Faculdade de Medicina de Ribeirão Preto
Programa de pós graduação em ginecologia e obstetrícia (PGGO)
RGO 5872- Introdução à Biologia Computacional e Análise Estatística em R para a Área da Saúde e Ciências Biológicas

"Análise estatística descritiva e inferencial do câncer de cólon (COAD)"

Componente: Anna Virgínia Bertelle Borges
Professor: Daniel Guimarães Tiezzi

Ribeirão Preto
2025

Análise Estatística do Câncer de Colorretal (COAD)

Através da base de dados do portal NIH-GDC-TCGA foram coletados dados públicos de expressão RNA-seq de pacientes com câncer de colorretal (COAD) com o objetivo de se realizar uma análise estatística descritiva e inferencial comparando amostras teciduais normais versus tumorais. Queremos analisar quais são os genes mais expressos e como a idade, o gênero e a expressão gênica influenciam o estágio do câncer e o risco de mortalidade deste câncer.

Foram selecionadas 82 amostras que possuíam tanto um tecido normal, quanto um tumoral. As variáveis clínicas de interesse foram: project.project_id, case.case_id, cases.submitter_id, demographic.age_at_index, demographic.gender, demographic.vital_status, diagnoses.ajcc_pathologic_stage

Para filtragem e remodelação do manifesto foi utilizado o pacote Pandas do Python e o GDC-Client-Tools. Para as análises estatísticas foram utilizados pacotes do R como o DESeq2 (para realizar **análises de expressão gênica diferencial** com dados de RNA-Seq), Bioconductor, dplyr, tidyverse, pheatmap, RColorBrewer, ggplot2 e readr.

A seguir uma tabela para ilustrar as análises estatísticas empregadas.

Variável	Tipo	Descrição	Exemplo de estatística
Idade	Quantitativa contínua	Idade ao diagnóstico	Média, mediana, desvio padrão
Sexo	Categórica nominal	Male/ Female	Frequência absoluta/relativa
Estágio do câncer	Categórica ordinal	Stage I/II/III/IV (A/B/C)	Distribuição de estágios
Status vital	Categórica nominal	Alive/ Dead	Proporções
Contagem de genes por amostra	Descritiva		Média, mediana de total counts
PCA de expressão gênica	Descritiva	Diferença entre grupos e influência de variáveis clínicas	Visualização da variância
Clustering hierárquico	Descritiva	Agrupamento por condição ou perfil	Heatmap, PCA
Top Genes	Descritiva	Ranking dos genes mais expressos	Heatmap, PCA e MA plot

Tumor x Normal (por gene)	Inferencial	Identificar genes diferencialmente expressos	DESeq2 (modelo GLM negativo binomial)
Idade x Expressão gênica	Inferencial	Avaliar o impacto da idade na expressão gênica	Correlação de Spearman/Pearson
Expressão x Estágio do câncer	Inferencial	Avaliar o gene mais expresso	ANOVA ou Kruskal-Wallis
Expressão vs. status vital	Inferencial	Diferença de expressão em vivos e óbitos	Teste t ou Mann-Whitney
Status vital vs idade vs gêneros	Inferencial	A idade e/ou o gênero interfere no status vital?	Histograma

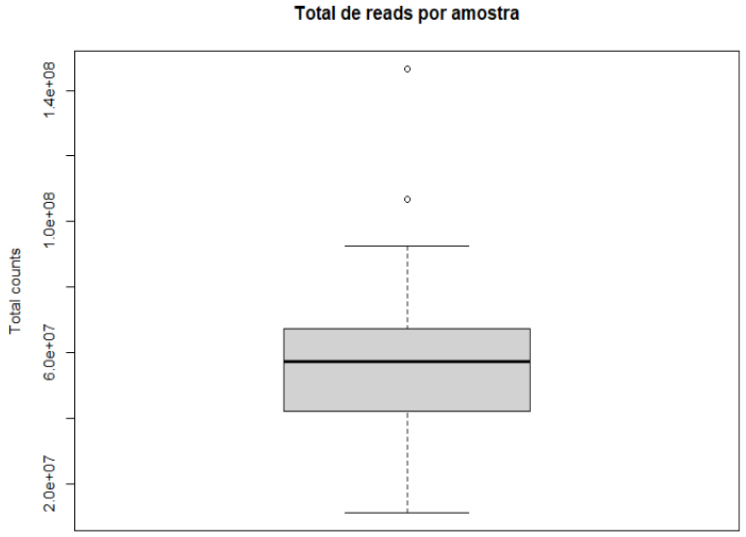
1. Resultados

A coorte selecionada teve como critério amostras que possuíam tanto um tecido normal, quanto um tumoral para fins de comparação.

