

# Sarah Hannah Alves - Especialista em Ciência de dados

---

- Formação Acadêmica

- Graduação engenharia metalúrgica UFRJ (2011): “Graph based optimization algorithm implementation and mathematical modeling”
- Mestrado em engenharia química PUC-Rio (2017) : “Neural networks on process mining optimization”
- Mestrado em engenharia elétrica PUC-Rio (2020) : “Explainable AI on synthetic data generation”
- Especialização em Data Science ITA (2022): “Classificador automático de medidores de energia elétrica baseado em Deep learning e XAI”

- Experiência profissional:

- Engineering background, DataScientist Specialist, Master's and Doctoral student in Artificial Intelligence. Also, Bike Entrepreneur.
- From 2008 to 2016 I worked with ETL data integration, graph-based optimization and time series algorithms in engineering and management areas to develop reports, dashboards, and analyses.
- From 2016 until now I've been working and studying other shallow and deep learning algorithms to solve different industries problems.

# Sarah Hannah Alves - Especialista em Ciência de dados

---



**Sarah Hannah Alves**  
Data Scientist Specialist



LinkedIn

People ▾

Sarah Hannah

Alves



**Sarah Hannah Alves**

Data Scientist Specialist  
Rio de Janeiro, Rio de Janeiro, Brazil  
861 followers · 500+ connections



Energisa



Rio de Janeiro State University

# Agrupamento fuzzy aplicado à integração de dados multiômicos

Sarah Hannah Alves



PONTIFÍCIA UNIVERSIDADE CATÓLICA  
DO RIO DE JANEIRO



# Resumo

---

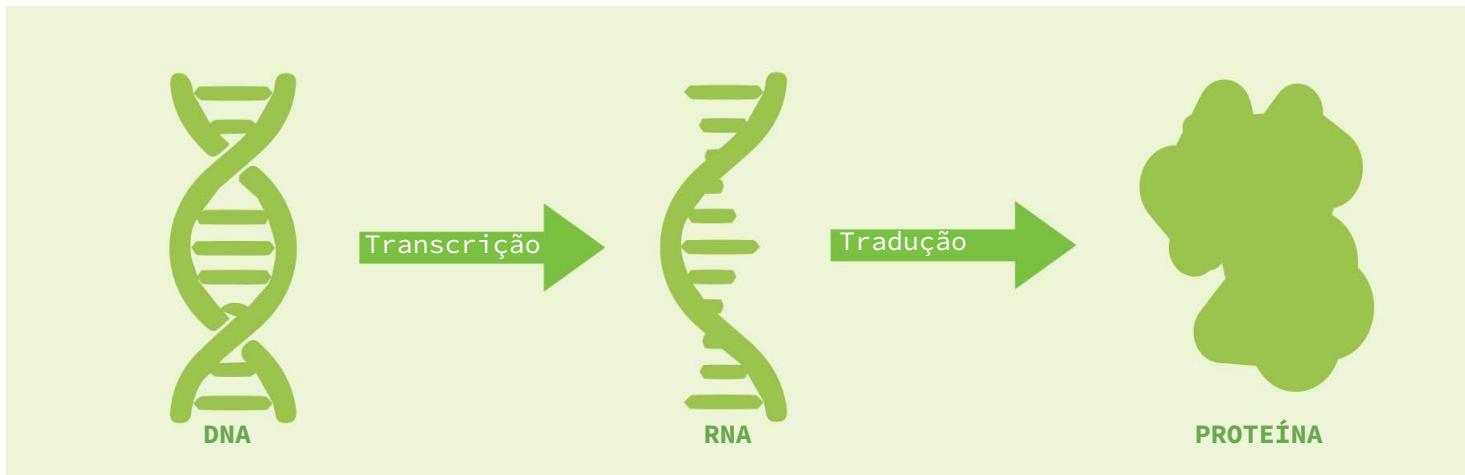
- Introdução
- Justificativa
- Objetivos
- Técnicas de integração de dados
- Análise computacional de dados biológicos
- Abordagem ômica x multiômica
- Seleção de atributos para dados multiômicos
- Agrupamento de dados multiômicos
- Estudo de caso

# Introdução

# Introdução

---

- Dogma central da biologia molecular

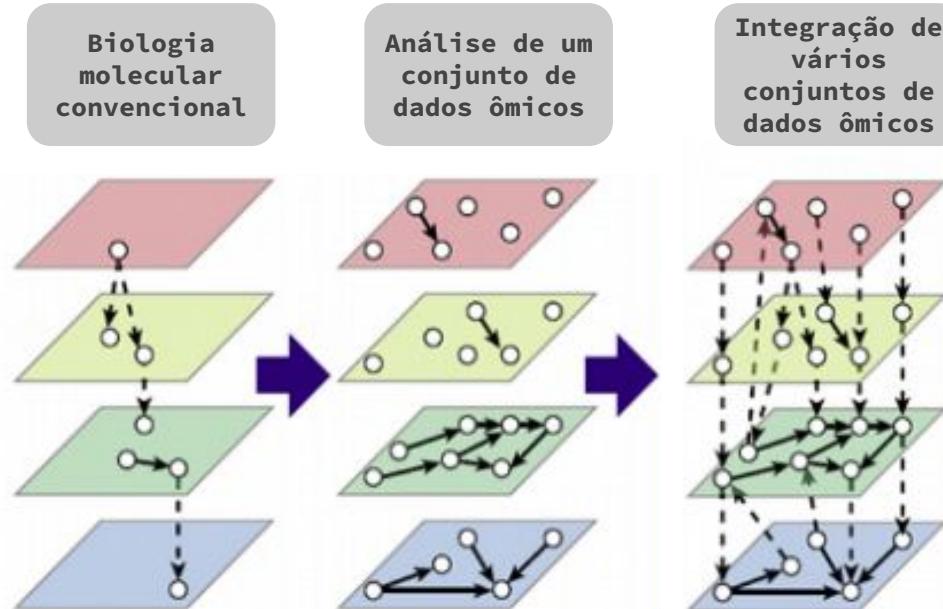


# Justificativa

# Justificativa

---

- Integração de diferentes conjuntos de dados



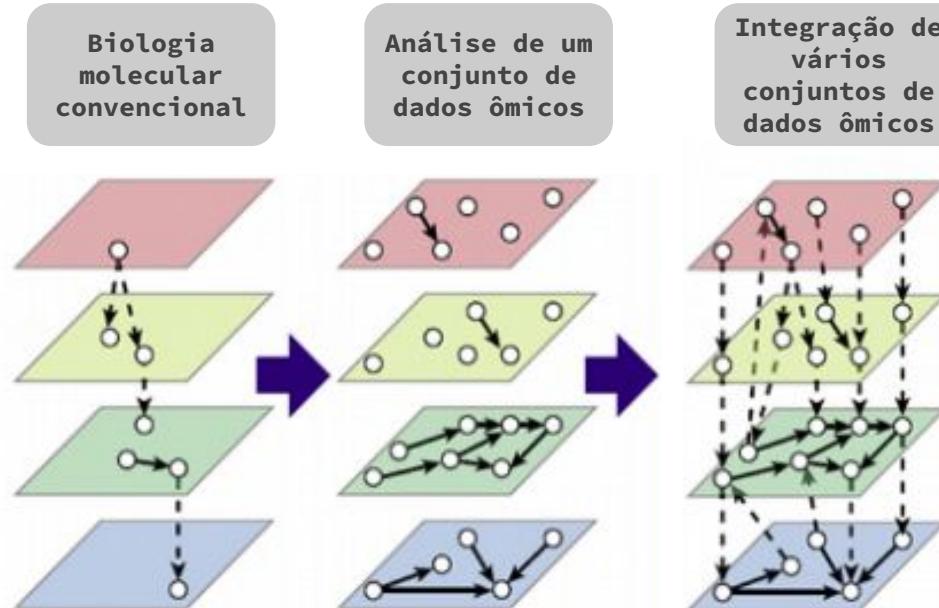
- Alta dimensionalidade
- Esparsidade

# Justificativa

---

- Integração de diferentes conjuntos de dados

THE CANCER GENOME ATLAS

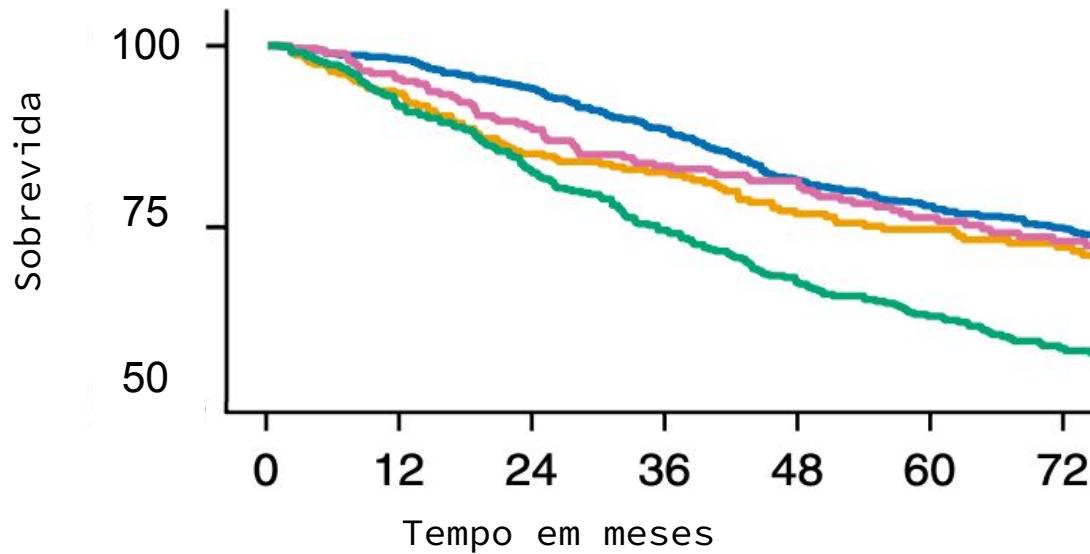


- Alta dimensionalidade
- Esparsidade

# Justificativa

---

- Caracterização de subtipos moleculares

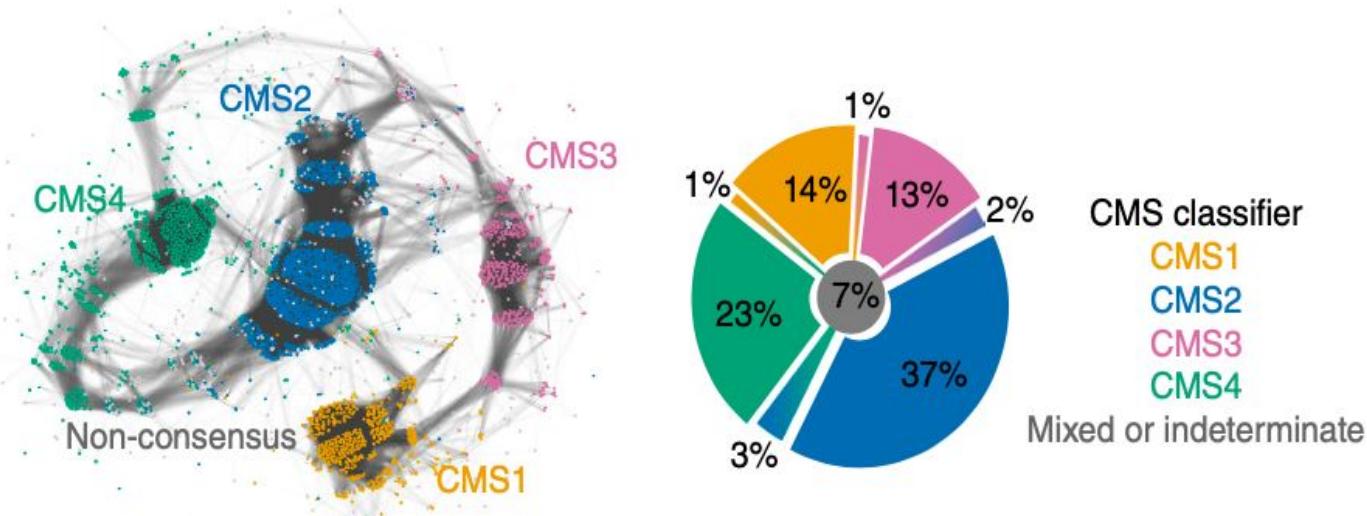


Fonte: Adaptado de The consensus molecular subtypes of colorectal cancer

# Justificativa

---

- Caracterização de subtipos moleculares



Fonte: Adaptado de The consensus molecular subtypes of colorectal cancer

# Justificativa

---

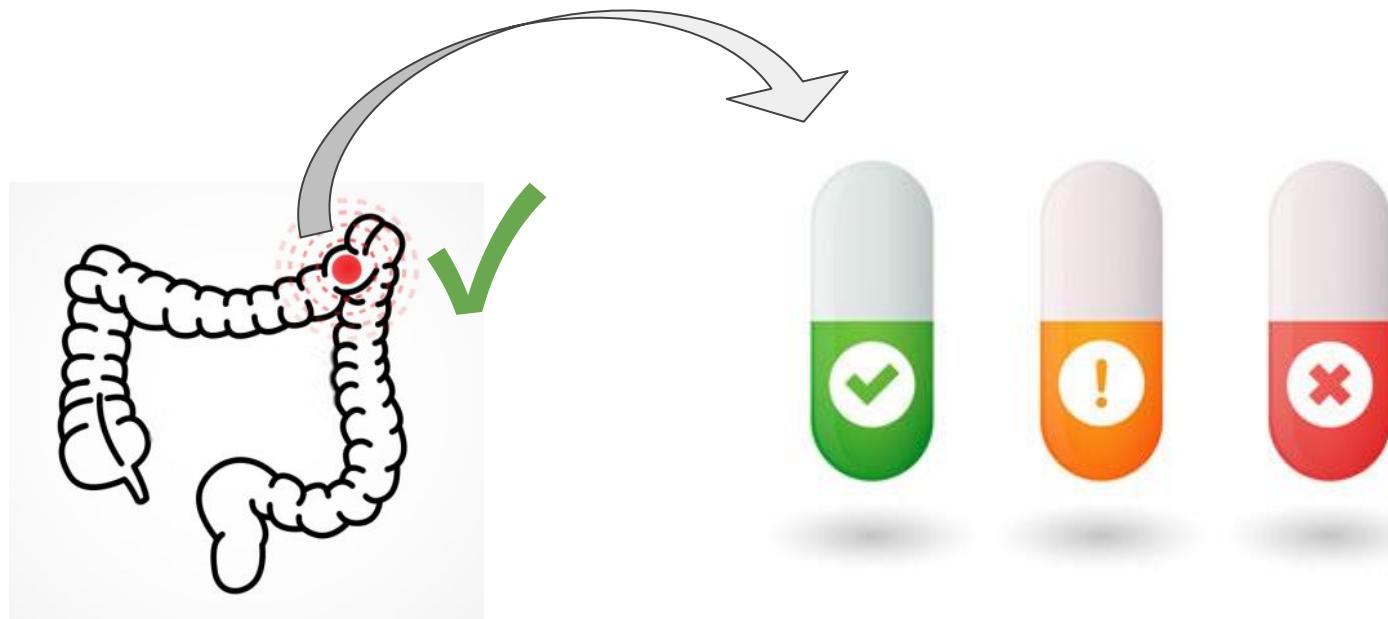
- Caracterização de subtipos moleculares



# Justificativa

---

- Caracterização de subtipos moleculares



# Objetivos

# Objetivos

---

- Geral:

Desenvolver metodologia capaz de relacionar dados ômicos e subtipos moleculares com maior precisão.

# Objetivos

---

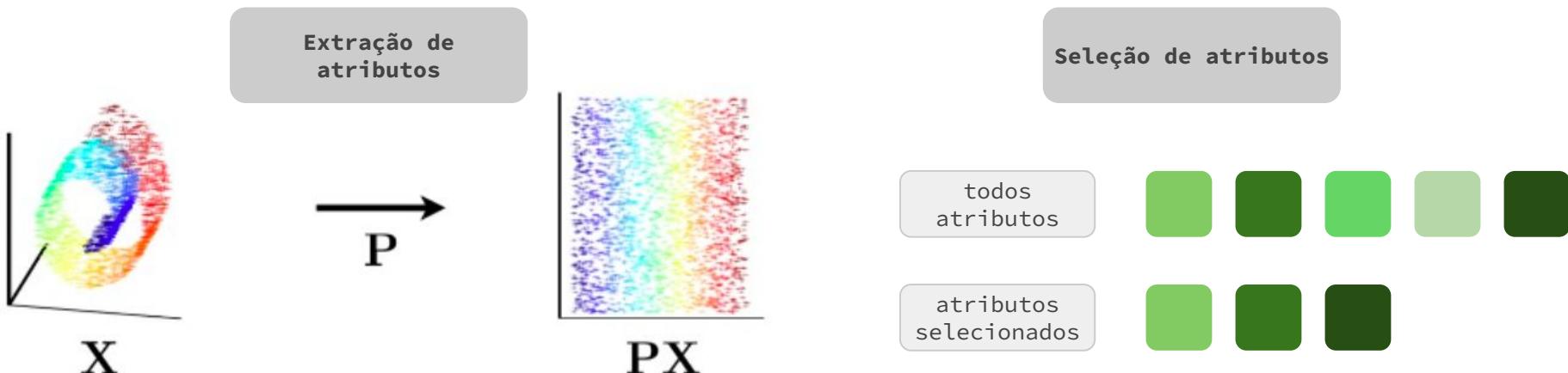
- Específicos:
  - Avaliar a **contribuição da adição de conjuntos de dados** à resultados da literatura e escolher quais devem ser analisados.
  - Analisar **métodos de seleção de atributos** adequados à dados multiômicos.
  - Caracterizar dados de pacientes com **gradações entre diferentes subtipos moleculares**
  - Identificar **outros perfis de subtipos moleculares**.
  - **Diminuir o compartilhamento** de características dos perfis de subtipos moleculares adicionando dados multiômicos.
  - **Validar os subtipos** moleculares identificados com relação às curvas de sobrevida de cada grupo, e demais características biológicas.
  - **Relacionar** os grupos deste trabalho com **grupos** já definidos na **literatura**

# Técnicas de integração de dados

# Técnicas de integração de dados

---

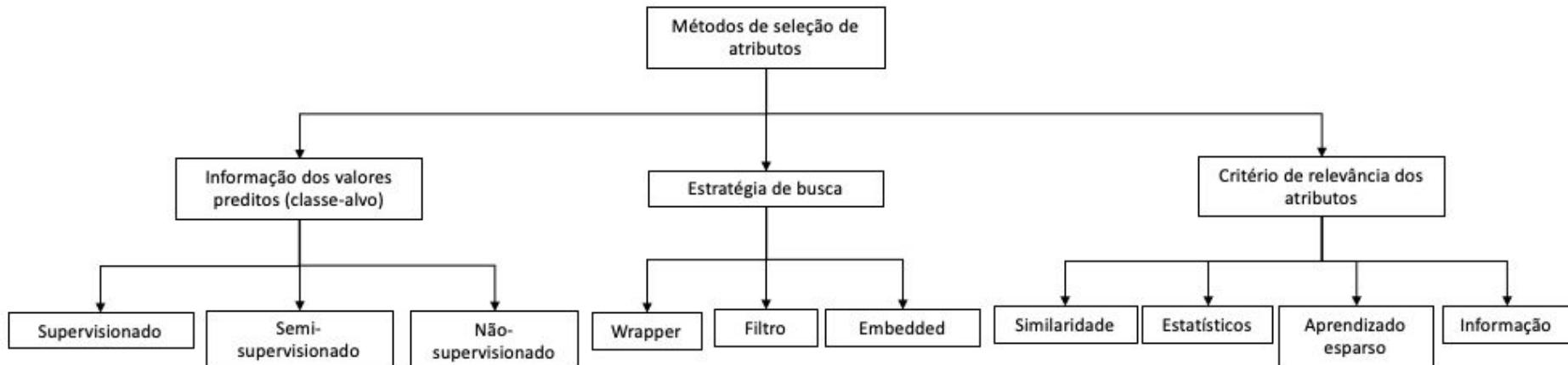
- Redução de dimensionalidade



# Técnicas de integração de dados

---

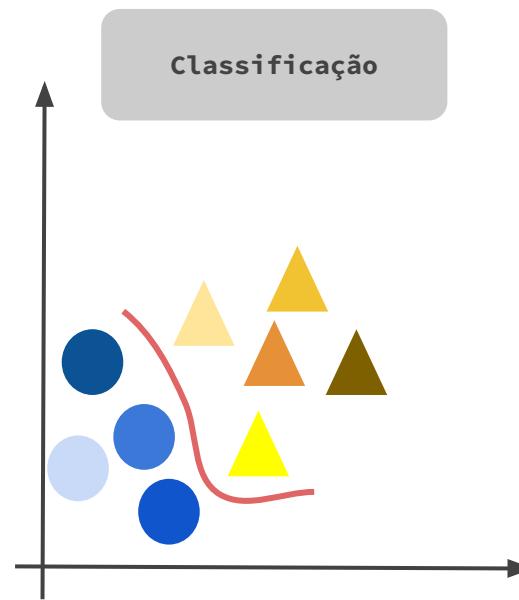
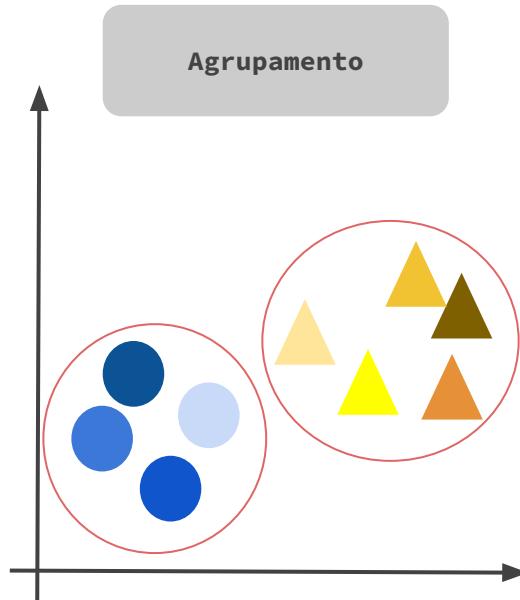
- Seleção de atributos



# Técnicas de integração de dados

---

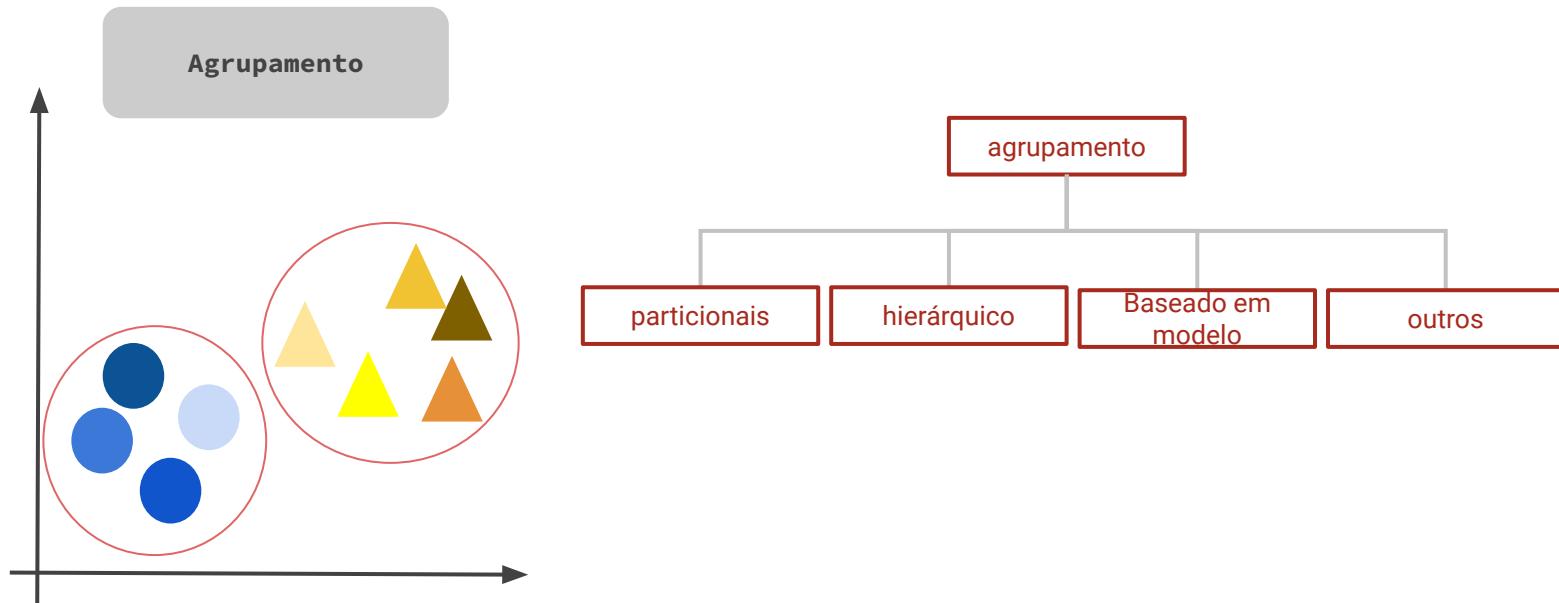
- Identificação de grupos ou classes



# Técnicas de integração de dados

---

- Identificação de grupos ou classes

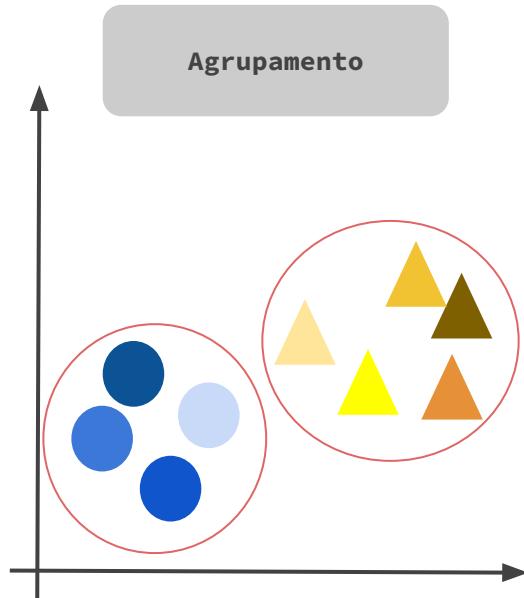


# Técnicas de integração de dados

---

particionais

- Identificação de grupos ou classes



Vantagens métodos tradicionais:

- 1) Escaláveis
- 2) Simplicidade de implementação
- 3) Boa performance de tempo de execução principalmente para conjuntos de dados limpos, pequenos e sintéticos

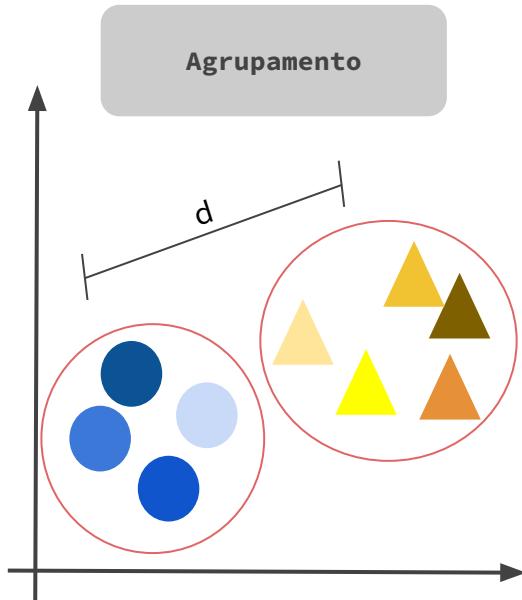
Desvantagens: utilização limitada para dados ruidosos e sobrepostos.

