Network Analysis of RNA Sequence and Drug Safety Data

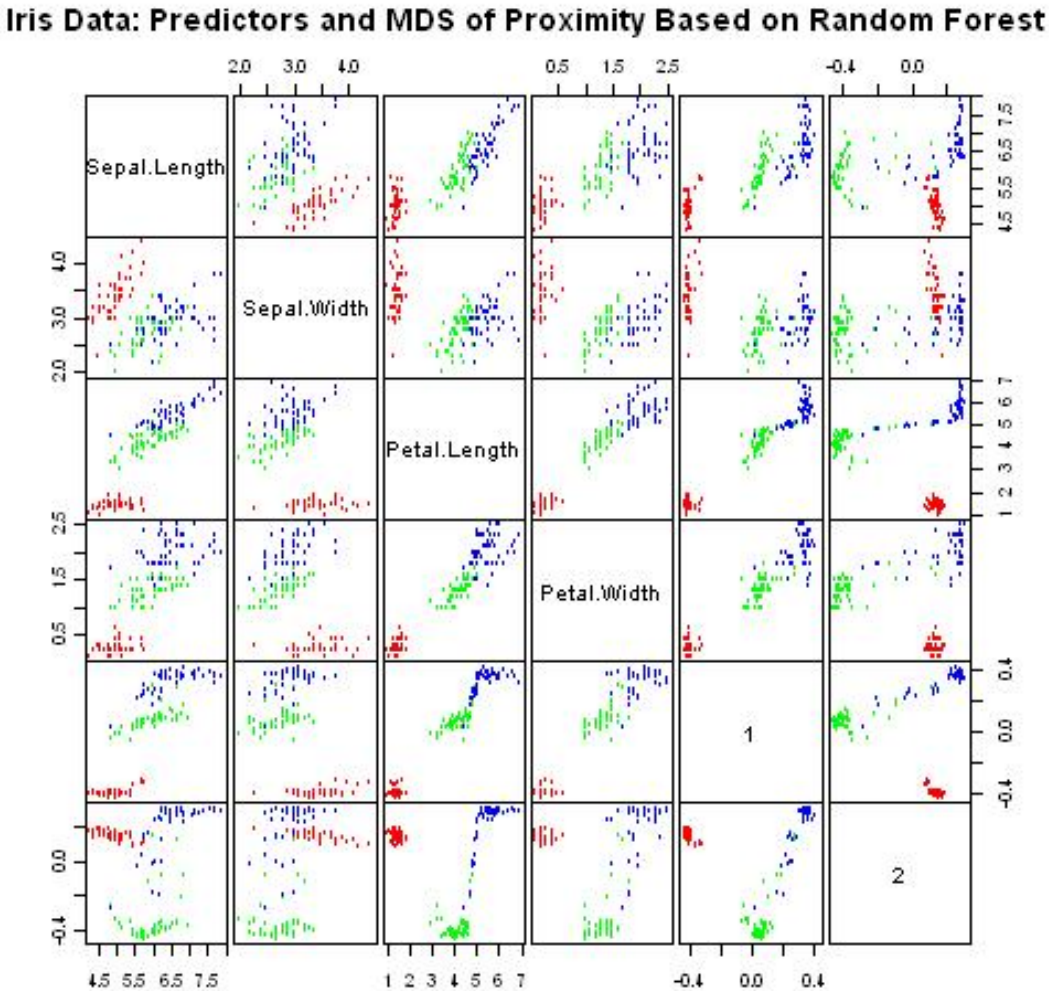
Karl Peace & Kao-Tai Tsai
JPHCOPH, GSU, GA & GBDS, BMS, NJ

Presented at the
Biopharmaceutical Applied Statistics Symposium (BASS)