1. 1 Resultados de Expressão gênica

1.1.1 Contagem total de reads por amostra

```
   Min.     1st Qu.    Median      Mean     3rd Qu.      Max.
11232972  43114606  57264629  56258701  67328376  146513462
```

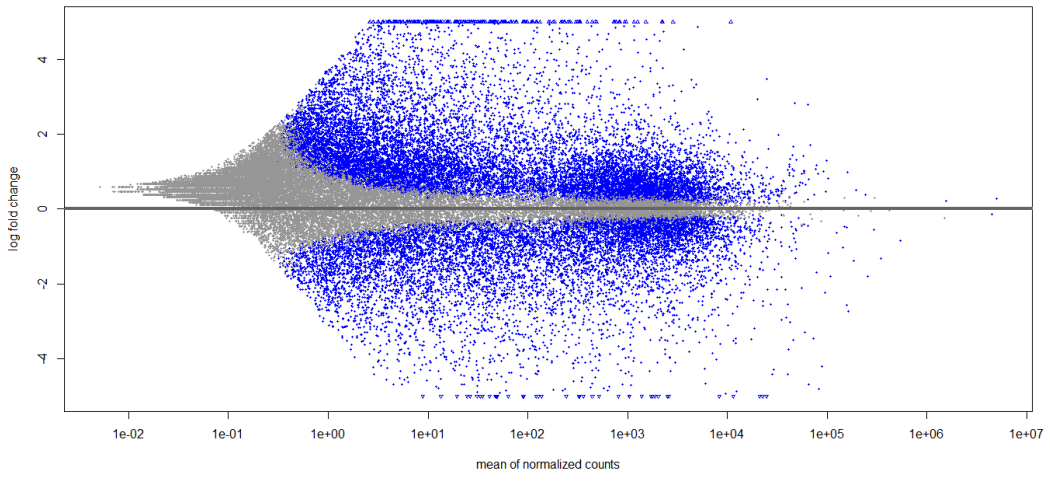


1.1.2 Média e mediana por gene

```
   Min. 1st Qu.  Median    Mean    3rd Qu.    Max.
      0       0       1     927      58     5040810
```

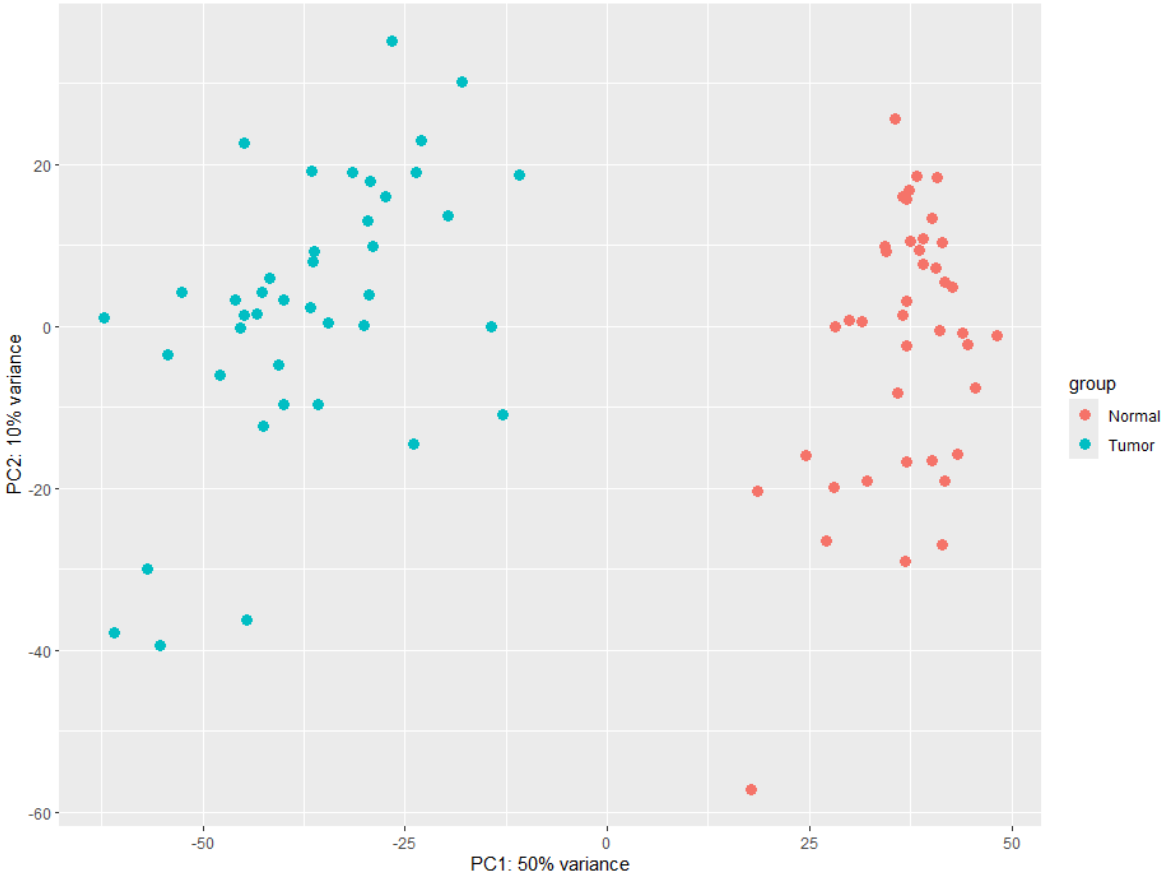
1.1.3 MA plot (expressão x log2FC)