# Técnicas de integração de dados

---

particionais

- Identificação de grupos ou classes



Etapas métodos tradicionais:

- Define centróide: representa características numéricas e/ou categóricas
- Define medida de distância que compatibilize características numéricas e/ou categóricas
- Define função de custo a ser minimizada

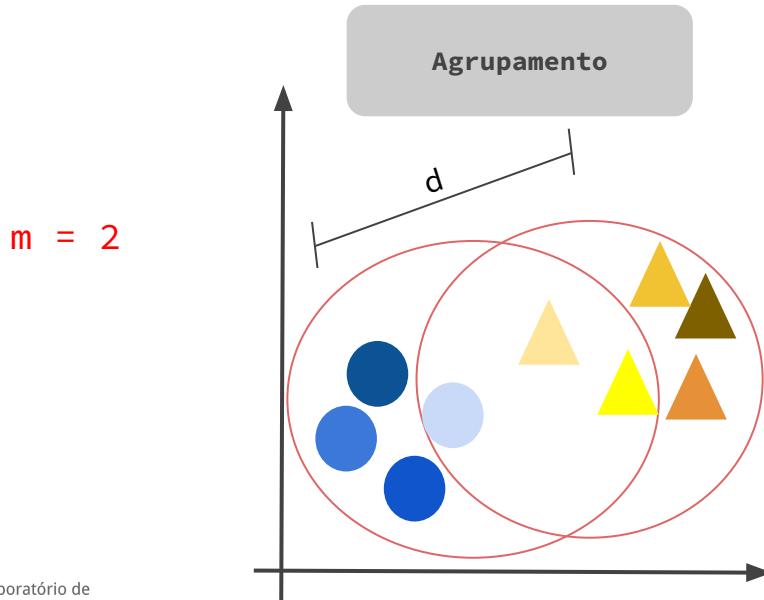
$$\sum_{i=1}^n \xi(d_i, C_i)$$

Exemplo: k-means

# Técnicas de integração de dados

---

- Identificação de grupos ou classes



Fuzzy c-means

$$J'(U, V) = \sum_{i=1}^C \sum_{j=1}^N u_{ij}^m \|X_j - V_i\|^2$$

Etapas agrupamento fuzzy c-means:

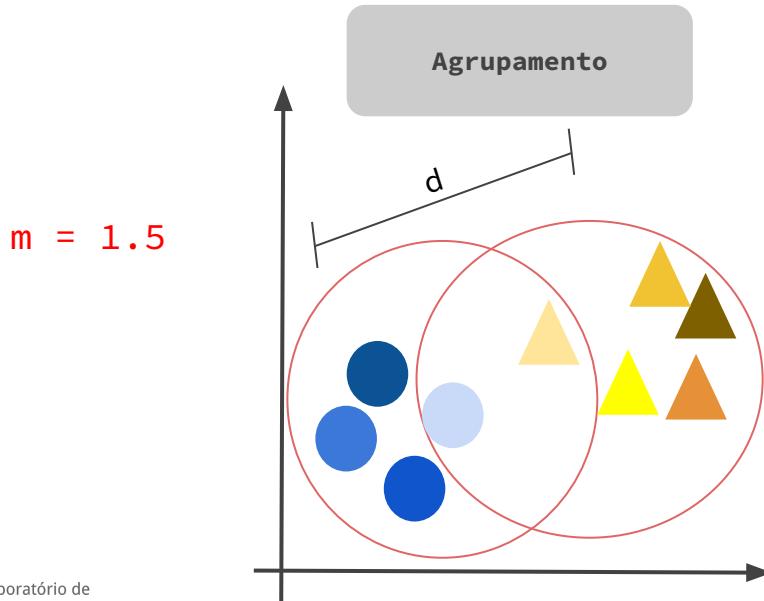
- 1) Escolher os centros C aleatoriamente (vetor V)
- 2) Calcular a matriz de partição U
- 3) Atualizar os centros  $v_i$
- 4) Computar a função objetivo J
- 5) Repetir as etapas 2 a 4 até a convergência ( $\|var(J)\| \leq \epsilon$ )

Vantagens: Para dados ruidosos e sobrepostos.

# Técnicas de integração de dados

---

- Identificação de grupos ou classes



Fuzzy c-means

$$J'(U, V) = \sum_{i=1}^C \sum_{j=1}^N u_{ij}^m \|X_j - V_i\|^2$$

Etapas agrupamento fuzzy c-means:

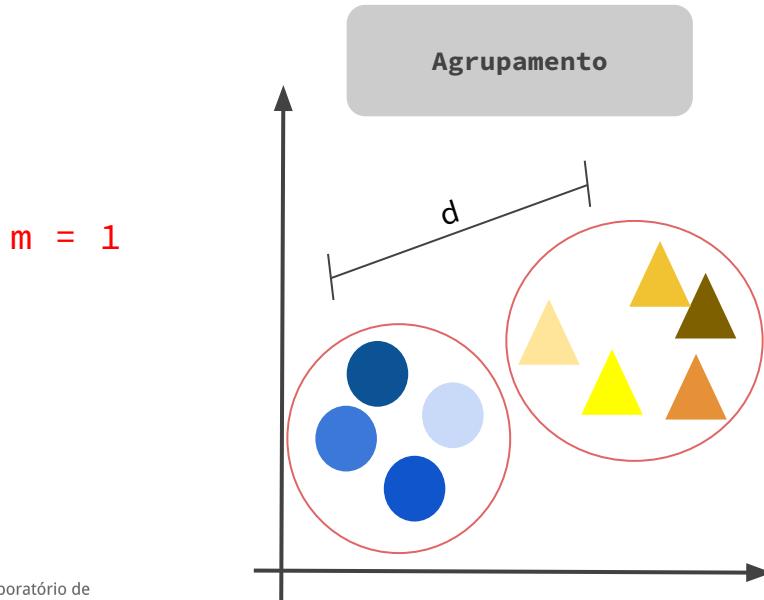
- 1) Escolher os centros C aleatoriamente (vetor V)
- 2) Calcular a matriz de partição U
- 3) Atualizar os centros  $v_i$
- 4) Computar a função objetivo J
- 5) Repetir as etapas 2 a 4 até a convergência ( $\|var(J)\| \leq \epsilon$ )

Vantagens: Para dados ruidosos e sobrepostos.

# Técnicas de integração de dados

---

- Identificação de grupos ou classes



Fuzzy c-means

$$J'(U, V) = \sum_{i=1}^C \sum_{j=1}^N u_{ij}^m \|X_j - V_i\|^2$$

Etapas agrupamento fuzzy c-means:

- Escolher os centros C aleatoriamente (vetor V)
- Calcular a matriz de partição U
- Atualizar os centros  $v_i$
- Computar a função objetivo J
- Repetir as etapas 2 a 4 até a convergência ( $\|var(J)\| \leq \epsilon$ )

Análogo: **ark-means** ruidosos e sobrepostos.

# Técnicas de integração de dados

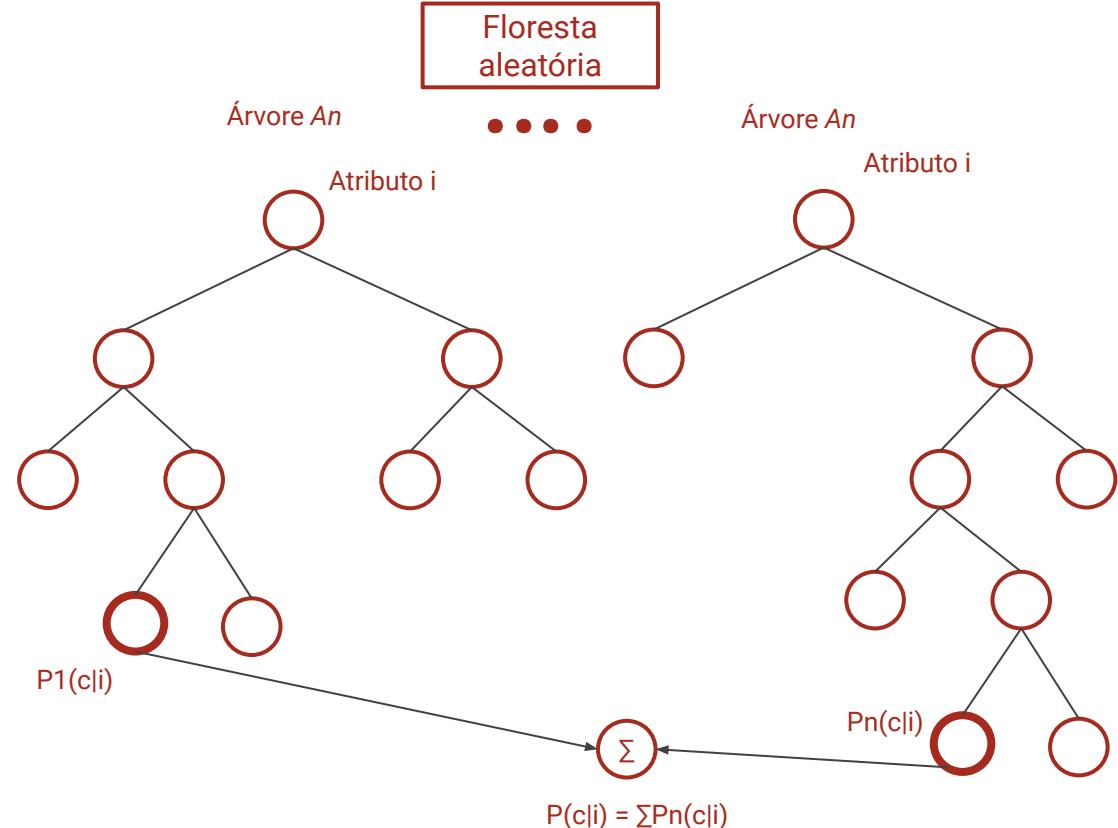
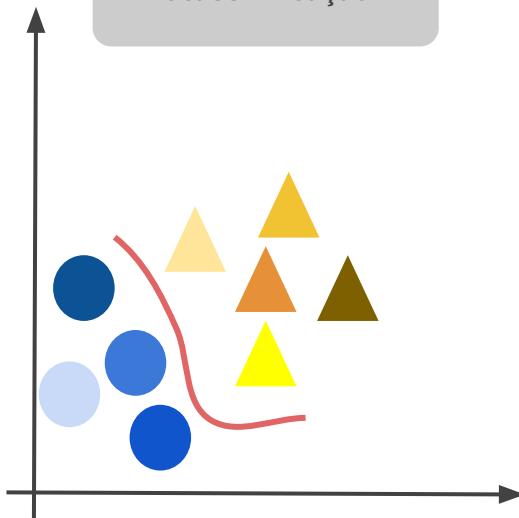
---

- Métrica de desempenho avaliam contribuição dos parâmetros e resultado:
  - **FS** : Representa melhor o número de grupos
  - **FPC** : Representa melhor o valor médio do grau de pertinência
  - **XB** : Pondera ambos valores do grau de fuzzificação e número de grupos

# Técnicas de integração de dados

- Identificação de grupos ou classes

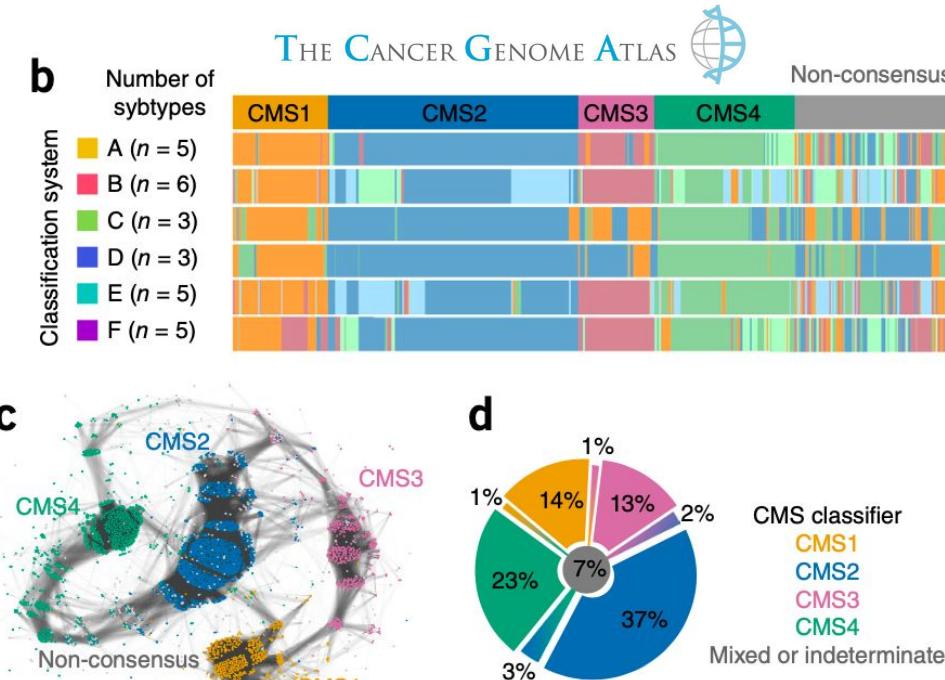
Classificação



# Análise computacional de dados biológicos

# Análise computacional de dados biológicos

- Guinney, 2015.

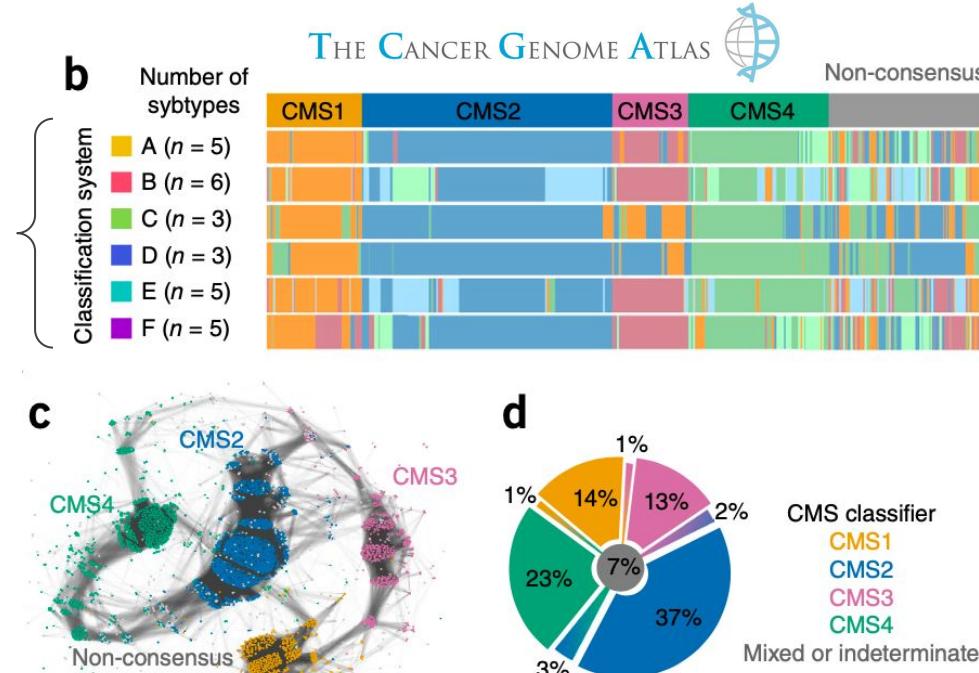


Fonte: Adaptado de The consensus molecular subtypes of colorectal cancer

# Análise computacional de dados biológicos

- Guinney, 2015.

Técnicas de agrupamento semelhantes



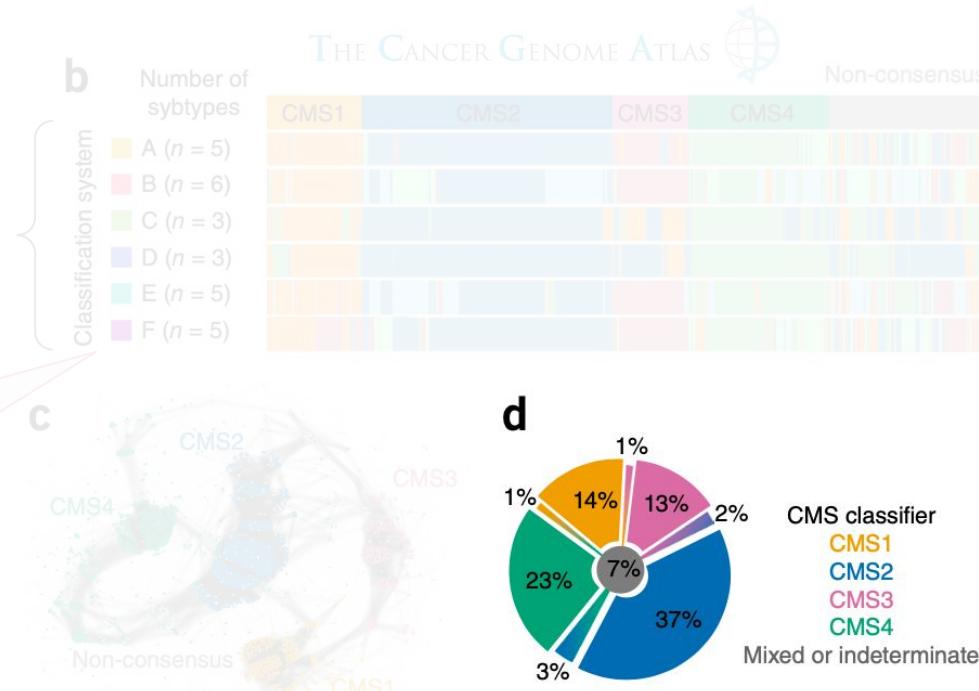
Fonte: Adaptado de The consensus molecular subtypes of colorectal cancer

# Análise computacional de dados biológicos

- Guinney, 2015.

Técnicas de agrupamento semelhantes

Amostras agrupadas em apenas um grupo, não refletem alto compartilhamento de características entre grupos

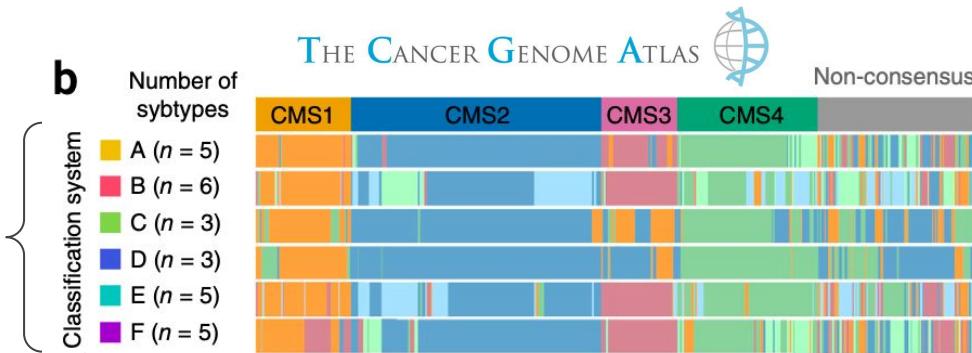


Fonte: Adaptado de The consensus molecular subtypes of colorectal cancer

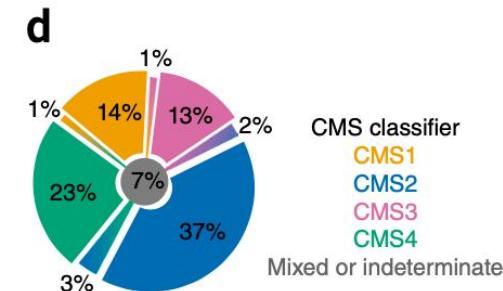
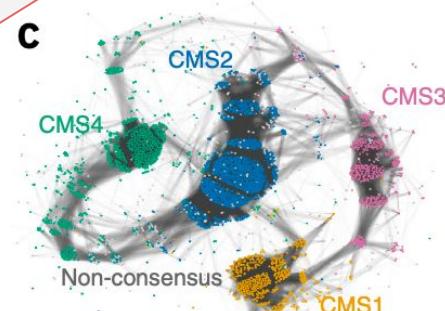
# Análise computacional de dados biológicos

- Guinney, 2015.

Técnicas de agrupamento semelhantes



Pacientes agrupados em apenas um grupo, não refletem alto compartilhamento de características entre grupos



Fonte: Adaptado de The consensus molecular subtypes of colorectal cancer

# Abordagem ômica x multiômica

# Objetivos

---

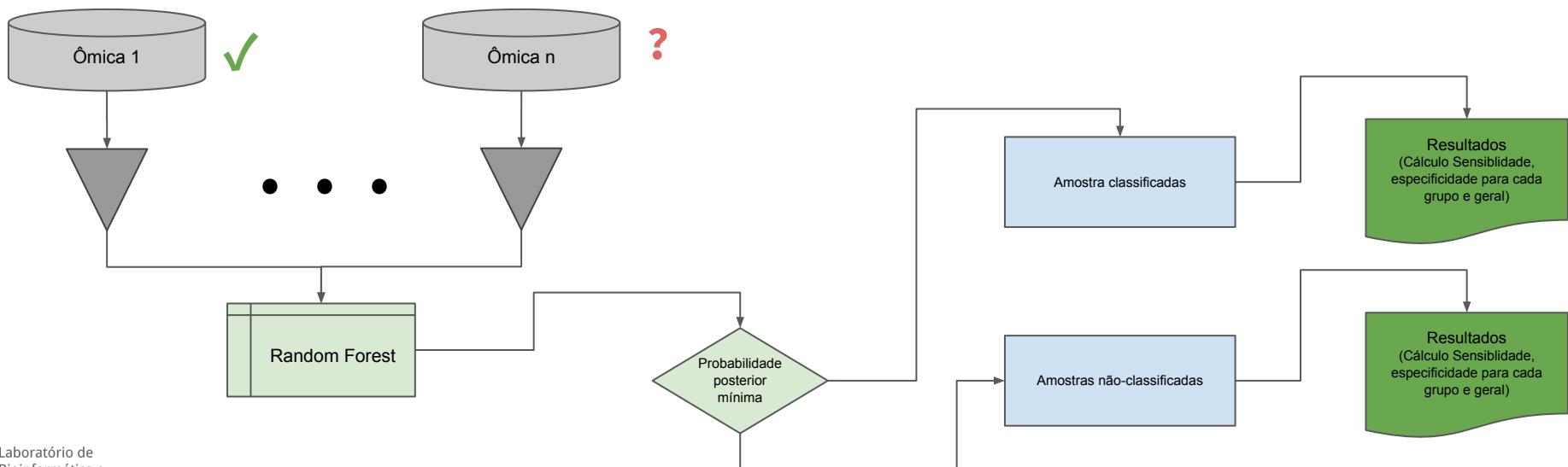
- Específicos:

- Avaliar a **contribuição da adição de conjuntos de dados** à resultados da literatura e escolher quais devem ser analisados.
- Analisar **métodos de seleção de atributos** adequados à dados multiômicos.
- Caracterizar dados de pacientes com **gradações entre diferentes subtipos moleculares**
- Identificar **outros perfis de subtipos moleculares**.
- **Diminuir o compartilhamento** de características dos perfis de subtipos moleculares adicionando dados multiômicos.
- **Validar os subtipos** moleculares identificados com relação às curvas de sobrevida de cada grupo, e demais características biológicas.
- **Relacionar os grupos** deste trabalho com **grupos** já definidos na **literatura**

# Abordagem ômica x multiômica

---

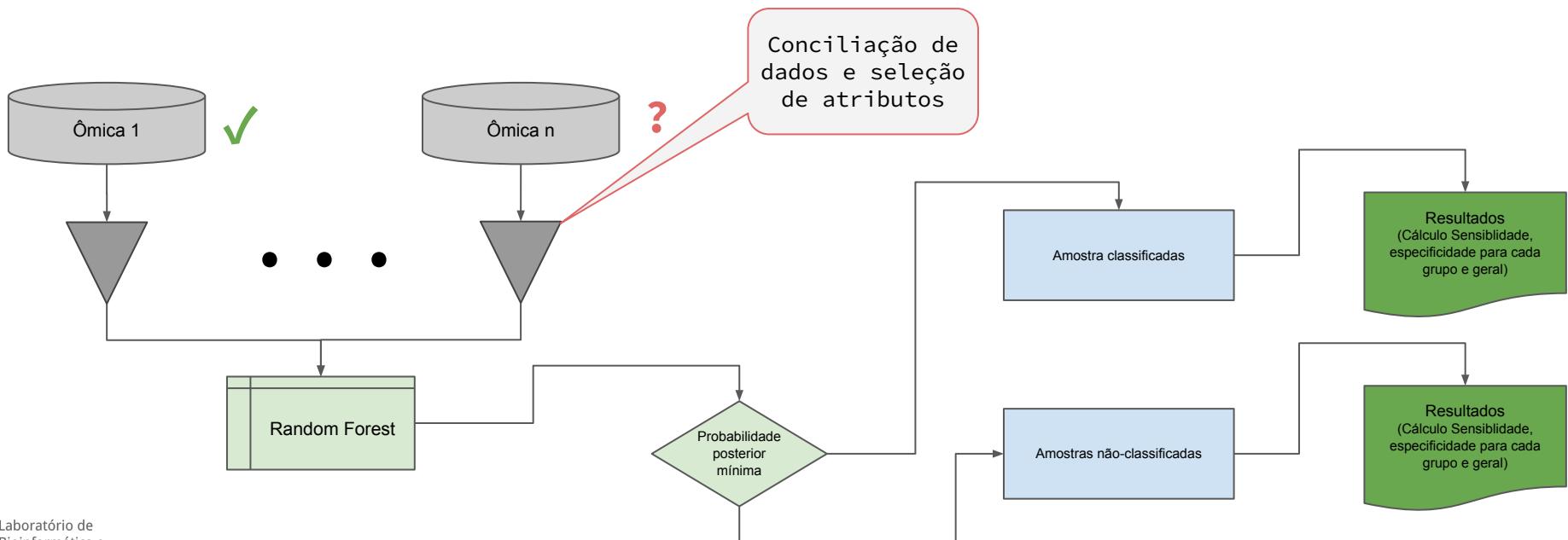
- Elaboração de classificador multiômico com base em trabalhos da literatura.