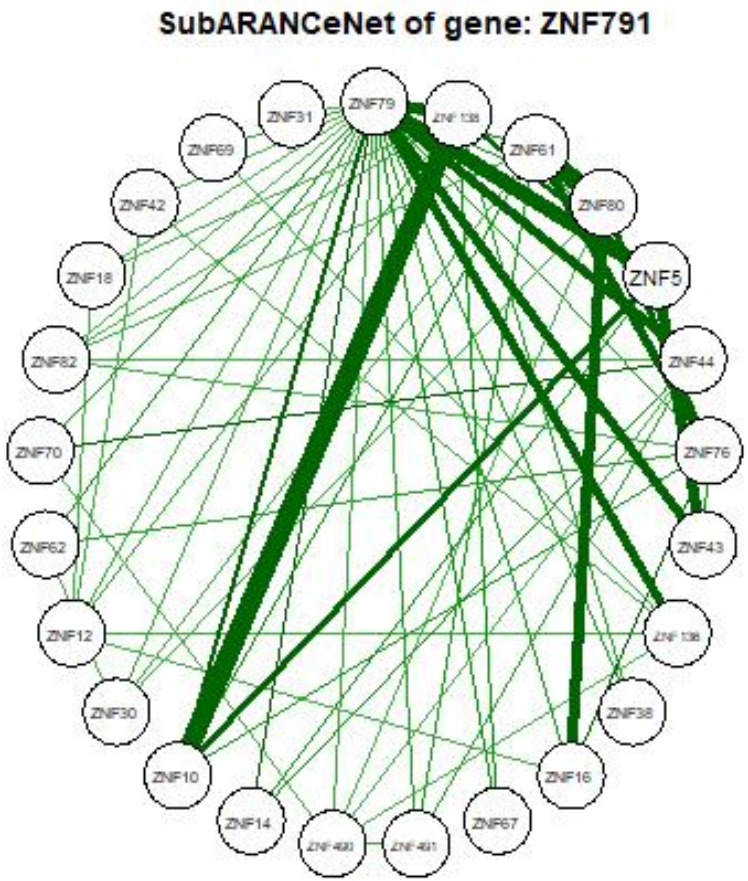
October 24, 2022

Outline of Presentation

Graphical exploration of big data - 1



Graphical exploration of big data - 2

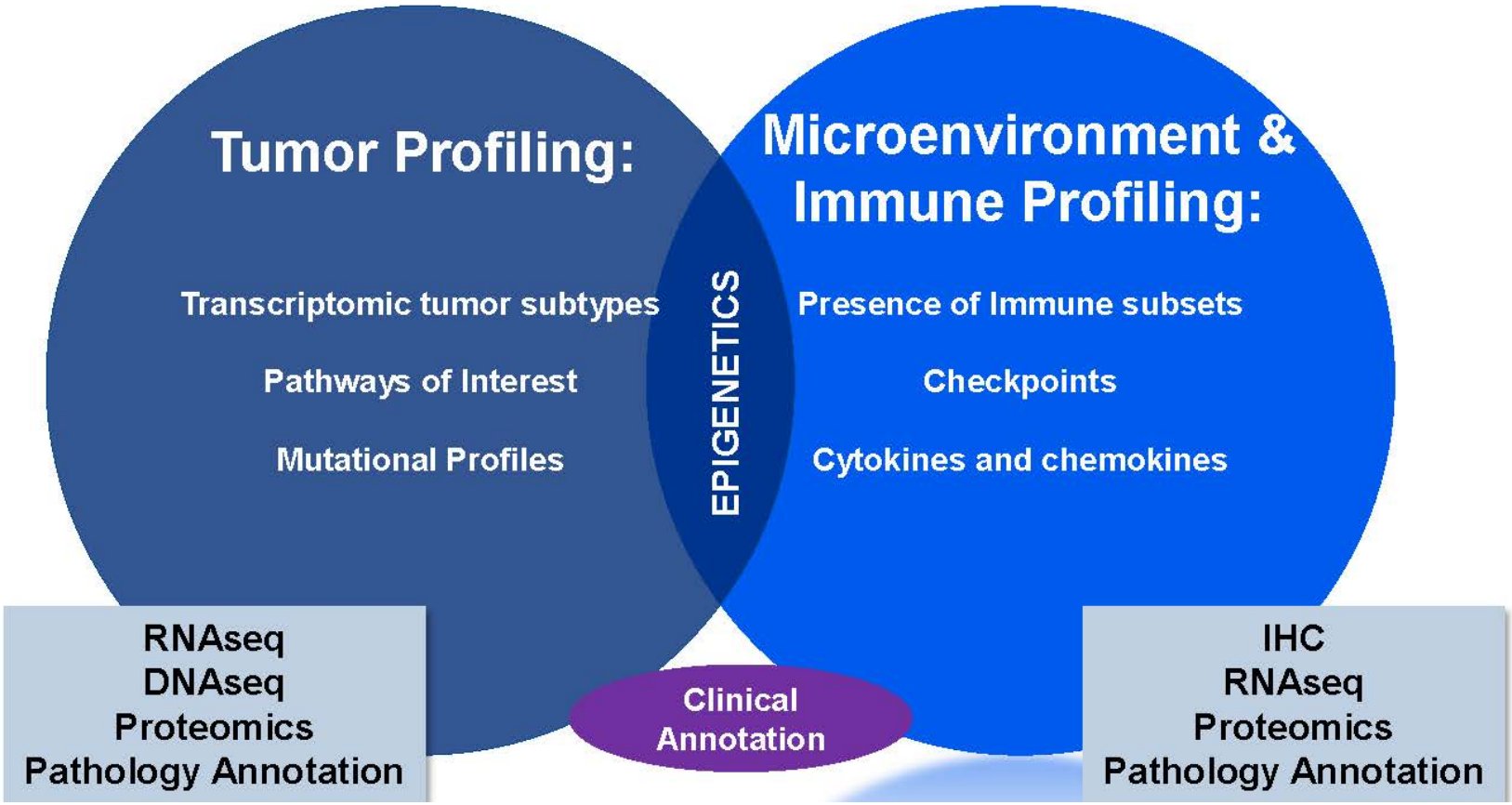


A Clinical Study on Pancreatic Cancer - 1

- A large Phase-III study was conducted to compare the efficacy of an experimental treatment (E) with standard treatment (C) as adjuvant therapy in subjects with surgically resected pancreatic adenocarcinoma (PDAC).
- Primary Objective: compare DFS and OS between treatment groups.
- Sample size: $N \approx 400$ for each treatment.
- Translational research was initiated to investigate and identify the Clinical/IHC/Genomics markers on treatment effects.

A Clinical Study on Pancreatic Cancer - 2

Biosample analysis: Goals and Data



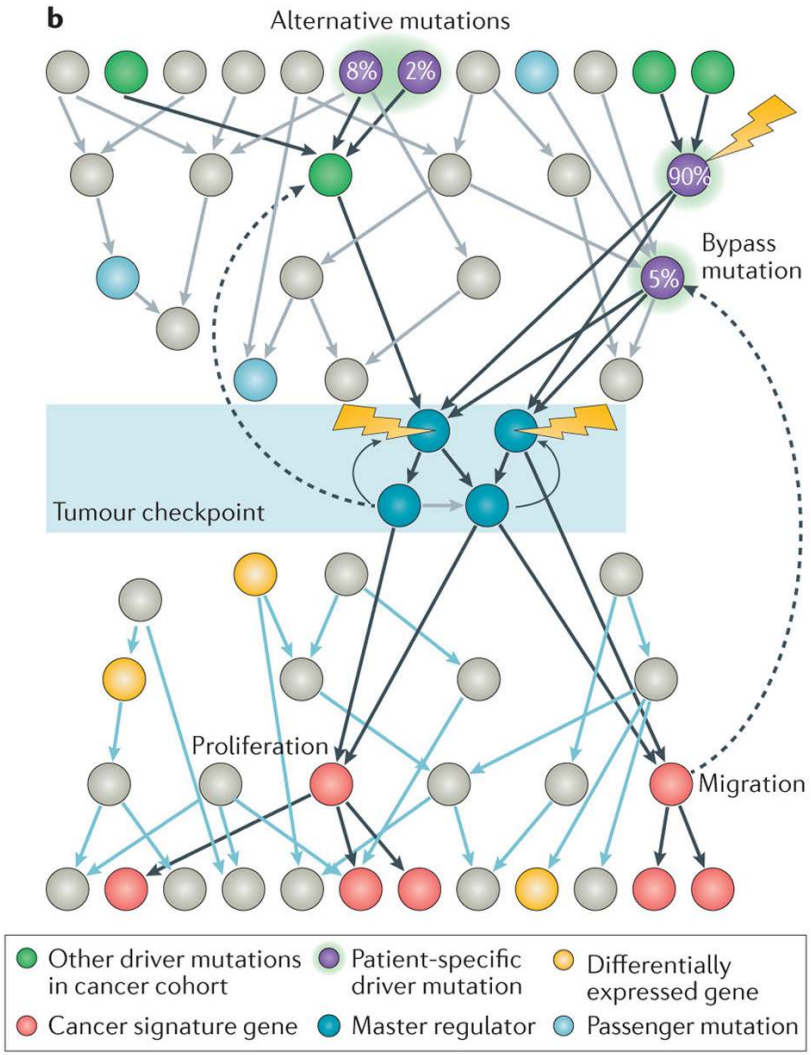
A Clinical Study on Pancreatic Cancer - 3

- From tumor samples, an extensive RNAseq data matrix $A_{(58685 \times 517)}$ was created, with $p = 58685$ genes, and $N = 517$ tumor samples (number of subjects)

Master regulators identification and checkpoint - 1

- In the paper *Califano and Alvarez, "The recurrent architecture of tumor initiation, progression and drug sensitivity" Nat Rev Cancer 2017*, they proposed a regulatory architecture implemented by master regulator (MR) proteins in tumor checkpoints.
- They explained the role of MR and its functionality.
- They also proposed an *Algorithm for the Accurate Reconstruction of Cellular Networks* (ARANCe) to identify the MRs and its regulon.