O p-value <0.1. Pode-se perceber que há uma maior quantidade de genes expressos em tecidos tumorais (log2FC >0)



1.1.4 PCA

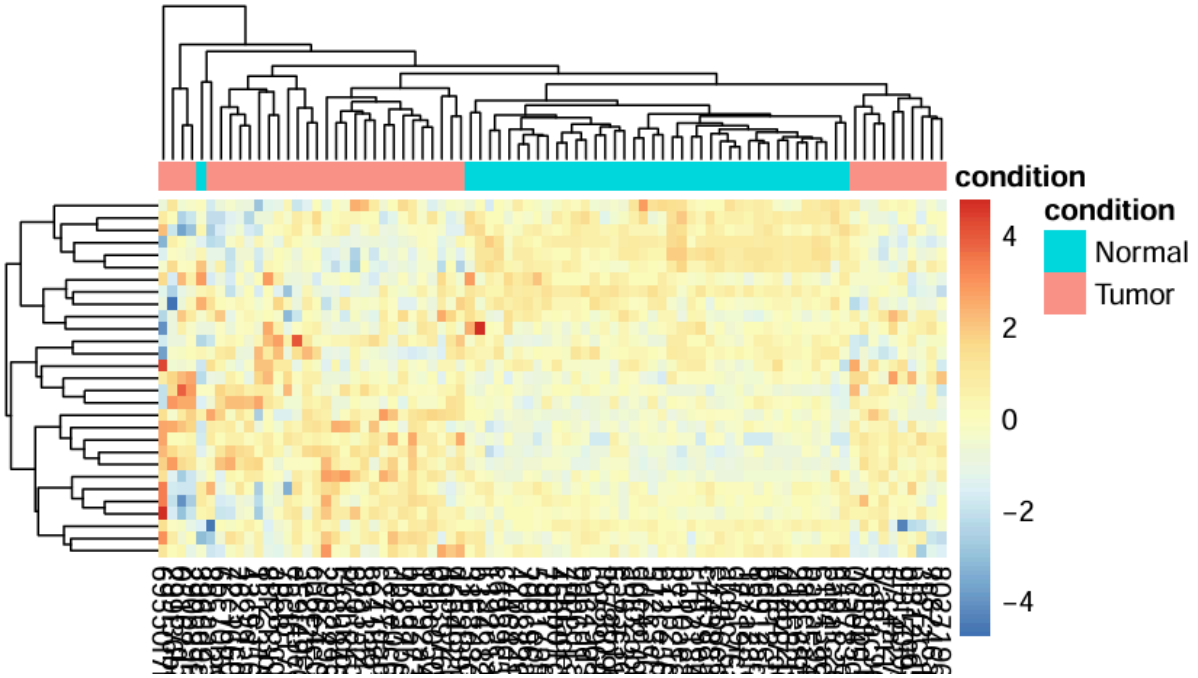
Gráfico representando o agrupamento das amostras

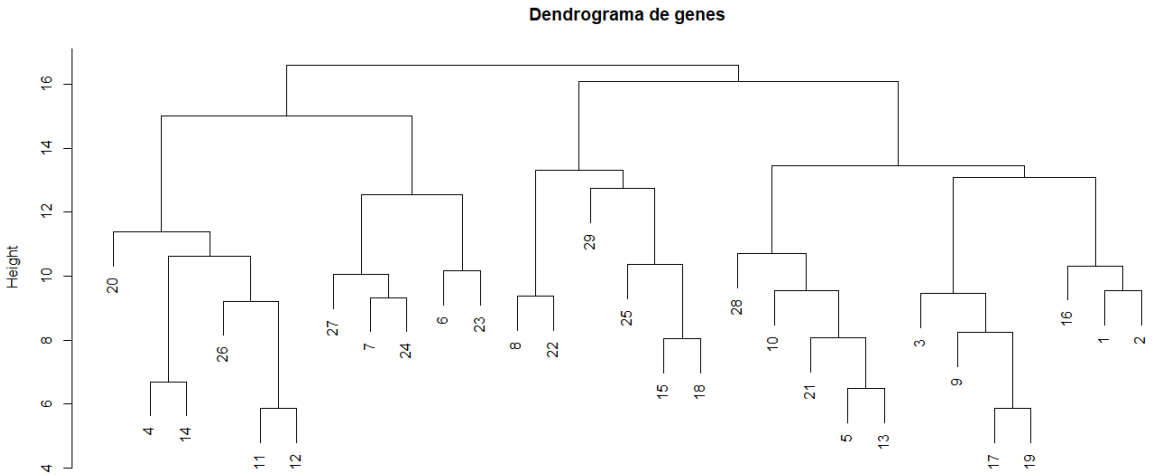


Pode-se observar que os grupos têm perfis de expressão gênica **distintos** e diferenças biológicas marcantes entre os tipos de amostra normal e tumoral.