# Abordagem ômica x multiômica

---

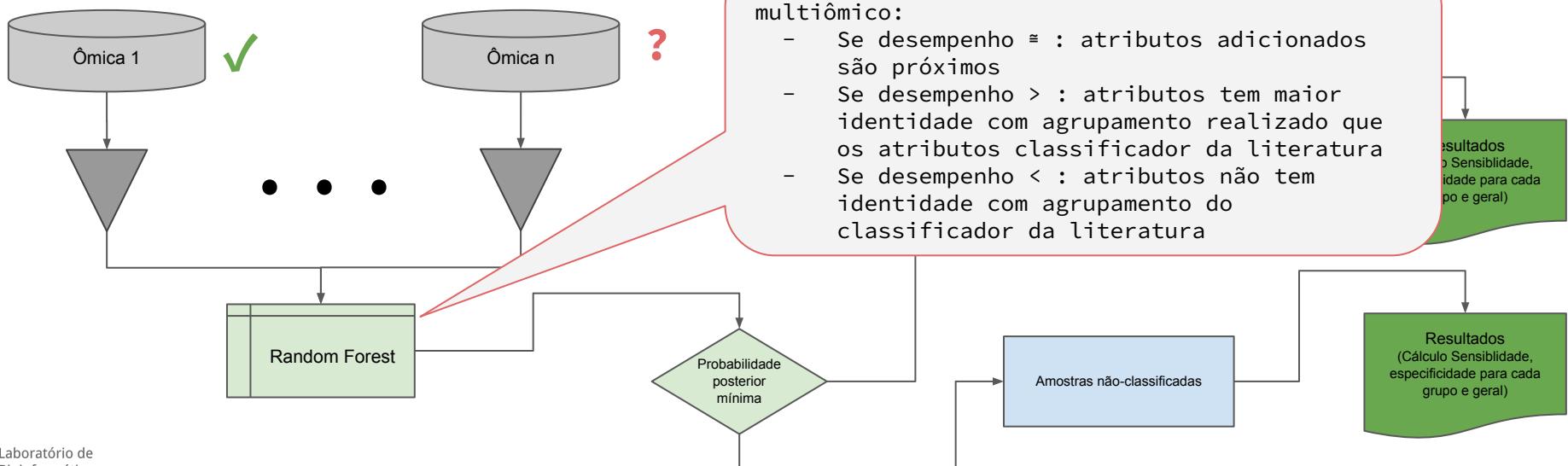
- Elaboração de classificador multiômico com base em trabalhos da literatura.



# Abordagem ômica x multiômica

---

- Elaboração de classificador multiômico com base em trabalhos da literatura.



# Seleção de atributos para dados multiônicos

# Objetivos

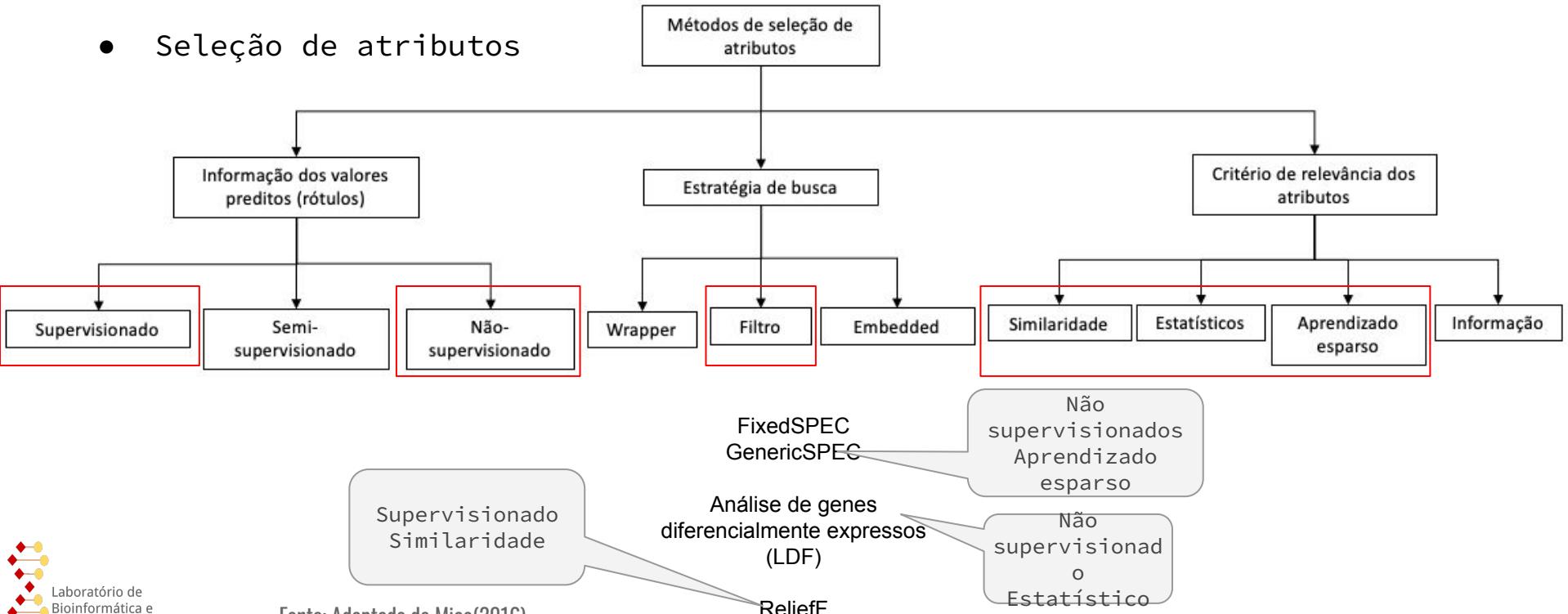
---

- Específicos:

- Avaliar a **contribuição da adição de conjuntos de dados** à resultados da literatura e escolher quais devem ser analisados.
- Analisar **métodos de seleção de atributos** adequados à dados multiônicos.
- Caracterizar dados de pacientes com **gradações entre diferentes subtipos moleculares**
- Identificar **outros perfis de subtipos moleculares**.
- **Diminuir o compartilhamento** de características dos perfis de subtipos moleculares adicionando dados multiônicos.
- **Validar os subtipos** moleculares identificados com relação às curvas de sobrevida de cada grupo, e demais características biológicas.
- **Relacionar os grupos** deste trabalho com **grupos** já definidos na **literatura**

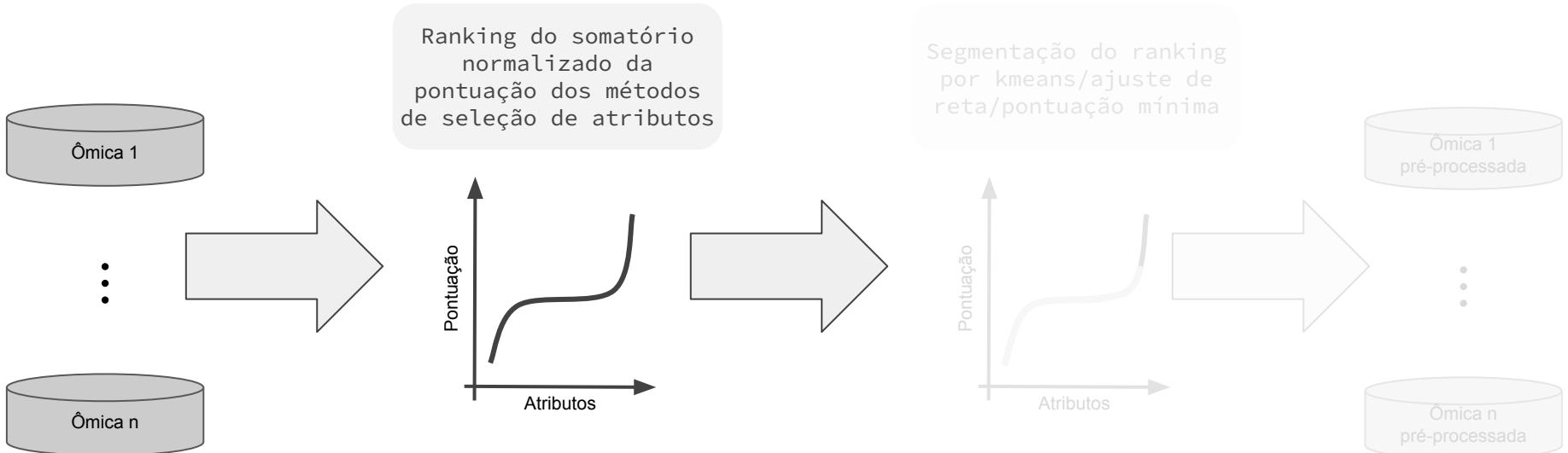
# Seleção de atributos para dados multiônicos

- Seleção de atributos



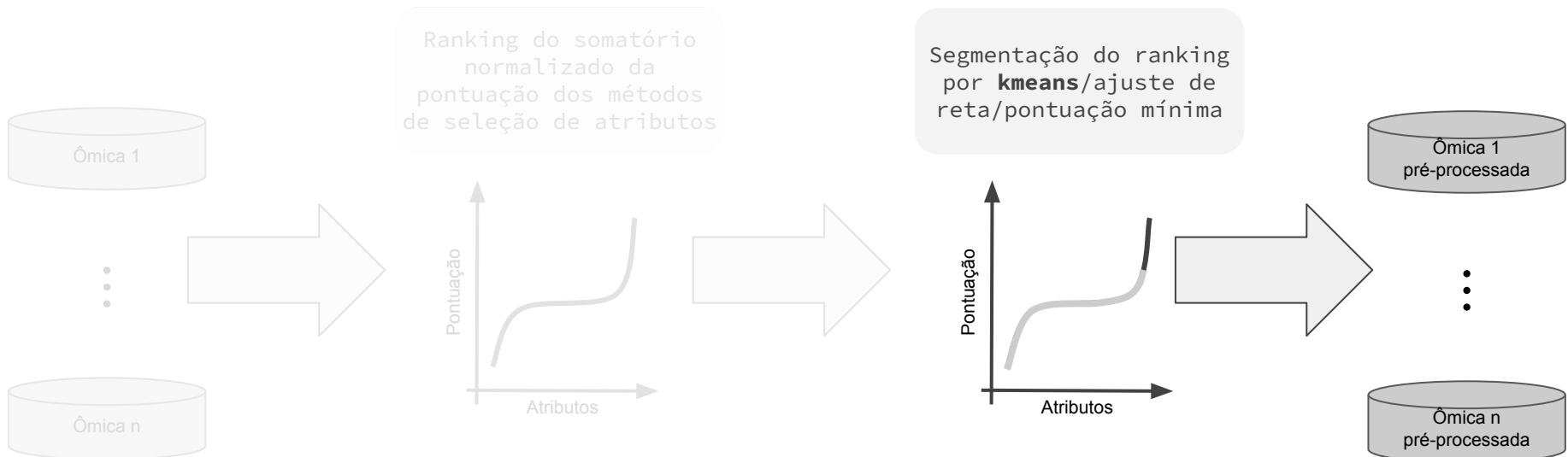
# Seleção de atributos para dados multiônicos

---



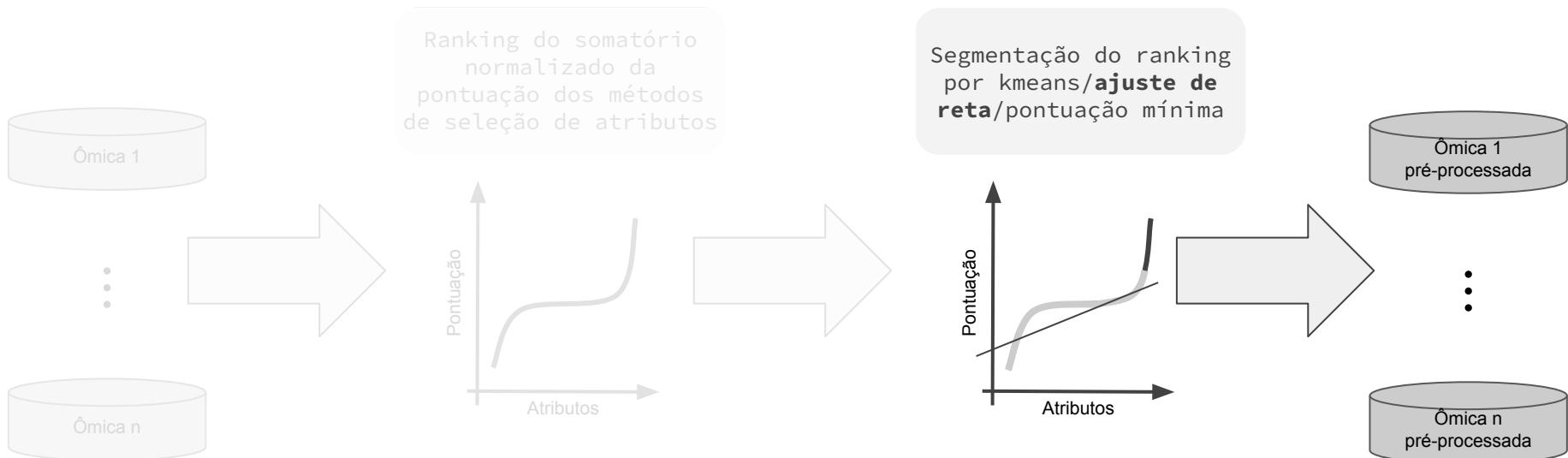
# Seleção de atributos para dados multiônicos

---



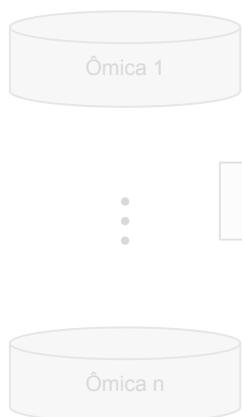
# Seleção de atributos para dados multiônicos

---

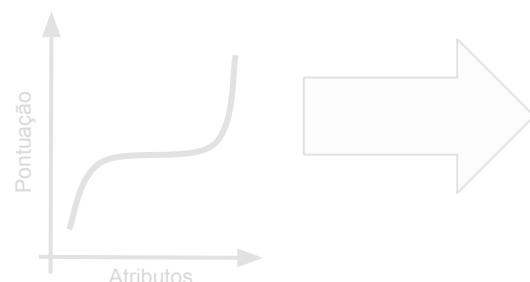


# Seleção de atributos para dados multiônicos

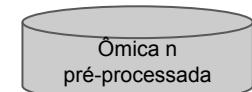
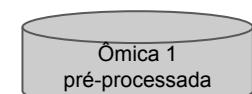
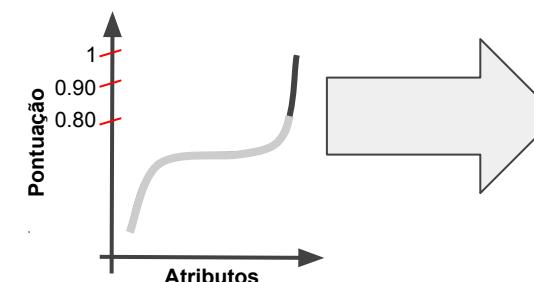
---



Ranking do somatório  
normalizado da  
pontuação dos métodos  
de seleção de atributos



Segmentação do ranking  
por kmeans/ajuste de  
reta/**pontuação mínima**

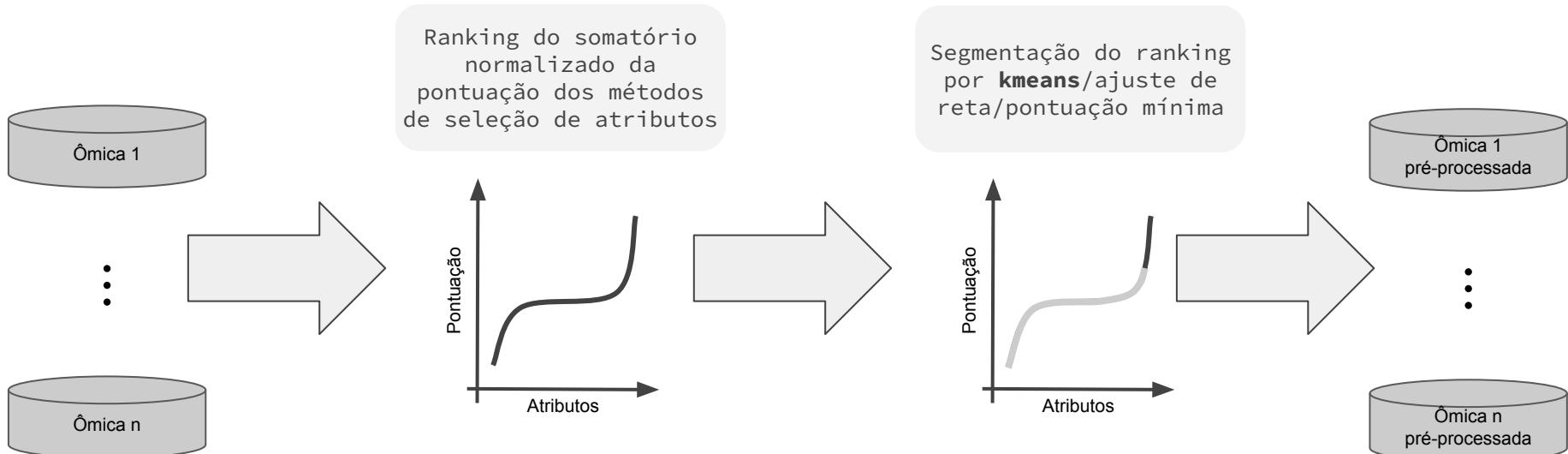


1º teste: Atributos com  
pontuação acima de  
0.5, 0.6, 0.7, 0.8 e 0.9.

2º teste: diminuir intervalo  
(por exemplo, 0.8, 0.825 e 0.85  
etc)

# Seleção de atributos para dados multiônicos

---



# Agrupamento de dados multiônicos

# Objetivos

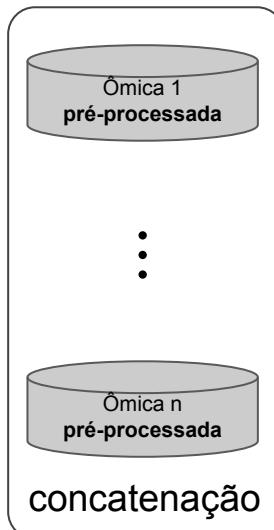
---

- Específicos:

- Avaliar a **contribuição da adição de conjuntos de dados** à resultados da literatura e escolher quais devem ser analisados.
- Analisar **métodos de seleção de atributos** adequados à dados multiômicos.
- Caracterizar dados de pacientes com **gradações entre diferentes subtipos moleculares**
- Identificar **outros perfis de subtipos moleculares**.
- **Diminuir o compartilhamento** de características dos perfis de subtipos moleculares adicionando dados multiômicos.
- **Validar os subtipos** moleculares identificados com relação às curvas de sobrevida de cada grupo, e demais características biológicas.
- **Relacionar os grupos** deste trabalho com **grupos** já definidos na **literatura**

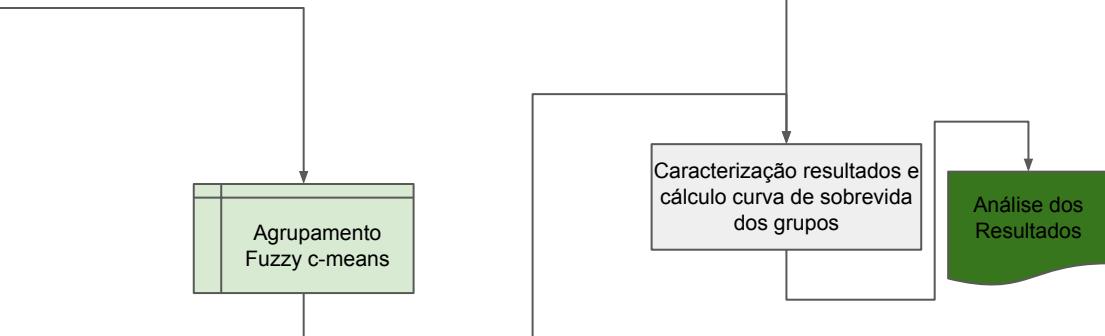
# Agrupamento de dados multiômicos

ômica 1			
	atributo A	atributo B	atributo C
paciente 1	...	...	...
paciente 3	...	...	...
paciente 4	...	...	...
paciente 5	...	...	...



ômica 2			
	atributo D	atributo E	atributo F
paciente 1	...	...	...
paciente 2	...	...	...
paciente 3	...	...	...
paciente 4	...	...	...

	ômica 1			ômica 2		
	atributo A	atributo B	atributo C	atributo D	atributo E	atributo F
paciente 1	...	...	...	...	...	...
paciente 3	...	...	...	...	...	...
paciente 4	...	...	...	...	...	...



# Agrupamento de dados multiômicos

---

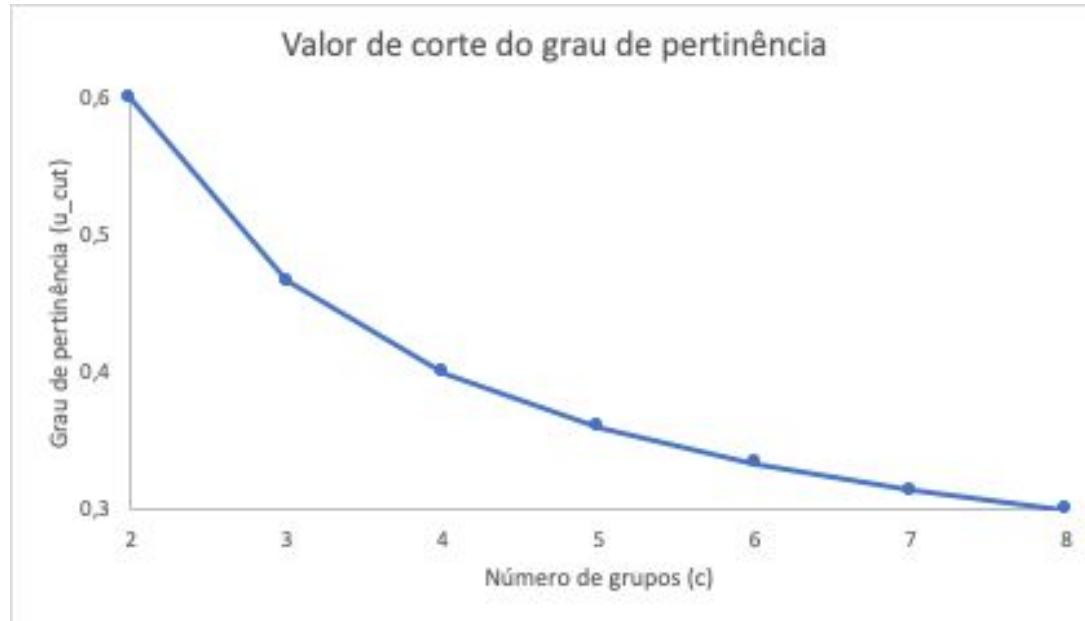
- Agrupamento fuzzy

Paciente	Grupo			
	1	2	3	4
TCGA-AG-3592	0.1	0.2	0.5	0.2
TCGA-AG-3999	0.3	0.2	0.3	0.2
TCGA-BM-6198	0.7	0.1	0	0.2
TCGA-AF-3400	0.3	0	0.2	0.5
⋮	⋮	⋮	⋮	⋮

# Agrupamento de dados multiômicos

---

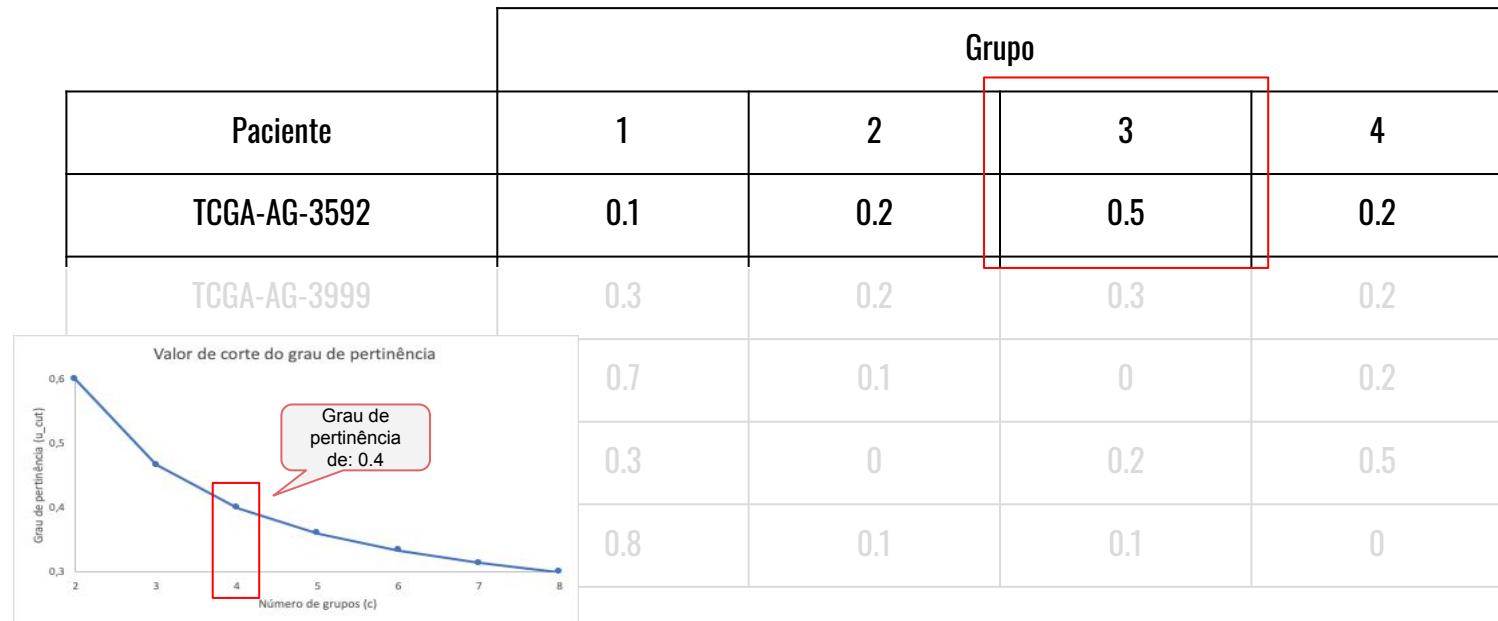
- Grau de pertinência: valor mínimo ajustado de acordo com número de grupos



