Modeling Dynamic Dependence Structure in Zero-Inflated Bivariate Count Data with Application to Single-Cell RNA Sequencing Data

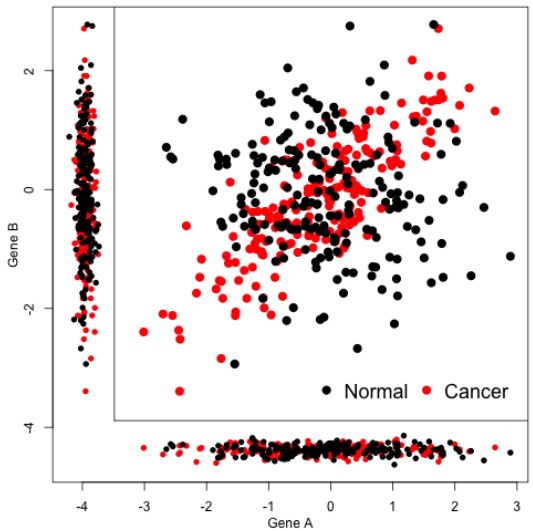
Yen-Yi Ho

Department of Statistics, College of Arts and Sciences
University of South Carolina

Outline

- Motivating Example
- The Data
- The ZENCO model
- Search Strategies
- Simulation Analyses
- Experimental Data Analysis
- Conclusion

Introduction and motivations



- Routine differential gene expression approaches ignore interactions between genes.
- Gene Co-expression analysis addresses this issue by evaluating whether there are correlated changes between pairs of genes across different modulating conditions.
- Genetic co-expression pattern can change dynamically in response to internal cellular signals or external stimuli.

Dynamic Coexpression

Dynamic coexpression changes: the coexpression of two genes, $\mathbf{X}_1$ and $\mathbf{X}_2$ can be mediated by a third variable, $\mathbf{X}_3$.

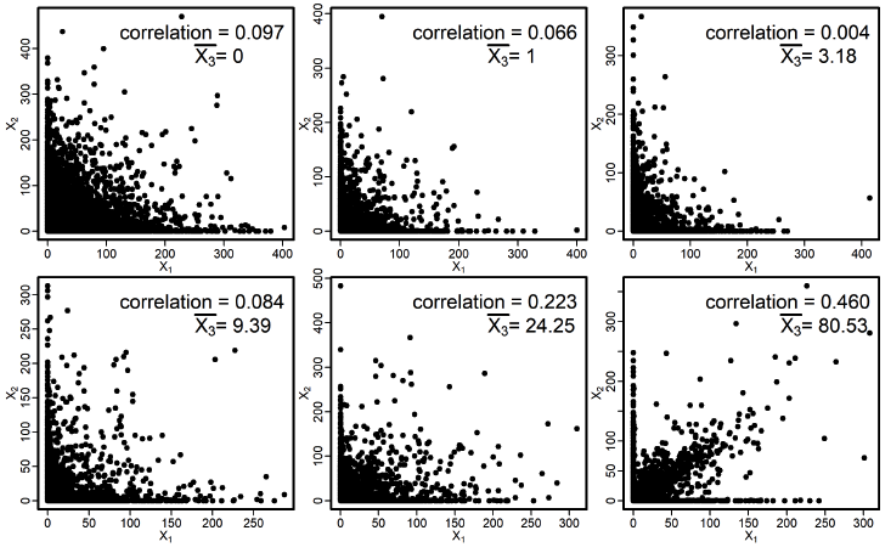


Figure: Simulated example of dynamic coexpression changes

- Single-cell RNA sequencing (scRNA-seq) data are count-based
- Zero-inflation

Motivating Example

- Biological pathways are highly dynamic. Cancer cells can acquire drug resistance by establishing alternative bypass signaling pathways after exposure to therapeutic agents.