1.1.5 Heatmap

O heatmap representa a expressão normalizada escalonada de cada gene (Z-score que varia de 4 a -4), os que estão com cores se aproximando do vermelho são os mais superexpressos, já os que estão se aproximando do azul são os mais subexpressões.





1. 2 Resultados de Dados clínicos

1.2.1 Dados clínicos do projeto COAD

1.2.1.1 Idade

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  40.00   64.00   74.00   70.27   80.00   89.00
```

1.2.1.2 Gênero

Frequência absoluta

```
female   male
    42     40
```

Frequência relativa

```
  female      male
51.21951  48.78049
```

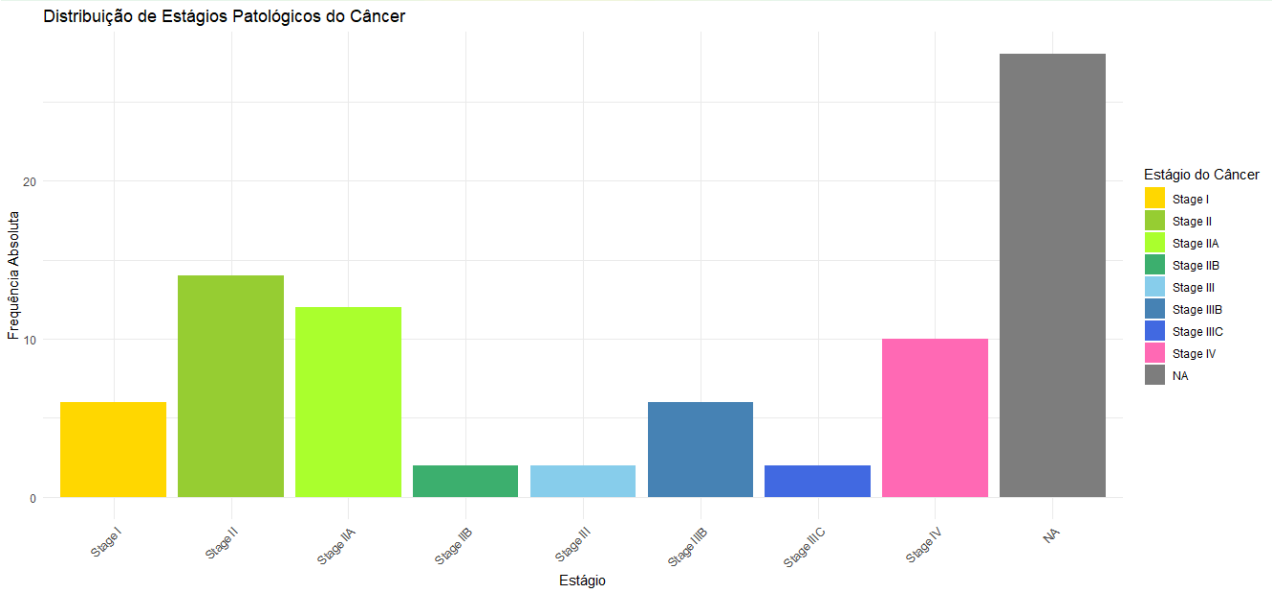
1.2.1.3 Estágio do câncer

Frequência absoluta

```
      '--    Stage I   Stage II    Stage IIA  Stage IIB  Stage III
        28          6         14          12          2          2
Stage IIIB   Stage IIIC   Stage IV
         6            2         10
```

Frequência relativa

```
      '--    Stage I   Stage II  Stage IIA
 34.146341   7.317073  17.073171  14.634146
 Stage IIB  Stage III Stage IIIB Stage IIIC
  2.439024   2.439024   7.317073   2.439024
  Stage IV
 12.195122
```



1.2.1.4 Status vital

Frequência absoluta