Master regulators identification and checkpoint - 2



Master regulators identification and checkpoint - 3

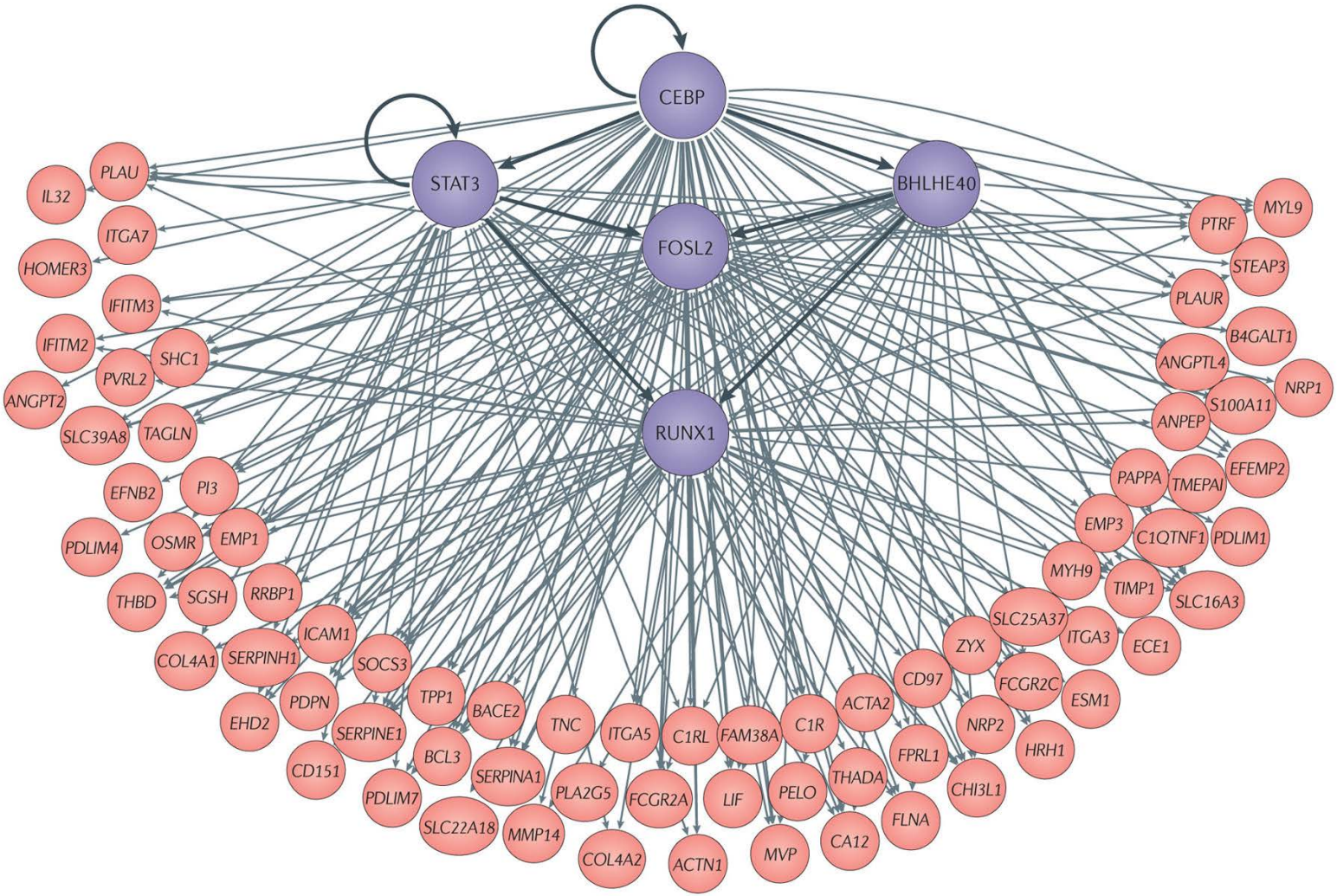


Figure 5. Tumour checkpoint architecture of the mesenchymal subtype of glioblastoma

General processes of network analysis - 1

Points to consider on genomics data analysis

- The functioning of complex biological processes very much based on the dependencies and interactions among the underlying genes and other factors.
- For example, cancer progression often involve the interaction of genomic and epigenetic factors with many external factors.
- Research has shown that genes can promote or inhibit tumor development within cell signaling pathways.
- These genes and their corresponding pathways form networks that regulate various cellular functions.

General processes of network analysis - 2

- Therefore the construction and exploration of the topology of such networks and their constituents is of great interest for developing and understanding the biological mechanisms behind disease development and progression.
- Hence it is critical to have a good measurement of the dependencies and interactions.

General processes of network analysis - 3

Measure of dependencies between genes

- Correlation: linear dependency
- Partial correlation: linear dependency excluding confounding effects
- Mutual information: general dependency based on Shannon differential entropy of multivariate joint density commonly used in information theory.
- *Note:* different measure of dependency may lead to different kind of network.

General processes of network analysis - 4

Analysis via Gaussian Graphical Model (Markov Random Field)

- For the multivariate data structure learning problem with the following multivariate Gaussian distribution:

$$p(x|\mu, \Sigma) \sim N(\mu, \Sigma), \tag{1}$$

- consider the corresponding precision matrix $Q = \Sigma^{-1}$ in equation (1), one can write

$$p(x|\mu, Q) = \frac{|Q|^{1/2}}{(2\pi)^{n/2}} \exp\Big(-\frac{1}{2}\mathbf{x}'\mathbf{Q}^{-1}\mathbf{x}\Big). \tag{2}$$

- If $Q_{ij} = 0 \Rightarrow X_i \perp X_j$, then there is no edge between nodes X_i and X_j.

General processes of network analysis - 5

Analysis via Mutual Information (MI):

- To estimate the dependence of i.i.d. samples, mutual information is a general approach and has broad applications in statistics, machine learning, and other computational sciences.
- In a typical scenario, we do not know the underlying joint distribution of the random variables and needs to be estimated.
- Based on Shannon differential entropy to estimate MI matrix:
  - *Kraskov, A., Stogbauer, H., and Grassberger, P. Estimating mutual information. Phys. Rev. E, 2004.*
  - *S. Gao, et al. Efficient estimation of mutual information for strongly dependent variables. JMLR, 2015..*

General processes of network analysis - 6

- Let $\mathbf{x} = (x_1, x_2, \cdots, x_d)$ be a continuous variable in $\boldsymbol{R}^d$
- with probability density function $p : \mathbf{R}^d \to \mathbf{R}$
- and marginal density for each x_j as $p_j : \mathbf{R} \to \mathbf{R}$ for $j = 1, 2, \cdots, d$,
- one can estimate the following Shannon Differential Entropy and Mutual Information as:

$$H(\mathbf{x}) = -\int_{R^d} \{\log p(x)\} p(x) dx, \tag{3}$$

$$I(\mathbf{x}) = \int_{R^d} \{\log \frac{p(x)}{\prod_{j=1}^{d} p_j(x_j)}\} p(x) dx, \tag{4}$$

respectively.

General processes of network analysis - 7

- Given an empirical sample and using k-nearest neighbors, the asymptotic unbiased estimate of entropy can be written as

$$\widehat{H}_{kNN,k}(\mathbf{x}) = -\frac{1}{n}\sum_{i=1}^{n}\log \hat{p}_k(x^i) - \gamma_k \tag{5}$$

- and asymptotic unbiased estimate of MI is

$$\begin{aligned}\hat{I}_{kNN,k} &= \sum_{i=1}^{n}\widehat{H}_{kNN,k}(x_i) - \widehat{H}_{kNN,k}(\mathbf{x}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\log\frac{\hat{p}_k(\mathbf{x}^{(i)})}{\hat{p}_k(\mathbf{x}_1^{(i)})\hat{p}_k(\mathbf{x}_2^{(i)})\cdots\hat{p}_k(\mathbf{x}_k^{(i)})} - (d-1)\gamma_k. \end{aligned} \tag{6}$$

where

$$\gamma_k = \frac{k^k}{(k-1)!}\int_0^\infty \log(x)x^{k-1}e^{-kx}dx.$$

General processes of network analysis - 8

- With the mutual information (MI) matrix, one can estimate the architecture of high-dimensional weighted networks (e.g., via the WGCNA[1] or ARANCe algorithm, etc.) and use `igraph` for visual presentation.
- One can also implement the MRNET [2] to keep the most relevant and minimum redundancy nodes (genes) and edges only.
- alternatively, apply `glasso` to create a sparse graph network (not recommended for highly correlated gene variables).
- *Note:* to avoid overly cluttered network graph, threshold is needed to eliminate the edges with low MI or correlation.