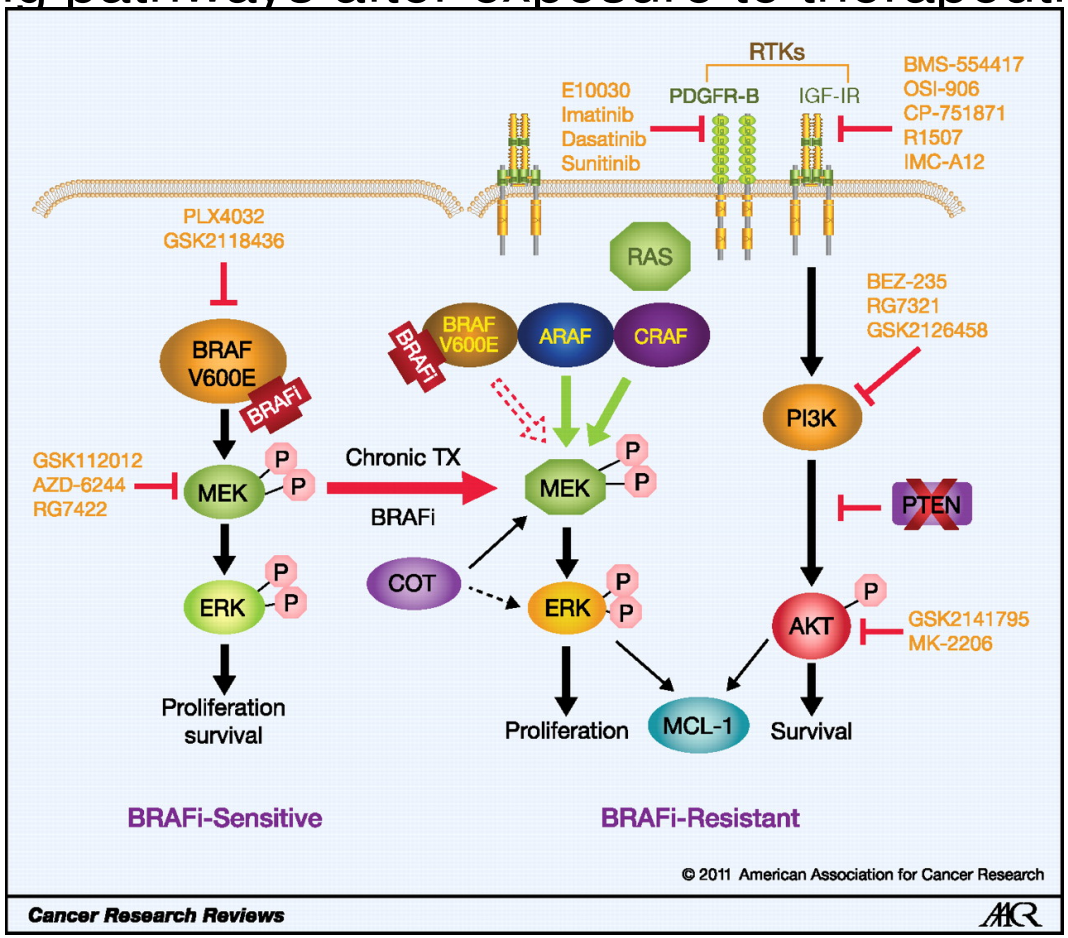
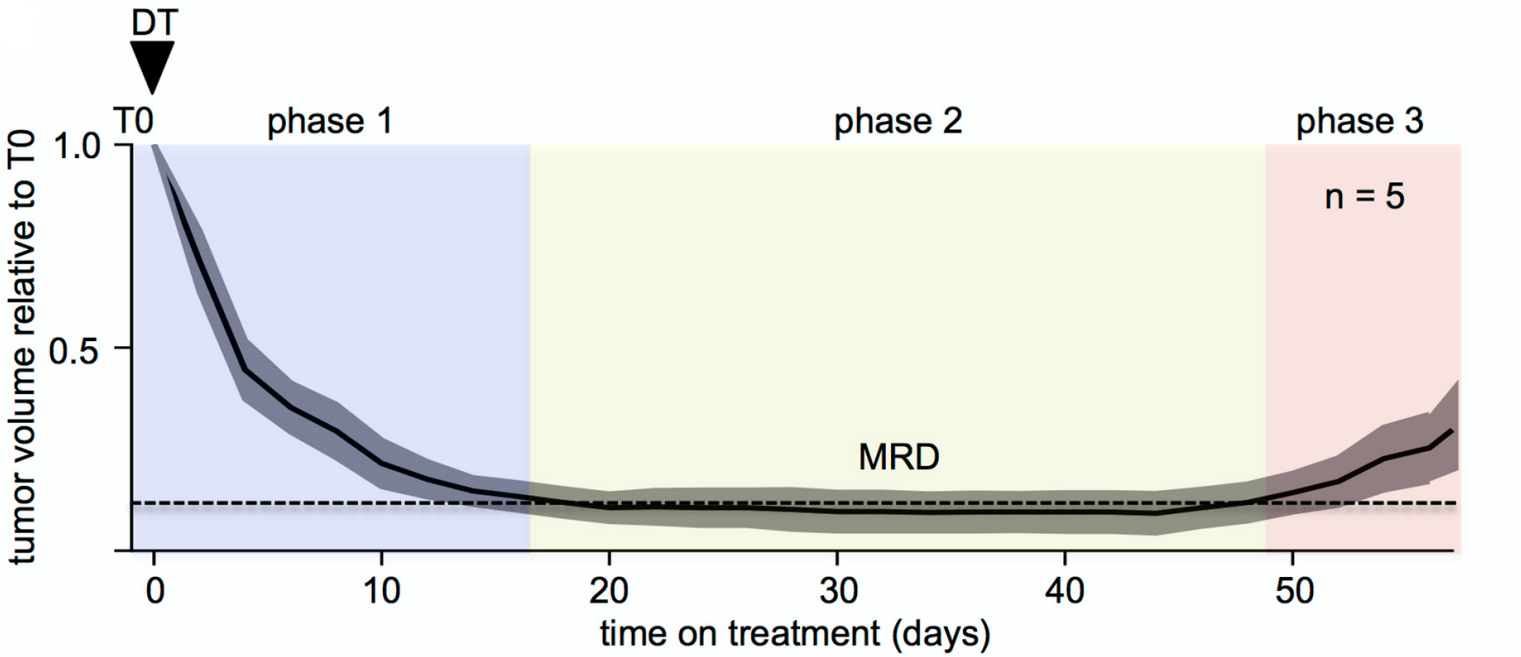


