Statistical Evaluation for Methods of Gene-set Analysis with Multivariate Non-normal Scenarios

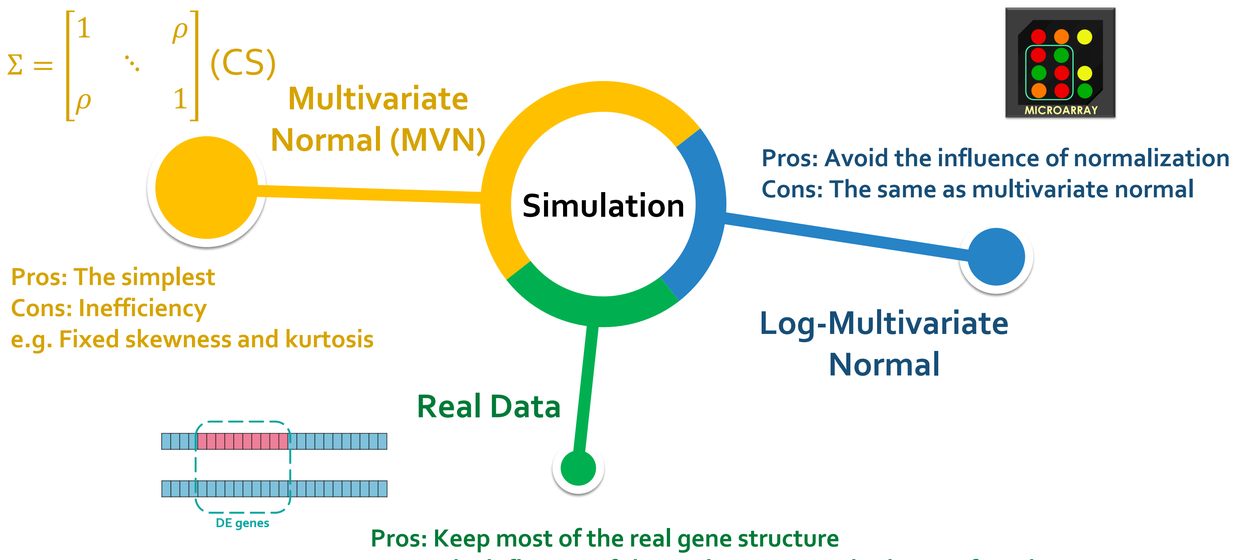
以非常態情境評估基因集合分析方法在真實基因資料下之表現研究

Presenter: Chi-Hsuan Ho Advisor: Prof. Chuhsing Kate Hsiao

Introduction

As the technology improves, more and more statistical methods for gene-set analysis (GSA) are developed to find pathogenic pathways and genes. In recent years, many studies use the multivariate normal (or lognormal) distribution in simulation studies to evaluate the performance of these GSA methods. However, the normality assumption for the gene expression data has been questionable. Therefore, here we first reveal the non-normality of the real data, and then propose a statistical evaluation method to compare five well-known GSA methods via radar plots.

