Dados de Câncer de Ovário

Dados de câncer de ovário simulados a partir do projeto TCGA (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga). Para as mesmas unidades amostrais, contém informações de 4 bancos de dados: CNV (Copy Number Variation), Metilação, Expressão Gênica e Proteica. As unidades amostrais correspondem a pacientes com tempo de sobrevivência inferior a 3 anos (casos – outcome = 1) e superior a 3 anos (controle – outcome = 0). Os dados foram gerados supondo a presença de um loco cromossômico (eQTM) com nível de metilação diferente para casos e controles, o qual influencia os genes LRIG1,TCEAL8 e MARCH9 (Figura 4 abaixo). A simulação dos casos e controles foi feita condicionalmente à expressão gênica destes três genes e do LRRN4. Foram simulados dados para 1000 pacientes, sendo 500 casos e 500 controles. Estão disponíveis 100 réplicas desse cenário de simulação, dentre as quais foi selecionada uma para cada aluno de MAE5776-2020.

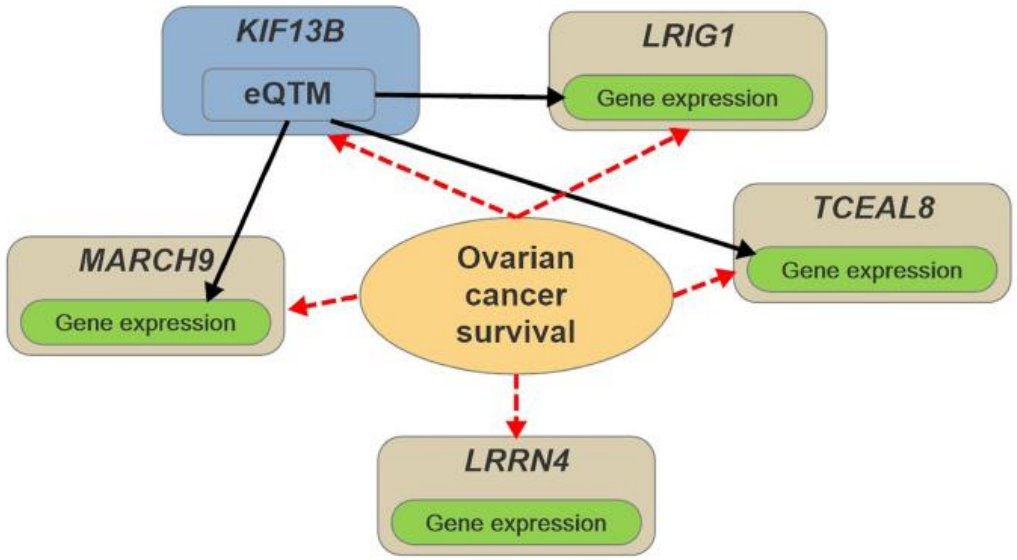

Figure 4: Hypothetical model for the survival time (short-term and long-term) of OV. The black solid arrows represent the regulatory effects of the eQTM on gene expression. The red dotted arrows represent the retrospective simulations of the methylation and gene expression levels conditional on the survival status.

- **CNV:** informação de CNV para 2884 regiões. As variáveis estão codificadas como -2, -1, 0, 1 e 2. Os valores negativos indicam a perda de duas ou uma cópia da região cromossômica, os valores positivos indicam o ganho de duas ou uma cópia e o valor nulo indica que a região cromossômica é normal.
- **Exp:** informação da intensidade de expressão gênica dos genes LRIG1, TCEAL8, MARCH9, LRRN4 e 2000 outros.
- **NorExp:** dados da intensidade de expressão gênica normalizados, em que foi eliminado ruídos aleatórios.
- **Methy:** dados de metilação de 2752 locais cromossômicos, além do eQTM. Os dados indicam o percentual de metilação em cada local.
- **Protein:** valores da expressão proteica normalizada para os mesmos genes apresentados na aba de expressão gênica.

Referência : Ren Hua Chung and Chen Yu Kang. A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification. *Giga Science* 8(**5**): 1–12, 2019. ISSN 2047217X. doi: 10.1093/gigascience/giz045.

Análise dos Dados de Câncer de Ovário

A visualização de dados é uma das etapas mais importantes, se não a mais importante, na análise de dados. Como é possível representar dados de alta dimensão em baixa dimensão (uni ou bidimensional)? Que propriedades ótimas das linhas bem como das colunas da matriz de dados estão preservadas nessas representações?

Considere a réplica dos dados de câncer de ovário fornecida e obtenha tais representações a partir das análises de:

1. Componentes Principais (ou Coordenadas Principais)

2. Análise Discriminante

3. Análise de Agrupamento (ou HeatMap)

4. Mínimos Quadrados Parciais (PLS)

5. Análise de Correlação Canônica

Em cada caso, selecione (livremente) uma partição dos dados para realizar a análise, justifique sua escolha e interprete os resultados. Os grupos de pacientes caso e controle foram bem discriminados na análise? Quais variáveis mais contribuíram na análise?

Produza um **Relatório Científico** da análise realizada.

Tabela com a definição das Réplicas a serem analisadas por cada aluno:

No. USP*	**Réplica do BD de Câncer de Ovário**
9756559	Réplica 1
11936752	Réplica 2
7550848	Réplica 3
11571801	Réplica 4
9812795	Réplica 5
11934872	Réplica 6
10380081	Réplica 7

*Os demais alunos podem escolher livremente uma réplica para analisar.