```
Alive  Dead
   58    24
```

Frequência relativa

```
  Alive     Dead
70.73171 29.26829
```

1.2.1.5 Tecido (Normal vs Tumoral)

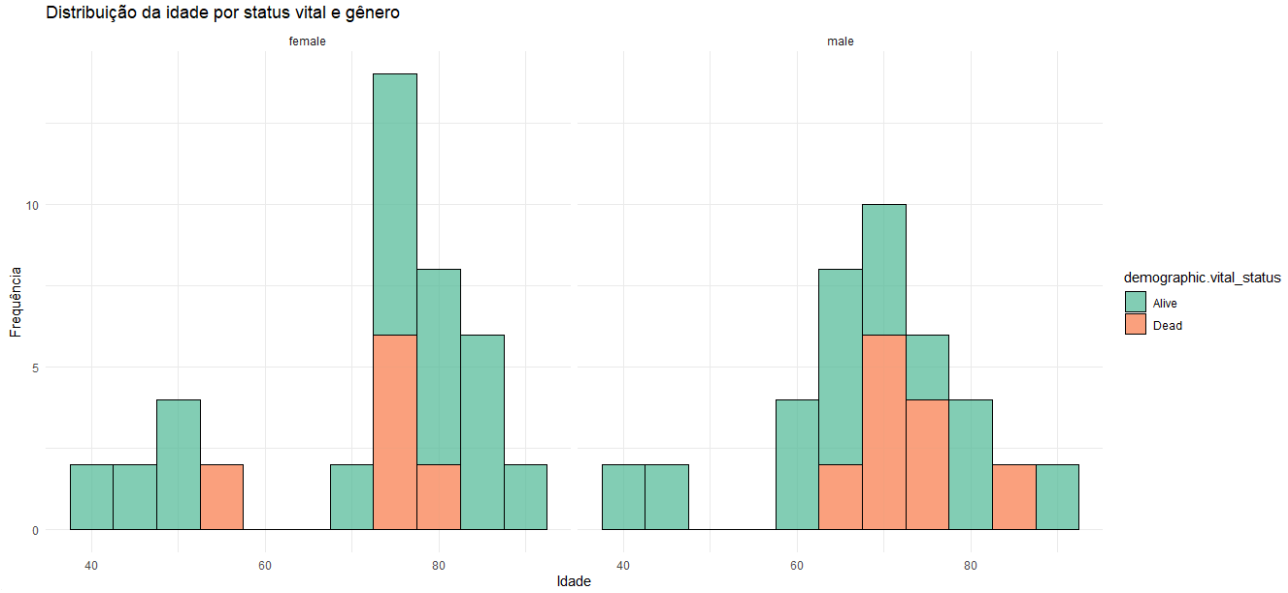
Frequência absoluta