# Agrupamento de dados multiômicos

---

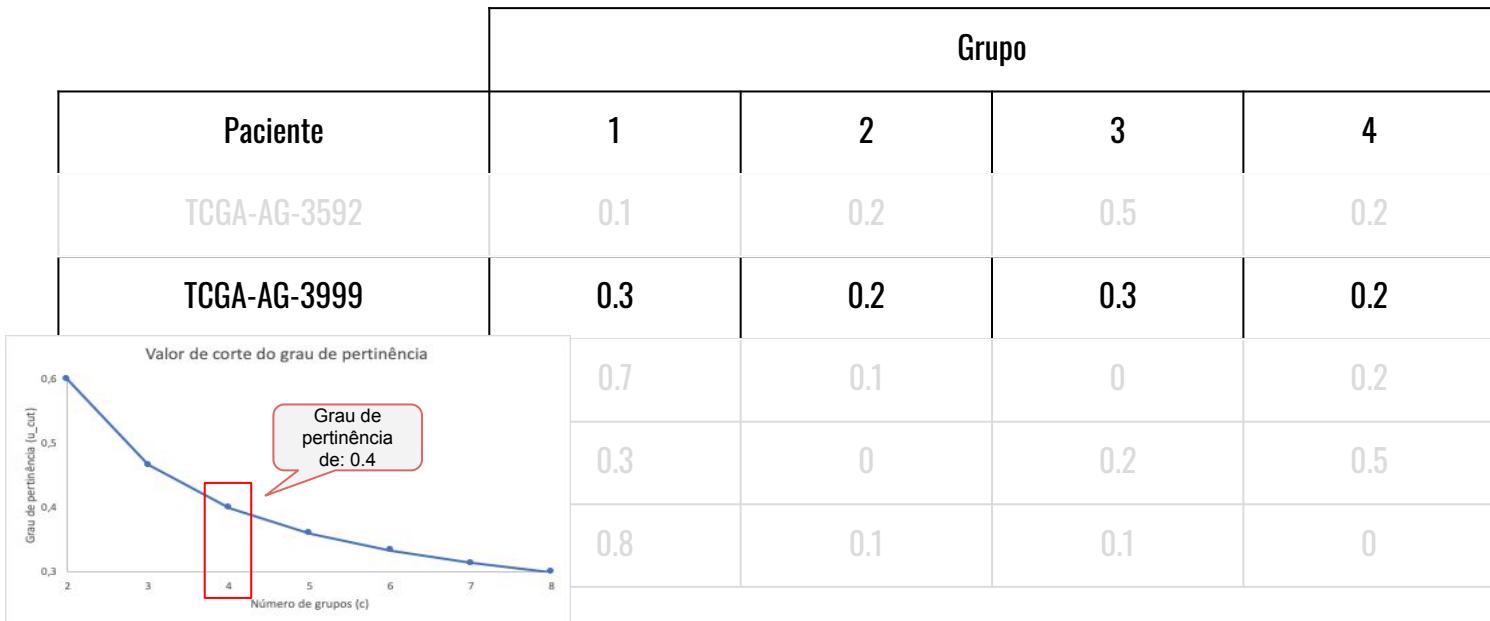
- Grau mínimo de pertinência: que decresce com aumento do número de grupos



# Agrupamento de dados multiômicos

---

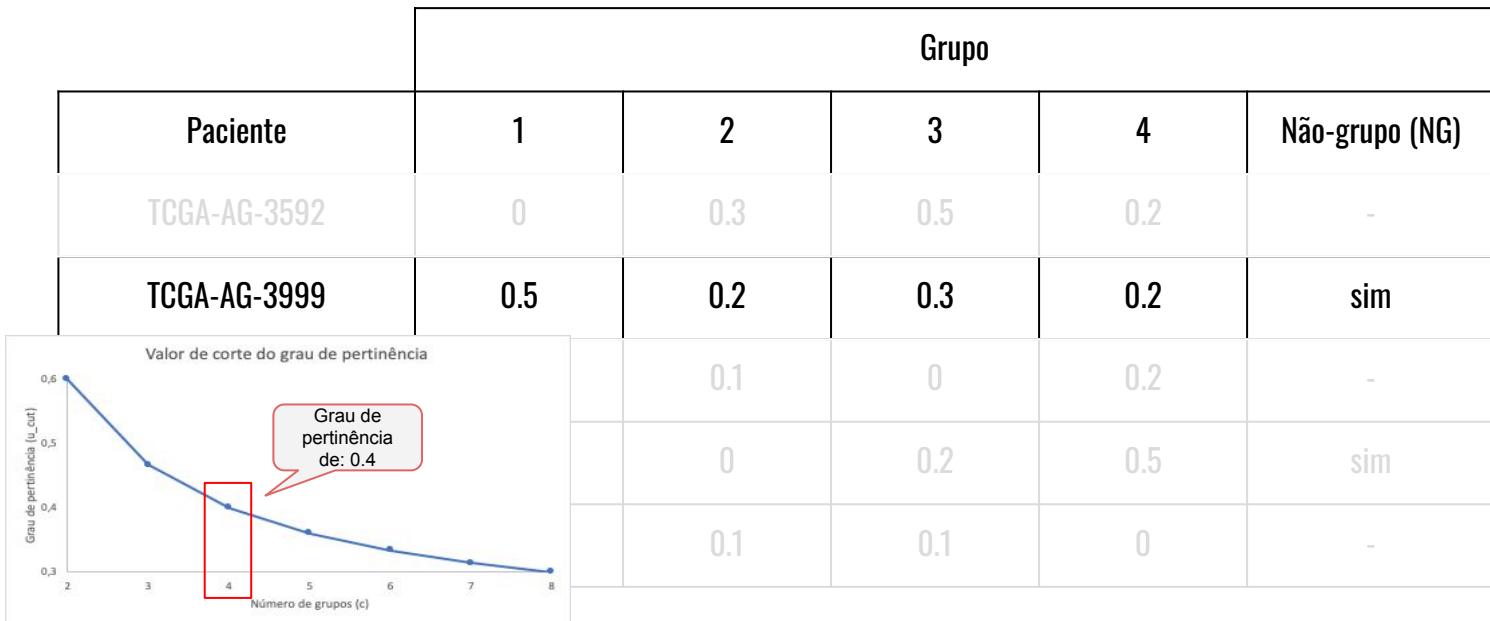
- Grau mínimo de pertinência: que decresce com aumento do número de grupos



# Agrupamento de dados multiômicos

---

- Grau mínimo de pertinência: que decresce com aumento do número de grupos



# Agrupamento de dados multiômicos

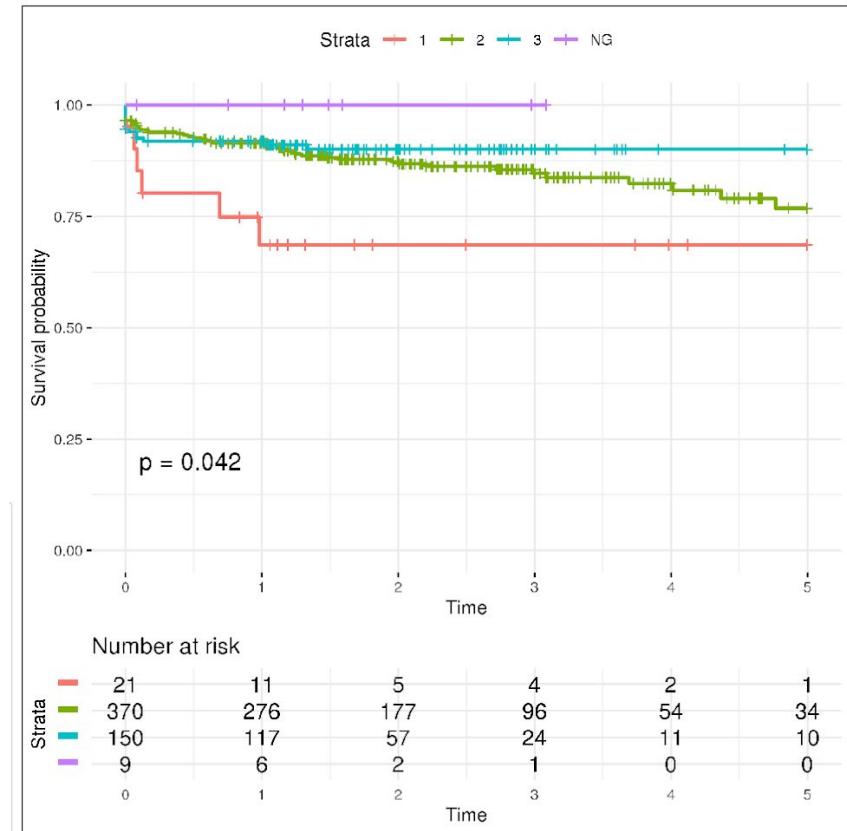
---

- Métrica de desempenho: FS, XB e **FPC** (que **usada para avaliar convergência do agrupamento fuzzy c-means**)
- Análise PCA (e %variância por componente):
  - reduzir dimensionalidade da base de dados para observar disposição dos grupos identificados pelo algoritmo fuzzy c-means
  - evr: verificar se as componentes identificadas pelo PCA tem significância em relação à variância que estas representam
- Mutação
  - Com relação à incidência de mutação nos grupos as amostras foram analisadas considerando-se todas as amostras e apenas as mutações drivers relacionadas ao fenótipo da doença, de acordo com a literatura
  - Também foram avaliadas as frequências de mutação de cada grupo

# Agrupamento de dados multiômicos

---

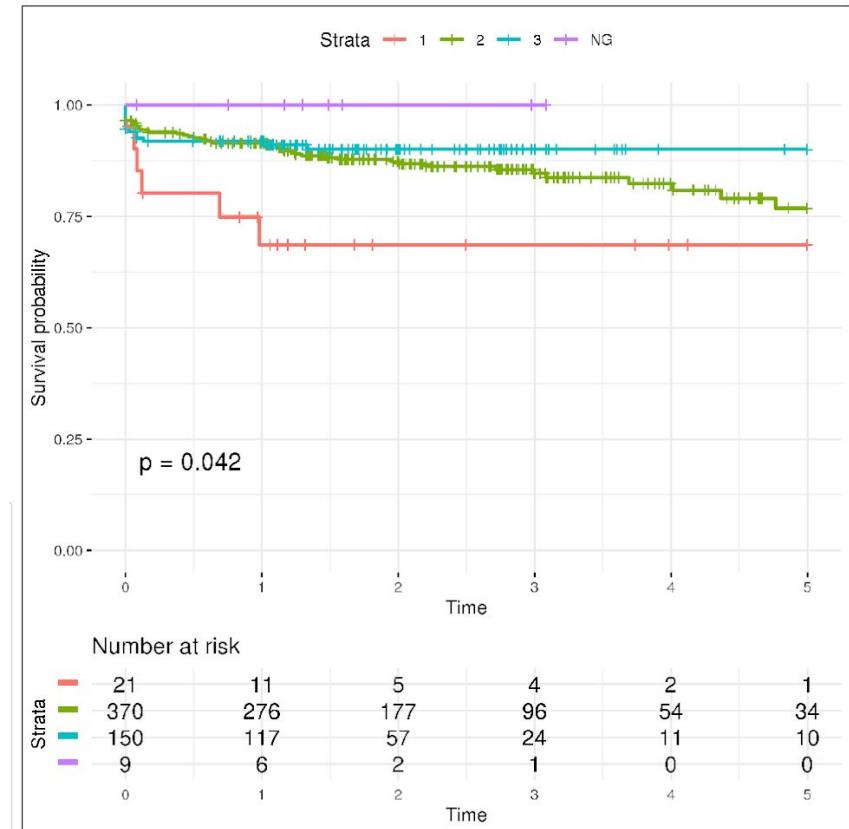
- Sobrevida
  - Elementos básicos



# Agrupamento de dados multiômicos

---

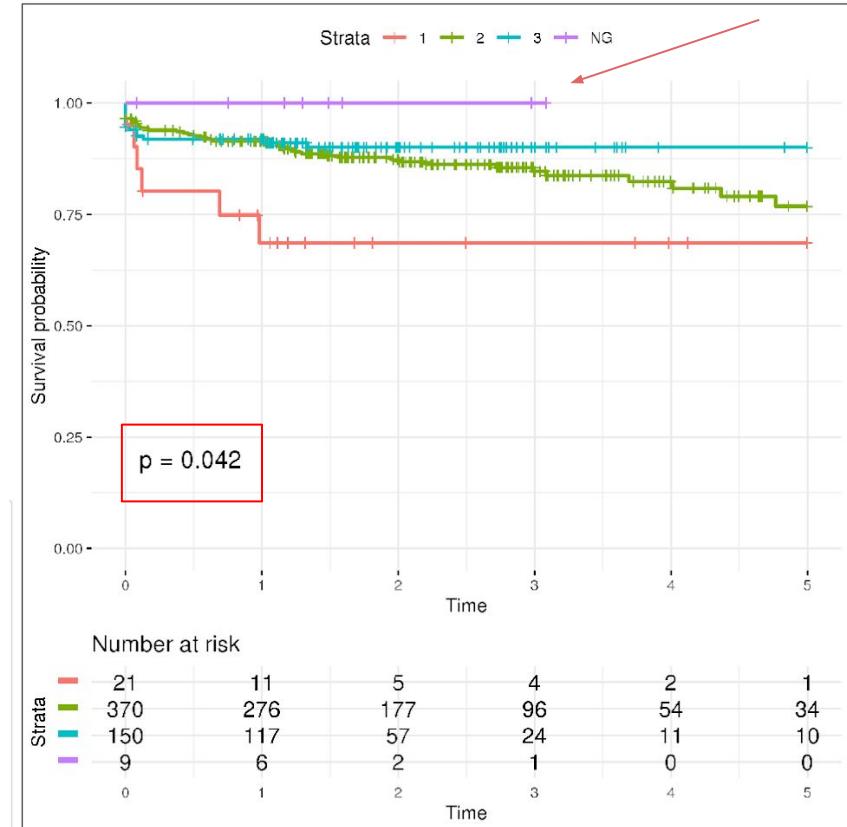
- Sobrevida
  - Elementos básicos:
    - Tempo de sobrevida: tempo até a ocorrência do evento (morte, recorrência da doença..)



# Agrupamento de dados multiômicos

- Sobrevida
  - Elementos básicos:
    - Censoreamento: fim do estudo ou perda de contato com paciente, permanece vivo ao final do estudo

Premissa: Tempo de sobrevida dos indivíduos censoreados é no mínimo maior que o tempo censoreado



# Estudo de caso

# Objetivos

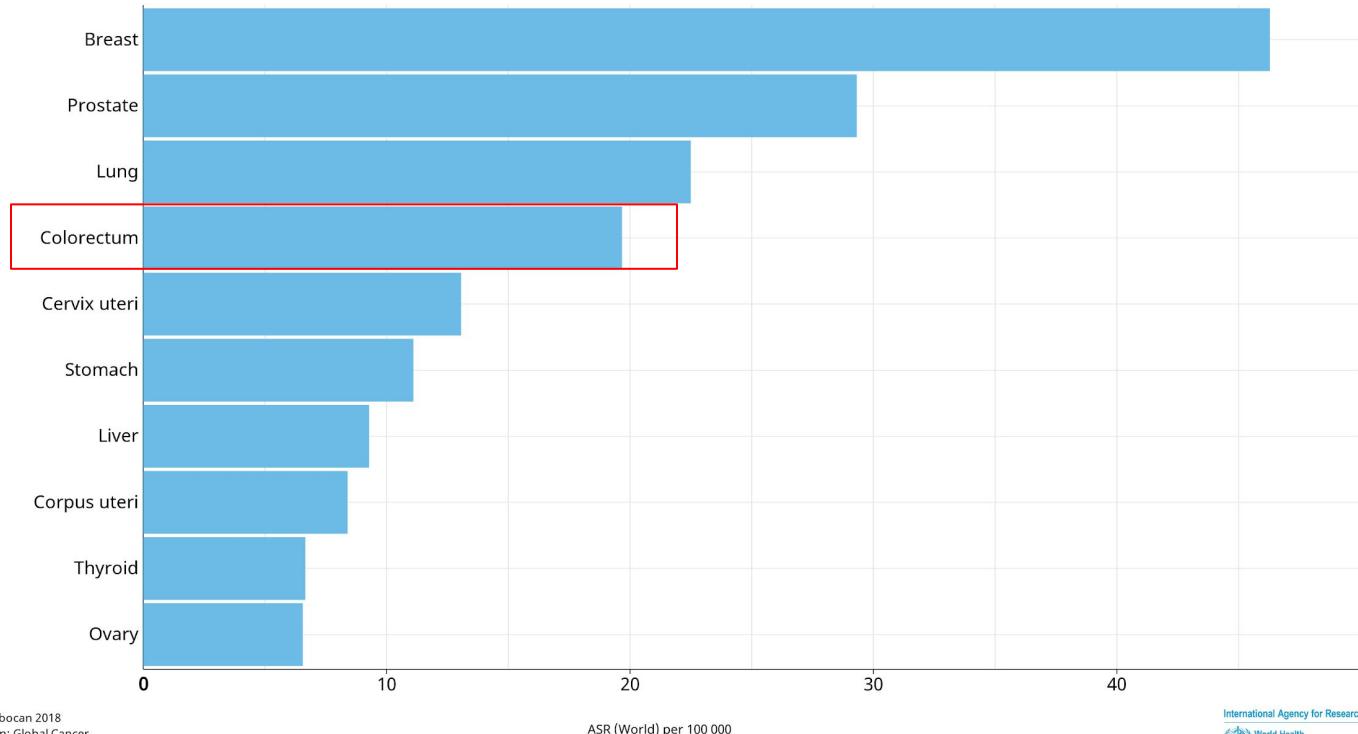
---

- Específicos:

- Avaliar a **contribuição da adição de conjuntos de dados** à resultados da literatura e escolher quais devem ser analisados.
- Analisar **métodos de seleção de atributos** adequados à dados multiômicos.
- Caracterizar dados de pacientes com **gradações entre diferentes subtipos moleculares**
- Identificar **outros perfis de subtipos moleculares**.
- **Diminuir o compartilhamento** de características dos perfis de subtipos moleculares adicionando dados multiômicos.
- **Validar os subtipos** moleculares identificados com relação às curvas de sobrevida de cada grupo, e demais características biológicas.
- **Relacionar** os grupos deste trabalho com **grupos** já definidos na **literatura**

# Estudo de caso

Incidência dos diferentes tipos de Câncer no mundo (2018)



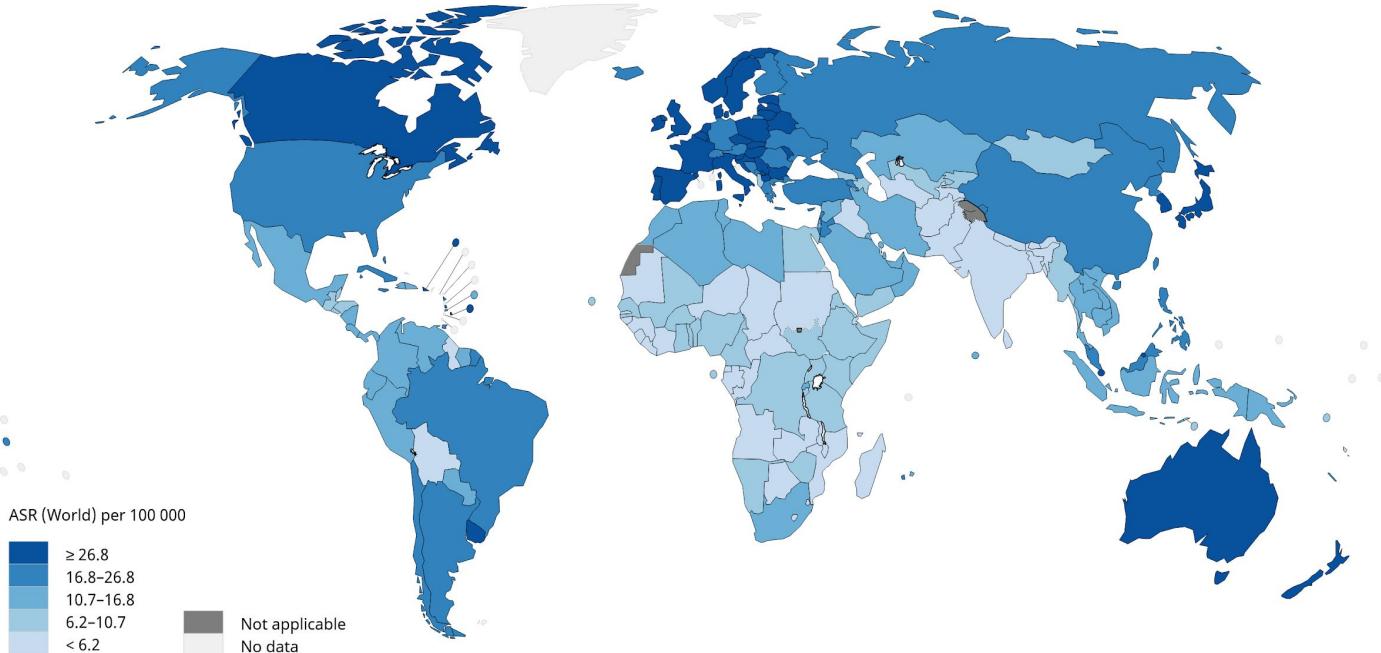
Data source: Globocan 2018  
Graph production: Global Cancer Observatory (<http://gco.iarc.fr>)

International Agency for Research on Cancer  
World Health Organization

# Estudo de caso

---

Distribuição geográfica da incidência do Câncer Colorretal no mundo (2018)

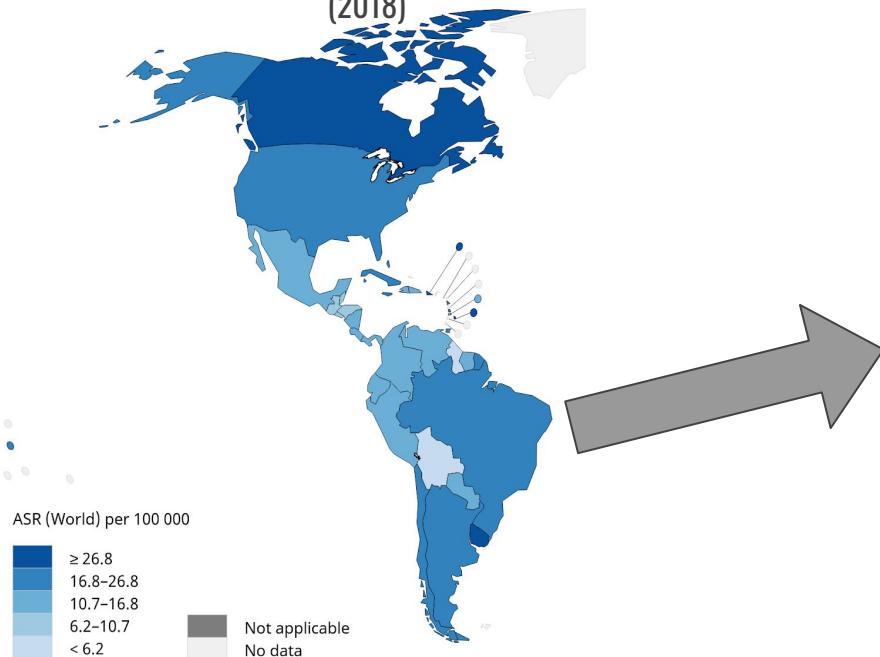


All rights reserved. The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the World Health Organization / International Agency for Research on Cancer concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate borderlines for which there may not yet be full agreement.

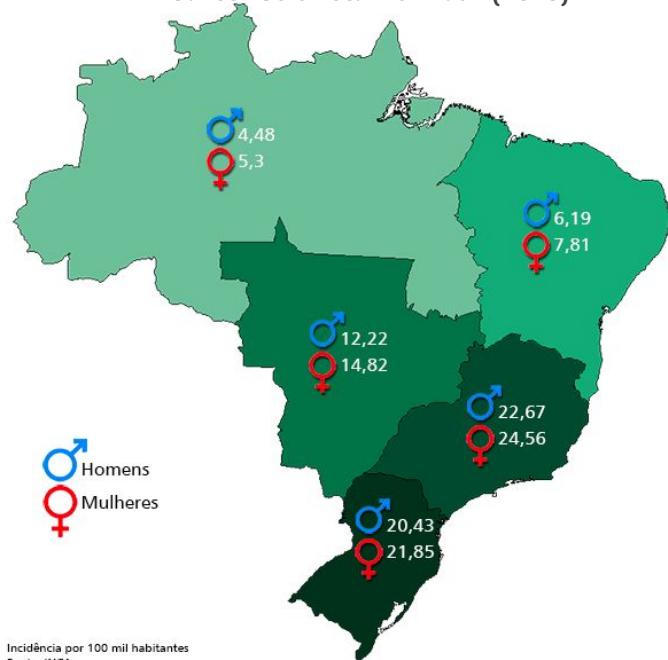
Data source: GLOBOCAN 2018  
Graph production: IARC  
(<http://gco.iarc.fr/today>)  
World Health Organization

# Estudo de caso

Distribuição geográfica da incidência do Câncer Colorretal no mundo  
(2018)



Distribuição geográfica da incidência do Câncer Colorretal no Brasil (2018)



All rights reserved. The designations employed and the presentation of the material in this publication do not i  
on the part of the World Health Organization / International Agency for Research on Cancer concerning the le  
or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines o  
which there may not yet be full agreement.

