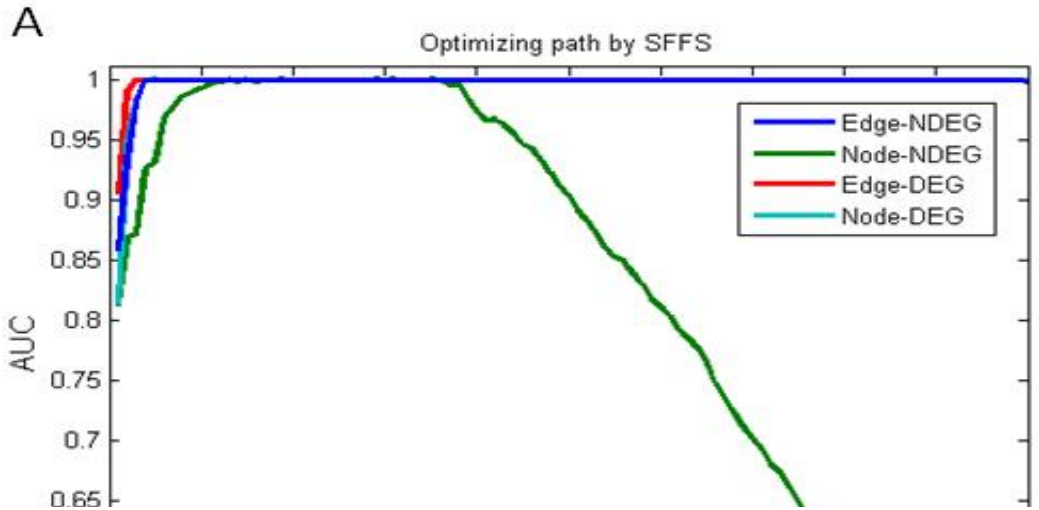
从不变 找到 变化

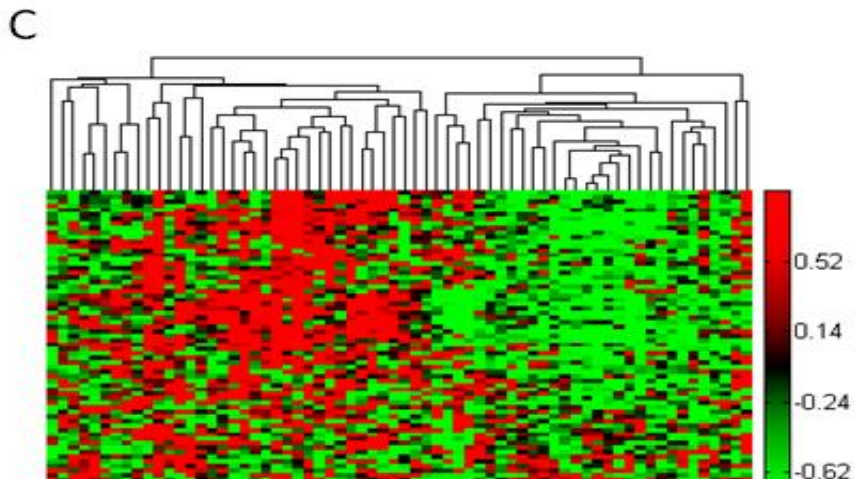
Choose only non-differential genes

1600 Non differential genes were selected with the criterion of P-value > 0.8

there are 365 Differential Correlation Pairs

Type I diabetes

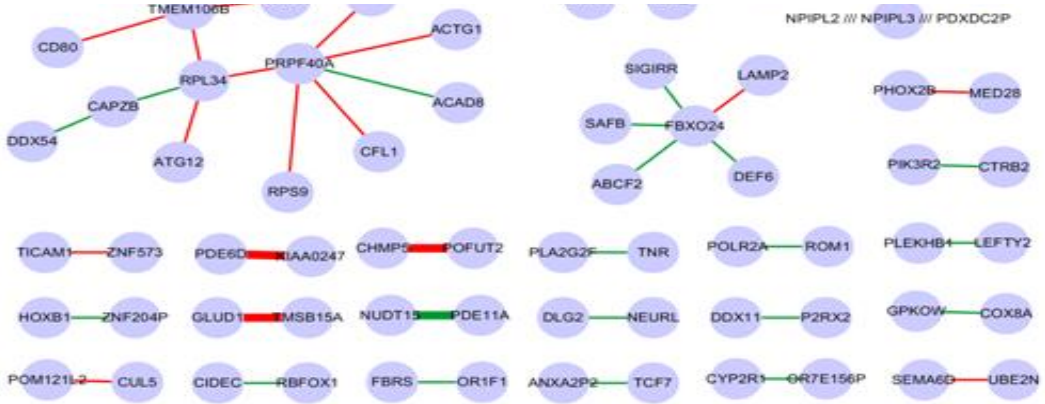


