



Introdução à linguagem R

Prof. Dr. Felipe Mesquita

Roteiro

- Compreender o sequenciamento de nova geração (Illumina)
- Classificação da análise de dados de sequenciamento
- Exemplos de dados genômicos, transcriptômicos e proteômicos
- Workflows e estrutura de dados
- Linguagem R
- Entendendo na prática

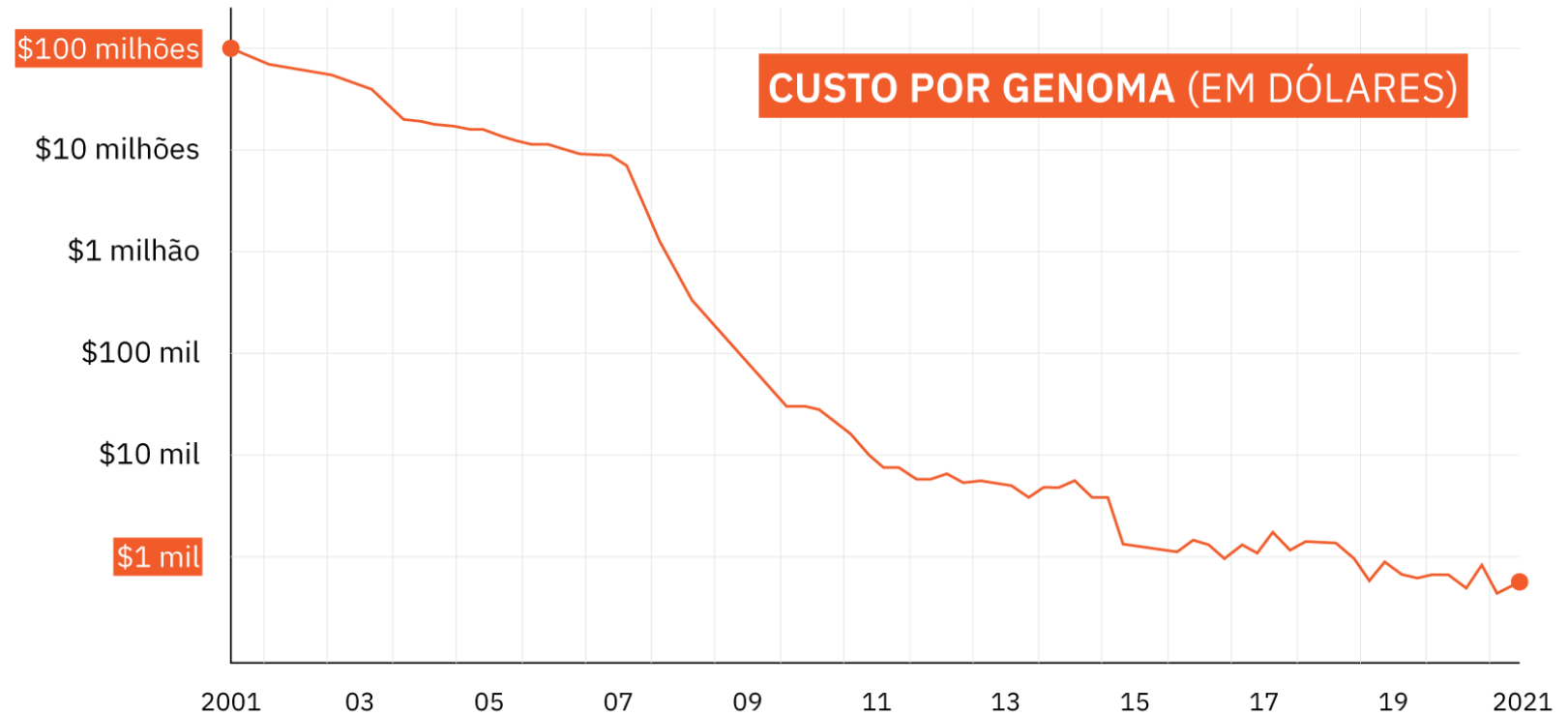
Sequenciamento de DNA/RNA

Geração	Princípio	Características	Exemplos
1ª geração	Terminação de cadeia e separação eletroforética	Alta precisão, leituras individuais e baixo rendimento	Sanger; Maxam–Gilbert, historicamente
2ª geração	Sequenciamento massivamente paralelo de fragmentos geralmente amplificados	Milhões de reads curtos, alto rendimento e baixo custo por base	Illumina , Ion Torrent, 454 e SOLiD
3ª geração	Sequenciamento de moléculas individuais em tempo real	Reads longos ou ultralongos; melhor resolução de variantes estruturais, regiões repetitivas, haplótipos e isoformas	PacBio SMRT/HiFi e Oxford Nanopore

4ª geração?