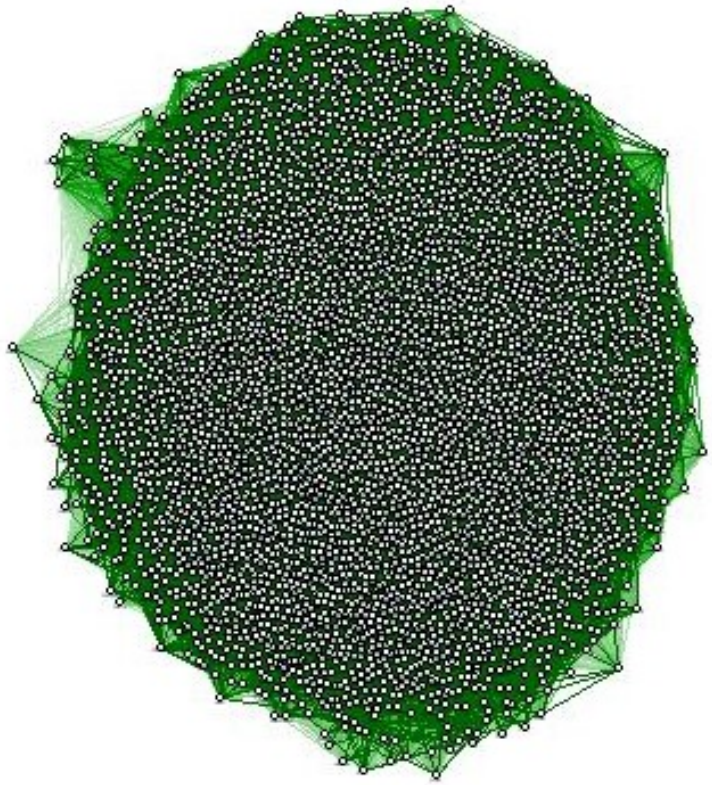
[1] WGCNA: weighted correlation network analysis

[2] H. Peng, et al., IEEE transaction 2005.

Graphical network of RNAseq data - 1

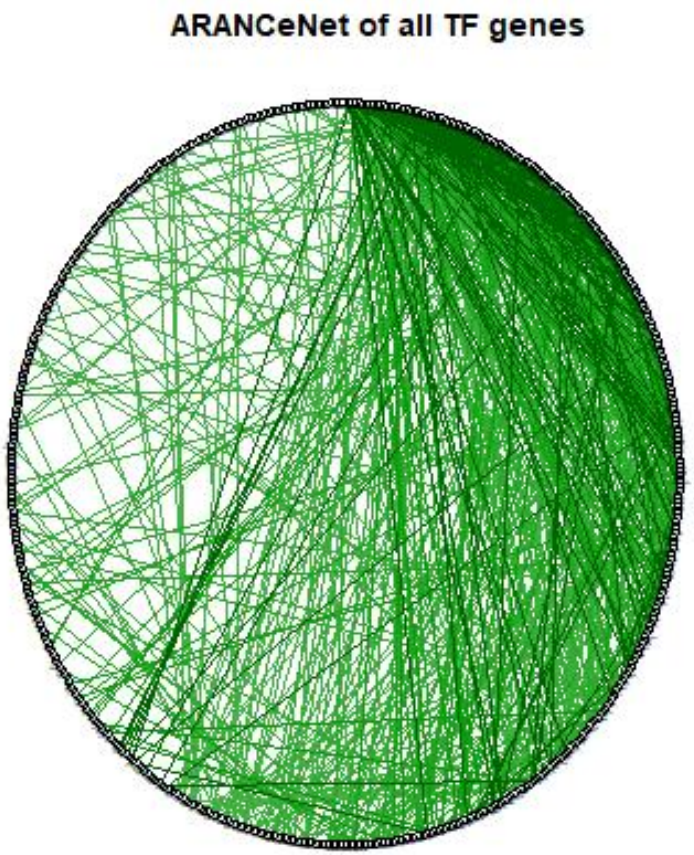
- The study RNAseq data matrix $\mathbf{A}_{(58685\times517)}$, namely, the number of genes is $p = 58685$ and $N = 517$ tumor samples (number of subjects).
- Since MRs are transcription factor (TF) (a total of 5051 through database search), the number of genes therefore was restricted to $p = 5051$. Computation is very time-consuming!

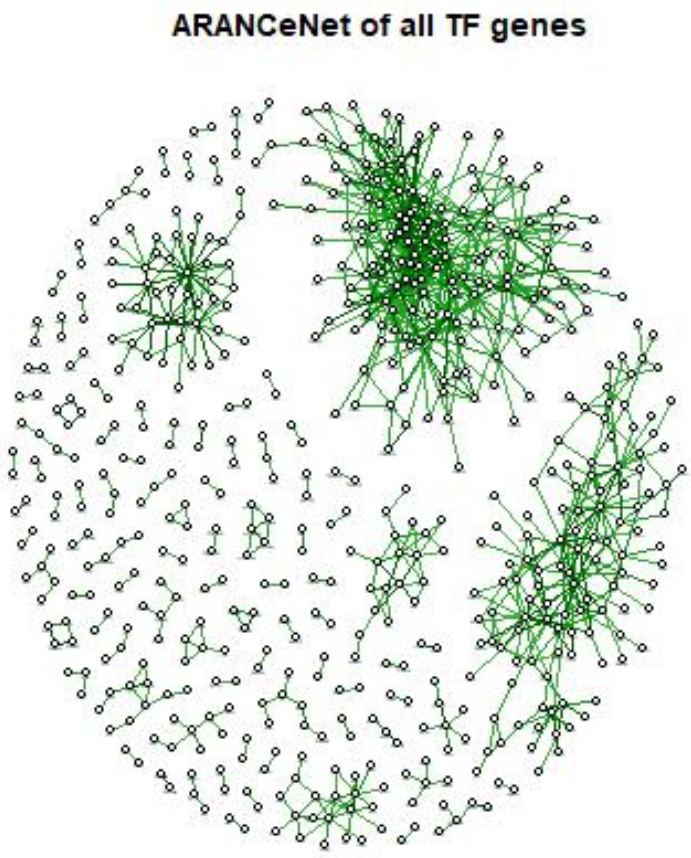
ARANCe for all master regulators and regulon