Fonte: <https://gco.iarc.fr/today/home>

# Estudo de caso

---

## Fatores de risco

Idade



Histórico



Cigarro



Alimentação



Álcool excessivo



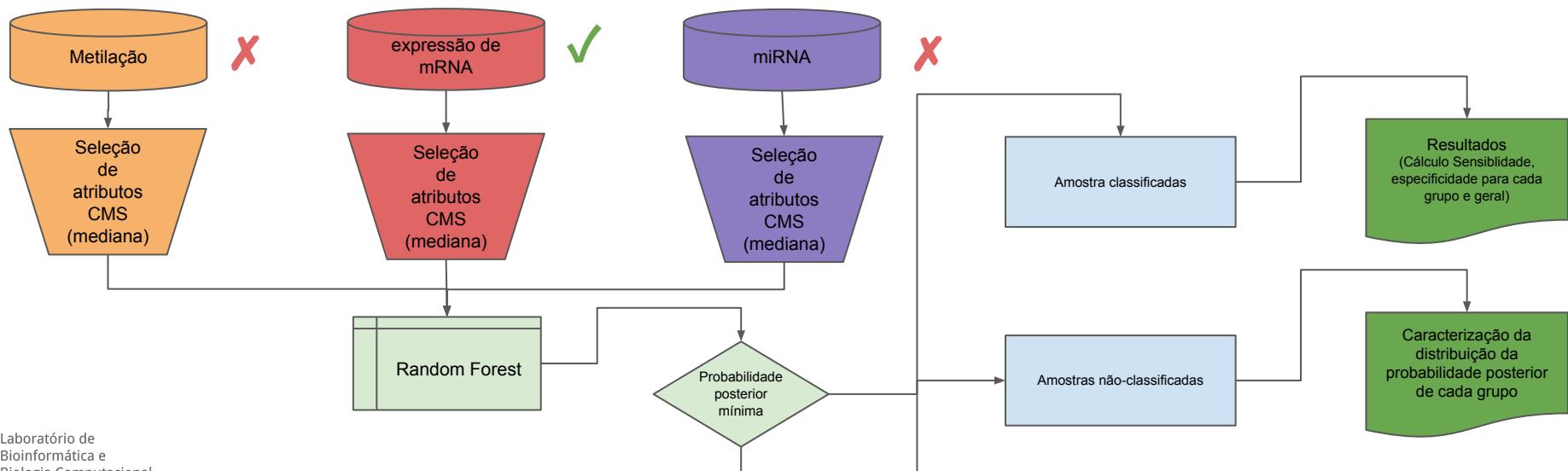
Estilo de vida sedentário



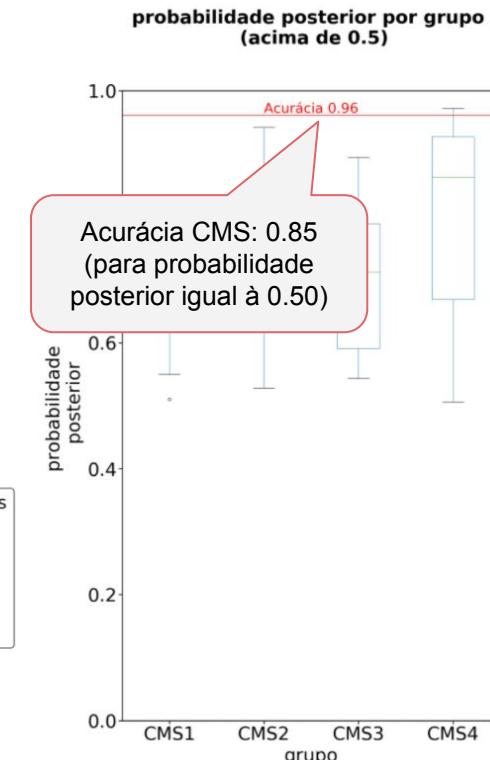
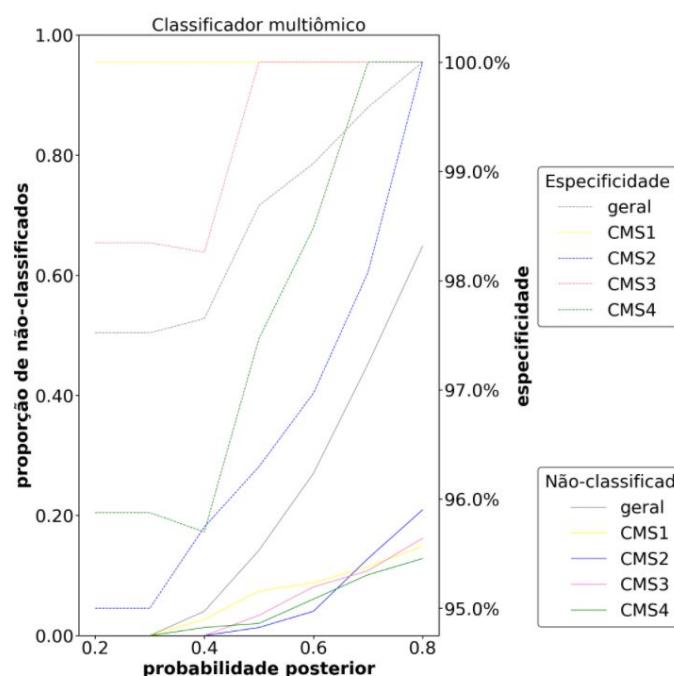
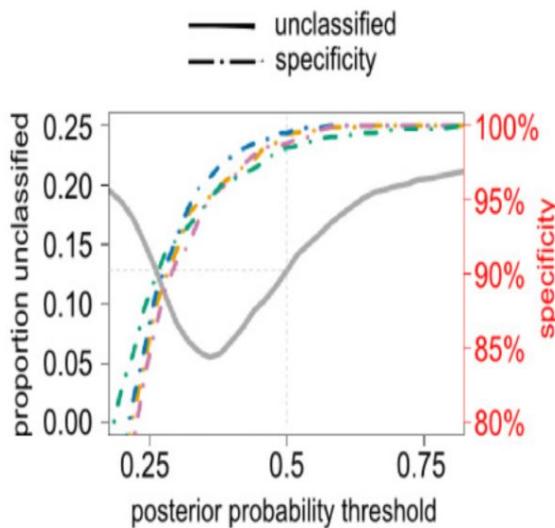
# Estudo de caso - Escolha dos conjuntos de dados

---

- Elaboração de classificador multiômico com base no classificador baseado em expressão de mRNA de Guinney, 2015 (CMS) e no trabalho de Liu, 2016.



# Estudo de caso - Escolha dos conjuntos de dados



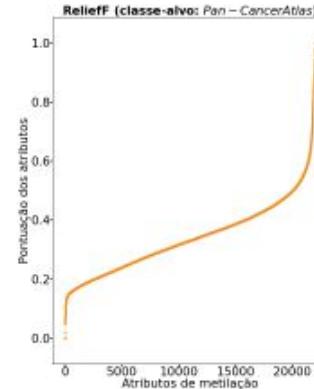
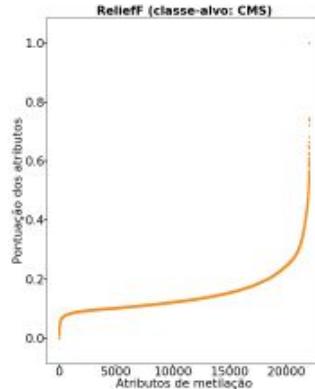
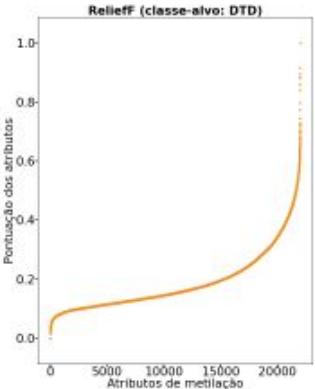
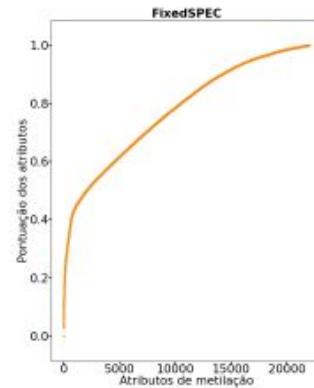
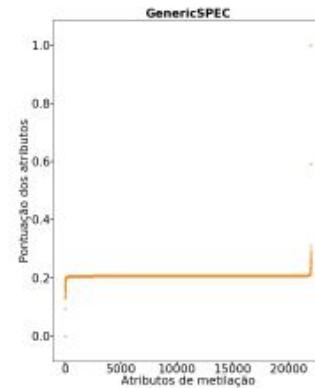
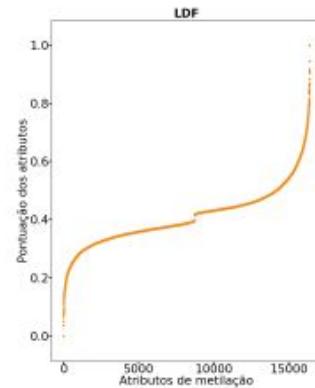
# Estudo de caso - Escolha dos conjuntos de dados

---

Probabilidade posterior acima de:	Especificidade geral	Especificidade CMS1	Especificidade CMS2	Especificidade CMS3	Especificidade CMS4	não-classificados geral	não-classificados CMS1	não-classificados CMS2	não-classificados CMS3	não-classificados CMS4
0.2	97.523	100.0	95.0	98.347	95.876	0.0	0.0	0.0	0.0	0.0
0.3	97.523	100.0	95.0	98.347	95.876	0.0	0.0	0.0	0.0	0.0
0.4	97.653	100.0	95.745	98.261	95.699	0.041	0.027	0.0	0.0	0.014
0.5	98.688	100.0	96.296	100.0	97.468	0.142	0.074	0.014	0.034	0.02
0.6	99.074	100.0	96.97	100.0	98.485	0.27	0.088	0.041	0.081	0.061
0.7	99.588	100.0	98.077	100.0	100.0	0.453	0.115	0.128	0.108	0.101
0.8	100.0	nan	100.0	100.0	100.0	0.649	0.149	0.209	0.162	0.128

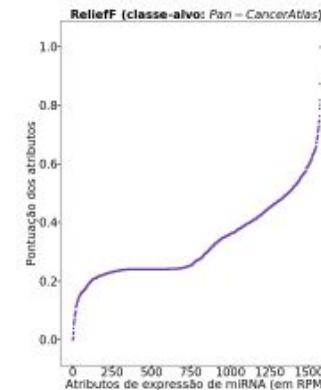
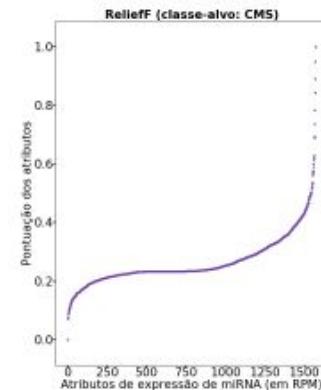
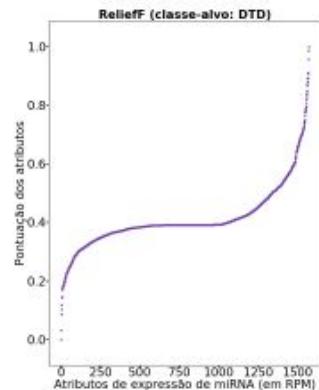
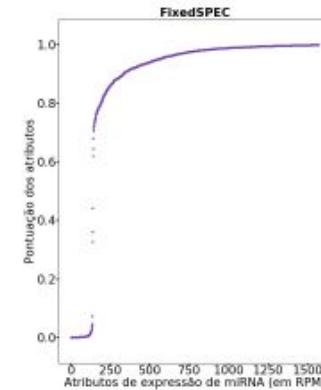
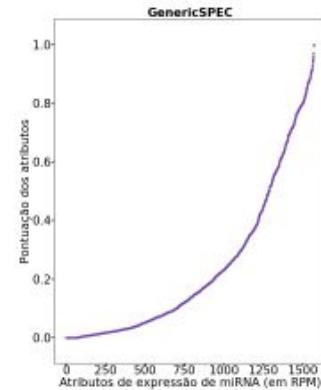
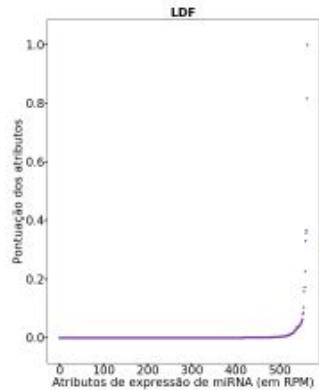
# Estudo de caso - Seleção de atributos preliminar

- metilação



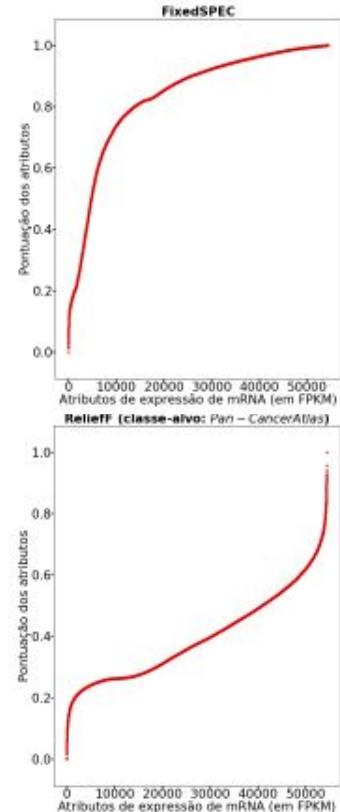
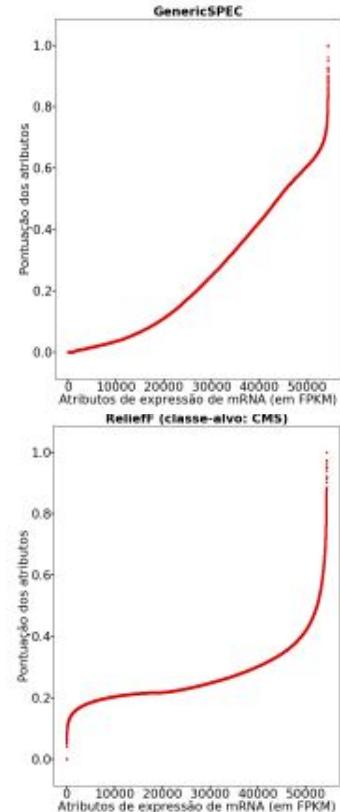
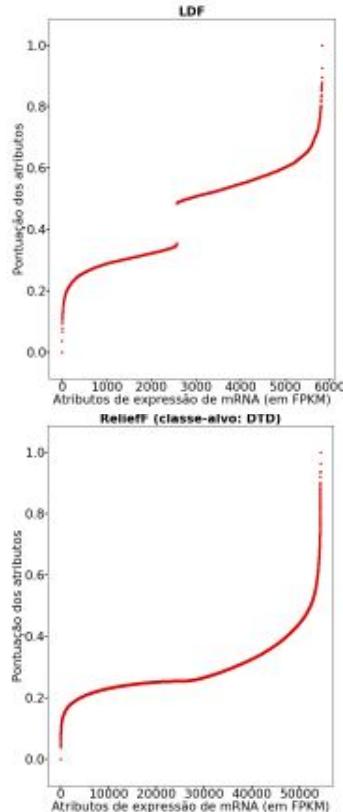
# Estudo de caso - Seleção de atributos preliminar

- mRNA

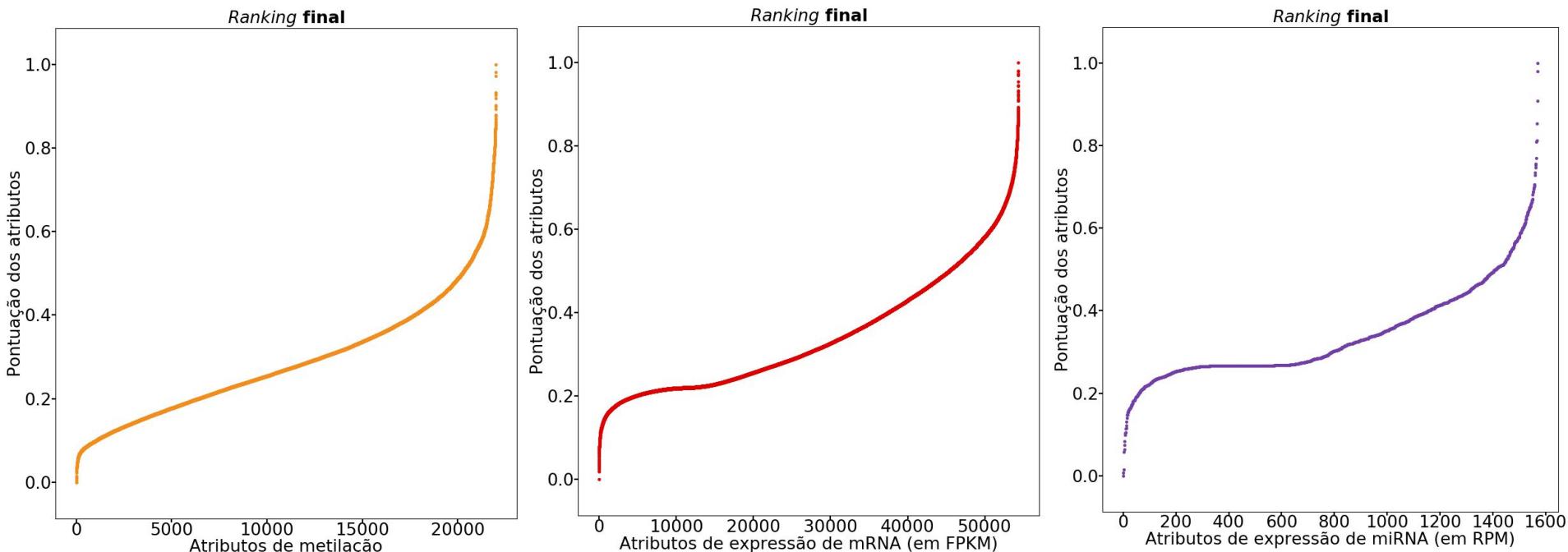


# Estudo de caso - Seleção de atributos preliminar

- miRNA



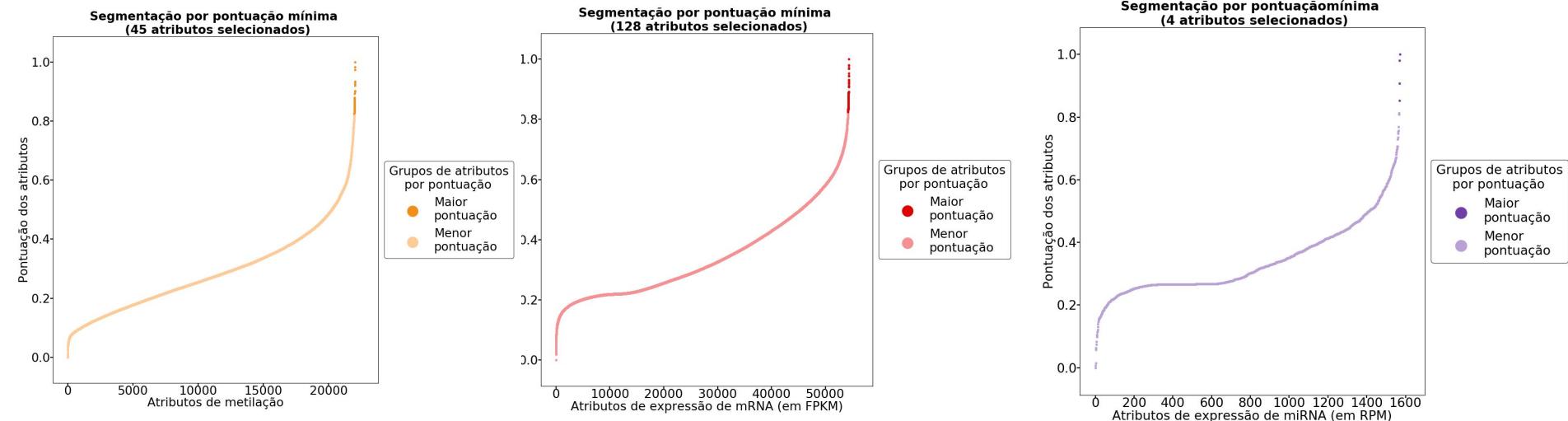
# Estudo de caso - Seleção de atributos



# Estudo de caso - Seleção de atributos

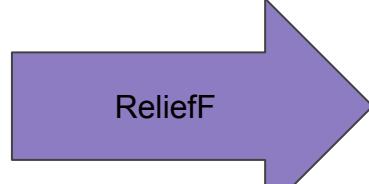
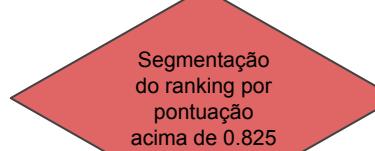
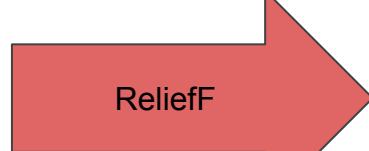
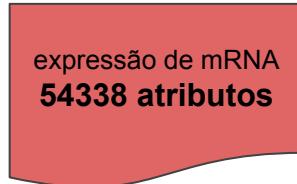
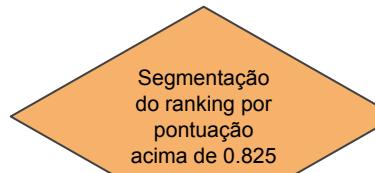
---

- Segmentação do ranking de atributos

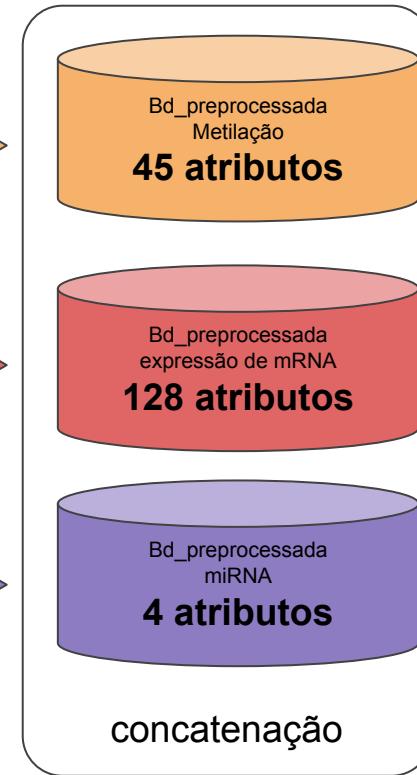


# Estudo de caso - Pré-processamento

---



 De total: **77926 atributos**



Redução de:

**99,79%**

**99,76%**

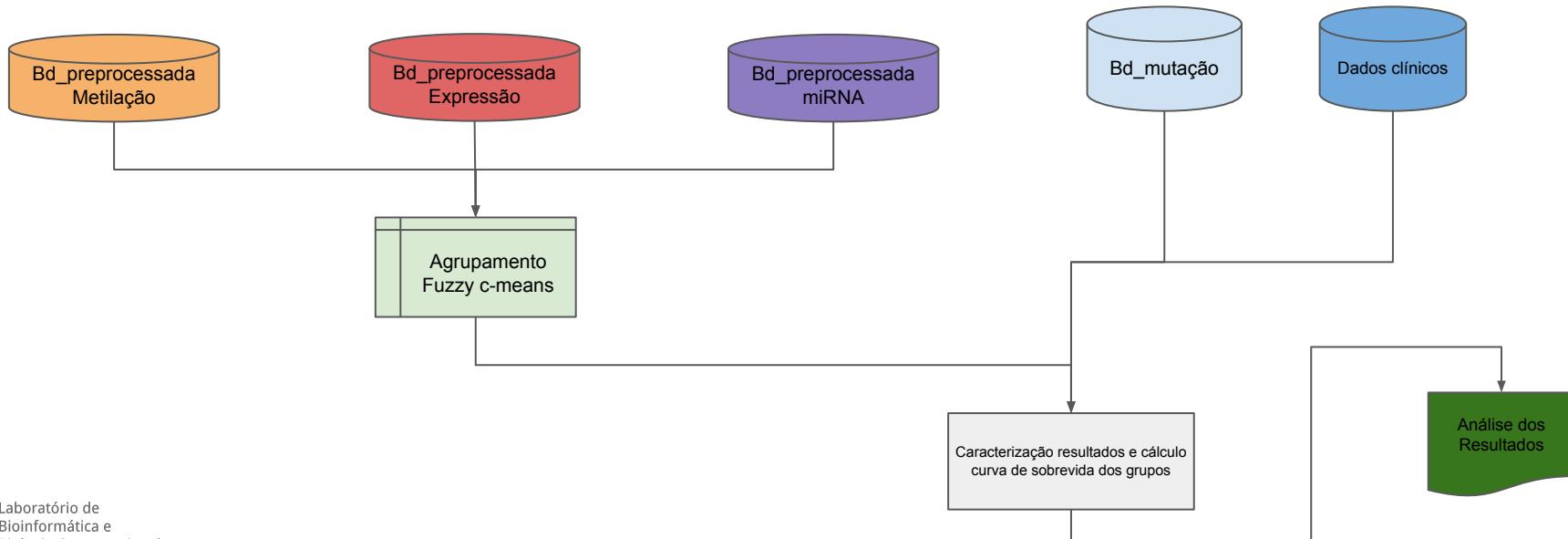
**99,74%**

Para total: **177 atributos**

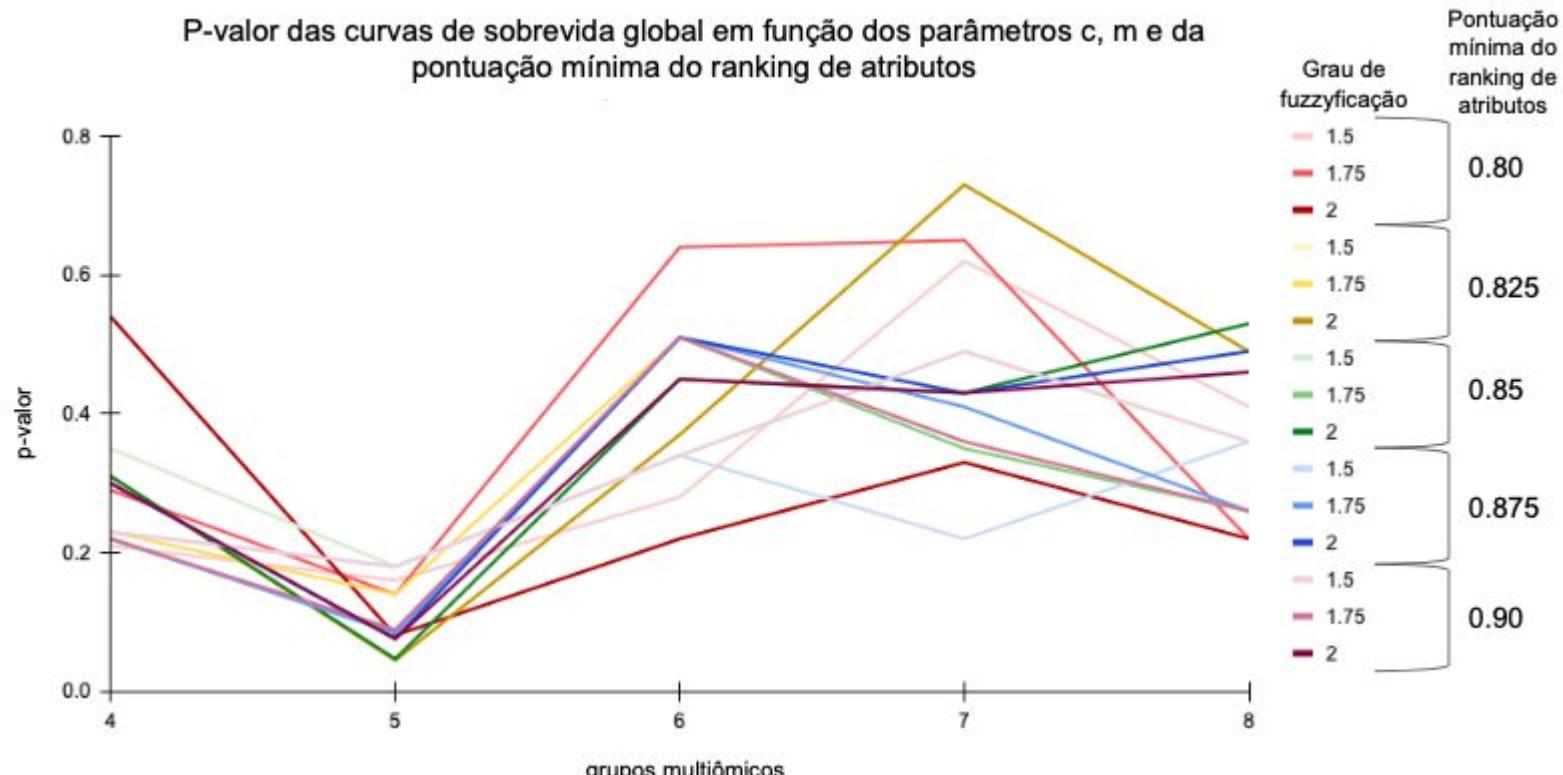
# Estudo de caso - Agrupamento dados concatenados