Sequenciamento de nova geração

- Custo



Fonte: www.genome.gov/sequencingcostsdata

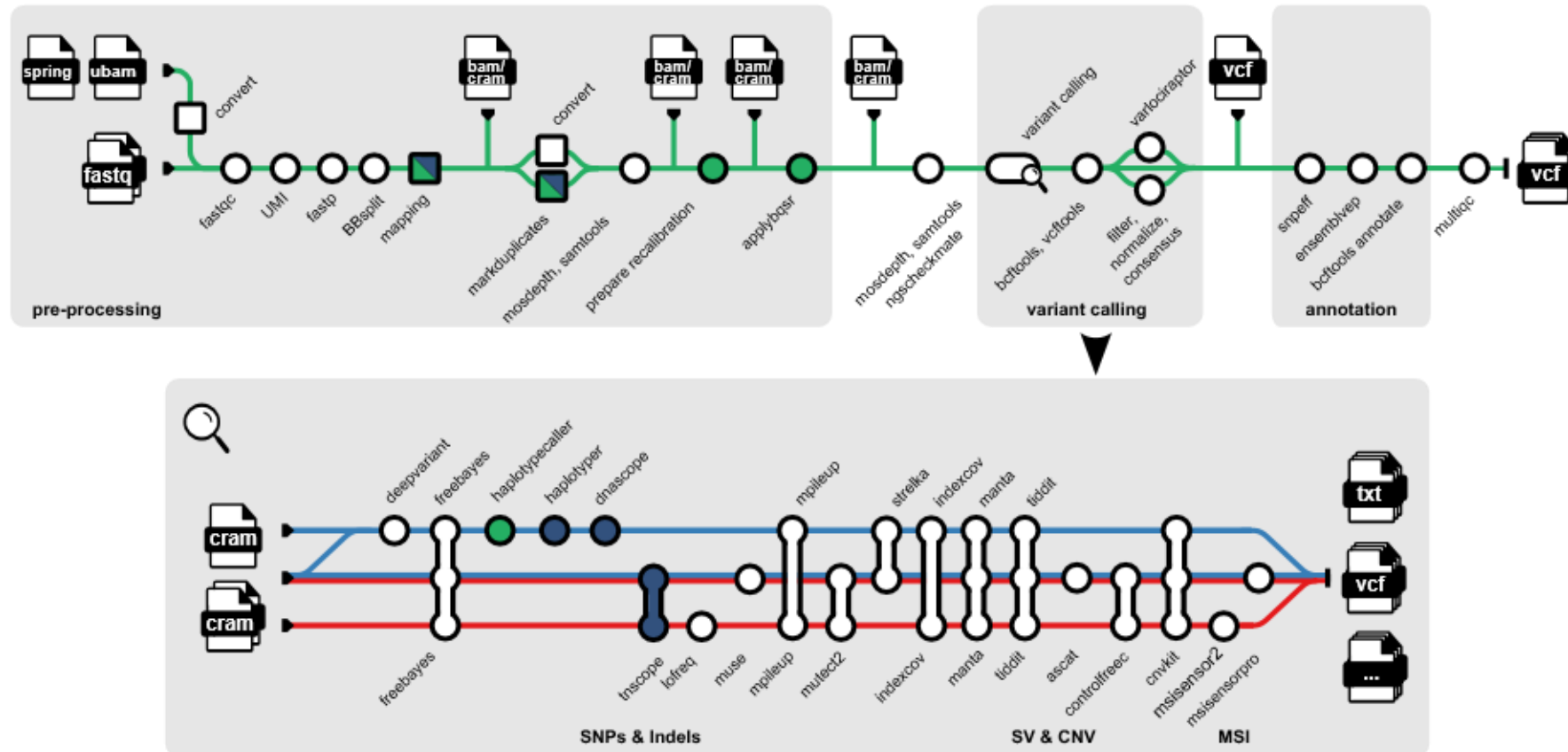
Classificação

- Análise primária
- Análise secundária
- Análise terciária

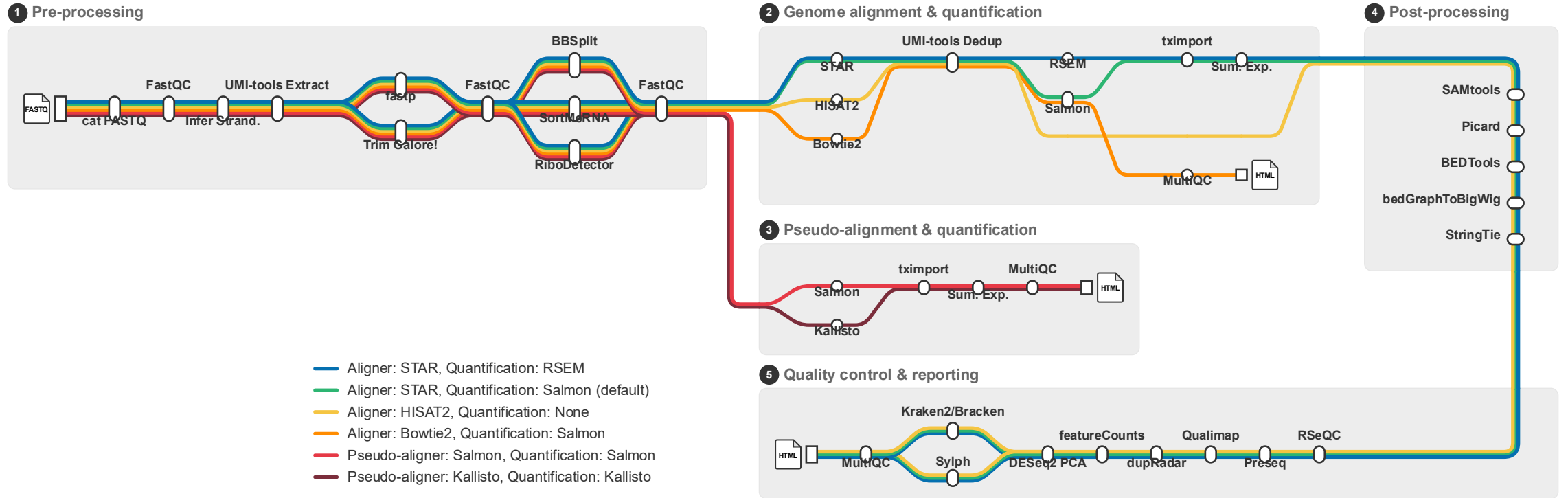
O que temos em mãos durante o processamento dos dados?