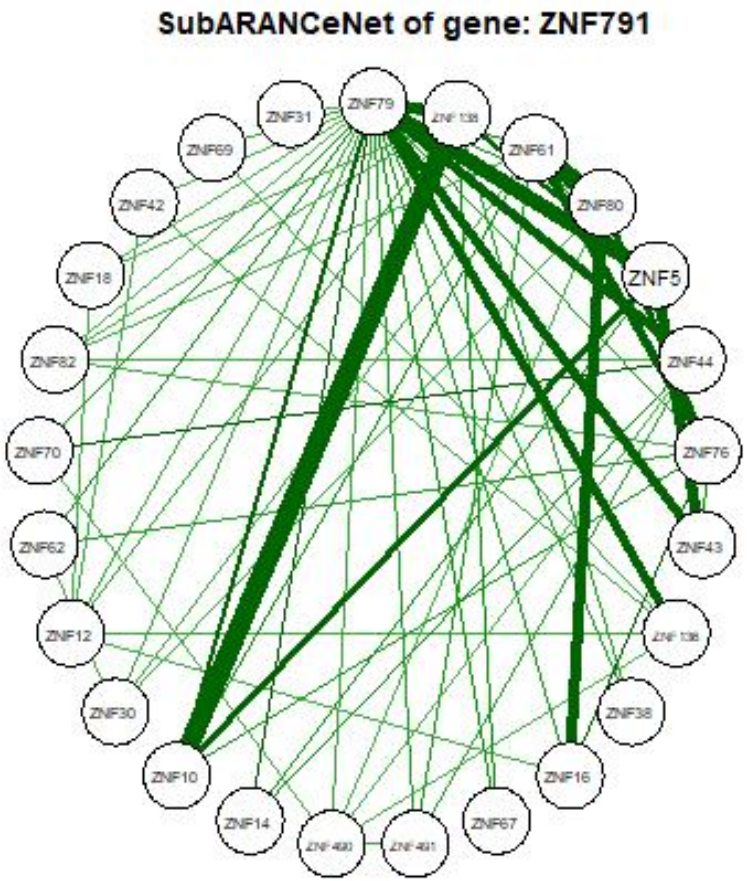
back

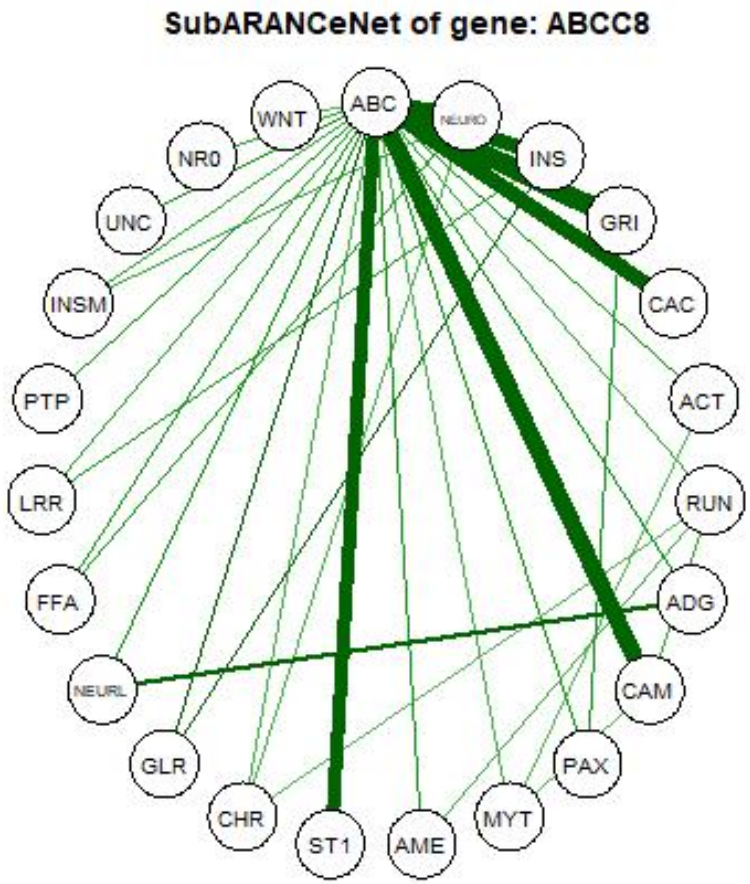
ARANNCe of all master regulators



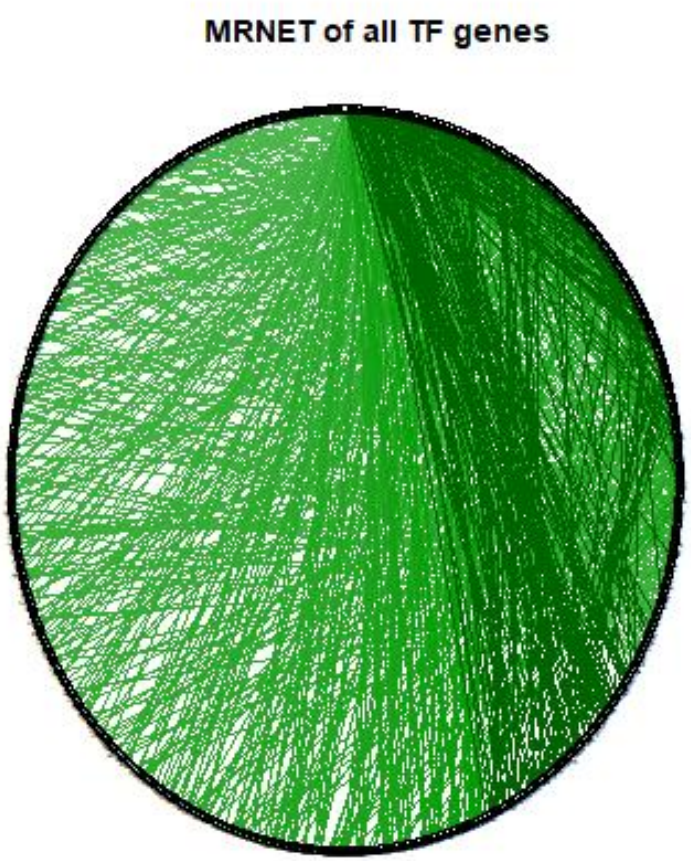


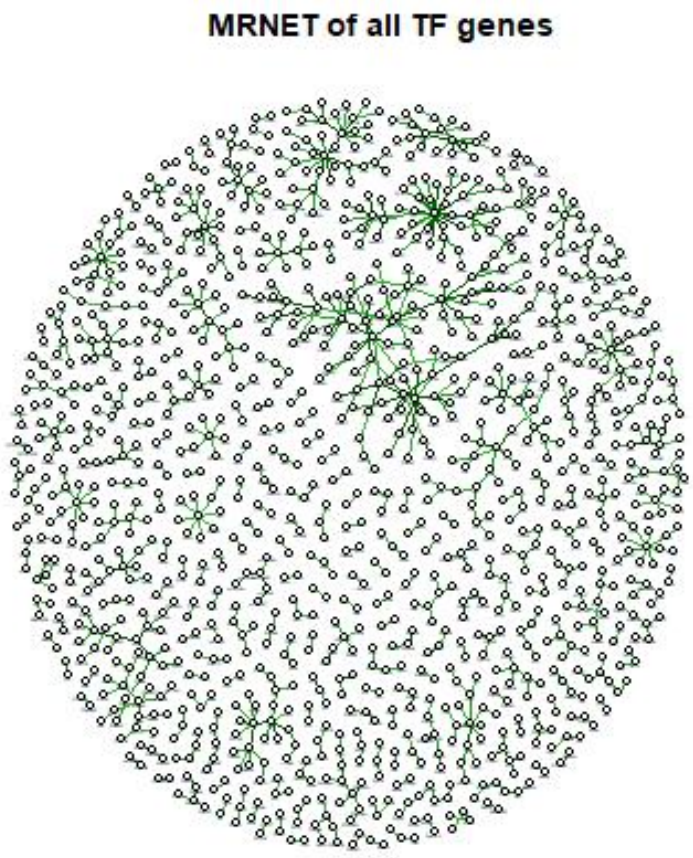
ARANCe subnetwork for master regulators and regulon