---

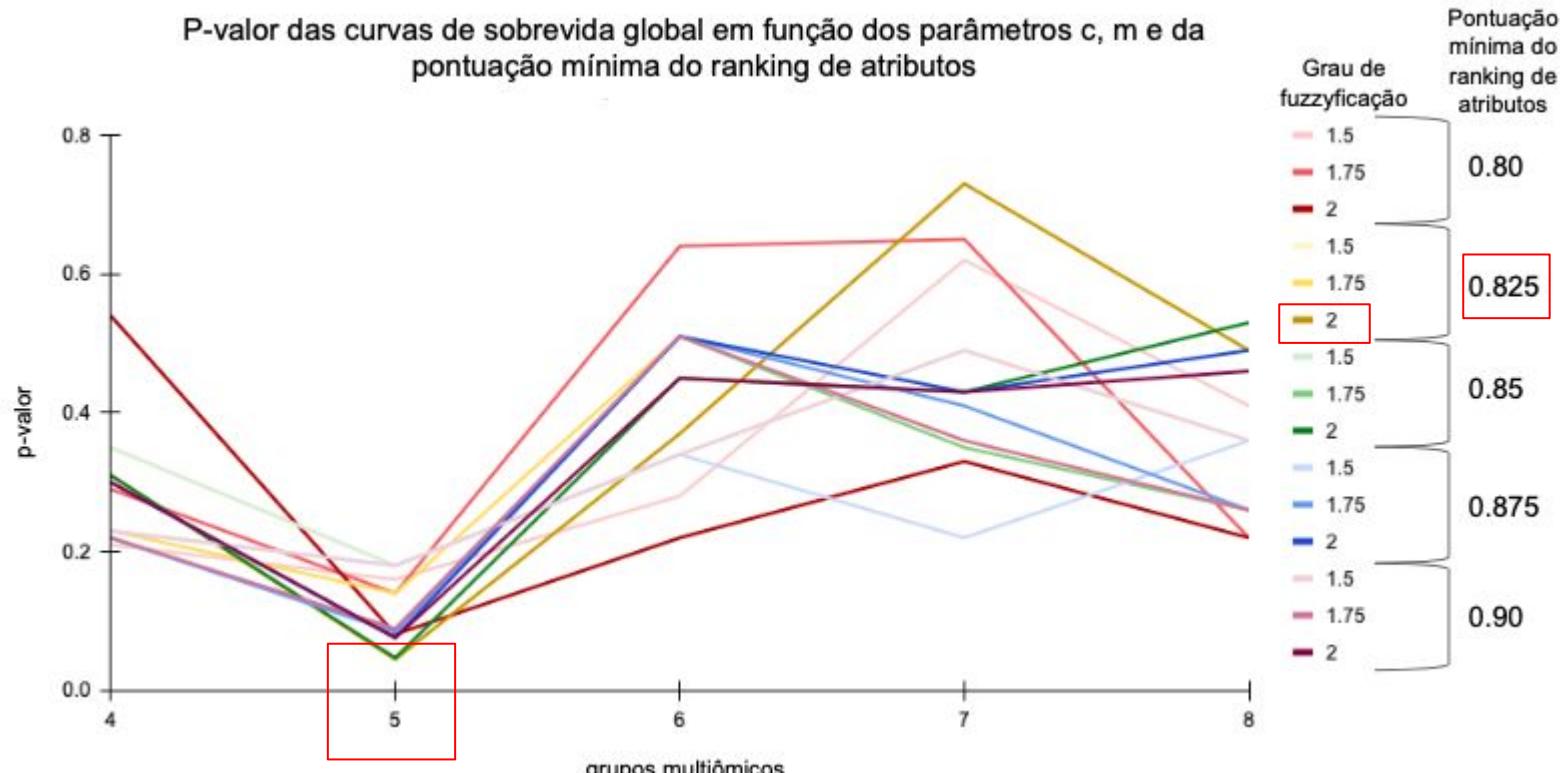
- Para cada par de valores dos parâmetros número de grupos, grau de fuzzificação e intervalo do ranking de atributos:



# Estudo de caso - Caracterização dos resultados

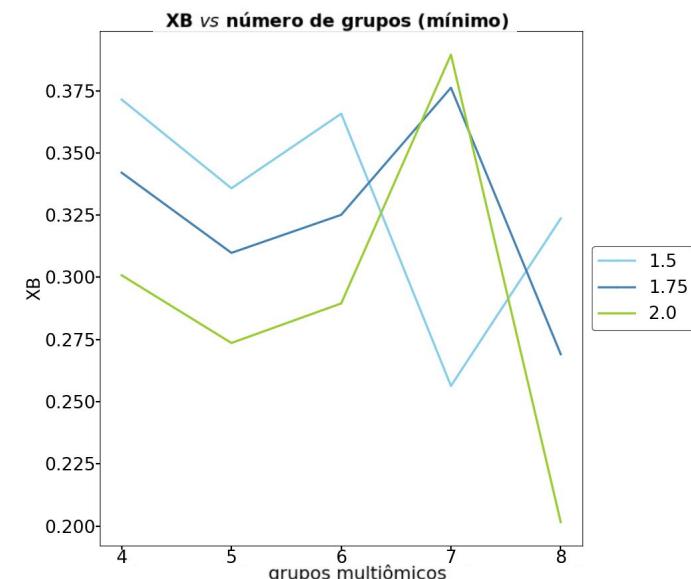
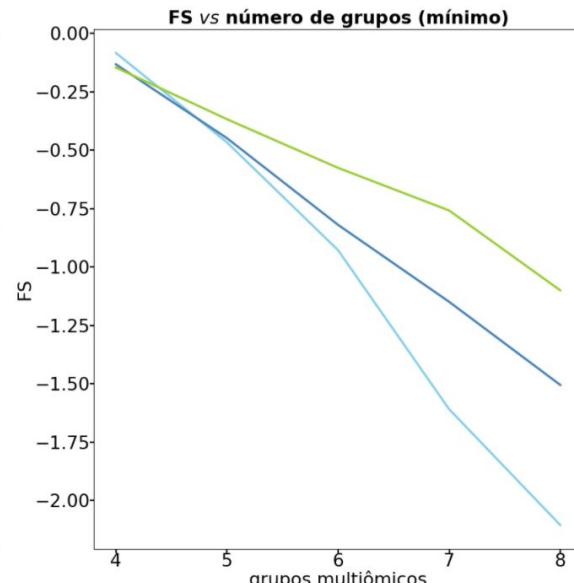
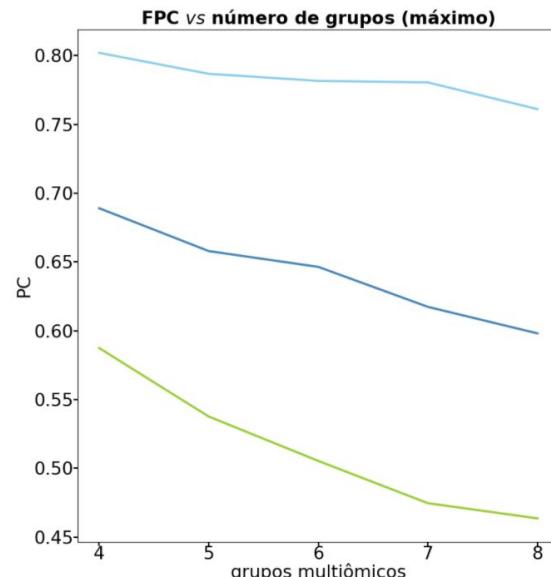


# Estudo de caso - Caracterização dos resultados



# Estudo de caso - Caracterização dos resultados

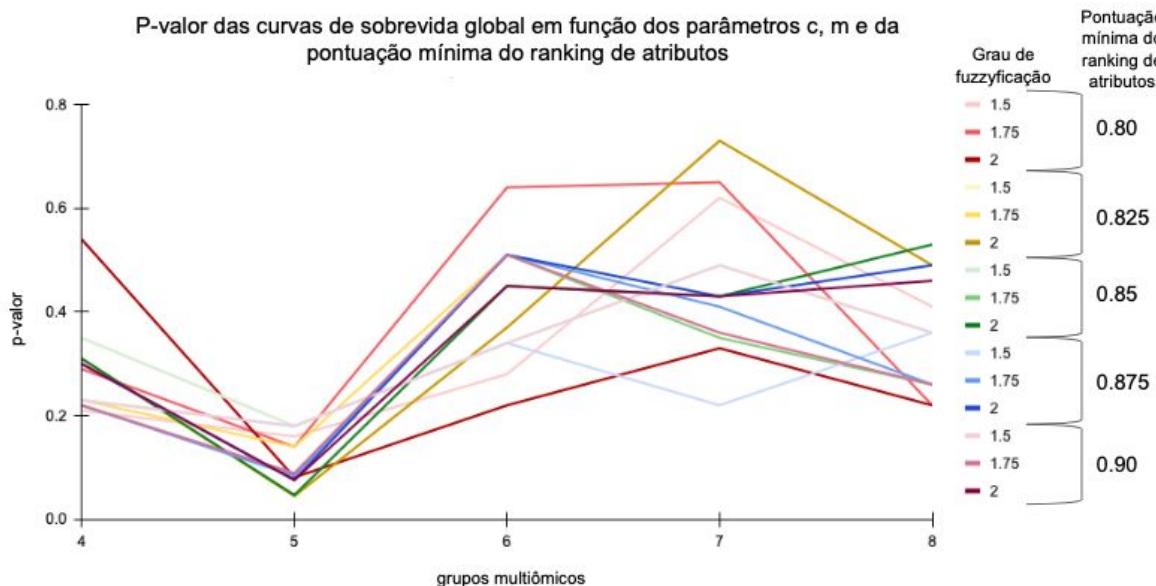
---



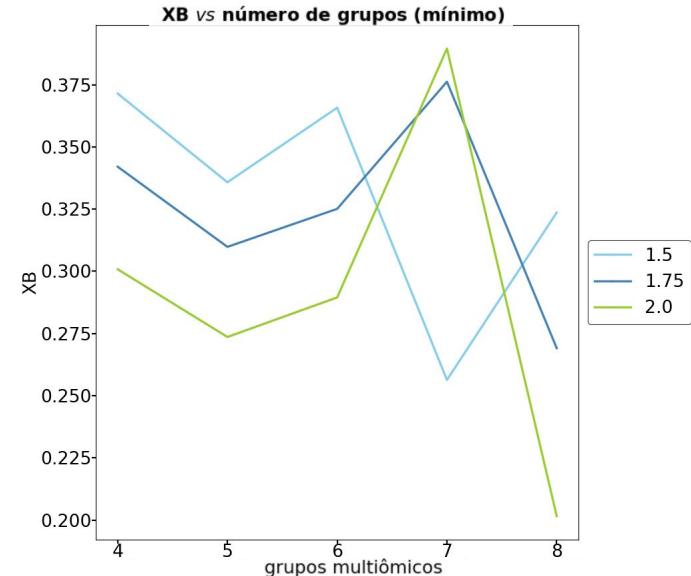
# Estudo de caso - Caracterização dos resultados

---

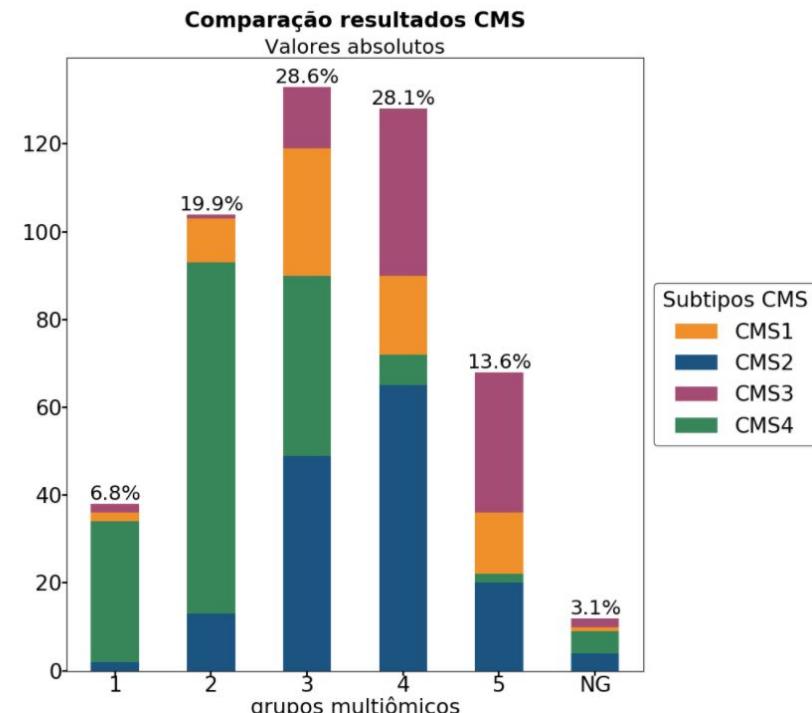
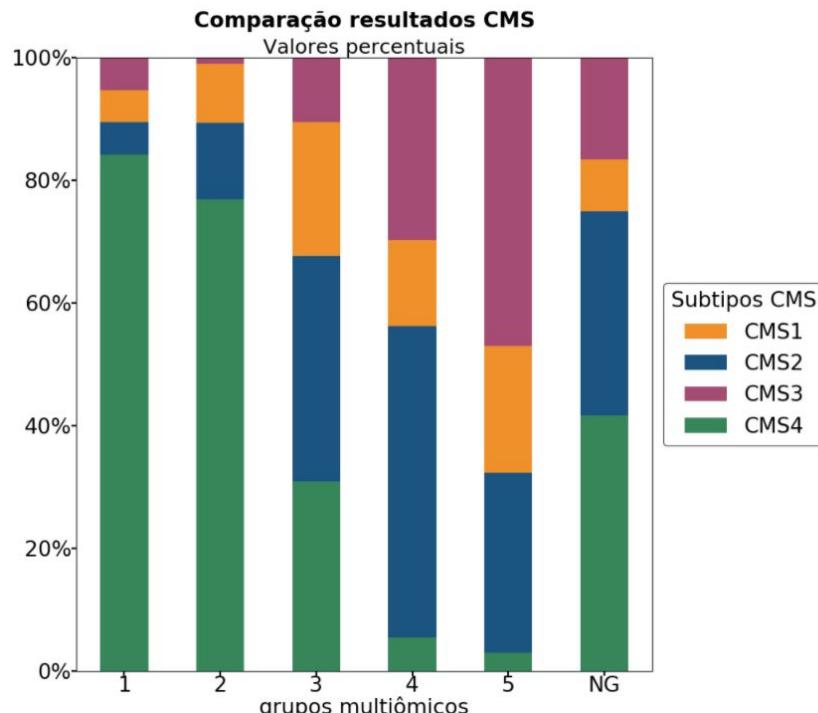
P-valor das curvas de sobrevida global em função dos parâmetros c, m e da pontuação mínima do ranking de atributos



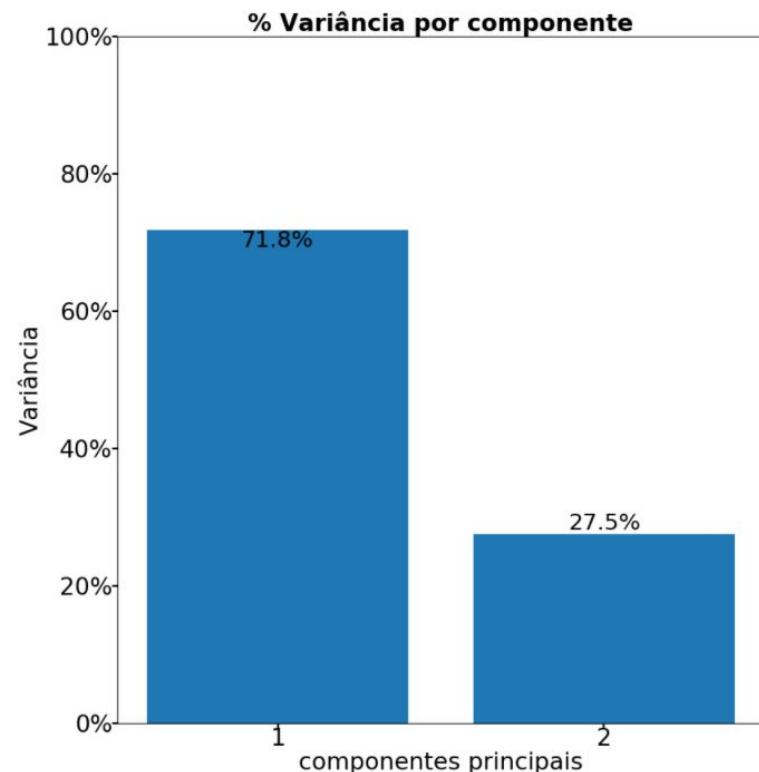
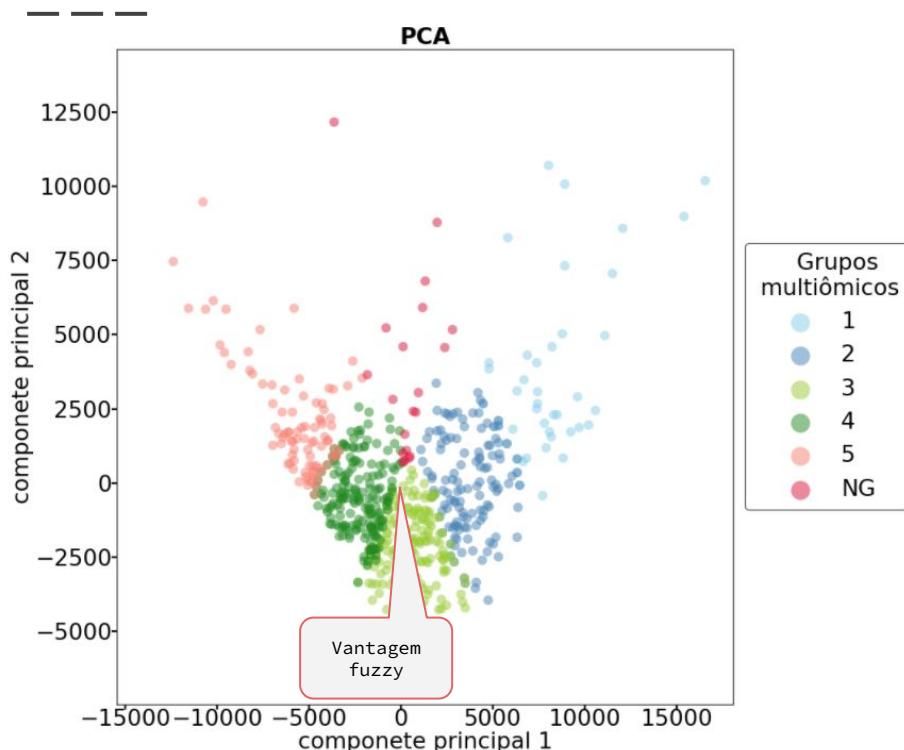
XB vs número de grupos (mínimo)



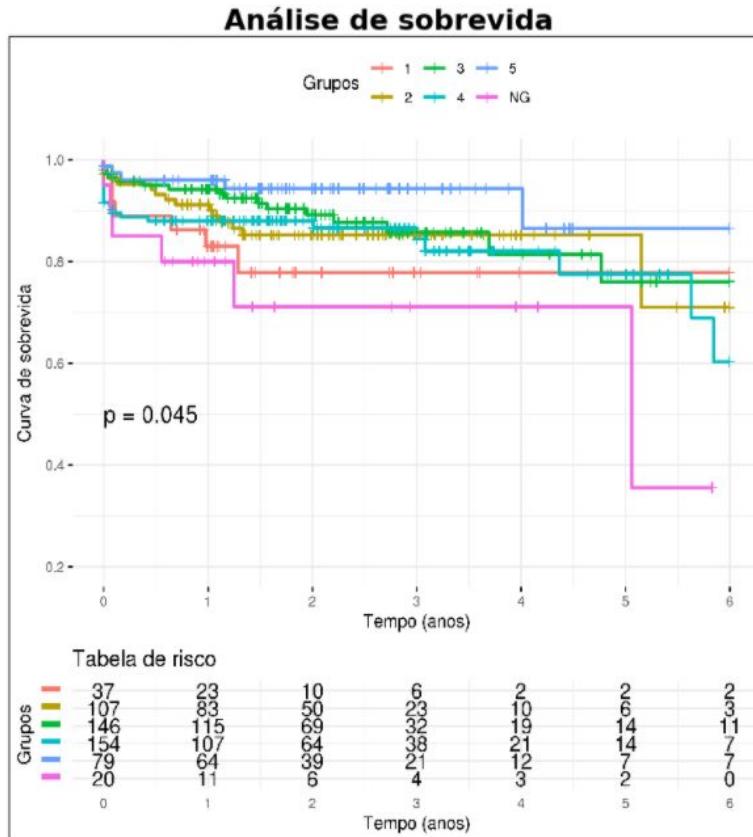
# Estudo de caso - Caracterização dos resultados



# Estudo de caso - Caracterização dos resultados



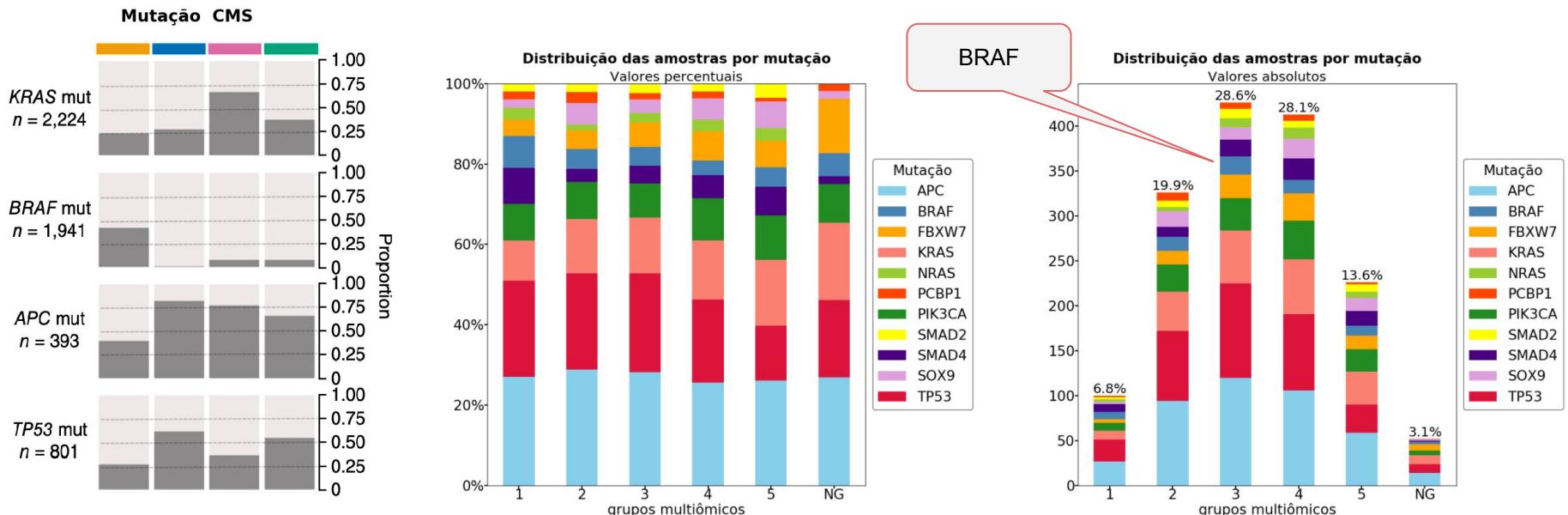
# Estudo de caso - Caracterização dos resultados



P valor par a par

	1	2	3	4	5
2	0.49	nan	nan	nan	nan
3	0.243	0.566	nan	nan	nan
4	0.631	0.753	0.413	nan	nan
5	0.115	0.203	0.334	0.13	nan
NG	0.49	0.203	0.115	0.23	0.04

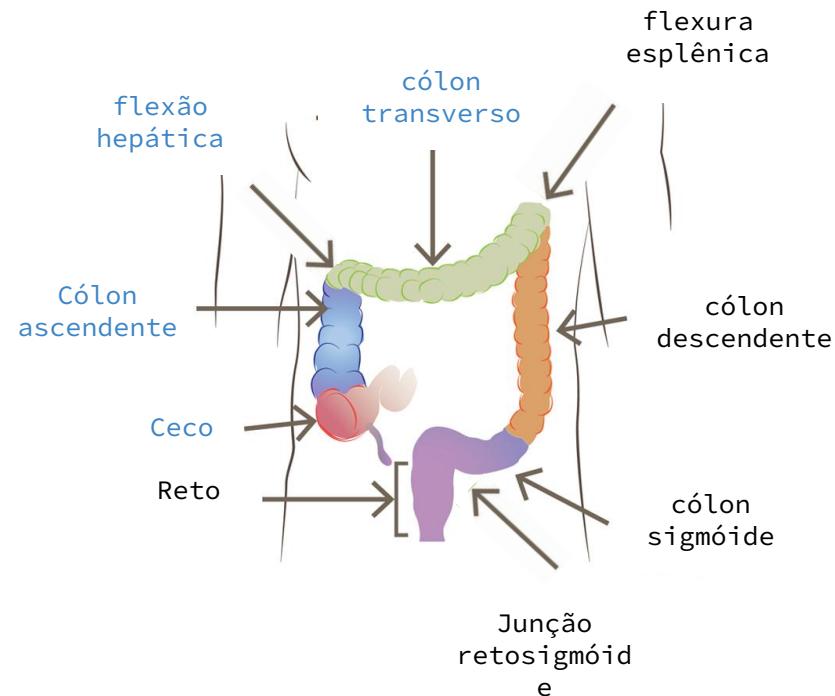
# Estudo de caso - Caracterização dos resultados