- Fastq
- BAM
- VCF (VCF anotado)
- FASTA
- TSV, TXT, MAF, CSV, ..

Analysis pipeline to detect germline or somatic variants (pre-processing, variant calling and annotation) from WGS / targeted sequencing



RNA sequencing analysis pipeline

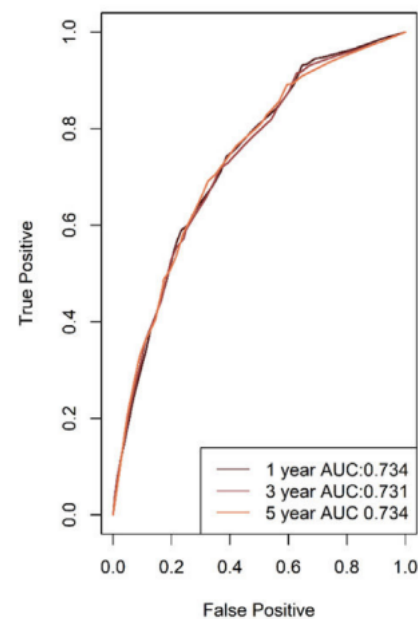
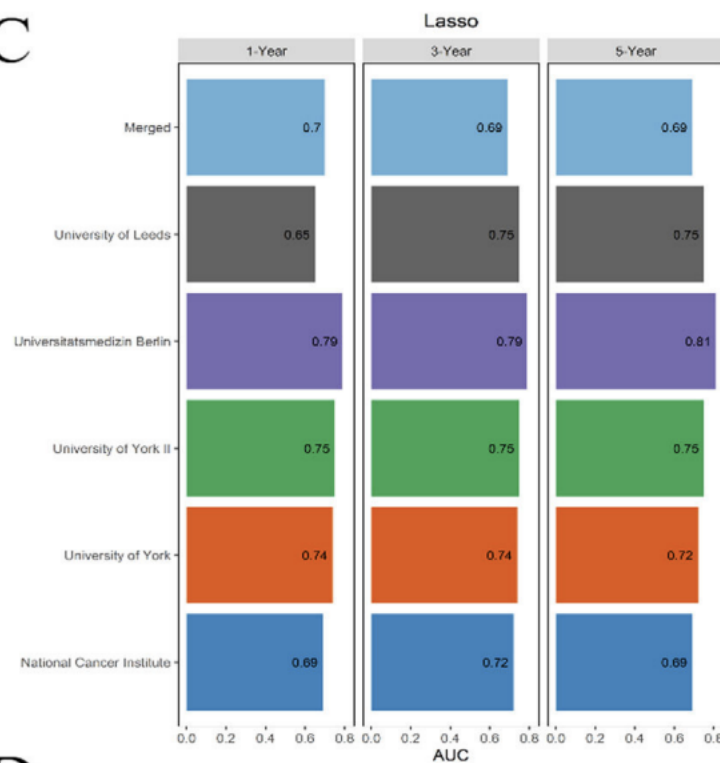


Por que a análise terciária é tão importante?

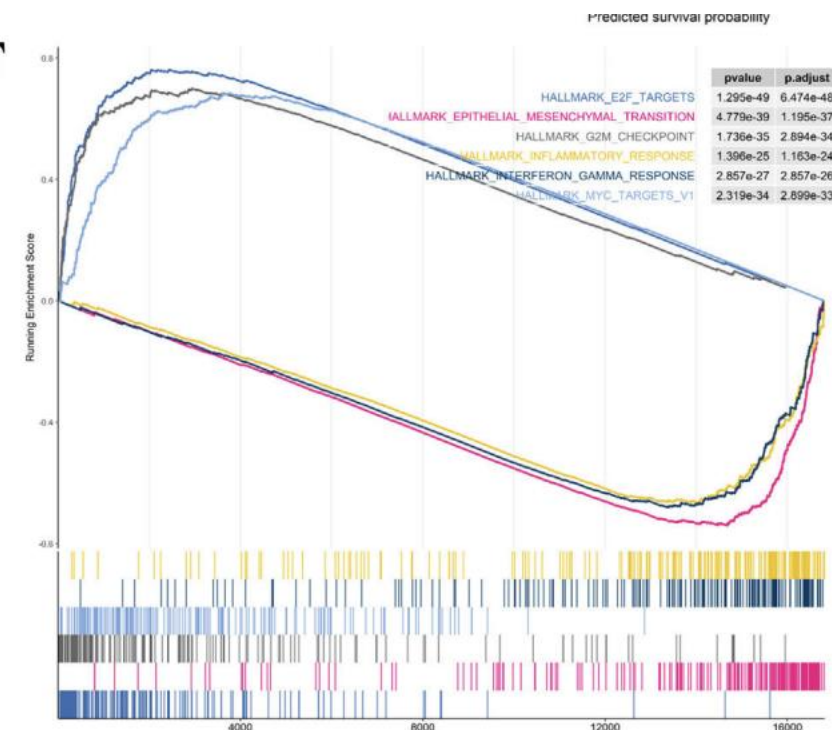
Letter to the Editor

Multi-cohort gene expression model enhances prognostic stratification in diffuse large B-cell lymphoma