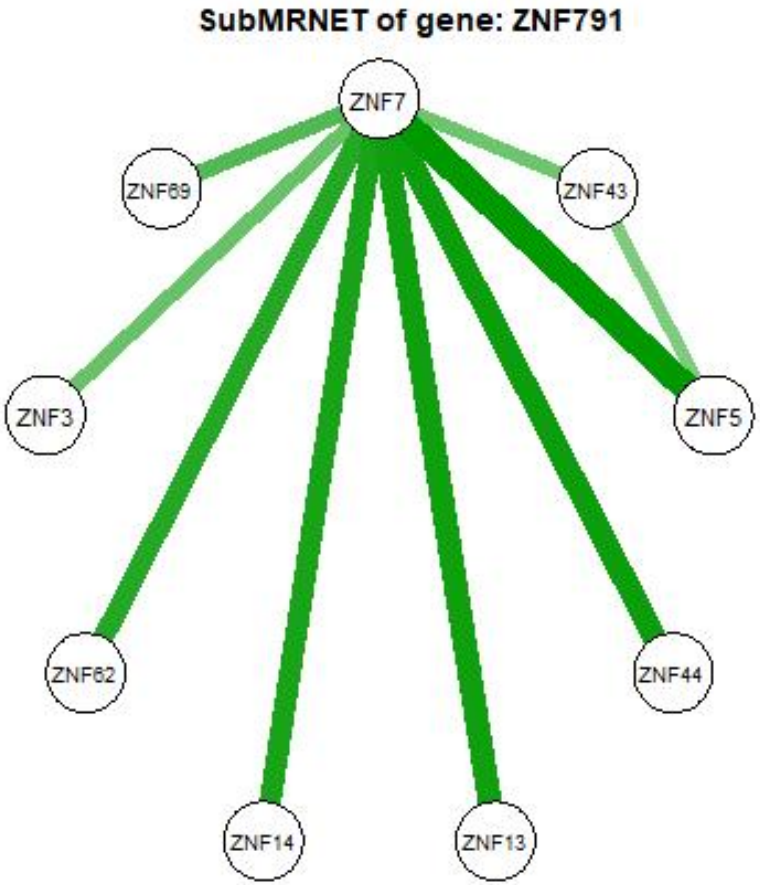


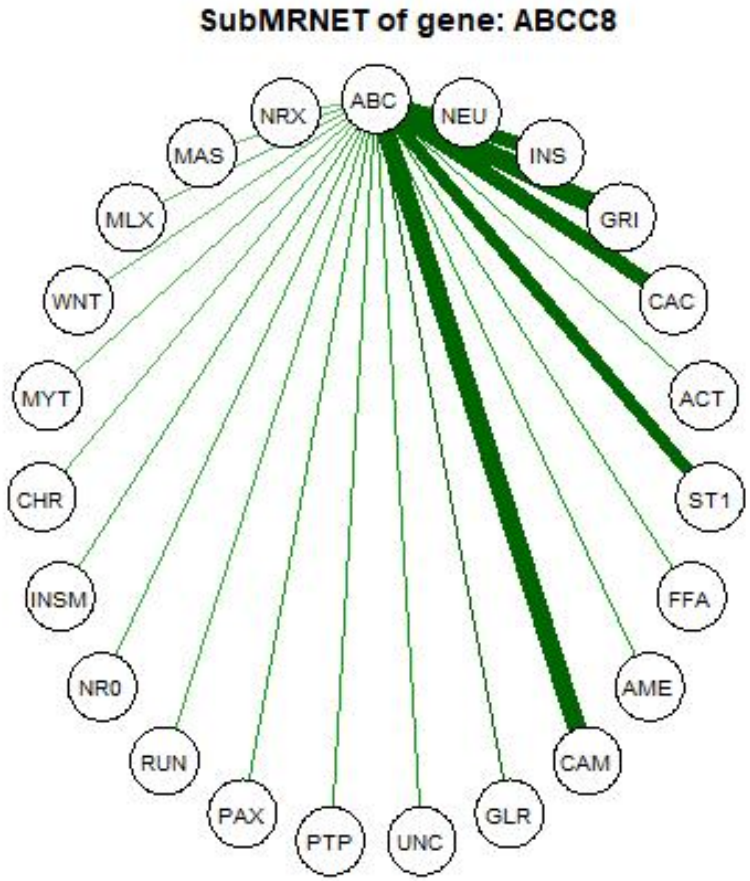
MRnet of all master regulators





MRnet subnetwork for master regulator and regulon





Analysis of Gene Sets on Overall Survival - 1

```
[1] "Hub gene is ABCC8"
             [,1]
NEUROD1 0.9349944
INS     0.9387978
GRIA2   1.0158681
CACNA1A 0.9432610
ACTL6B  0.7241400
RUNDC3A 0.6253913
ADGRV1  0.7719110
CAMK2B  0.9839202
PAX6    0.7166507
MYT1    0.5478607
AMER3   0.7109090
ST18    0.9097229
CHRNB2  0.5563474
GLRA1   0.7887109
NEURL1  0.7515493
```

```
FFAR1   0.7276207
LRRTM3  0.6695446
PTPRN   0.7235296
INSM1   0.6229934
UNC13A  0.7298692
NR0B1   0.6233546
WNT4    0.5045686
```

Analysis of Gene Sets on Overall Survival - 2

```
## Cox regression with all genes in sunetwork

            coef exp(coef) se(coef)      z Pr(>|z|)
ABCC8    -0.18615   0.83015  0.09039 -2.059  0.03946 *
NEUROD1  -0.04958   0.95163  0.16018 -0.310  0.75692
INS       0.04496   1.04598  0.07282  0.617  0.53695
GRIA2    -0.15184   0.85913  0.16441 -0.924  0.35574
CACNA1A  -0.07602   0.92680  0.12393 -0.613  0.53963
ACTL6B   -0.28628   0.75105  0.38965 -0.735  0.46251
RUNDC3A   0.15265   1.16492  0.21538  0.709  0.47847
ADGRV1    0.09139   1.09569  0.08004  1.142  0.25357
CAMK2B    0.61092   1.84213  0.18652  3.275  0.00105 **
PAX6      0.07137   1.07398  0.07382  0.967  0.33359
MYT1      0.10571   1.11150  0.13512  0.782  0.43400
AMER3     0.09423   1.09881  0.44913  0.210  0.83382
ST18     -0.24070   0.78608  0.16478 -1.461  0.14408
CHRNB2   -0.17969   0.83553  0.20133 -0.893  0.37210
GLRA1     0.33959   1.40437  0.15809  2.148  0.03171 *
NEURL1   -0.08629   0.91733  0.11179 -0.772  0.44019
FFAR1    -0.39707   0.67229  0.21424 -1.853  0.06383 .
```

Analysis of Gene Sets on Overall Survival - 3

LRRTM3	-0.20119	0.81776	0.18278	-1.101	0.27102
PTPRN	0.07447	1.07731	0.08492	0.877	0.38052
INSM1	-0.07349	0.92915	0.34037	-0.216	0.82905
UNC13A	0.09205	1.09642	0.09163	1.005	0.31510
NR0B1	0.04498	1.04601	0.15718	0.286	0.77473
WNT4	0.02661	1.02697	0.08548	0.311	0.75554