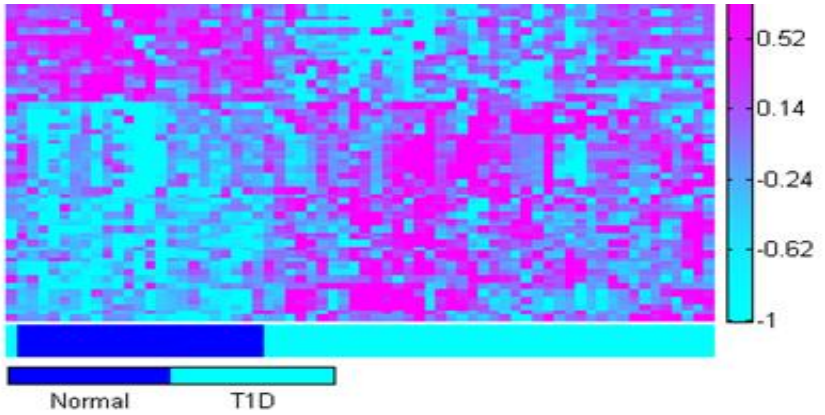


(A) For NDEGs and DEGs, the curves by SFFS on the top 100 edge features and the involved genes are plotted. (B) The discriminative edge features are

No n -differential genes achieve the same or better effect than differential genes ! ! !