C

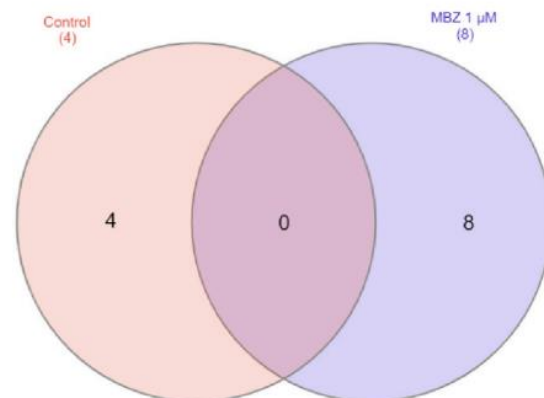


F





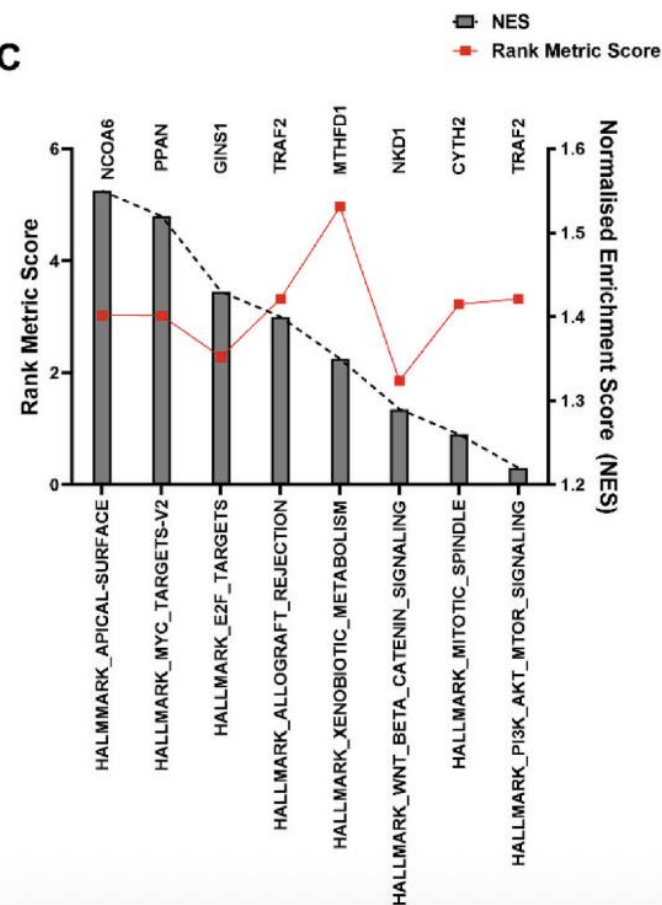
A



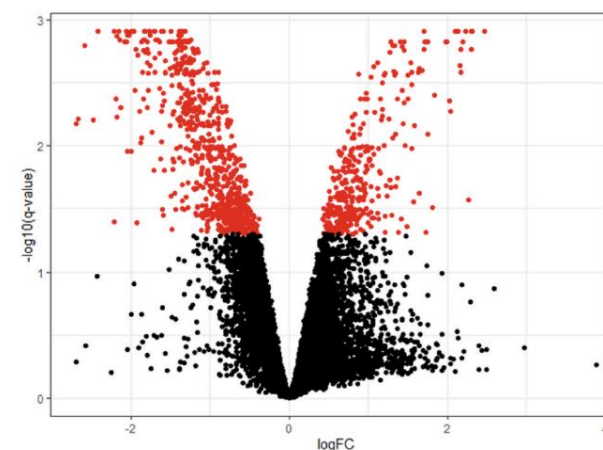
Transcriptome analysis displays new molecular insights into the mechanisms of action of Mebendazole in gastric cancer cells

Emerson Lucena da Silva^{a,1}, Felipe Pantoja Mesquita^{a,1}, Laine Celestino Pinto^{b,1}, Bruna Puty Silva Gomes^c, Edivaldo Herculano Correa de Oliveira^c, Rommel Mario Rodríguez Burbano^{d,e}, Maria Elisabete Amaral de Moraes^a, Pedro Filho Noronha de Souza^{a,f,g}, Raquel Carvalho Montenegro^{a,h,*}