- **Simulation methods used in current GSA**



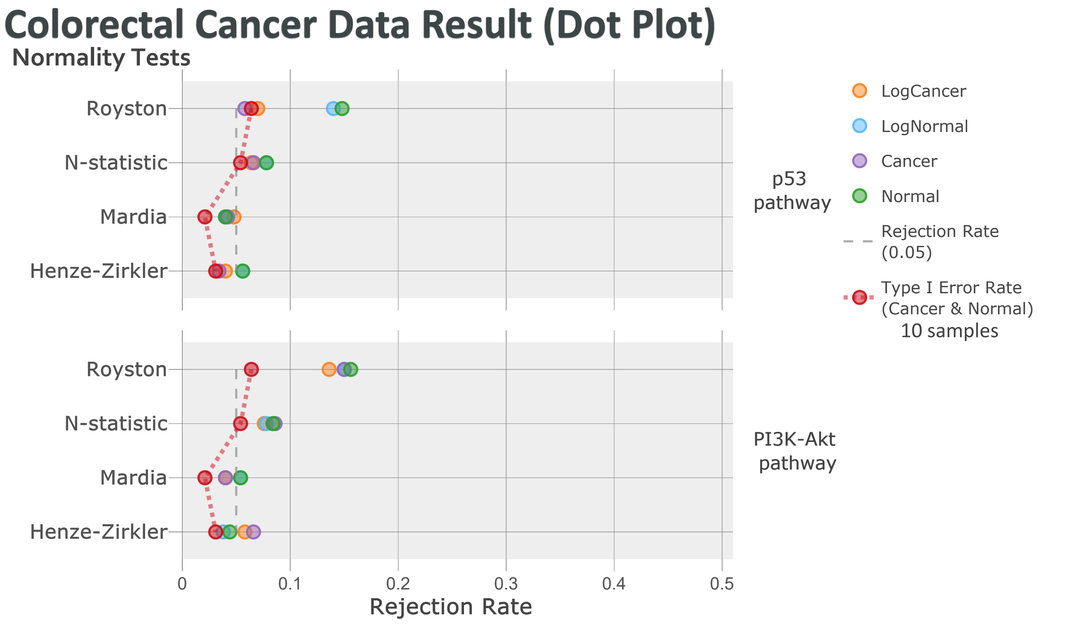
Non-normality of Real Data

- **Data Processing**

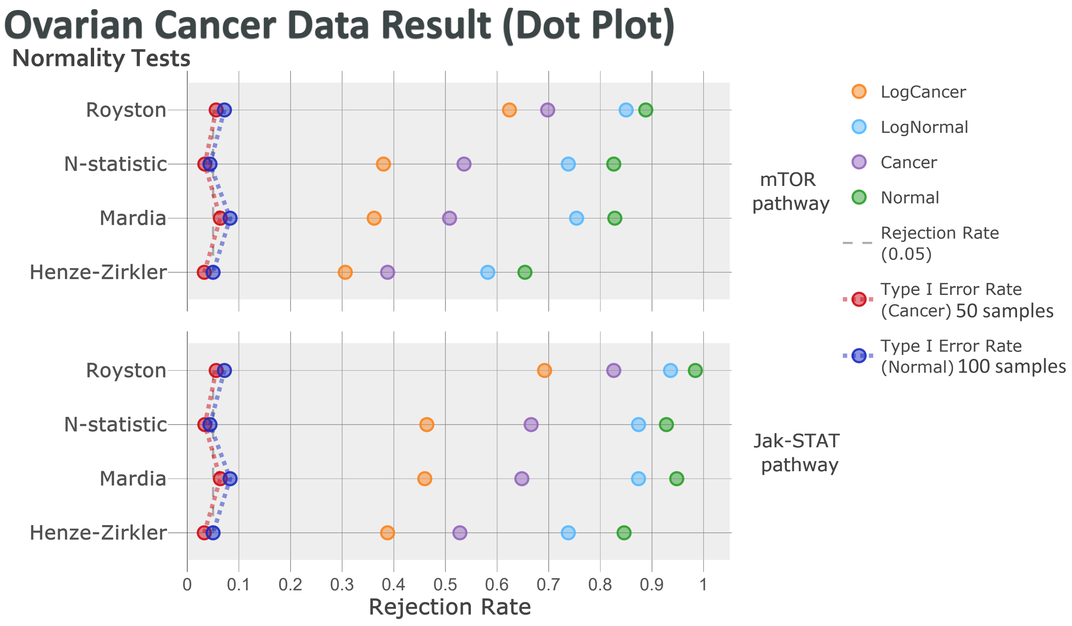
Raw Data (NCBI GEO/TCGA)
↓
RMA (Robust Multi-array Average)
↓
Quantile Normalization
↓
Disease-related Pathways Data Selection (GSA)
↓
Homogeneous Clinical Data Management

Colorectal Cancer	Ovarian Cancer
•CS/ADJ	•CS/CN
•Size: 11	•Size:47/116
•Pathway: 5	•Pathway: 4
•Wnt	•PI3K-Akt
•RAS	•mTOR
•PI3K-AKT	•Jak-STAT
•P53	•p53
•TGF-β	
Breast Cancer	Lung Cancer
	Leukemia