```
Normal  Tumor
    41     41
```

Frequência relativa

```
Normal  Tumor
    50     50
```

1.2.1.6 Comparação entre Idade x Status Vital x Gênero

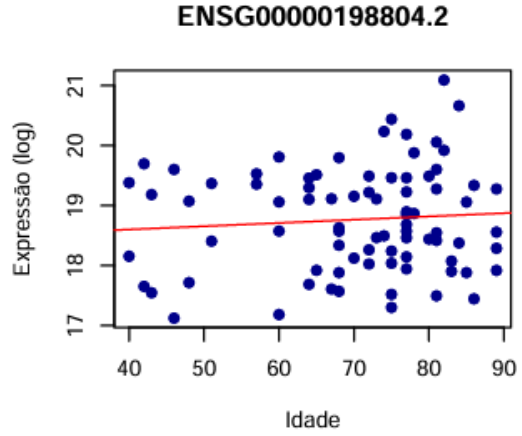


1. 3 Resultados da comparação entre Dados clínicos e Expressão gênica

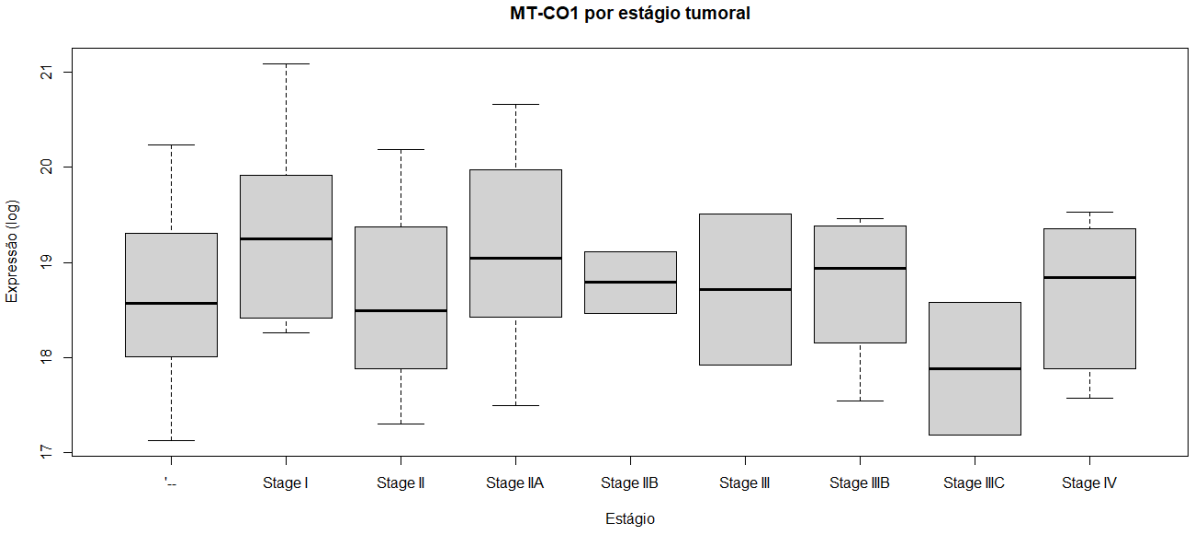
Correlacionando a expressão gênica com os dados clínicos da coorte selecionada, foram obtidos os seguintes resultados

O gene com a maior média de expressão gênica foi o ENSG00000198804.2 (MT-CO1), então será usado para as análises estatísticas a seguir

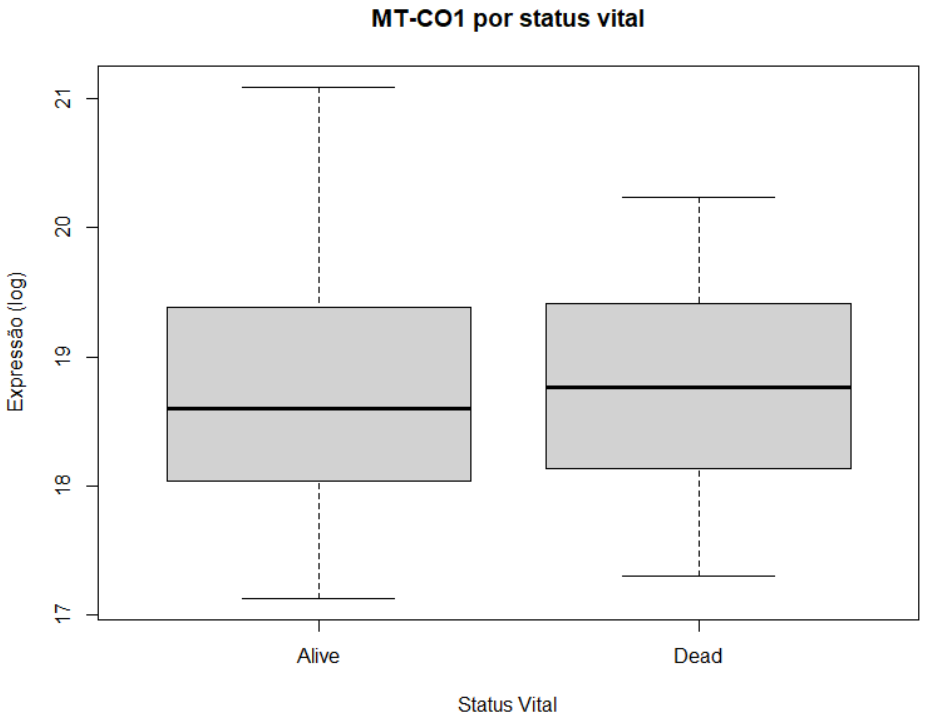
1.3.1 Idade vs Expressão gênica



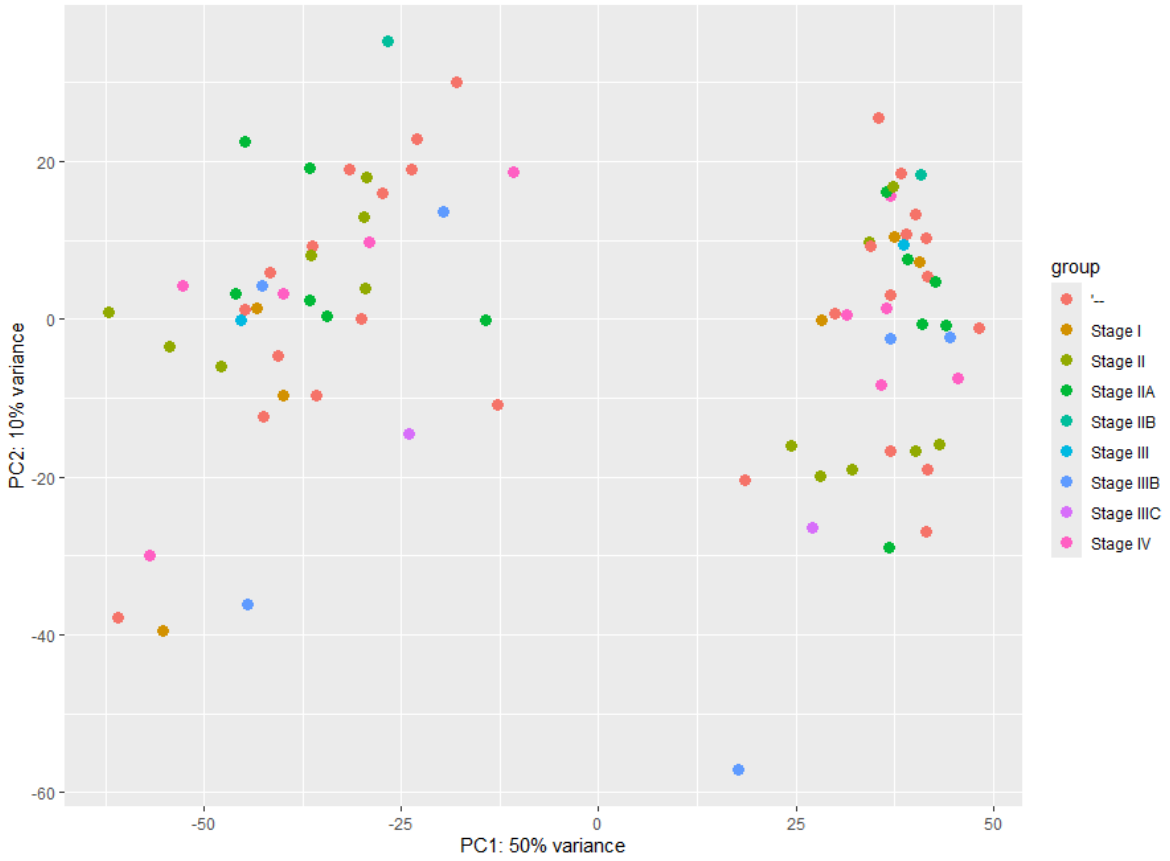
1.3.2 Estágio do câncer vs Expressão gênica



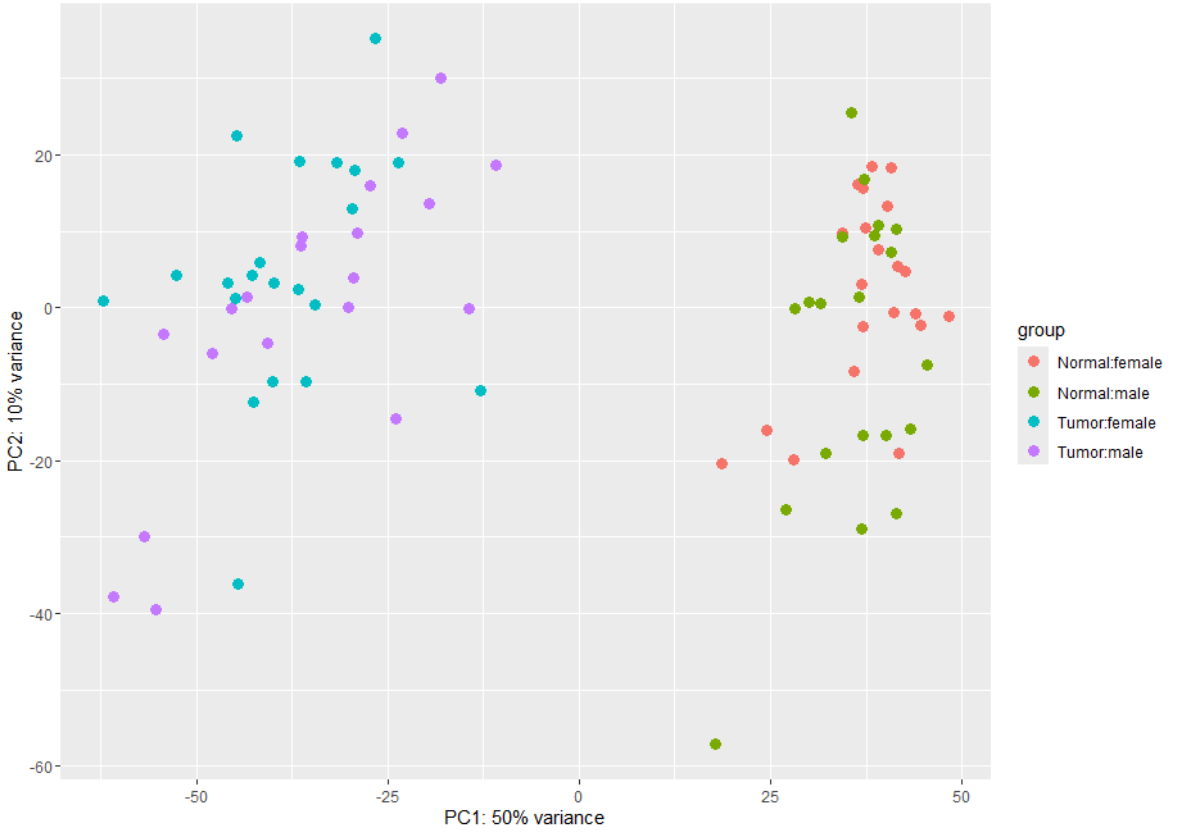
