UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

R para Ciências da Vida (BCM13065)
Aula 6

PPGBCM - UFRGS

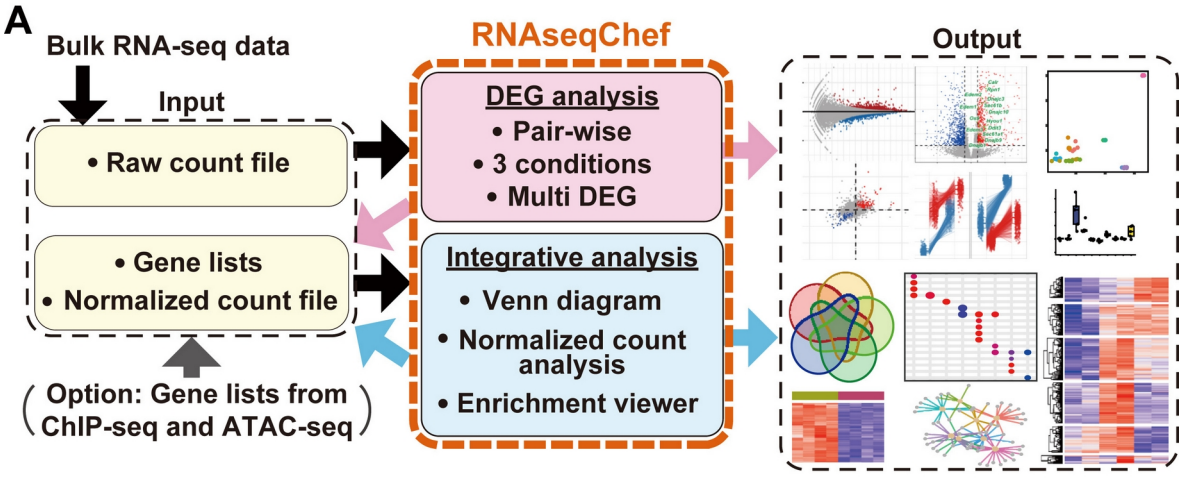
Diego Bonatto

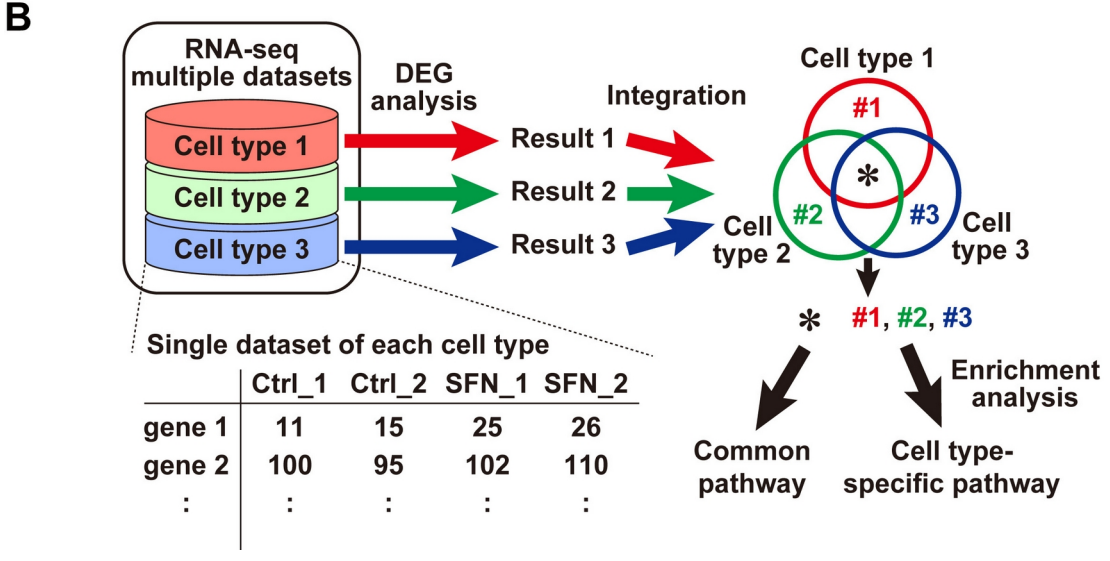
2024/2

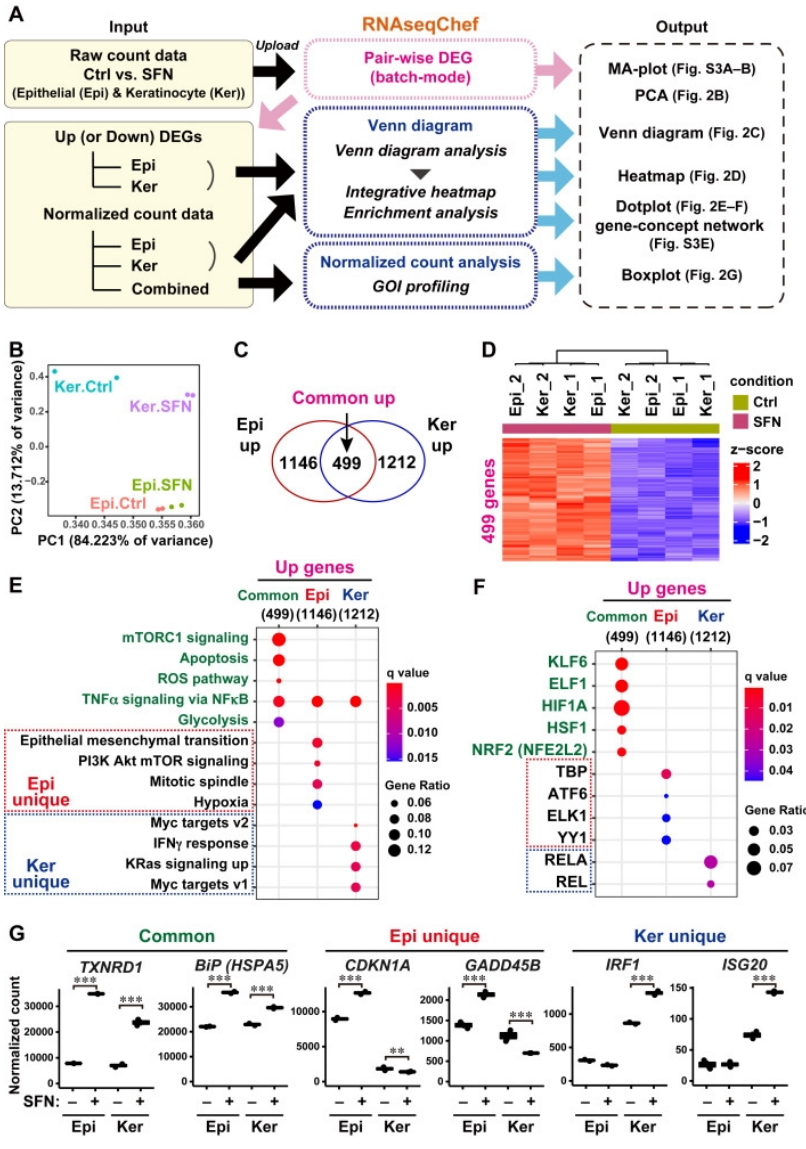
Data mining versus knowledge discovery

- **O que é data mining?**
  - **Foco:** Extração de padrões e relações a partir de grandes conjuntos de dados.
  - **Técnicas:** Utiliza algoritmos de aprendizado de máquina como árvores de decisão, redes neurais, clustering e associação de regras.
  - **Objetivo:** Encontrar informações relevantes e úteis nos dados.