- **Colorectal Cancer Data Result (Dot Plot)**



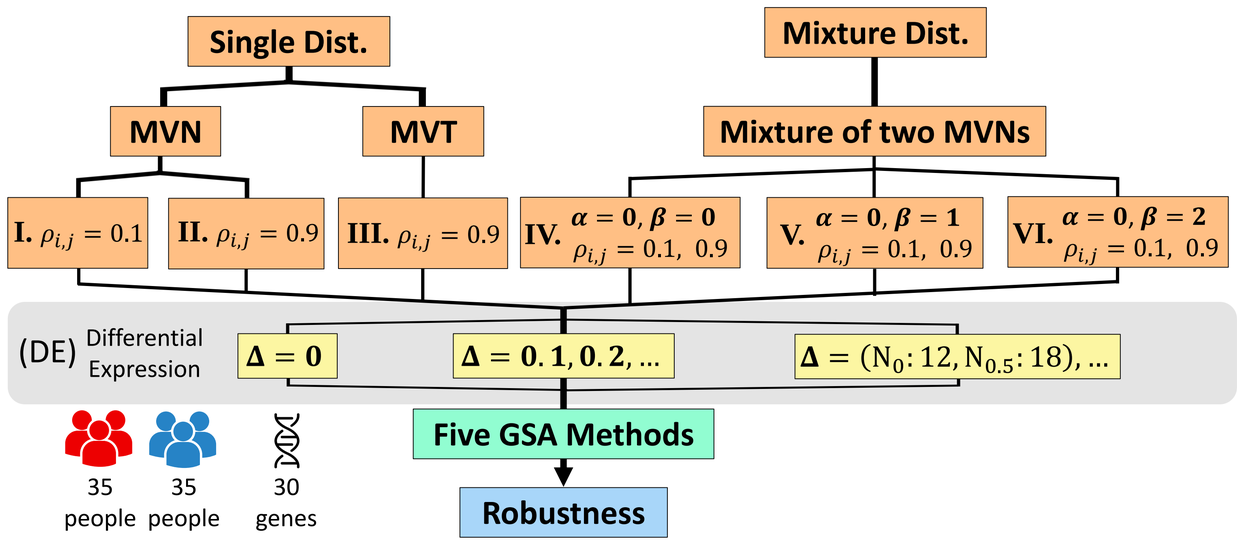
- **Ovarian Cancer Data Result (Dot Plot)**



- We sampled 4 genes from each pathway to test normality (repeat 500 times).
- Gene expression data are unlikely normally (log-normally) distributed.
- The test result of colorectal cancer data set is influenced by sample size.