Figure adapted from [**?**]

scRNA-seq Data

- BRAF mutant patient-derived xenograft (PDX) melanoma cohorts [**?**].
- Once the PDX tumors grew to comparable size, mice were treated with concurrent RAF/MEK-inhibition
- The data contain information for 57,445 transcripts from 675 melanoma cells from all phases.
- The three phases are: drug-sensitive, minimum residual disease (MRD), drug-resistance



The ZEro-inflated Negative binomial dynamic COrrelation (ZENCO) model

- Let X_{ij} denote the transcript counts for the i-th gene in the j-th cell and $\mathbf{X}_i$ represents the gene expression count for the i-th gene. The distribution of $\mathbf{X_i}$ is modelled as:

$$\mathbf{X}_i \sim \begin{cases} Poisson(\lambda_0), & \text{with probability } p_i; \\ NB(\mu_i, \phi_i), & \text{with probability } 1 - p_i. \end{cases}$$

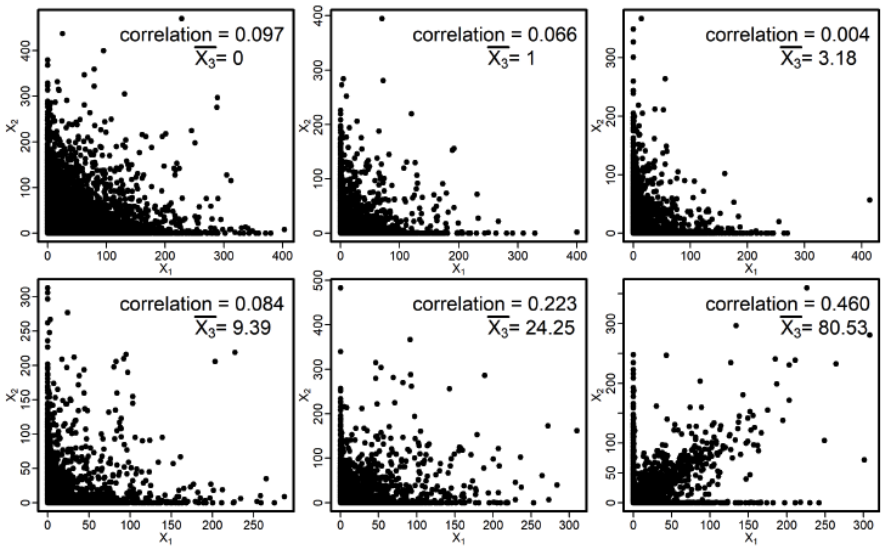
- p_i is the dropout rate of $\mathbf{X}_i$ and is modelled as a function of μ_i: $p = \frac{e^{(b_0 + b_1 \mu)}}{1 + e^{(b_0 + b_1 \mu)}}$.