# Estudo de caso - Caracterização dos resultados

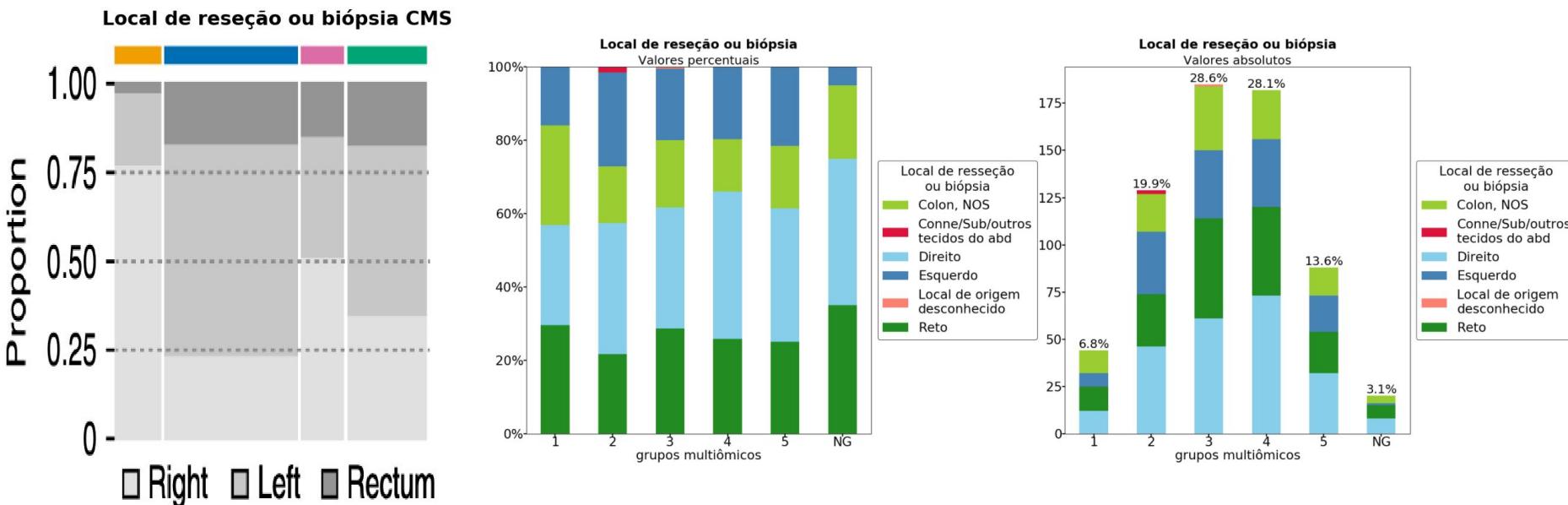
---

- Local de ressecção ou biópsia
  - direito (ceco, cólon ascendente, flexão hepática e cólon transverso)
  - esquerdo (flexura esplênica, colon sigmóide e descendente)
  - Reto
  - Cólon, Nos
  - Conne/sub/outros tecidos do abd
  - Local de origem desconhecido



Fonte: <https://www.google.com.br/imghp>

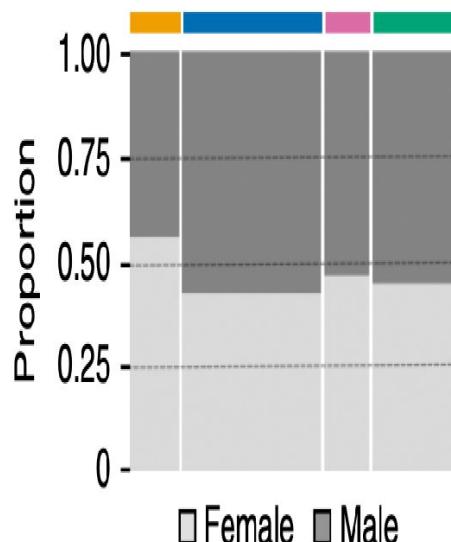
# Estudo de caso - Caracterização dos resultados



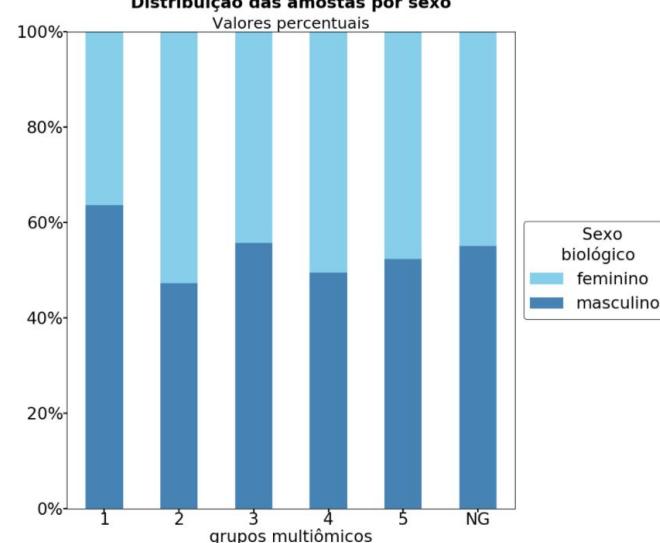
# Estudo de caso - Caracterização dos resultados

---

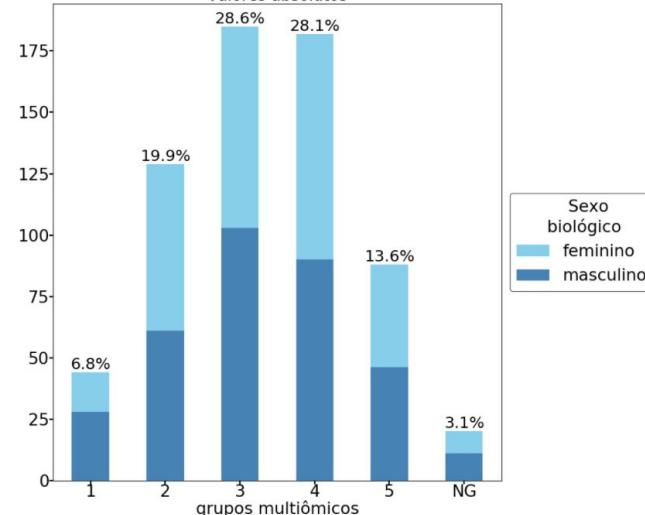
Distribuição das amostras por sexo CMS



Distribuição das amostras por sexo  
Valores percentuais

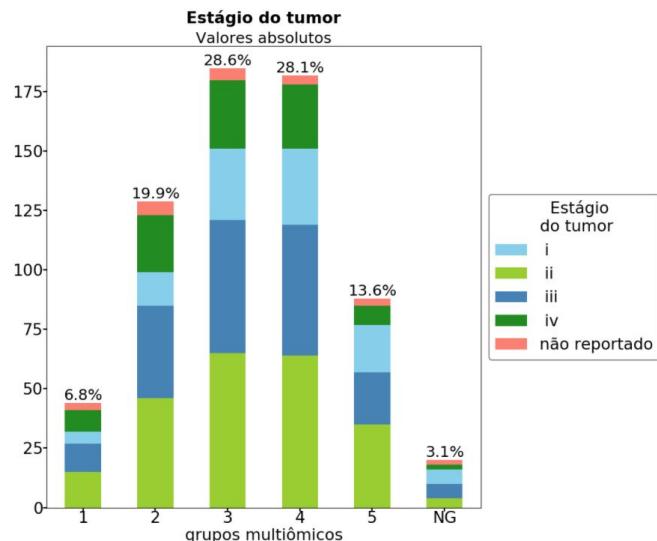
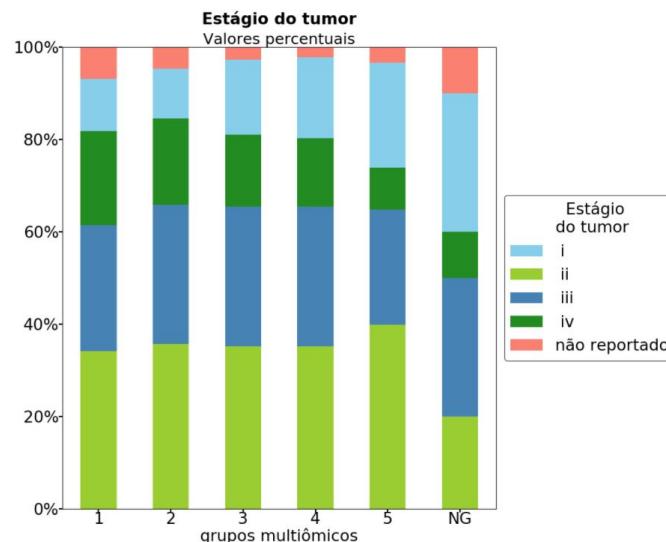
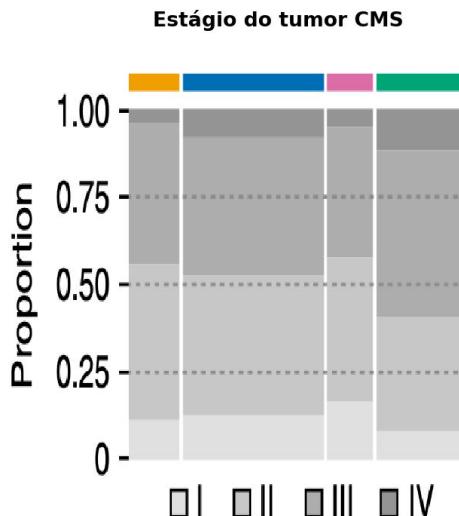


Distribuição das amostras por sexo  
Valores absolutos



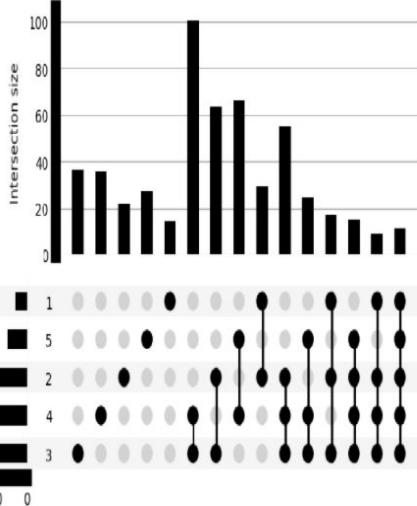
# Estudo de caso - Caracterização dos resultados

---

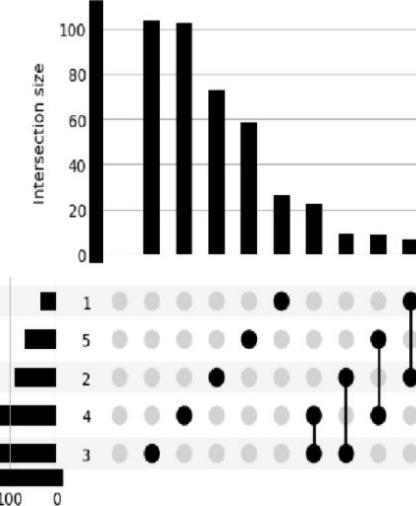


# Estudo de caso - Caracterização dos resultados

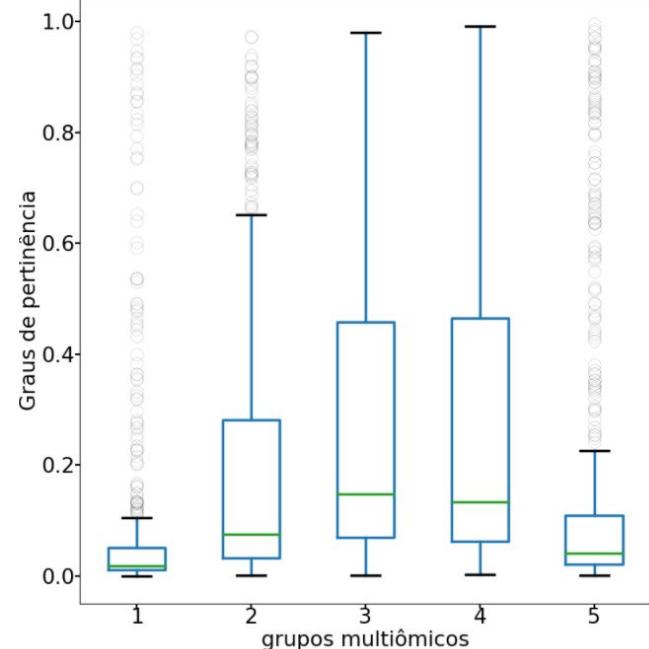
Graus de pertinência (mínimo: 0.1)



Graus de pertinência (mínimo: 0.36)



Graus de pertinência (com outliers)



# Conclusões

---

- A metodologia proposta mostrou ser capaz de verificar de forma satisfatória a contribuição de cada conjunto de dados ômicos.
- Apenas o algoritmo ReliefF se mostrou capaz de separar todos os conjuntos de dados ômicos de forma satisfatória, no entanto, cada método de seleção de atributos testado mostrou-se mais adequado para um cada conjunto de dados em específico, o que mostra que as características dos conjuntos de dados são bastante diferentes e que é possível que estes sejam refinados adicionando-se ao ranking os resultados dos métodos que forem satisfatórios mesmo que para apenas um dos conjuntos de dados.
- Observou-se que a métrica de desempenho XB tem comportamento semelhante à variação do p-valor global considerando-se alterações nos valores dos parâmetros do agrupamento fuzzy c-means e da quantidade de atributos selecionados de cada conjunto de dados ômicos.
- Observou-se que, para o fenótipo analisado, o delineamento de um novo perfil molecular em comparação aos resultados da literatura. No entanto, este perfil, assim como os demais, ainda apresenta alto compartilhamento de características.

# Trabalhos futuros

---

- Testar outras funções, por exemplo a métrica XB, como função de avaliação do modelo fuzzy.
- Escolher os métodos de seleção de atributos para cada conjunto de dados.
- Analisar métodos de seleção de atributos que possam considerar as variações de características biológicas (como mutação e local de origem do tumor, por exemplo) na escolha dos atributos.

# Obrigada!

[shalves@aluno.puc-rio.br](mailto:shalves@aluno.puc-rio.br)



PONTIFÍCIA UNIVERSIDADE CATÓLICA  
DO RIO DE JANEIRO



iINCA  
INSTITUTO NACIONAL DE CÂNCER

# Referências

---

- Yugi, Katsuyuki & Kubota, Hiroyuki & Hatano, Atsushi & Kuroda, Shinya. (2016). Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple ‘Omic’ Layers. *Trends in Biotechnology*. 34. 10.1016/j.tibtech.2015.12.013.
- Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, Bot BM, Morris JS, Simon IM, Gerster S, Fessler E, De Sousa E Melo F, Missiaglia E, Ramay H, Barras D, Homicsko K, Maru D, Manyam GC, Broom B, Boige V, Perez-Villamil B, Laderas T, Salazar R, Gray JW, Hanahan D, Tabernero J, Bernards R, Friend SH, Laurent-Puig P, Medema JP, Sadanandam A, Wessels L, Delorenzi M, Kopetz S, Vermeulen L, Tejpar S. The consensus molecular subtypes of colorectal cancer. *Nat Med*. 2015 Nov;21(11):1350-6. doi: 10.1038/nm.3967. Epub 2015 Oct 12. PMID: 26457759; PMCID: PMC4636487.
- Jianyu Miao e Lingfeng Niu. “A Survey on Feature Selection”. Em: Procedia Computer Science 91 (2016). Promoting Business Analytics and Quantitative Management of Technology: 4th International Conference on Information Technology and Quantitative Management (ITQM 2016), pp. 919–926. issn: 1877-0509. doi: <https://doi.org/10.1016/j.procs.2016.07.111>. url: <http://www.sciencedirect.com/science/article/pii/S1877050916313047>

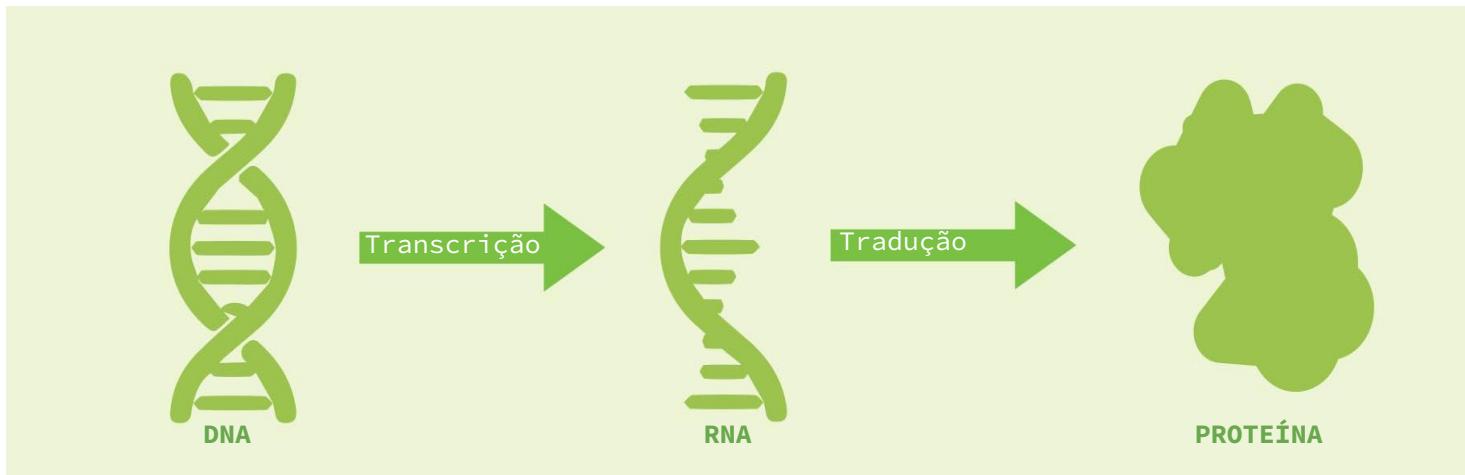
---

# PRÉVIA

# Introdução

---

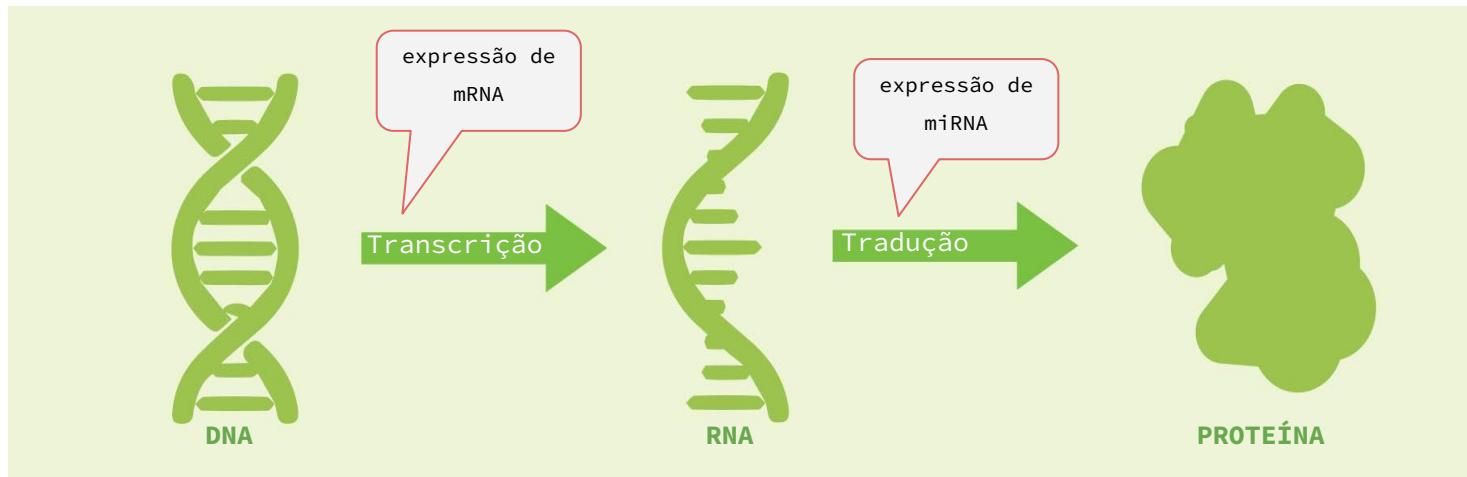
- Dogma central da biologia molecular



# Introdução

---

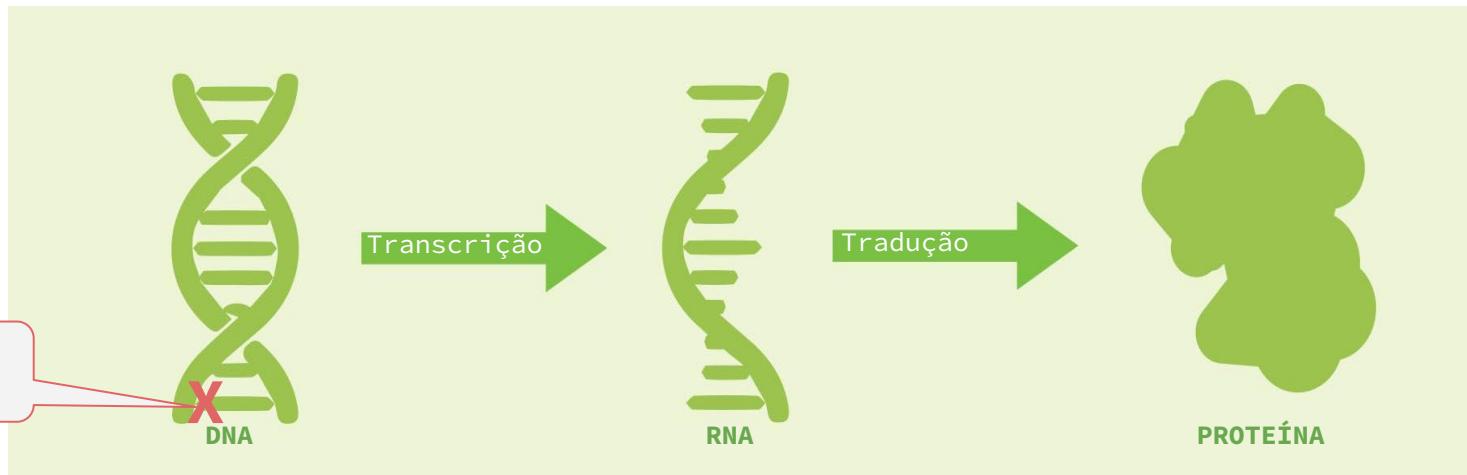
- Dogma central da biologia molecular



# Introdução

---

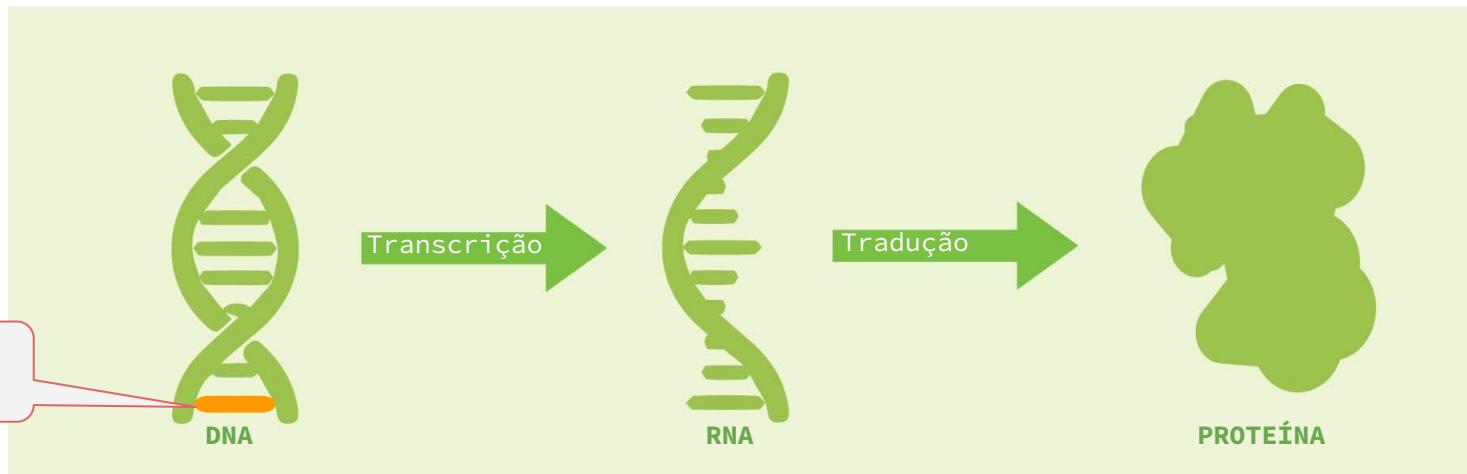
- Dogma central da biologia molecular



# Introdução

---

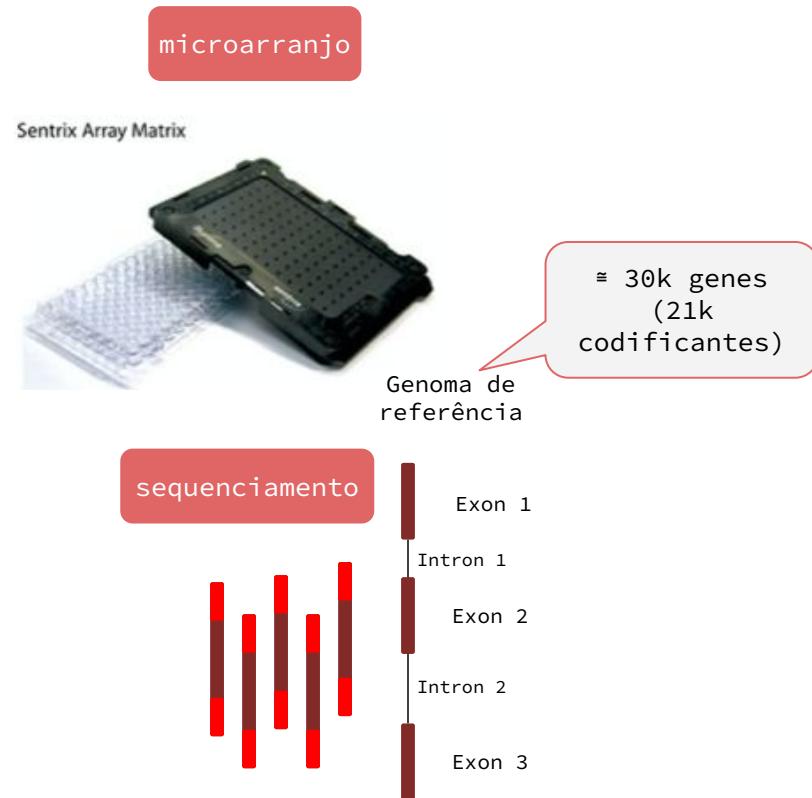
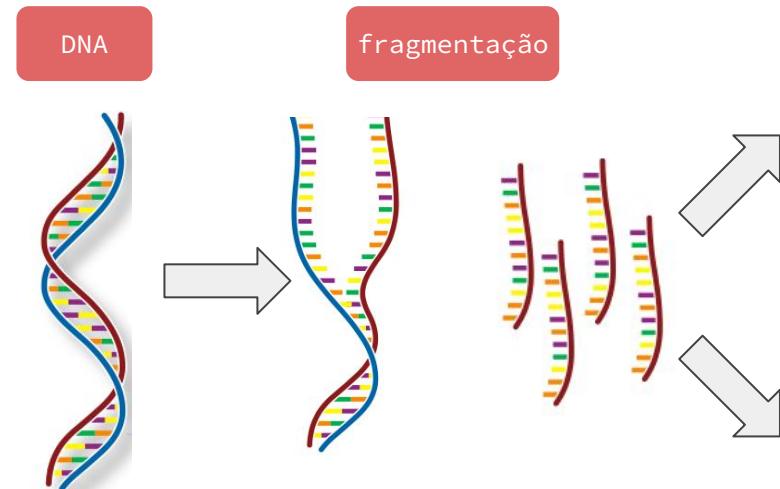
- Dogma central da biologia molecular



