IBI5086 Métodos Estatísticos em Bioinformática **2º Sem/2023**

LISTA #01 e LISTA #02

Dados de Câncer de Ovário

Dados de câncer de ovário foram simulados a partir do projeto TCGA (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga).

No link (https://figshare.com/articles/Ovarian_cancer_profile_for_OmicsSIMLA/7763627) é posível baixar os dados.

No link (https://omicssimla.sourceforge.io/simuomicsTCGA.html) há uma explicação do que significa cada coluna de cada banco de dados.

São dados simulados: há 50 réplicas dos mesmos dados (em *.batch1 estão as réplicas de 01 a 15; em *.batch2 estão as restantes, de26 a 50). Detalhes sobre o simulador de dados usado estão em Chung e Kang (2019). Para as mesmas unidades amostrais, estão disponíveis informações de 4 bancos de dados: CNV (Copy Number Variation), Metilação, Expressão Gênica (dados originais e normalizados) e Proteína. As unidades amostrais correspondem a pacientes com tempo de sobrevida inferior a 3 anos (casos – outcome = 1) e superior a 3 anos (controle – outcome = 0). Os dados foram gerados supondo a presença de uma região cromossômica (eQTM) com nível de metilação diferente para casos e controles, a qual influencia os genes LRIG1,TCEAL8 e MARCH9 (Figura 4 abaixo extraída do artigo de Chung e Kang, 2019). A simulação dos casos e controles foi feita condicionalmente à expressão gênica destes três genes e do LRRN4. Foram simulados dados para 1000 pacientes (amostras balanceadas, sendo 500 casos e 500 controles). Estão disponíveis 50 réplicas do mesmo cenário de simulação, dentre as quais foi selecionada uma a ser analisada por cada grupo de IBI5086-2023.

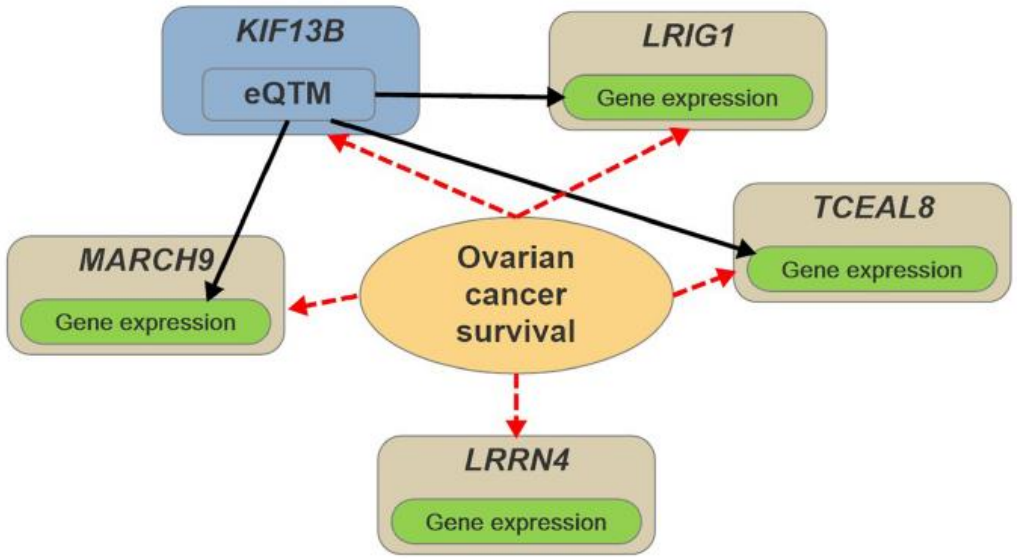

Figure 4: Hypothetical model for the survival time (short-term and long-term) of OV. The black solid arrows represent the regulatory effects of the eQTM on gene expression. The red dotted arrows represent the retrospective simulations of the methylation and gene expression levels conditional on the survival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

A seguinte notação foi adotada aos bancos de dados:

- **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 de 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 foram eliminados 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 considerados nos dados 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.

Tabela com a definição das Réplicas a serem analisadas por cada grupo*:

No. Grupo*	Réplica do BD de Câncer de Ovário
1	Réplica 10
2	Réplica 21
3	Réplica 13
4	Réplica 9
5	Réplica 42
6	Réplica 11
7	Réplica 32

*A composição dos grupos é livre, sendo recomendado entre 5 a 8 alunos.

Análise dos Dados de Câncer de Ovário

LISTA #01- Análise de uma Variável Resposta Quantitativa: Expressão Gênica

1. Sem considerar a separação dos grupos caso e controle, construa um gráfico com os **boxplots** dos dados de expressão gênica (Expr) de cada fragmento avaliado. Faça o mesmo para representar os dados de expressão normalizados (NorExpr). Comente sobre o efeito da normalização na distribuição das respostas.

2. Para cada fragmento genômico, compare as médias das respostas de expressão (Expr) de acordo com os grupos caso e controle (use, por exemplo, testes t ou de Wilcoxon, justificando). Apresente os resultados em um **gráfico vulcão** e em um **gráfico Manhattan**. Faça o mesmo para as respostas de expressão normalizadas (NorExpr). Há evidência de fragmentos com expressão diferencial significante entre os grupos? Justifique. Adote algum tipo de correção para múltiplos testes.

Gráfico vulcão é um gráfico de dispersão da estimativa da diferença entre as médias, no eixo das abscissas, pela significância dessa diferença na escala –log10(valor p), no eixo das ordenadas).

Gráfico Manhattan é um gráfico de dispersão, com -log10(valor-p), na ordenada, versus o índice do fragmento, na abscissa.

Análise dos Dados de Câncer de Ovário

LISTA #02- Análise de uma Variável Resposta Binária: Grupos Caso e Controle

Considere os dados da Réplica analisada pelo seu grupo na Lista #01.

1. Escolha uma dentre as 2004 expressões gênicas, digamos a MARCH9. Codifique essa variável como, MARCH9_cat=0 se MARCH9<Mediana(MARCH9), e MARCH9_cat=1 se MARCH9≥Mediana(MARCH9). Construa a tabela de contingência 2x2 com a distribuição de Casos e Controles de acordo com MARCH9_cat.

(i) Calcule (à mão) a estatística razão de chances (odds ratio, OR) e interprete.

(ii) Realize o teste Qui-Quadrado de associação entre a expressão gênica categorizada e os grupos Caso e Controle. Há evidência de associação significante?

(ii) Ajuste um modelo de regressão logística para análise dos dados na correspondente tabela 2x2 construída. Construa o intervalo de 95% de confiança para o OR e discuta a significância da associação.

2. Faça a mesma análise considerando todas as variáveis de expressão gênica avaliadas. Apresente os resultados em um **gráfico vulcão** e em um **gráfico Manhattan**. Há evidência de associação entre os grupos Caso e Controle e as variáveis de expressão (categorizadas)? Justifique. Adote algum tipo de correção para os múltiplos testes.

Gráfico vulcão, neste caso, é um gráfico de dispersão da estimativa de e^{β} (β é o coeficiente de regressão associado ao efeito da expressão gênica categorizada no logito), no eixo das abscissas, pela significância dessa diferença na escala –log10(valor p), no eixo das ordenadas.

Gráfico Manhattan é um gráfico de dispersão, com -log10(valor-p), na ordenada, versus o índice do fragmento de expressão categorizada, na abscissa.

3. Discuta e compare as análises realizadas na Lista#01 com a Lista#02.