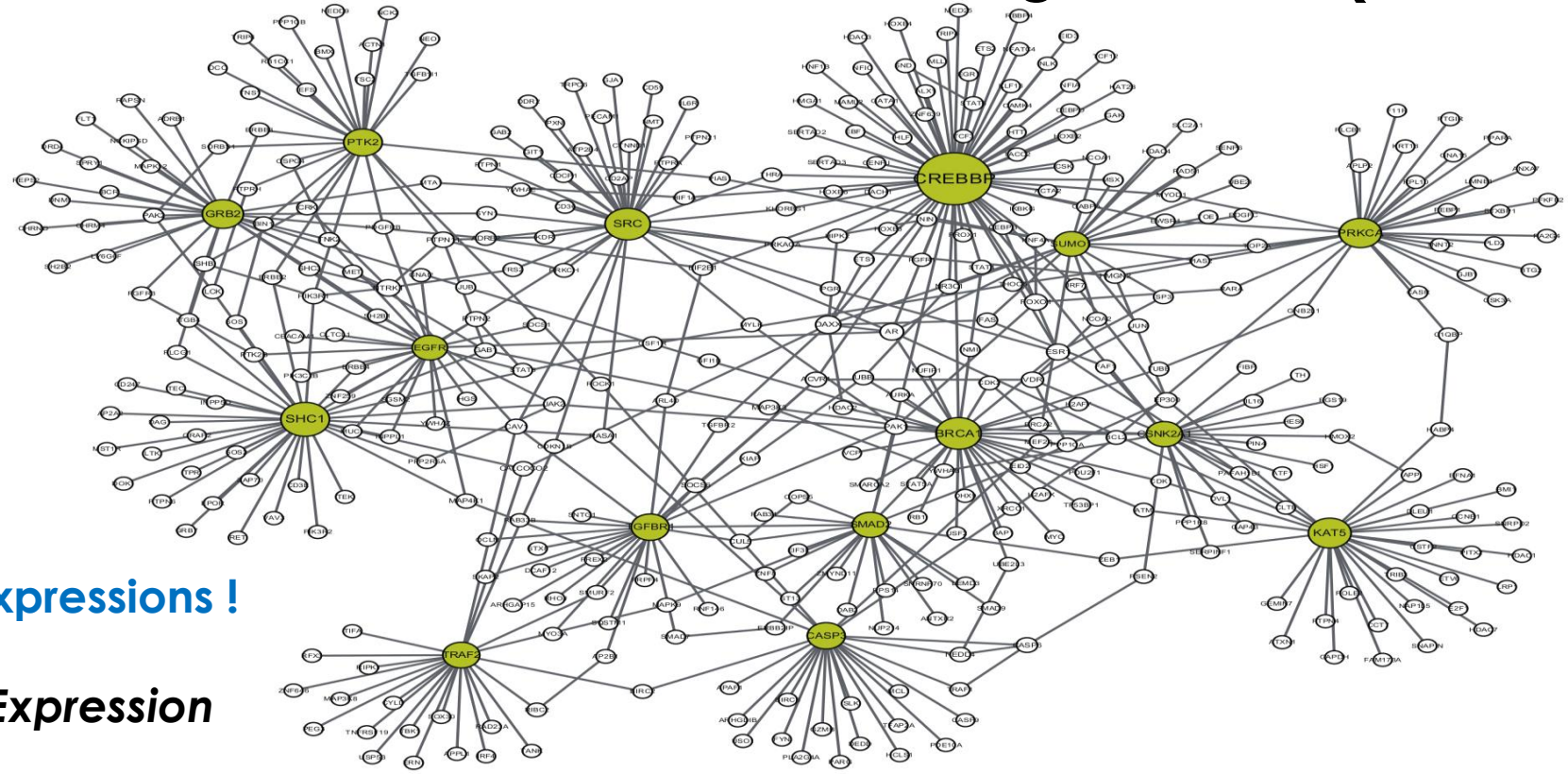
not in gene level but in network level





Heatmap of the edge data by (2).

A Drug Resistance Gene Prediction for Lung Cancer (PC9 and PC9-DR)



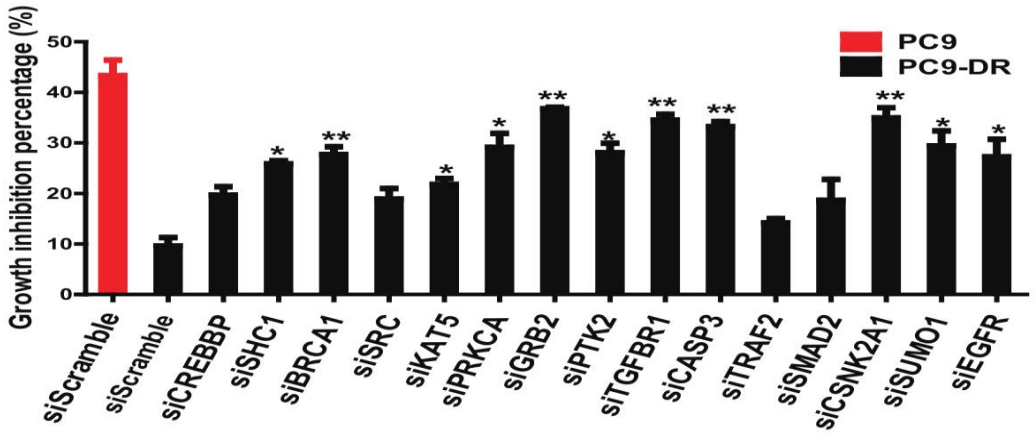
No differential expressions !