Statistical Evaluation for GSA Methods

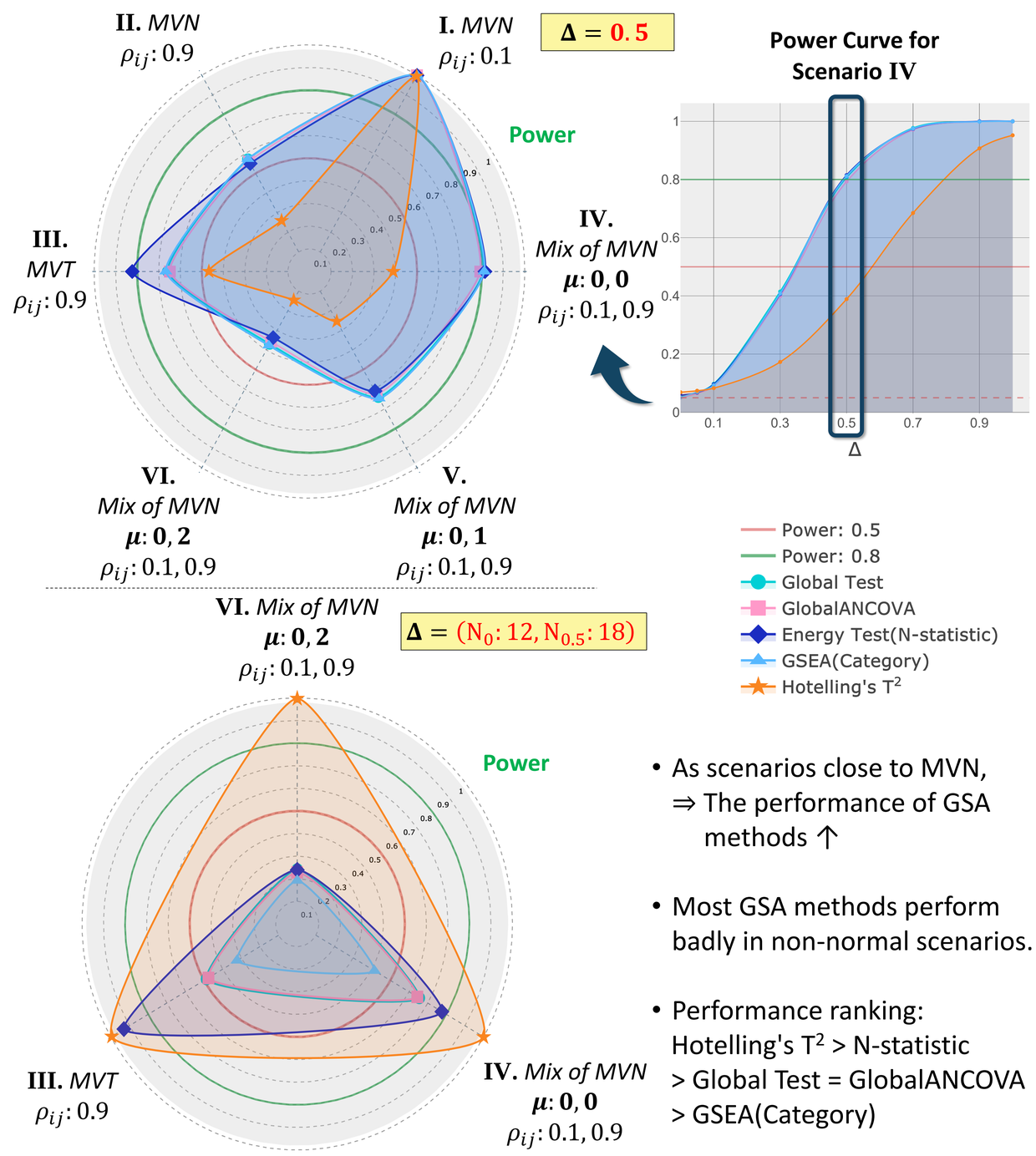
- **Evaluation Procedure:**



- **Methods of Gene-set Analysis (GSA Methods)**

Distance-based	Regression-based	Other
$D(dist._{CS},\ dist._{CN})$	• **The Global test**	• **GSEA (Category)**
• **Hotelling's T^2**	→ Score test for detecting the random effect of each gene (Y: Phenotype, X: Genes)	→ Self-contained version of GSEA (Running-sum statistic: Per-gene t-statistic)
→ Multidimensional version of two-sample t-test (Mahalanobis distance)	• **The GlobalANCOVA**	
• **Energy Test (N-statistic)**	→ Linear regression model for each gene (RSS) (Y: Genes, X: Confounder)	
→ Detecting the Euclidean distance		

- **Evaluation Results (Radar Plots)**



- As scenarios close to MVN,
  ⇒ The performance of GSA methods ↑
- Most GSA methods perform badly in non-normal scenarios.
- Performance ranking:
  Hotelling's T^2 > N-statistic > Global Test = GlobalANCOVA > GSEA(Category)

Conclusions & Discussion

- Three elements in our evaluation method are important:
  1. Correlations between genes (ρ_{ij})
  2. The choice of scenarios
  3. The structure of difference vector ($\boldsymbol{\Delta}$)
- Some advice for researchers to develop GSA methods:
  1. Distribution-free methods are preferred.
  2. Binary phenotypes: Multivariate distance-based methods perform better than regression-based methods.
- Some settings can be considered in the future:
  1. Real data scenarios
  2. Copula-based scenarios
  3. DE ⇒ Correlation matrix or distribution