1.3.3 Status vital vs Expressão gênica



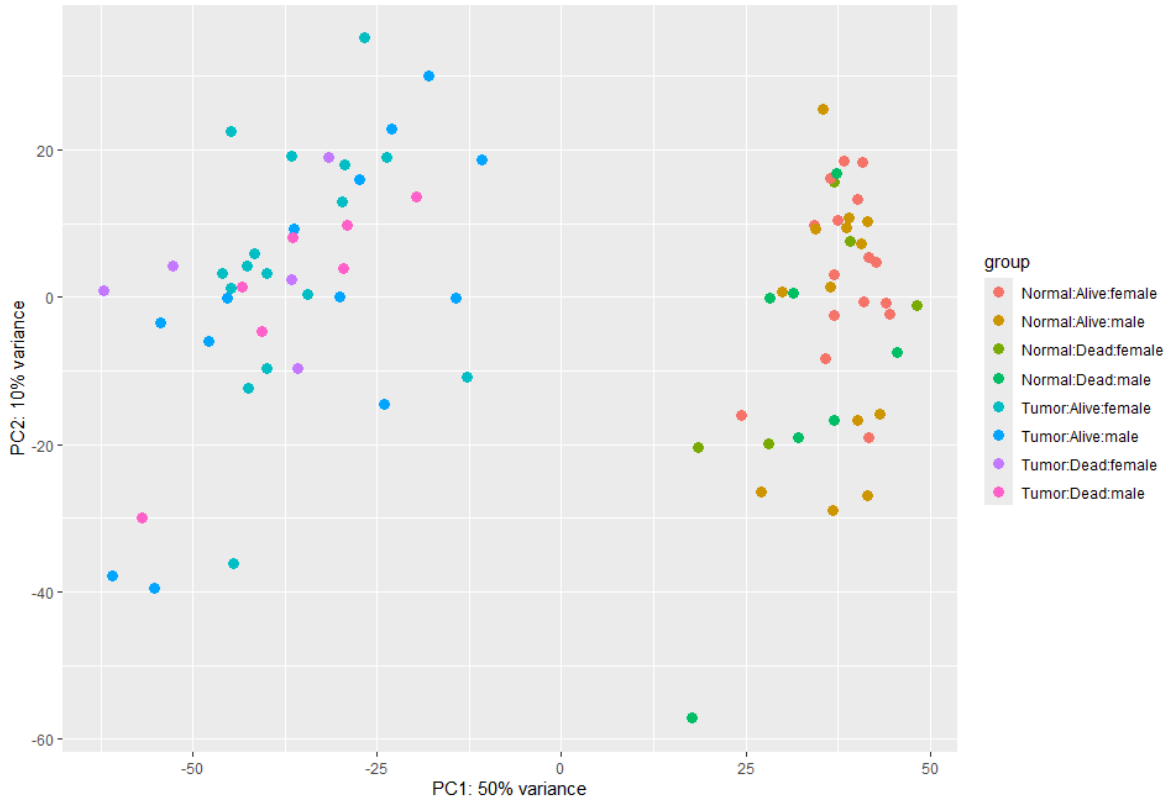
1.3.4.1 PCAs com dados clínicos por estágio do câncer



1.3.4.2 PCAs com dados clínicos por sexo



1.3.4.3 PCAs com dados clínicos por status vital e gênero



2. Metodologia empregada

1) Baixar o manifesto (Filtrar por RNA-seq e STAR-Counts) {GDC portal}
2) Baixar os dados clínicos e criar uma tabela com as colunas de interesse (clinical_data.R)
3) Baixar o manifesto com o tipo de amostra e barcode (generate_manifest.py)
Executar no terminal (exemplo):
python C:\\Users\\Usuário\\Downloads\\TCGA\\generate_manifest.py -i C:\\Users\\Usuário\\Downloads\\TCGA\\gdc_manifest_coad_one.txt
4) Filtrar o manifesto, pareando uma amostra normal com uma tumoral e excluindo as demais (filtered_manifest.py)
5) Excluir as colunas sample_type, id e barcode (rebuild_manifest.py)
6) Baixar os dados de expressão gênica
Executar no terminal:
.\gdc-client download -m manifesto_coad_reb.txt
7) Matriz de Contagem RNA-Seq, de modo que os id's estejam alinhados ao manifesto (build_matrix.R)
Deve gerar o arquivo TCGA_COAD_RNASEQ_ORDERED
8) mapear os IDs do manifesto para os IDs da matriz de contagem, criar a matriz sample_conditions para análise diferencial (build_matrix.R)
9) Análises estatísticas inferenciais com DESeq (build_matrix.R)
10) Análises estatísticas descritivas e inferenciais com os dados clínicos (clinical_data.R)
11) Correlação entre a expressão gênica e os dados clínicos (clinical_data.R)