Dark Matter in Expression

B

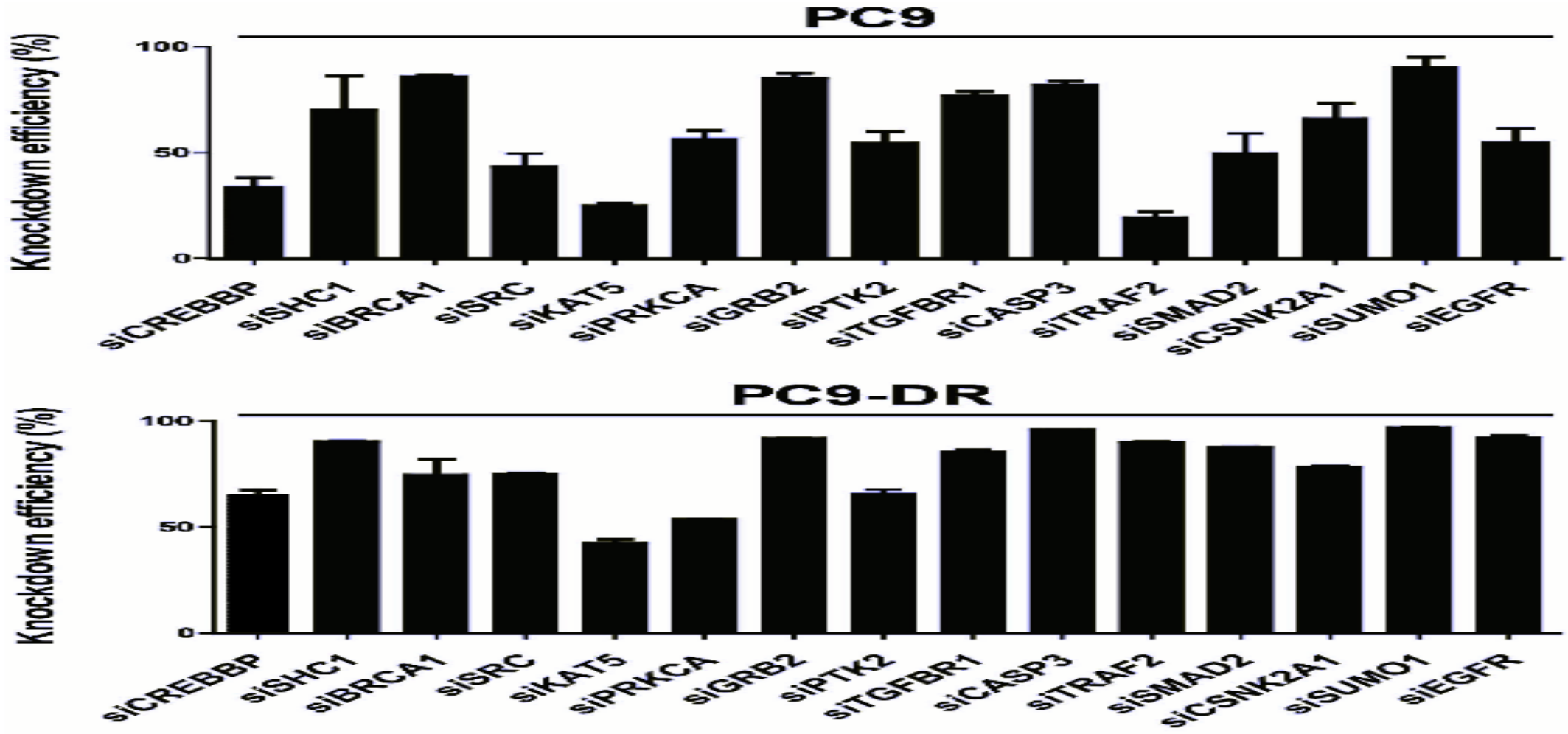
Gene symbols	Degree	Fold change	Significance to drug resistance
CREBBP	68	0.754549298	No
SHC1	37	1.489521613	Yes*
BRCA1	33	0.838699498	Yes**
SRC	33	1.232327829	No
KAT5	30	1.273340765	Yes*
PRKCA	29	0.746732494	Yes*
GRB2	28	1.30560704	Yes**
PTK2	25	0.761979505	Yes*
TGFBR1	24	0.712501428	Yes**
CASP3	24	0.876307508	Yes**
TRAF2	23	1.218907247	No
SMAD2	22	0.805846744	No
CSNK2A1	22	1.147045792	Yes**
SUMO1	21	0.694003697	Yes*
EGFR	21	0.897537327	Yes*