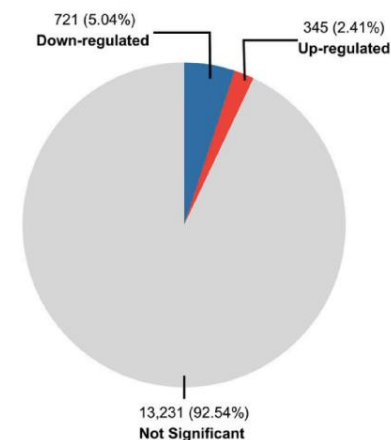
C



A



B





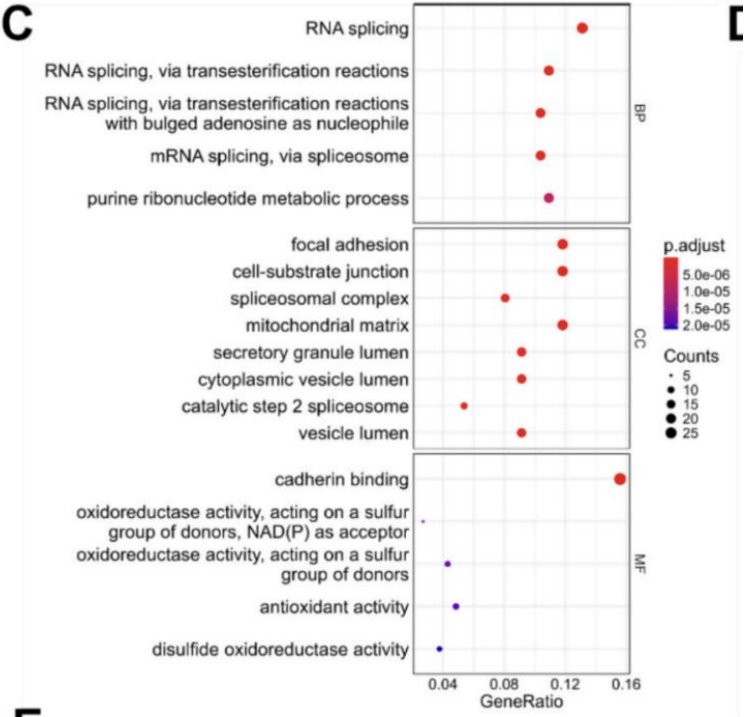
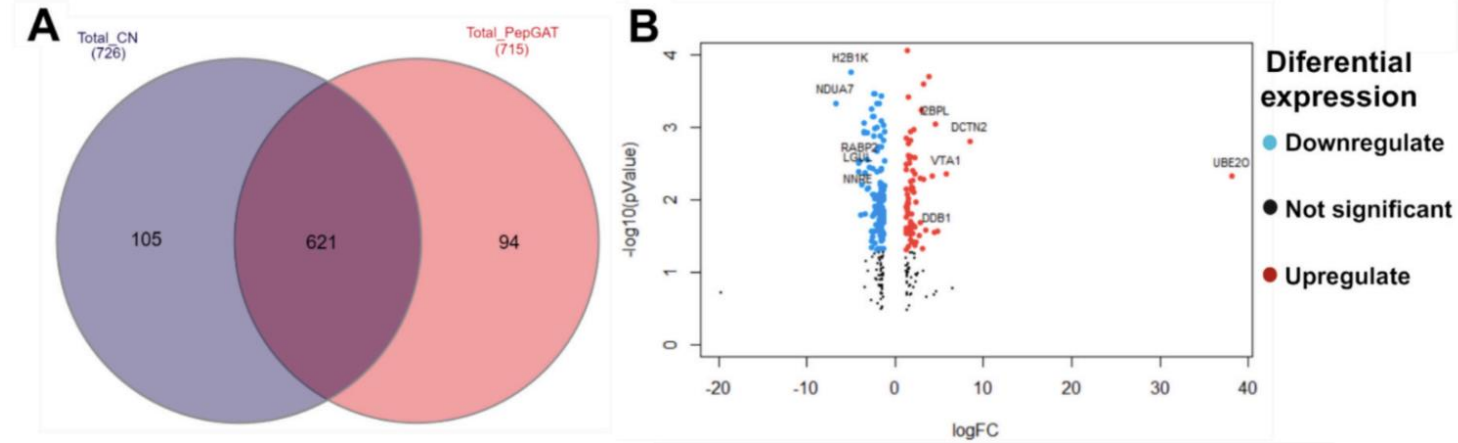
PepGAT, a chitinase-derived peptide, alters the proteomic profile of colorectal cancer cells and perturbs pathways involved in cancer survival

Pedro Filho Noronha Souza^{a,b,c,*}, Elmer Adilson Espino Zelaya^{b,1}, Emerson Lucena da Silva^b, Laís Lacerda Brasil-Oliveira^b, Francisco Laio de Oliveira^b, Maria Elisabete Amaral de Moraes^b, Raquel Carvalho Montenegro^b, Felipe Pantoja Mesquita^{a,b,**}

^a Laboratory of Bioinformatics Applied to Health, Drug Research and Development Center (NPDM), Federal University of Ceará, Fortaleza, CE 60430-275, Brazil

^b Pharmacogenetics Laboratory, Drug Research and Development Center (NPDM), Federal University of Ceará, Fortaleza, CE 60430-275, Brazil

^c Cearense Foundation to Support Scientific and Technological Development, Brazil