Poisson-Gamma mixture with random effects

- The correlation of a gene pair: $\mathbf{X}_1$ and $\mathbf{X}_2$ can be observed when both genes are observed in the j-th cell.
- Poisson-Gamma mixture

$$X_{ij} \sim Poisson(u_{ij}\mu_i), u_{ij} \sim Gamma(\alpha_i, \alpha_i).$$

- Integrate out u_{ij}, $X_{ij} \sim NB(\mu_i, \phi_i = \frac{1}{\alpha_i})$
- u_{ij} can be considered as the cell-specific random effect



Modeling correlation structure in count data

- Let the latent variable $\mathbf{Z_j} = (Z_{1j}, Z_{2j})'$ be a bivariate normal variable that

$$\mathbf{Z_j} \sim N_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_j \\ \rho_j & 1 \end{bmatrix} \right).$$

- The correlation, ρ_j, of (Z_{1j}, Z_{2j}) is specified as

$$\log\left(\frac{1+\rho_j}{1-\rho_j}\right) = \tau_0 + \tau_1 X_{3j}.$$

- Plug-in $\mathbf{Z_j}$ into u_{ij}, we have

$$X_{ij} \sim Poisson[F_{\alpha_i}^{-1}\{\Phi(Z_{ij})\}\mu_i],$$

where $F_{\alpha_i}(\cdot)$ is the cumulative distribution function of a $Gamma(\alpha_i, \alpha_i)$ distribution with $\alpha_i = 1/\phi_i$ and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution.

- The joint distribution of $\mathbf{X}_1$ and $\mathbf{X}_2$ can be specified using:

$$x_{ij} \sim \begin{cases} Poisson(\lambda_0), & \text{with probability } p_i; \\ Poisson[F_{1/\phi_i}^{-1}\{\Phi(z_{ij})\}\mu_i], & \text{with probability } 1-p_i. \end{cases}$$

Search Strategies

- For a given pair of genes ($\mathbf{X}_1$, $\mathbf{X}_2$), screen the whole-genome to identify a third modulator gene.
- For a given modulator variable ($\mathbf{X}_3$), screen the whole-genome to identify a pair of genes that are modulated by $\mathbf{X}_3$ ($\binom{m}{2}$, m is the total number of genes).
- If no prior information about $\mathbf{X}_3$ or ($\mathbf{X}_1$, $\mathbf{X}_2$) is available, screen the relevant pathways or the whole genome to identify potential gene triplets $\binom{m}{3}$.
- When the number of genes under considerations is large (for example $\approx$ 20,000). Pre-screening is beneficent such as [**?**] or the screening statistic (ζ) introduced in [**?**].

Simulation Analyses

$$\log\left(\frac{1+\rho_j}{1-\rho_j}\right) = \tau_0 + \tau_1 X_{3j}.$$

- Under the hypotheses:

$$H_0 : \tau_1 = 0 \text{ versus } H_1 : \tau_1 \neq 0,$$

Table: Coverage probability (CP) of 95% credible interval (CI) and interval lengths based on 1,000 MCMC simulations ($\tau_0 = 0.01$, $\tau_1 = 0.05$)

		Without Zero-inflation		With Zero-inflation	
	Parameter	CP	CI length	CP	CI length
$N = 200$	τ_0	0.997	0.455	1.000	0.541
	τ_1	0.170	0.042	0.942	0.111
$N = 500$	τ_0	0.985	0.288	1.000	0.342
	τ_1	0.009	0.022	0.950	0.064
$N = 1,000$	τ_0	0.955	0.204	1.000	0.242
	τ_1	0.000	0.014	0.951	0.043

Simulation Analyses

Table: Mean square errors (MSE) and mean bias errors (MBE) based on 1,000 MCMC simulations ($\tau_0 = 0.01$, $\tau_1 = 0.05$). MBE= $\frac{1}{N}\sum_{i=1}^{N}(\widehat{\beta}_i - \beta)$.