Exemplo de data mining biológico -10.1016/j.jbc.2023.104810

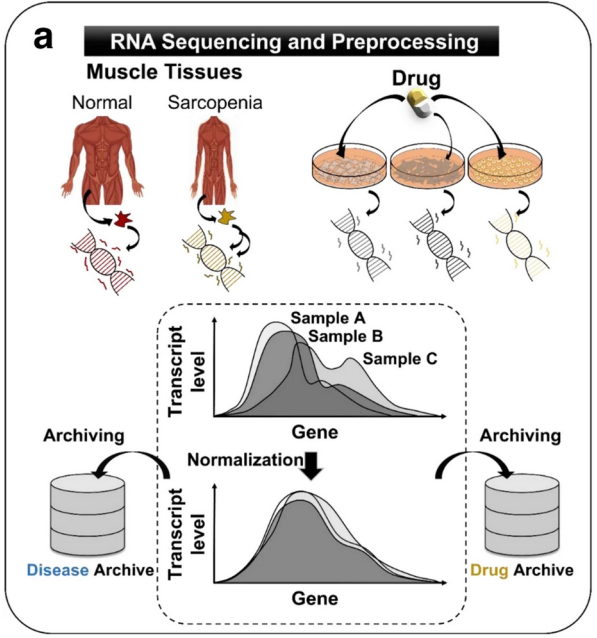


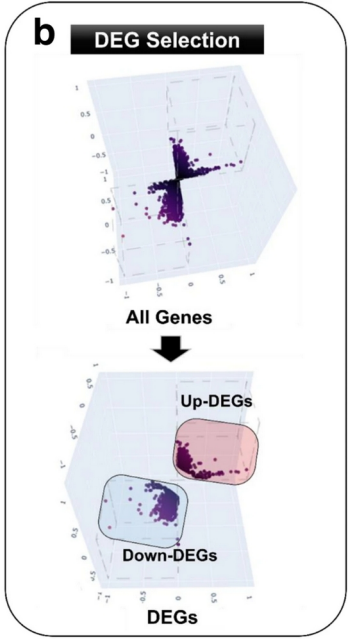


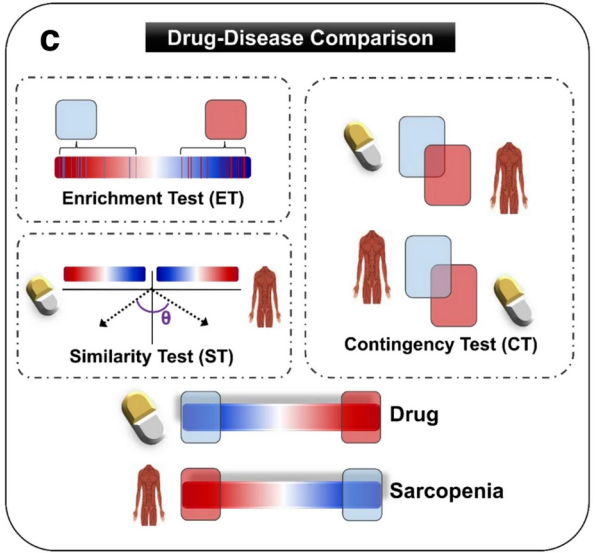


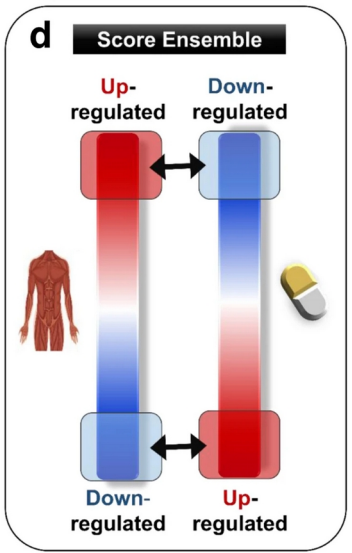
Data mining versus knowledge discovery

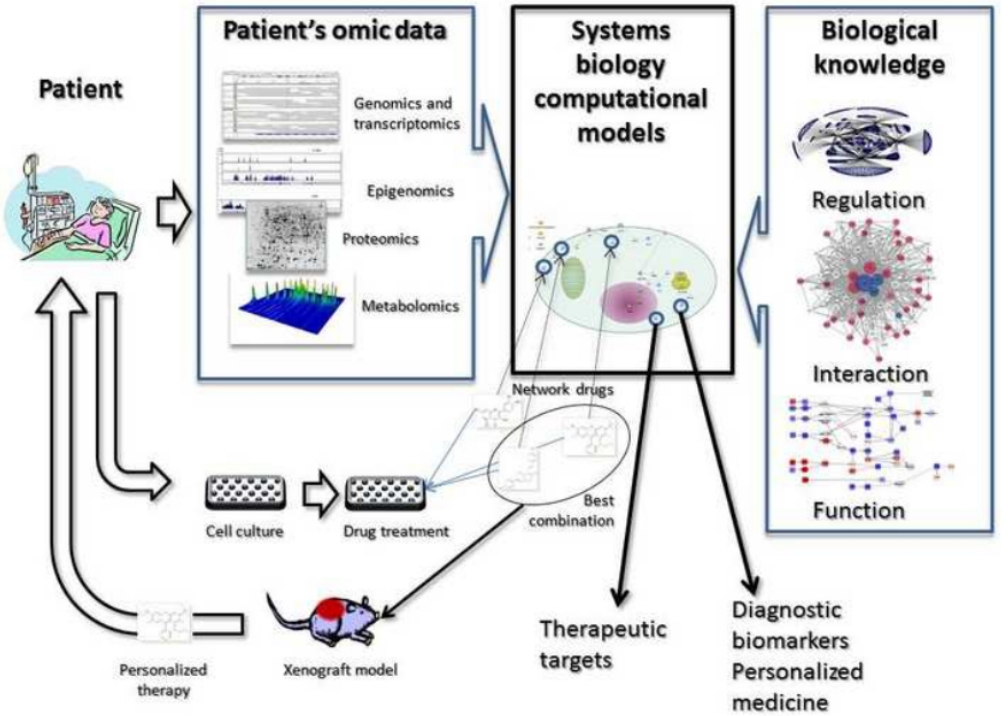
- **O que é knowledge discovery?**
  - **Foco:** Processo mais amplo que inclui todas as etapas para transformar dados brutos em conhecimento acionável.
  - **Etapas:** Seleção de dados, pré-processamento, data mining, interpretação, visualização e geração de conhecimento.
  - **Objetivo:** Não apenas encontrar padrões, mas também entender seu significado e como utilizá-los para tomada de decisões.