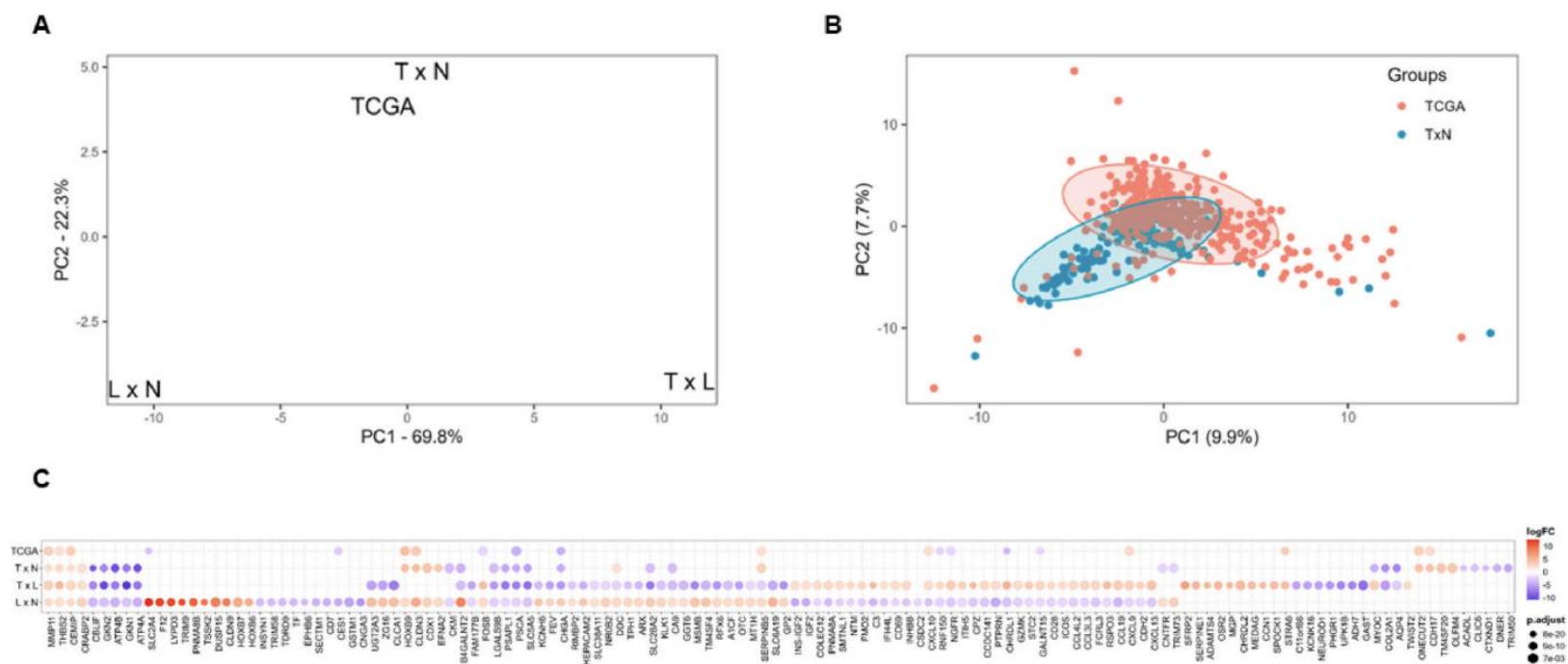


Molecular characterization and biomarker discovery in gastric cancer progression through transcriptome meta-analysis

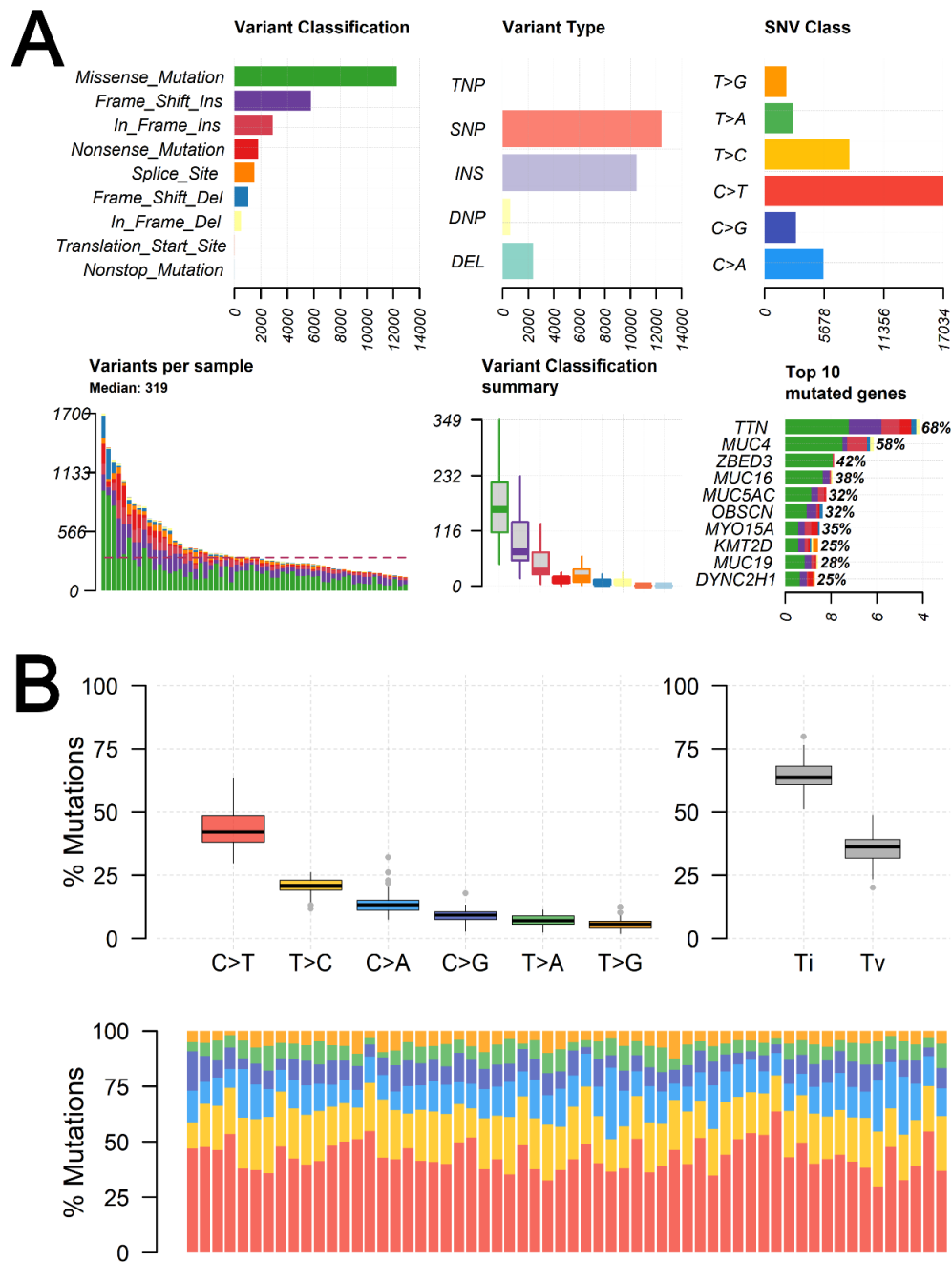
Thiago Loreto Matos^a, Pedro Filho Noronha Souza^a, Maria Elisabete Amaral de Moraes^a,
Silvia Helena Barenhorst^b, Felipe Pantoja Mesquita^{a,*}, Raquel Carvalho Montenegro^a