		Without Zero-inflation		With Zero-inflation	
	Parameter	MSE	MBE	MSE	MBE
$N = 200$	τ_0	0.008	0.044	0.001	-0.008
	τ_1	0.002	-0.039	0.001	-0.001
$N = 500$	τ_0	0.006	0.051	0.000	-0.008
	τ_1	0.002	-0.040	0.000	0.001
$N = 1,000$	τ_0	0.005	0.051	0.000	-0.009
	τ_1	0.002	-0.041	0.000	0.001

Power Comparison to existing methods

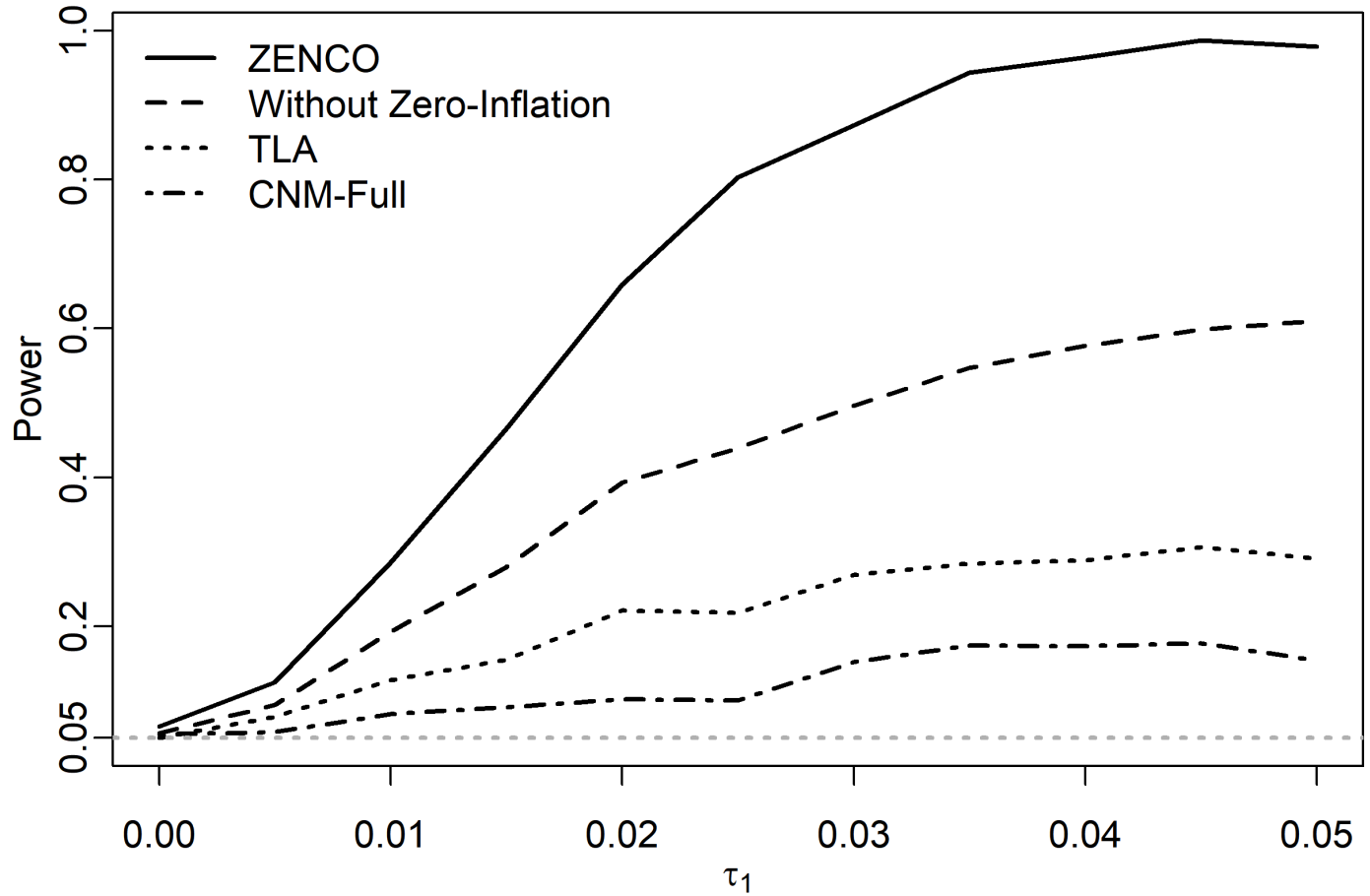
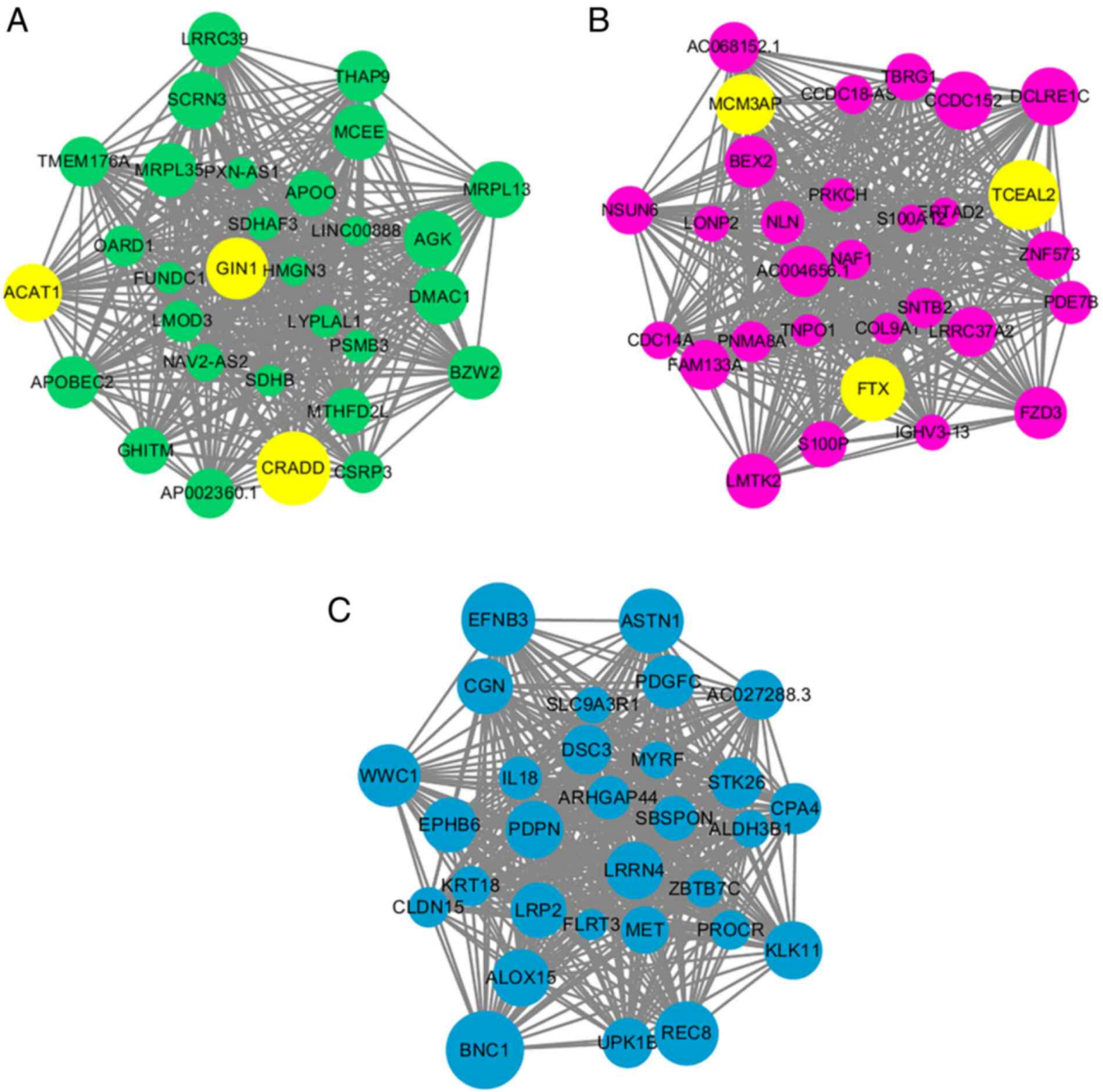


Figure: Power curves comparing various methods. Both TLA and CNM-Full approaches are Gaussian-based models [**?**, **?**].