10.1016/j.drudis.2013.06.003
10.1038/s12276-024-01189-z

Data mining versus knowledge discovery

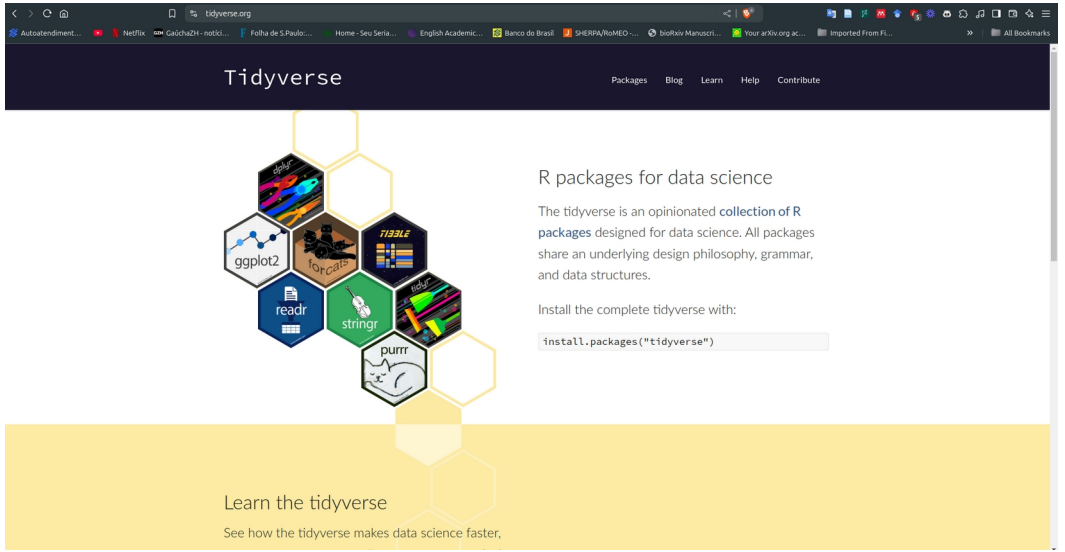
- **Em resumo:**
  - Data Mining é uma parte do processo de Knowledge Discovery.
  - Data Mining se concentra na extração de padrões, enquanto Knowledge Discovery engloba todo o processo de transformar dados em conhecimento útil.
- **Pontos-chave:**
  - Data Mining é como minerar um terreno em busca de ouro: você utiliza ferramentas para encontrar algo valioso.
  - Knowledge Discovery é como transformar esse ouro em joias: você refina, molda e dá um significado ao que foi encontrado.
  - Ambos são importantes: O data mining fornece os dados brutos, enquanto o knowledge discovery transforma esses dados em insights valiosos.