^a Pharmacogenetics Laboratory, Drug Research and Development Center, Department of Physiology and Pharmacology, Federal University of Ceará, Fortaleza, 60430-160, Brazil

^b Molecular Genetics Laboratory, Department of Pathology and Legal Medicine, Federal University of Ceará, Fortaleza, 60430-160, Brazil



The Distinct Mutational Landscape of Gastric Cancer in Northeastern Brazil Identifies *MUC4* and epigenetic regulators as a Candidate Driver.



O  é um forte aliado nas análises terciárias de dados ômicos

História



A linguagem mantém a herança do S, mas seu desenvolvimento é coletivo e aberto.

R é linguagem, ambiente e sistema de execução

LINGUAGEM

Sintaxe para expressar cálculos, decisões, funções e gráficos.

AMBIENTE

Objetos, pacotes e ferramentas para análise.

RUNTIME

O motor que interpreta expressões e mantém a sessão em memória.

R não é apenas um programa de estatística

Comprehensive R Archive Network (CRAN)

- Uma rede mundial de servidores espelho que mantém versões sincronizadas de código e documentação.
- <https://cran.r-project.org/>

INSTALAR DO
CRAN

```
install.packages("ggplot2")
```

CRAN ≠ pacote

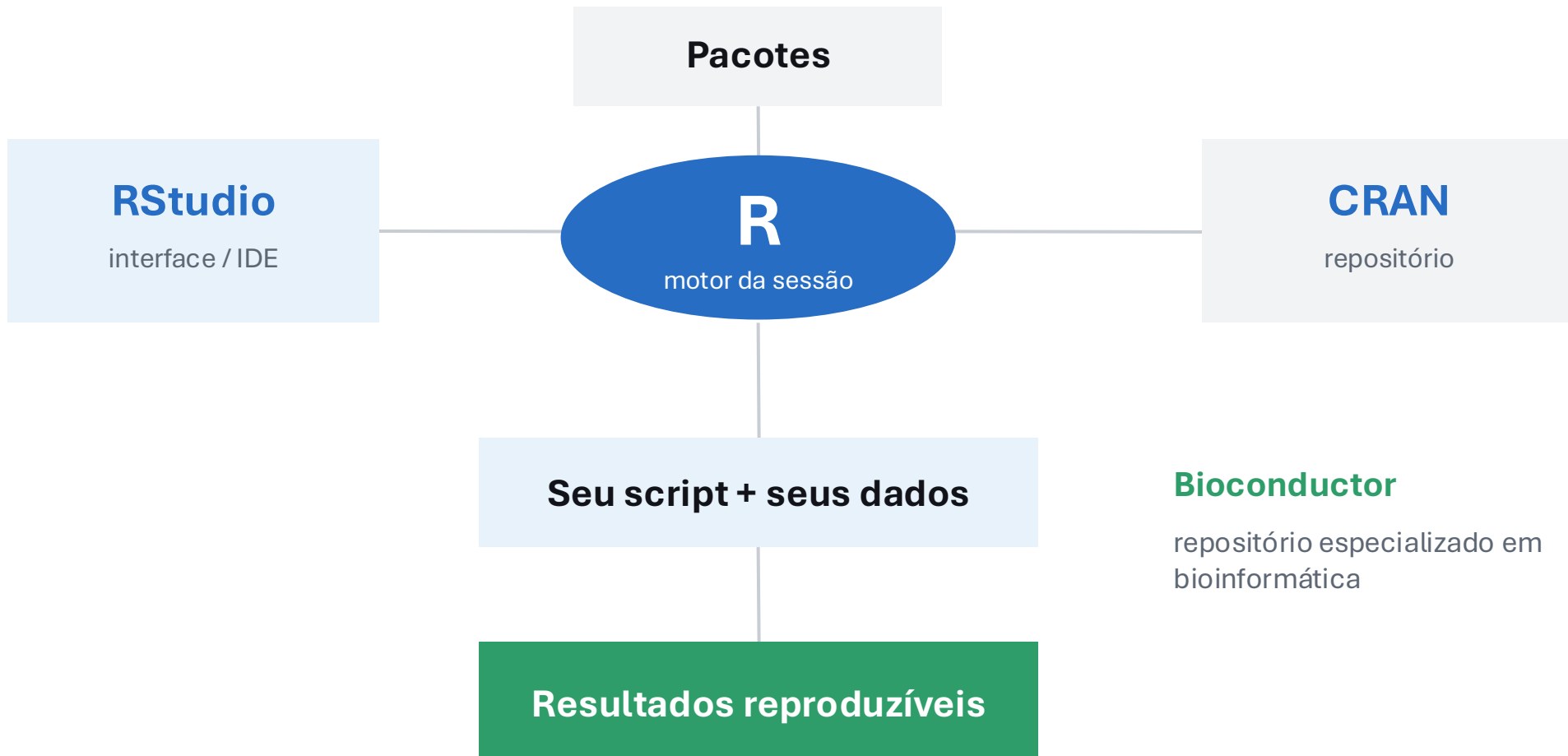
-> CRAN é o repositório

-> ggplot2 é um pacote hospedado nele.