WGCNA


A
LRRC39
THAP9
SCRN3
MCEE
TMEM176A
MRPL35
PXN-AS1
APOO
MRPL13
SDHAF3
LINC00888
AGK
OARD1
GIN1
HMGN3
FUNDC1
ACAT1
DMAC1
LMOD3
LYPLAL1
PSMB3
NAV2-AS2
BZW2
APOBEC2
SDHB
MTHFD2L
GHITM
CRADD
CSRP3
AP002360.1
B
AC068152.1
TBRG1
MCM3AP
CCDC18-AS1
CCDC152
DCLRE1C
BEX2
TCEAL2
PRKCH
NSUN6
LONP2
NLN
S100A12
ERTAD2
ZNF573
NAF1
AC004656.1
PDE7B
SNTB2
CDC14A
PNMA8A
TNPO1
COL9A1
LRRC37A2
FAM133A
FTX
FZD3
IGHV3-13
S100P
LMTK2
C
EFNB3
ASTN1
CGN
PDGFC
AC027288.3
SLC9A3R1
DSC3
MYRF
WWC1
IL18
STK26
ARHGAP44
CPA4
EPHB6
PDPN
SBSPON
ALDH3B1
LRRN4
KRT18
ZBTB7C
CLDN15
LRP2
FLRT3
MET
PROCR
KLK11
ALOX15
BNC1
UPK1B
REC8

Experimental Data Analysis

- We use BRAF gene expression count as $\mathbf{X}_3$ and screen all gene-pair combinations in the KEGG melanoma pathway.

Table: Top table of dynamic correlations differences. $\Delta\tau_1$ is the difference between τ_1 estimates in Phase 3 (P3) and Phase 1 (P1).

Gene 1	Gene 2	$\tau_1(P1)$	$\tau_1(P3)$	$\Delta\tau_1$
PDGFC	FGFR1	0.084 (0.045,0.120)	0.000 (-0.006,0.007)	-0.084
BAX	POLK	0.053 (0.023,0.085)	0.000 (-0.007,0.005)	-0.054
AKT1	ARAF	-0.024 (-0.046,-0.004)	0.019 (0.000,0.039)	0.043
AKT1	MAPK1	0.004 (-0.008,0.015)	0.043 (0.020,0.060)	0.039
AKT3	MAP2K2	0.033 (0.017,0.048)	-0.003 (-0.010,0.002)	-0.037
AKT1	BAK1	-0.027 (-0.053,-0.004)	0.008 (-0.003,0.030)	0.035
MAP2K2	FGFR1	0.031 (-0.001,0.081)	-0.003 (-0.009,0.003)	-0.033
BAX	MDM2	0.032 (0.005,0.059)	-0.001 (-0.007,0.005)	-0.033
AKT1	AKT2	0.003 (-0.009,0.014)	0.031 (0.003,0.050)	0.029
MAP2K2	BAX	0.035 (-0.006,0.075)	0.006 (-0.003,0.016)	-0.029

Conclusion

- The results from the simulation analysis indicates that our proposed ZENCO model outperforms other existing Gaussian-based approaches due to the fact our model accounts for zero-inflation, over-dispersion in scRNAseq data
- We used the expression level of BRAF as the modulator variable $\mathbf{X}_3$. In other applications, $\mathbf{X}_3$ can be easily modified to represent other conditions such as tumor status, degree of inflammation, or cell types, ...etc.
- In this work, our focus is on the change of co-expression patterns between a gene pair. It's plausible that higher-order interactions between genes exist, a generalization of our approach to higher dimension is feasible. However, special treatments need to be consider to ensure the positive definiteness of the variance covariance matrix in higher-dimension.

References I