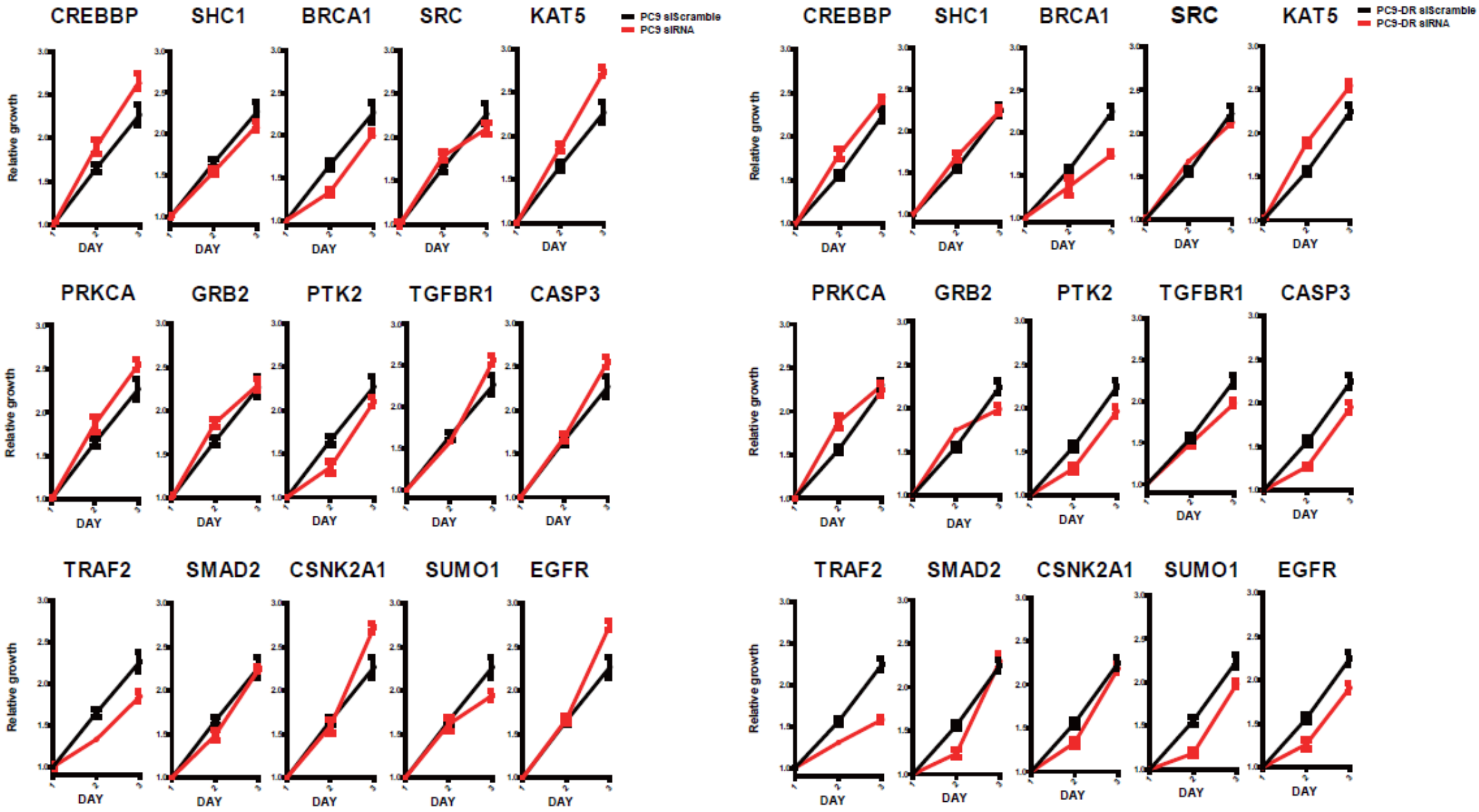
C



Dark Matter

- Universe → 90% Dark Matter in Universe
- Non-coding RNA → 90% Dark Matter in sequence
- Non-differential expression → 90% Dark Matter in expression


PC9
Knockdown efficiency (%)
100
50
0
siCREBBP
siSHC1
siBRCA1
siSRC
siKAT5
siPRKCA
siGRB2
siPTK2
siTGFBR1
siCASP3
siTRAF2
siSMAD2
siCSNK2A1
siSUMO1
siEGFR
PC9-DR
Knockdown efficiency (%)
100
50
0
siCREBBP
siSHC1
siBRCA1
siSRC
siKAT5
siPRKCA
siGRB2
siPTK2
siTGFBR1
siCASP3
siTRAF2
siSMAD2
siCSNK2A1
siSUMO1
siEGFR


CREBBP
SHC1
BRCA1
SRC
KAT5
PC9 siScramble
PC9 siRNA
PC9-DR siScramble
PC9-DR siRNA
PRKCA
GRB2
PTK2
TGFBR1
CASP3
TRAF2
SMAD2
CSNK2A1
SUMO1
EGFR
Relative growth
DAY

Liu et al. NAR 2016

Conclusion

All SSNs are comparable, given a reference network

- Construct Single Sample Network
- Single-cell network is single-sample network
- Replace molecular biomarkers by network biomarkers

Cross species, Cross tissues, Cross platforms

- Predict key or driver genes even without differential expressions
- Achieve network biomarker for personalized medicine
- Single Sample DNB and DNB landscape

Acknowledgments

- Xiaoping Liu,
  Chinese Academy of Sciences
- Kazuyuki, Aihara,
  The University of Tokyo
- Hongbin Ji, Yuetong Wang
  Chinese Academy of Sciences

THANK YOU!