Analysis of Gene Sets on Overall Survival - 4

```
## Cox regression using lasso with all genes in subnetwork

ABCC8   -0.02998745
NEUROD1  .
INS      .
GRIA2    .
CACNA1A  .
ACTL6B   .
RUNDC3A  .
ADGRV1   .
CAMK2B   .
PAX6     .
MYT1     .
AMER3    .
ST18    -0.01583754
CHRNB2   .
GLRA1    .
NEURL1   .
FFAR1    .
LRRTM3   .
```

Analysis of Gene Sets on Overall Survival - 5

```
PTPRN     .
INSM1     .
UNC13A    .
NROB1     .
WNT4      .
```

Analysis of Gene Sets on Overall Survival - 6

Gene set variation analysis (GSVA) for RNA-Seq data analysis

- To investigate whether the genes in a gene set are more likely to be highly expressed, therefore have higher ranks among genes, and can be found at either tail of the rank distribution, the Kolmogorov-Smirnov type of random walk statistic is used to construct the measures.
- Let x_{ij} be gene i of the jth sample from the RNA gene sequence profile, define the normalized scores as

$$z_{ij} = (1/n) \sum_{k=1}^{n} \int_{-\infty}^{(x_{ij}-x_{ik})/h_i} (1/\sqrt{2\pi}) \exp(-t^2/2) dt.$$

Analysis of Gene Sets on Overall Survival - 7

- Convert z_{ij} to ranks $z_{(i)j}$ for each sample j and normalize further $r_{ij} = |p/2 - z_{(i)j}|$ to make the ranks symmetric around zero.
- Define the Kolmogorov-Smirnov type of random walk statistic for the kth gene set as

$$w_{jk}(l) = \frac{\sum_{i=1}^{l} |r_{ij}|^{\tau} I(g_{(i)} \in \gamma_k)}{\sum_{i=1}^{p} |r_{ij}|^{\tau} I(g_{(i)} \in \gamma_k)} - \frac{\sum_{i=1}^{l} I(g_{(i)} \notin \gamma_k)}{(p - |\gamma_k|)} \qquad (7)$$

where

  - τ is a parameter describing the weight of the tail in the random walk,
  - $I(.)$ is the indicator function, γ_k is the kth gene set,
  - $|\gamma_k|$ is the number of genes in the kth gene set, and
  - p is the size of the gene population.

Analysis of Gene Sets on Overall Survival - 8

- The enrichment score (GSVA) is defined as the maximum deviation from zero (since the ranks are centralized at the mid-rank) of the random walk of the jth sample with respect to the kth gene set:

$$ES_{\max}(jk) = \max_{l=1,\cdots,p} \{w_{jk}(l)\}.$$

- Conceptually, GSVA transforms a p-gene by n-sample gene expression matrix into a g-gene set by n-sample pathway enrichment matrix. This facilitates many forms of statistical analysis in the 'space' of pathways rather than genes, providing a higher level of interpretability.

Analysis of Gene Sets on Overall Survival - 9

```
## Cox regression using MR only

Call: coxph(formula = Surv(AVAL, CNSR) ~ xdat4$ABCC8, data = xdat4)
n= 515, number of events= 311
                coef exp(coef) se(coef)      z Pr(>|z|)
xdat4$ABCC8 -0.07011   0.93229  0.02324 -3.017  0.00255
            exp(coef) exp(-coef) lower .95 upper .95
xdat4$ABCC8    0.9323      1.073    0.8908    0.9757

------------------------------------------------------------
## Cox regression using GSVA score of all genes in subnetwork

Call: coxph(formula = Surv(AVAL, CNSR) ~ t(zzgsva_es), data = xdat4)
n= 515, number of events= 311
                 coef exp(coef) se(coef)      z Pr(>|z|)
t(zzgsva_es) -0.18442   0.83158  0.08805 -2.095   0.0362
             exp(coef) exp(-coef) lower .95 upper .95
t(zzgsva_es)    0.8316      1.203    0.6998    0.9882
```

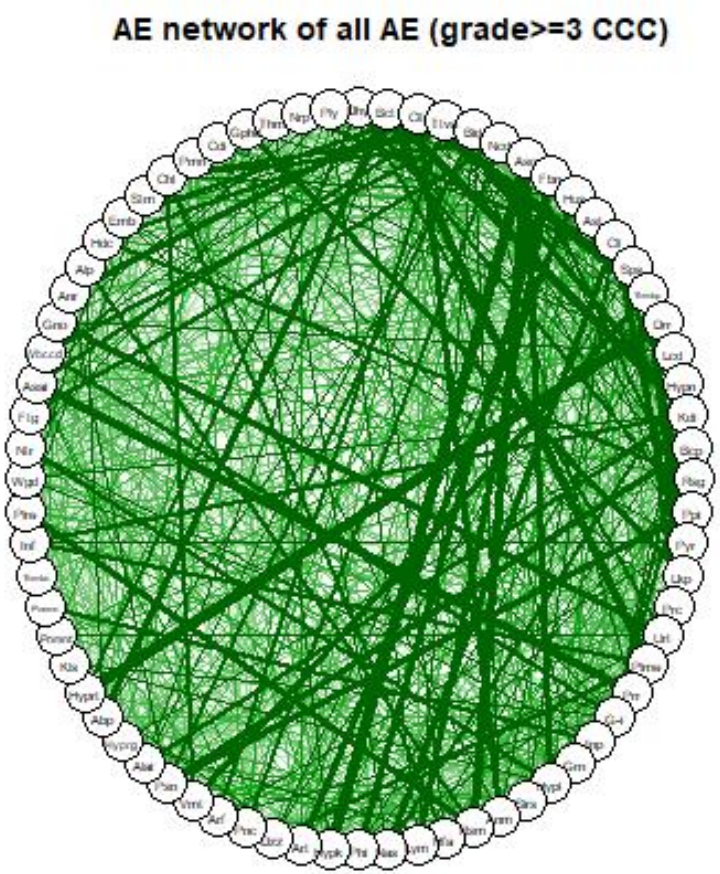
Gene signature selections for clinical development - 1