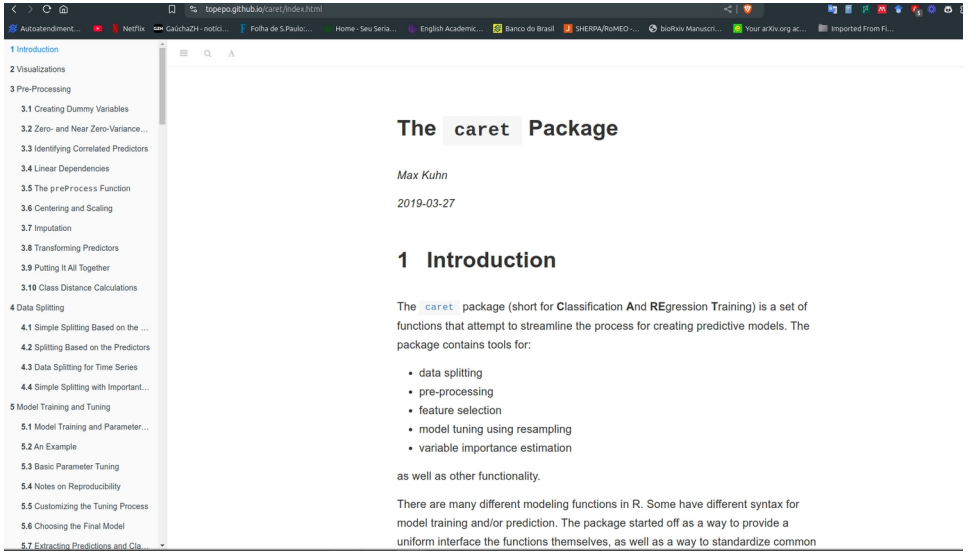
Característica	Data Mining	Knowledge Discovery
Foco	Identificação de padrões e relações em dados de expressão gênica.	Transformação desses padrões em conhecimento biológico útil para a prospecção de drogas.
Técnicas	Algoritmos de clustering, classificação, associação de regras.	Análise estatística, bioinformática, integração de dados de diversas fontes.
Objetivo	Descobrir biomarcadores e genes-alvo.	Descobrir novas drogas, entender mecanismos moleculares e desenvolver diagnósticos.
Exemplo	Identificar um grupo de genes que são consistentemente sobrexpressos em tumores malignos.	Investigar a função desses genes, suas interações com outras proteínas e seu potencial como alvos para novas drogas.
Resultado	Lista de genes ou grupos de genes com padrões de expressão interessantes.	Hipóteses sobre os mecanismos moleculares da doença e novas estratégias terapêuticas.

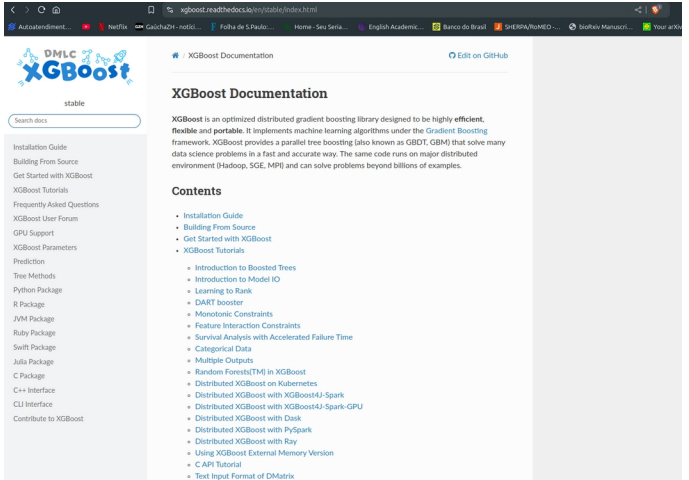
Por que usar R para data mining e knowledge discovery biológicos?

- O R oferece uma vasta gama de pacotes para realizar diversas tarefas de data mining e knowledge discovery.
- A escolha do pacote ideal dependerá da natureza dos dados, do tipo de análise que se deseja realizar e dos objetivos propostos.