RStudio é a bancada de trabalho

- **IDE** = ambiente integrado de desenvolvimento
- O RStudio organiza edição, execução, objetos, arquivos e gráficos em uma interface.





Fluxo de execução

1

Escrever

2

Executar

3

Avaliar

4

Armazenar

5

Produzir

Pacotes

- Ampliação do vocabulário do R
- Um pacote reúne código, doc e recursos em uma unidade instalável.
 - Funções (operações automatizadas)
 - Dados (exemplos e refs)
 - Ajuda (help?)
 - Código compilado (C, C++, etc)

Base R continua funcionando sem pacotes de forma "básica", mas os pacotes adicionam **especialização**.

Origem dos pacotes?

- Base
- CRAN (<https://cran.r-project.org/>)
- Bioconductor (<https://www.bioconductor.org/>)
- Git (<https://github.com/>)

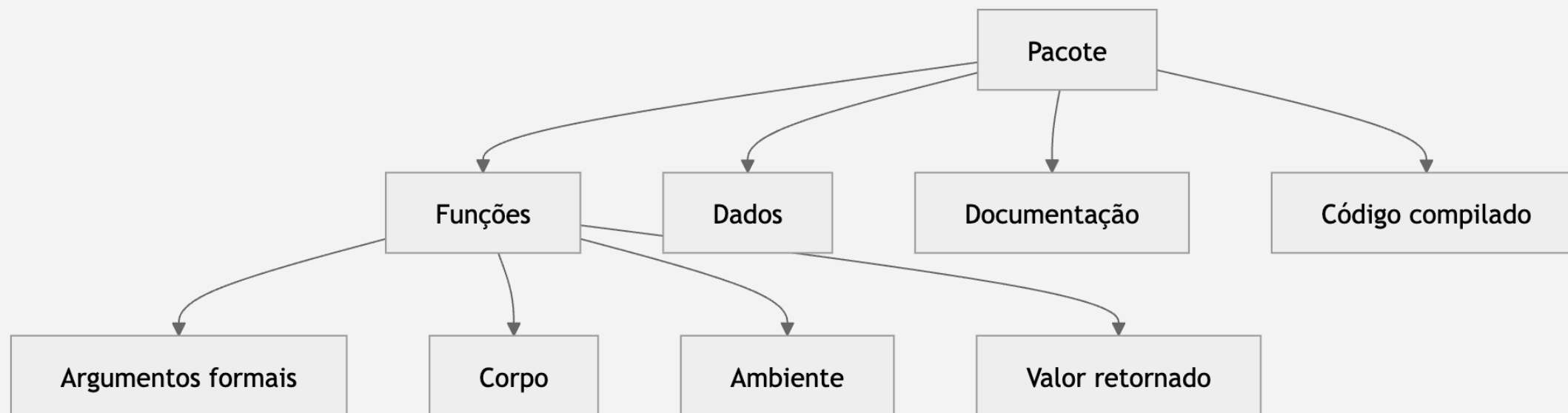
Instalar e carregar pacotes

UMA VEZ

```
install.packages("dplyr")
```

A CADA NOVA SESSÃO

```
library(dplyr)
```



O que é função e argumento?

Função `mean`

```
mean(x, na.rm = TRUE)
```

Função `round`

```
round(x = 3.14159, digits = 2)
```

Variáveis e objetos

TRÊS
LIGAÇÕES

```
genes <- c("TP53", "EGFR", "BRCA1")  
n <- length(genes)  
ativo <- TRUE  
Expressao <- c(8.2, 12.5, 6.7)
```

Elemento	O que representa
expressao	Variável, nome ou símbolo
<-	Operador de atribuição
c(8.2, 12.5, 6.7)	Objeto: vetor numérico

Tipos de vetores atômicos

- Um único valor também é um vetor
- ```
idade <- 25
```

| Tipo      | Exemplo | typeof()  | Uso típico              |
|-----------|---------|-----------|-------------------------|
| logical   | TRUE    | logical   | condições               |
| integer   | 42L     | integer   | contagens               |
| double    | 42.0    | double    | medidas                 |
| character | "TP53"  | character | texto / identificadores |

# Os vetores se organizam de maneiras diferentes

- Estrutura de dados

| Estrutura            | Dimensão  | Tipos internos          | Uso principal                               |
|----------------------|-----------|-------------------------|---------------------------------------------|
| <b>Vetor atômico</b> | 1D        | Um único tipo           | Valores homogêneos                          |
| <b>Lista</b>         | 1D        | Tipos diferentes        | Objetos heterogêneos                        |
| <b>Matriz</b>        | 2D        | Geralmente um tipo      | Dados numéricos tabulares                   |
| <b>Array</b>         | 3 ou mais | Geralmente um tipo      | Dados multidimensionais                     |
| <b>Fator</b>         | 1D        | Categorias predefinidas | Variáveis categóricas                       |
| <b>Data frame</b>    | 2D        | Um tipo por coluna      | Tabelas de dados                            |
| <b>Tibble</b>        | 2D        | Um tipo por coluna      | Versão moderna do data frame, pacote tibble |