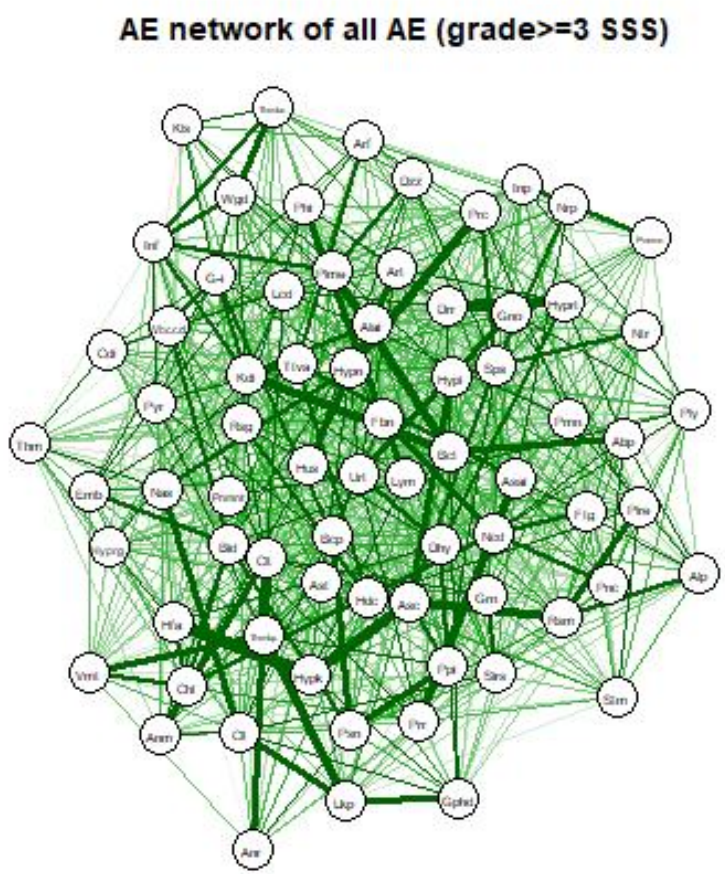
- A total of about 1200 MR with their regulons were identified based on the MI network analysis.
- Each MR (or MR+regulons) were further analyzed based on their potential causal-effect on clinical outcomes.
- A candidate set of genes were selected (using R-package `randomForest` or something similar) for further test and development.

Network analysis of AEs - 1

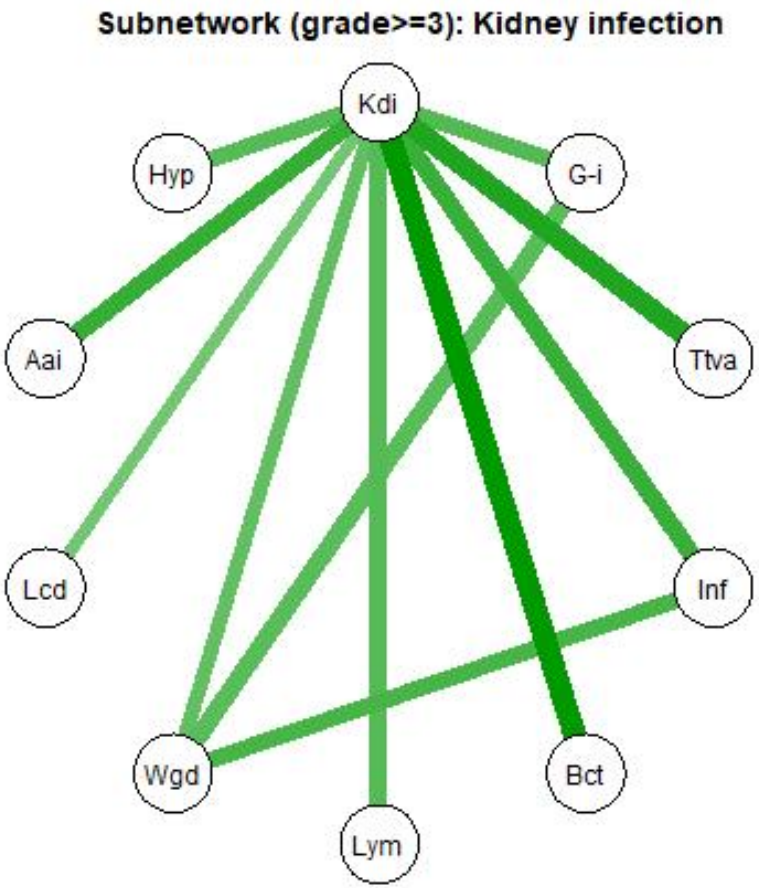
- Among all the study subjects, 318 subjects had SAE ($Grade \geq 3$) with 73 different SAEs.
- Conventional analysis of AE data is table summary of all SOC and PT without consideration of correlation between AEs.
- Network analysis provides a convenient approach to aggregate and show the correlation between AEs.
- The display of networks provides a convenient and insightful way to understand the prevalence and relations between AEs.

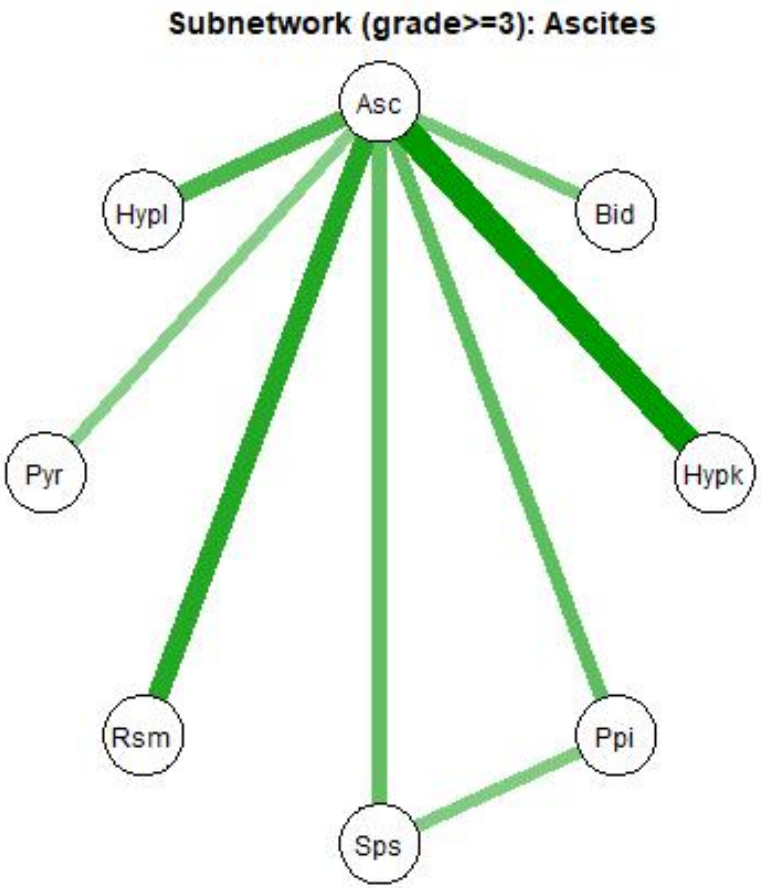
Network for all SAEs





Subnetworks for Kidney Infection and Ascites





Network analysis for adverse events - 1

```
[1] "Hub gene is Kidney infection"
Gamma-glutamyltransferase increased 0.1219713
Toxicity to various agents          0.1589390
Infection                           0.1411710
Bacteraemia                         0.1816202
Lymphopenia                         0.1206374
Weight decreased                    0.1117944
Lymphocyte count decreased          0.1000401
Alanine aminotransferase increased  0.1426728
Hyperglycaemia                      0.1196140

[1] "Hub gene is Ascites"
Blood iron decreased      0.1156233
Hypokalaemia              0.2136131
Post procedural infection 0.1348318
Sepsis                    0.1312286
Rash macular              0.1837402
Pyrexia                   0.1007323
Hypoalbuminaemia          0.1514079
```

Summary/Discussion - 1

- To achieve the vision of Precision Medicine, gene-based drug discovery is critical.
- Even though RNA sequence provides huge amount of data, how to extract useful information therein is still a big challenge.
- Network analysis is one of the many techniques within the Machine Learning framework to conduct EDA to better understand the data and its inherited complex structure, much work are still needed.
- In order to enhance our knowledge and conduct more fruitful drug discoveries, more collaboration between biologists and statisticians are desperately needed.