Pacotes do R


tidyverse.org
Tidyverse
Packages
Blog
Learn
Help
Contribute
ggplot2
forcats
tibble
readr
stringr
tidyr
purrr
R packages for data science
The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.
Install the complete tidyverse with:
install.packages("tidyverse")
Learn the tidyverse
See how the tidyverse makes data science faster,


topepo.github.io/caret/index.html
1 Introduction
2 Visualizations
3 Pre-Processing
3.1 Creating Dummy Variables
3.2 Zero- and Near Zero-Variance...
3.3 Identifying Correlated Predictors
3.4 Linear Dependencies
3.5 The preProcess Function
3.6 Centering and Scaling
3.7 Imputation
3.8 Transforming Predictors
3.9 Putting It All Together
3.10 Class Distance Calculations
4 Data Splitting
4.1 Simple Splitting Based on the ...
4.2 Splitting Based on the Predictors
4.3 Data Splitting for Time Series
4.4 Simple Splitting with Important...
5 Model Training and Tuning
5.1 Model Training and Parameter...
5.2 An Example
5.3 Basic Parameter Tuning
5.4 Notes on Reproducibility
5.5 Customizing the Tuning Process
5.6 Choosing the Final Model
5.7 Extracting Predictions and Cla...
The caret Package
Max Kuhn
2019-03-27
1 Introduction
The caret package (short for Classification And REgression Training) is a set of functions that attempt to streamline the process for creating predictive models. The package contains tools for:
data splitting
pre-processing
feature selection
model tuning using resampling
variable importance estimation
as well as other functionality.
There are many different modeling functions in R. Some have different syntax for model training and/or prediction. The package started off as a way to provide a uniform interface the functions themselves, as well as a way to standardize common


xgboost.readthedocs.io/en/stable/index.html
DMLC XGBoost
stable
Search docs
Installation Guide
Building From Source
Get Started with XGBoost
XGBoost Tutorials
Frequently Asked Questions
XGBoost User Forum
GPU Support
XGBoost Parameters
Prediction
Tree Methods
Python Package
R Package
JVM Package
Ruby Package
Swift Package
Julia Package
C Package
C++ Interface
CLI Interface
Contribute to XGBoost
XGBoost Documentation
Edit on GitHub
XGBoost Documentation
XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way. The same code runs on major distributed environment (Hadoop, SGE, MPI) and can solve problems beyond billions of examples.
Contents
Installation Guide
Building From Source
Get Started with XGBoost
XGBoost Tutorials
Introduction to Boosted Trees
Introduction to Model IO
Learning to Rank
DART booster
Monotonic Constraints
Feature Interaction Constraints
Survival Analysis with Accelerated Failure Time
Categorical Data
Multiple Outputs
Random Forests(TM) in XGBoost
Distributed XGBoost on Kubernetes
Distributed XGBoost with XGBoost4J-Spark
Distributed XGBoost with XGBoost4J-Spark-GPU
Distributed XGBoost with Dask
Distributed XGBoost with PySpark
Distributed XGBoost with Ray
Using XGBoost External Memory Version
C API Tutorial
Text Input Format of DMatrix

Integração de pacotes do R para data mining e knowledge discovery para dados biológicos

- A integração de pacotes R é fundamental para realizar análises complexas e abrangentes em dados biológicos.
- Ao combinar diferentes pacotes, pode-se construir pipelines de análise robustas e eficientes, desde a limpeza dos dados até a interpretação dos resultados.

Tipos de integrações de pacotes do R

- **Sequencial:**
  - Descrição: A execução de um pacote após outro, com a saída de um sendo a entrada do próximo.
  - Exemplo: Limpeza dos dados com o pacote tidyverse, seguido da análise de expressão gênica com o pacote DESeq2 e, finalmente, a visualização dos resultados com ggplot2.
- **Em Paralelo:**
  - Descrição: Múltiplos pacotes são executados simultaneamente em diferentes partes dos dados.
  - Exemplo: Utilizar o pacote foreach para paralelizar cálculos intensivos, como a análise de diferentes conjuntos de genes.

Tipos de integrações de pacotes do R

- **Dentro de Funções Personalizadas:**
  - Descrição: Criação de funções que combinam a funcionalidade de diversos pacotes.
  - Exemplo: Criar uma função que realize a normalização de dados, a seleção de genes e a análise de enriquecimento genético, utilizando pacotes como limma, DESeq2 e clusterProfiler.
- **Utilizando Pipelines:**
  - Descrição: Construção de pipelines de análise que definem uma sequência de passos a serem executados, com a possibilidade de parametrização e reutilização.
  - Exemplo: Utilizar o pacote magrittr para criar pipelines concisos e legíveis, ou o pacote drake para gerenciar dependências e reprodutibilidade.

Exemplos de Integração de Pacotes