# Introdução

---

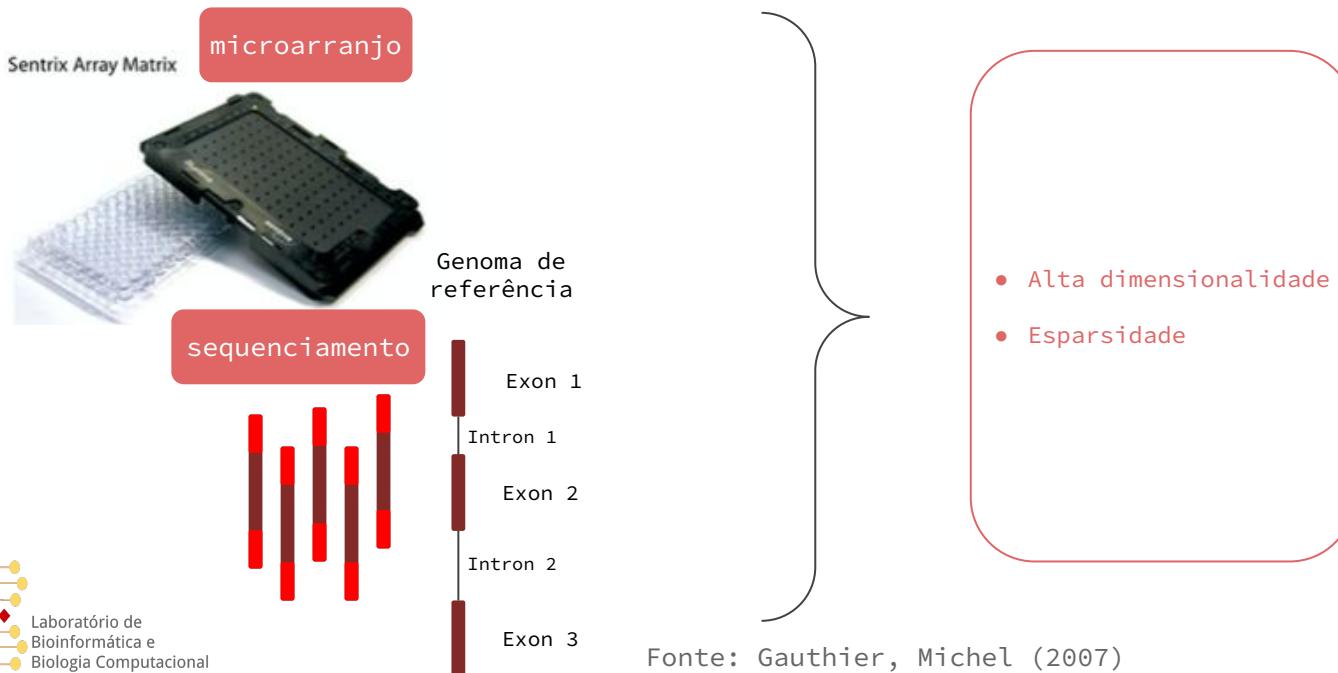
- Tecnologias de alto rendimento



# Introdução

---

- Tecnologias de alto rendimento



# Introdução

---

- Tecnologias de alto rendimento



<ul style="list-style-type: none"><li>• Menor custo</li><li>• Não requer preparação de bibliotecas de sequenciamento de DNA</li></ul>	<ul style="list-style-type: none"><li>• Viés ligações não-específicas entre os fragmentos (contornado adequando-se o número de leituras)</li></ul>
<ul style="list-style-type: none"><li>• Maior precisão</li><li>• Preparação de bibliotecas de referência</li></ul>	<ul style="list-style-type: none"><li>• Maior custo</li></ul>

Fonte: Gauthier, Michel (2007)

# Introdução

---

- Tecnologias de alto rendimento



THE CANCER GENOME ATLAS



Fonte: Gauthier, Michel (2007)

# Revisão bibliográfica

# Introdução

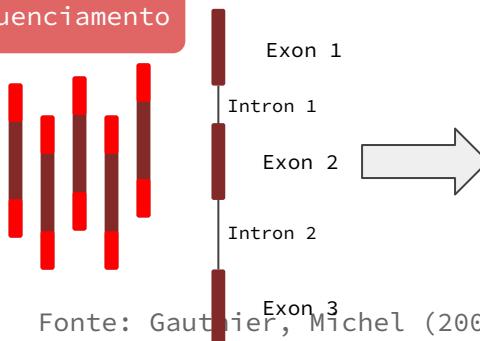
- Tecnologias de alto rendimento

microarranjo

Sentrix Array Matrix

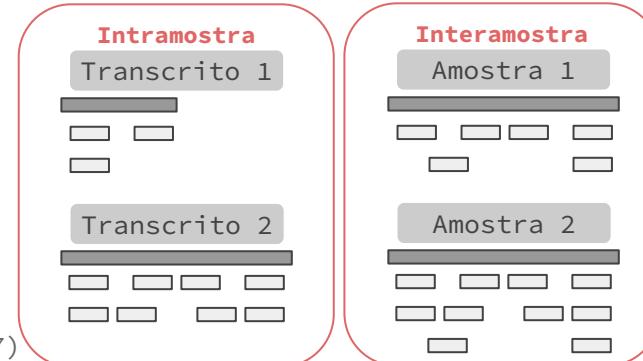


sequenciamento



		$\beta$ ou M ( $\log\beta$ )			
		gene 1	gene 2	gene 3	...
paciente 1	...	...	...	...	...
	...	...	...	...	...
paciente 2	...	...	...	...	...
paciente 3	...	...	...	...	...
...	...	...	...	...	...

miRNA apenas normaliza por leituras por milhão de leituras (tam pequeno)



mRNA

FPKM =  $\frac{\text{fragmentos}}{\text{kilobase de transrito por milhão de leituras mapeadas}}$

# Revisão bibliográfica

---

- Xie-Beni (XB)

$$XB = \frac{J_2}{d_{min}} \quad \left\{ \begin{array}{l} J_m = \frac{1}{N} \sum_{i=1}^c \sum_{j=1}^N u_{ij}^m d^2(X_j, c_i) \\ d_{min} = \min_{i,j} [d^2(c_i, c_j)] \end{array} \right.$$

Relativo à compacidade do cluster

- Fukuyama-Sugeno Index (FS)

$$J_m = \sum_{j=1}^n \sum_{i=1}^c (\mu_{ij})^m \|x_j - V_i\|^2 - \sum_{i=1}^c \|V_i - \bar{V}\|^2$$

where  $\bar{V} = \frac{1}{c} \sum_{i=1}^c V_i$ .

Relativo à distância de cada cluster e a média do centróide de cada cluster

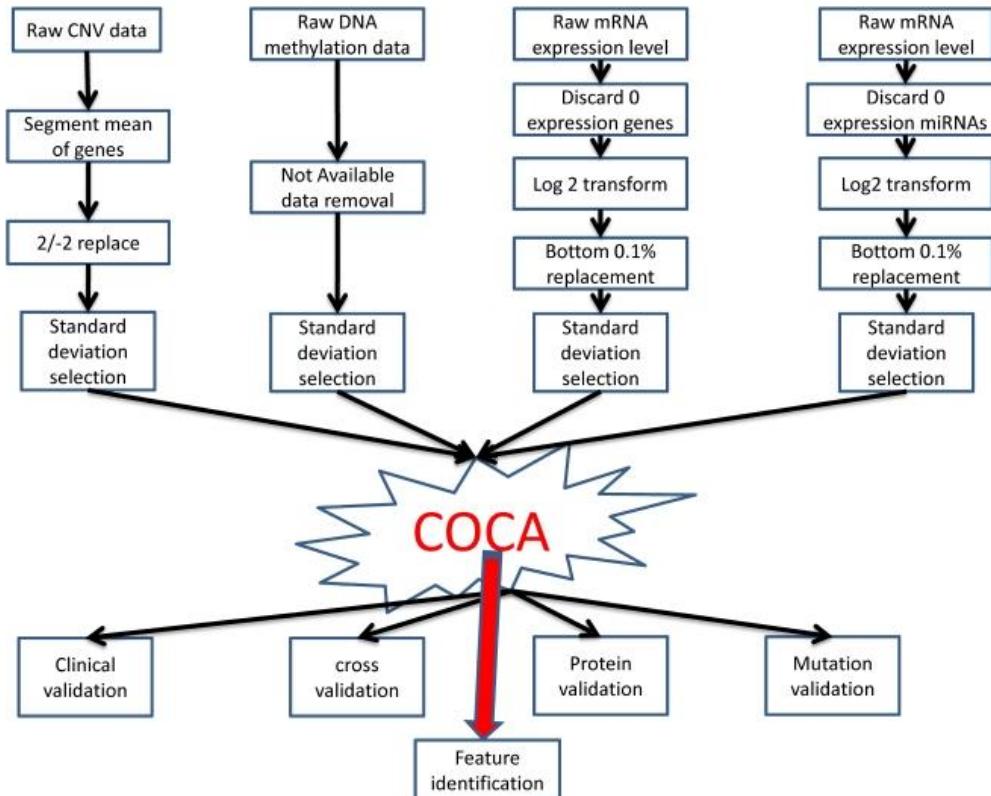
Medida de compacidade

- Fuzzy Partition coefficient (FPC)

$$F = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n (\mu_{ij})^2$$

Valor médio dos graus de pertinência

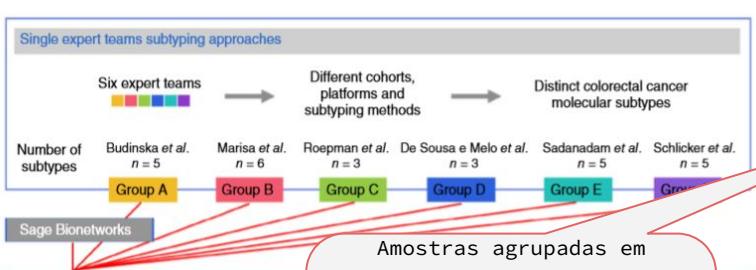
# Revisão bibliográfica



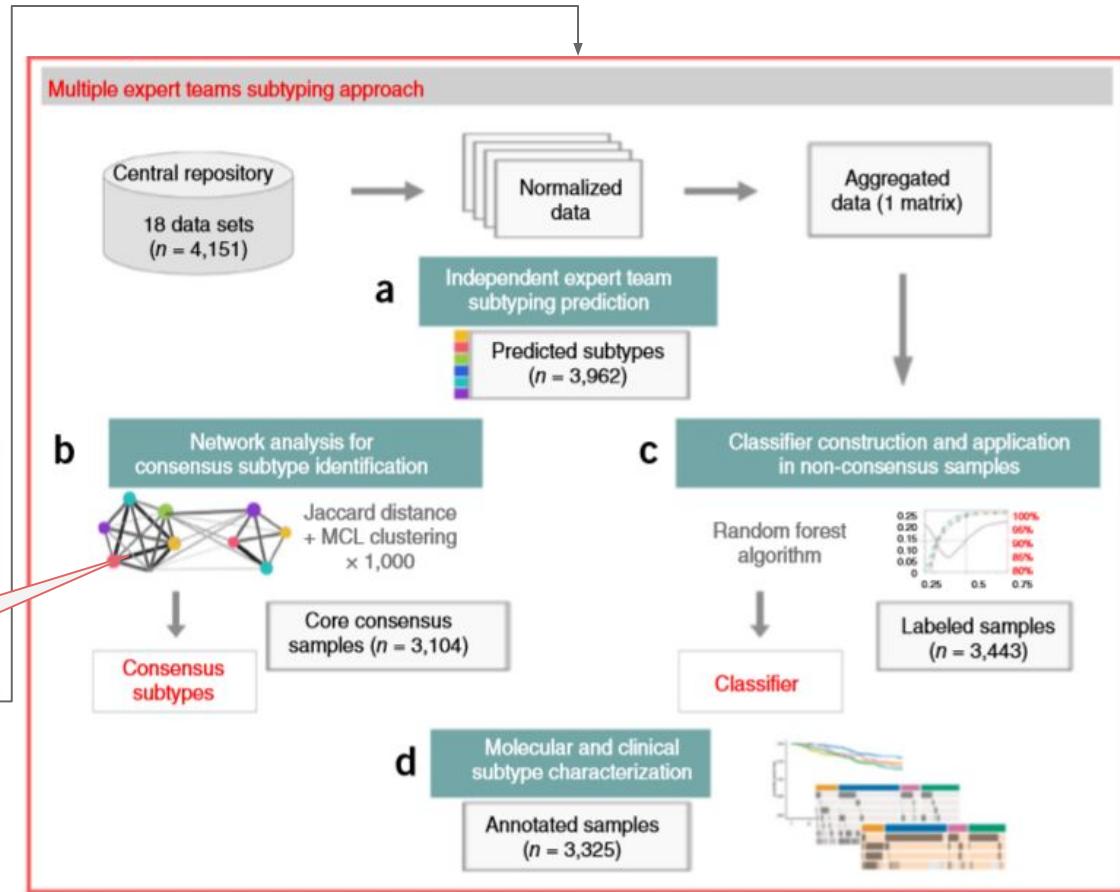
Fonte: Integrated Multiple “-omics” Data Reveal Subtypes of Hepatocellular Carcinoma

# Revisão bibliográfica

## Técnicas de agrupamento semelhantes

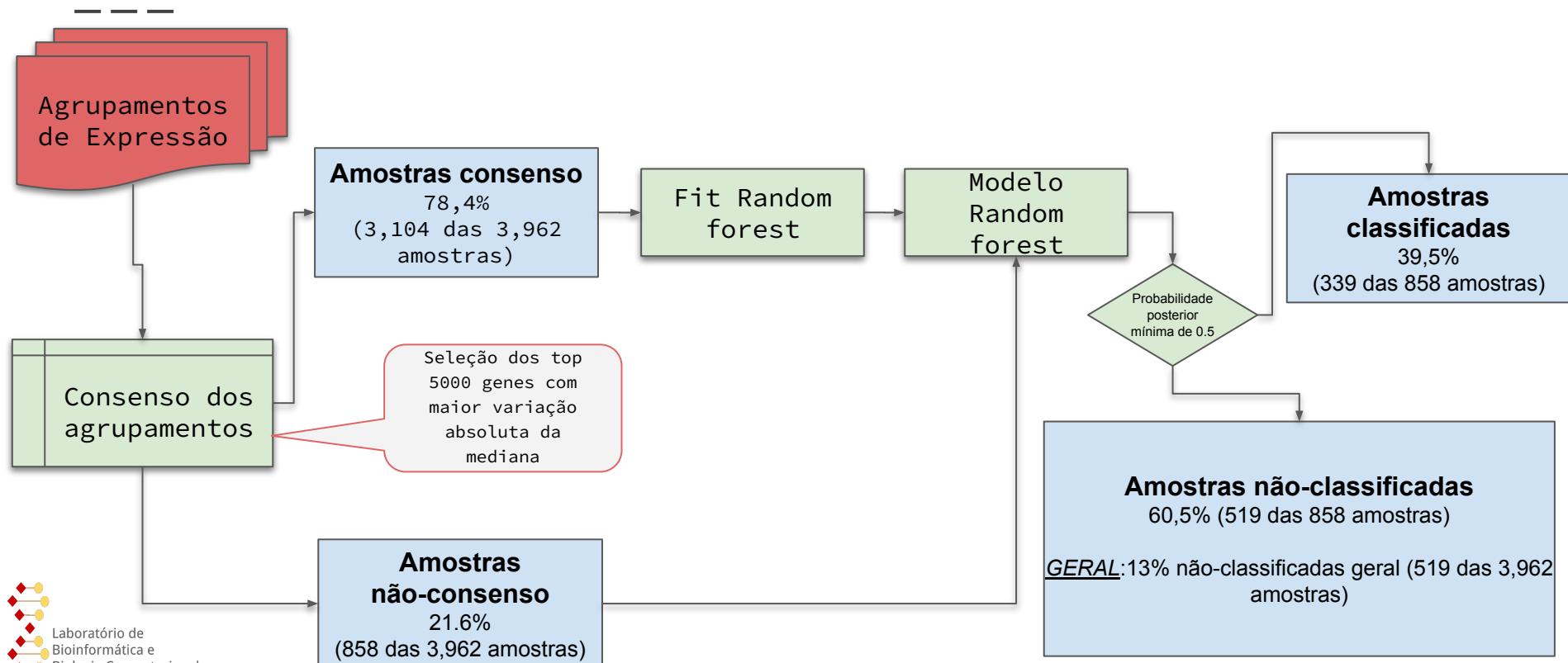


Amostras agrupadas em apenas um grupo, não refletem alto compartilhamento de características entre grupos



Fonte: The consensus molecular subtypes of colorectal cancer

# Revisão bibliográfica - Agrupamento e Floresta aleatória CMS

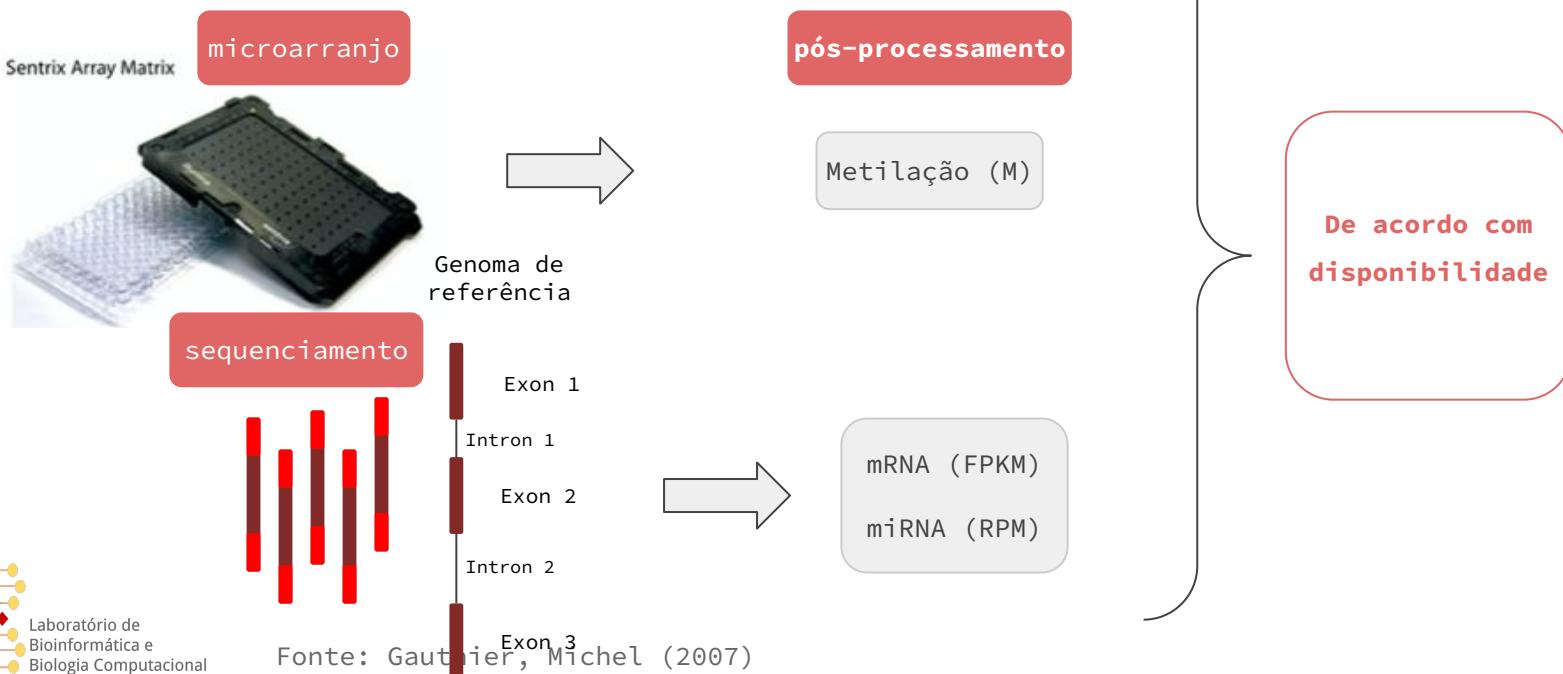


# Metodologia

# Metodologia - Extração e normalização de dados biológicos

---

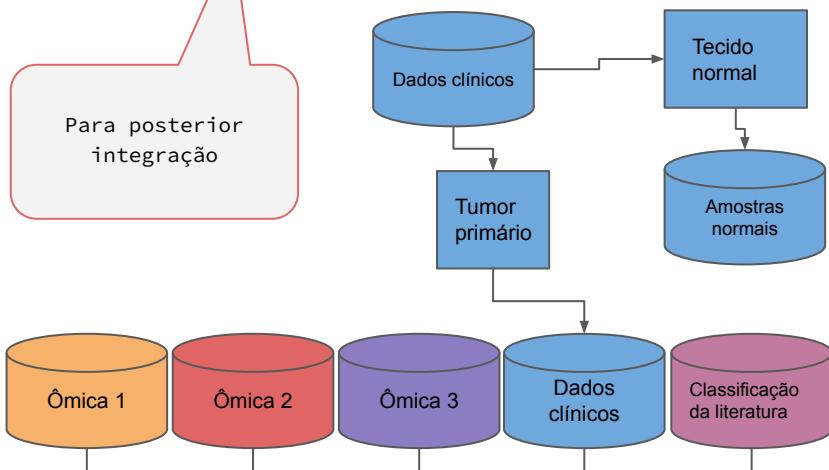
- Tecnologias de alto rendimento



# Metodologia - Conciliação e limpeza de dados

## Conciliação entre bases de dados

Para posterior  
integração

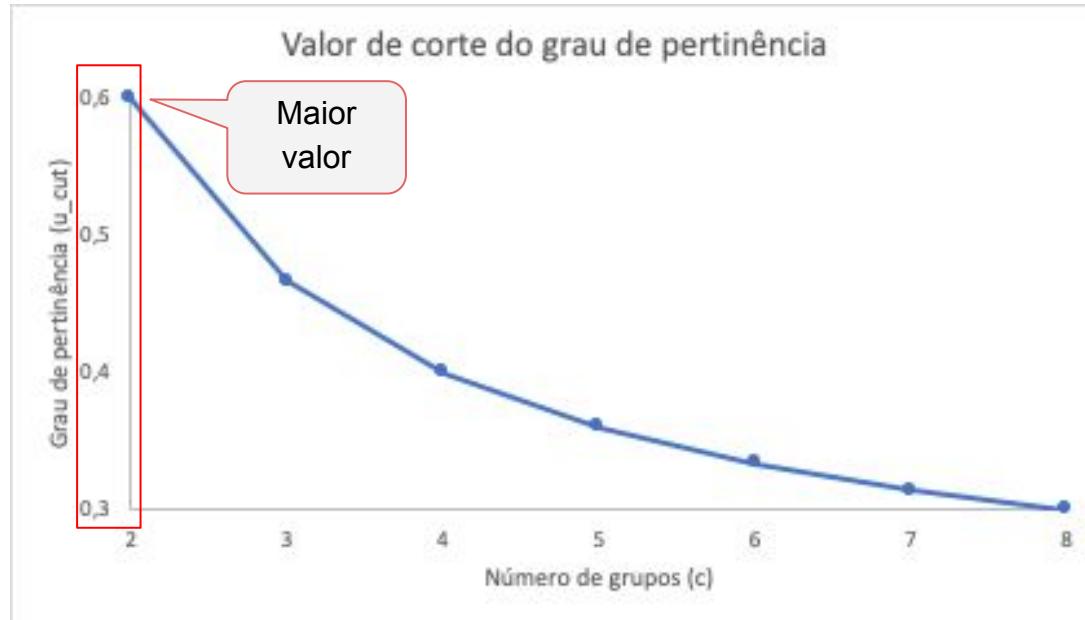


## Limpeza dados

# Metodologia - Grau de pertinência mínimo

---

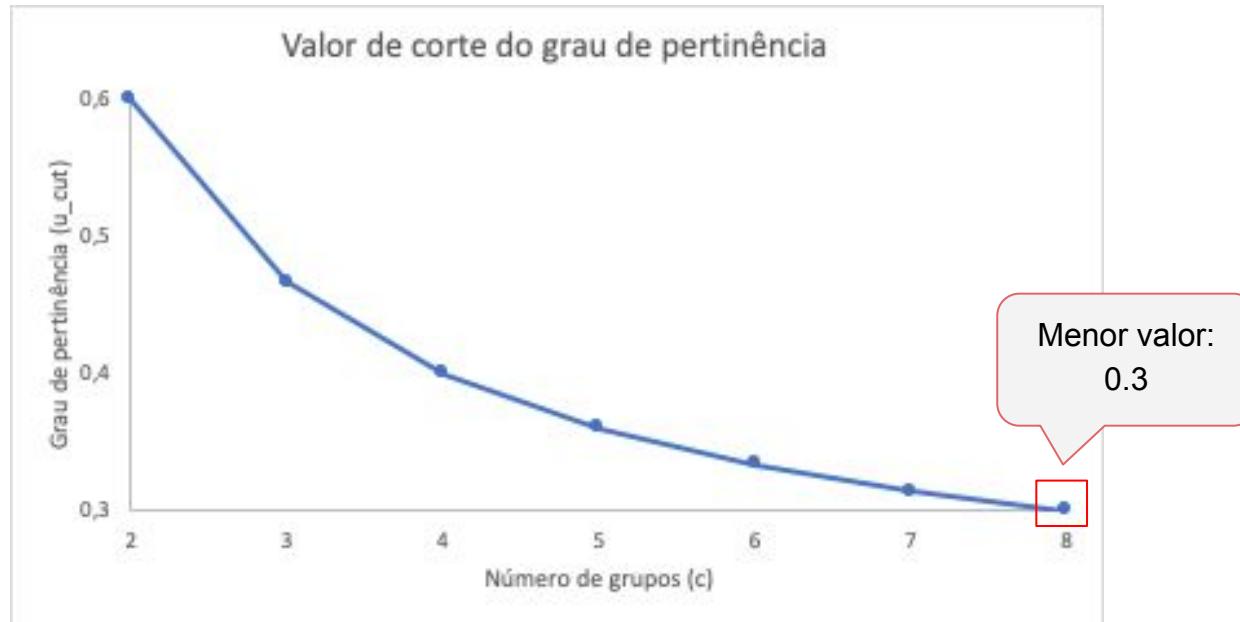
- Grau de pertinência: valor mínimo ajustado de acordo com número de grupos



# Metodologia - Grau de pertinência mínimo

---

- Grau de pertinência: valor mínimo ajustado de acordo com número de grupos



# Metodologia - Grau de pertinência de maior valor

---

- Grau mínimo de pertinência (**u fuzzy**): maior valor do grau de pertinência

Paciente	Grupo							
	1	2	3	4	5	6	7	8
TCGA-AG-3592	0.1	0.23	0.17	0.05	0	0.2	0.15	0.1
TCGA-AG-3999	0.1	0.1	0.1	0.1	0.2	0.2	0.1	0.1
TCGA-BM-6198	0.3	0.1	0.1	0.1	0.1	0.1	0.1	0.1
TCGA-AF-3400	0.1	0.2	0.2	0.1	0.1	0.1	0.1	0.1
TCGA-AF-2692	0.1	0.1	0.3	0.1	0.1	0.1	0.1	0.1

# Metodologia - Grau de pertinência de maior valor

---

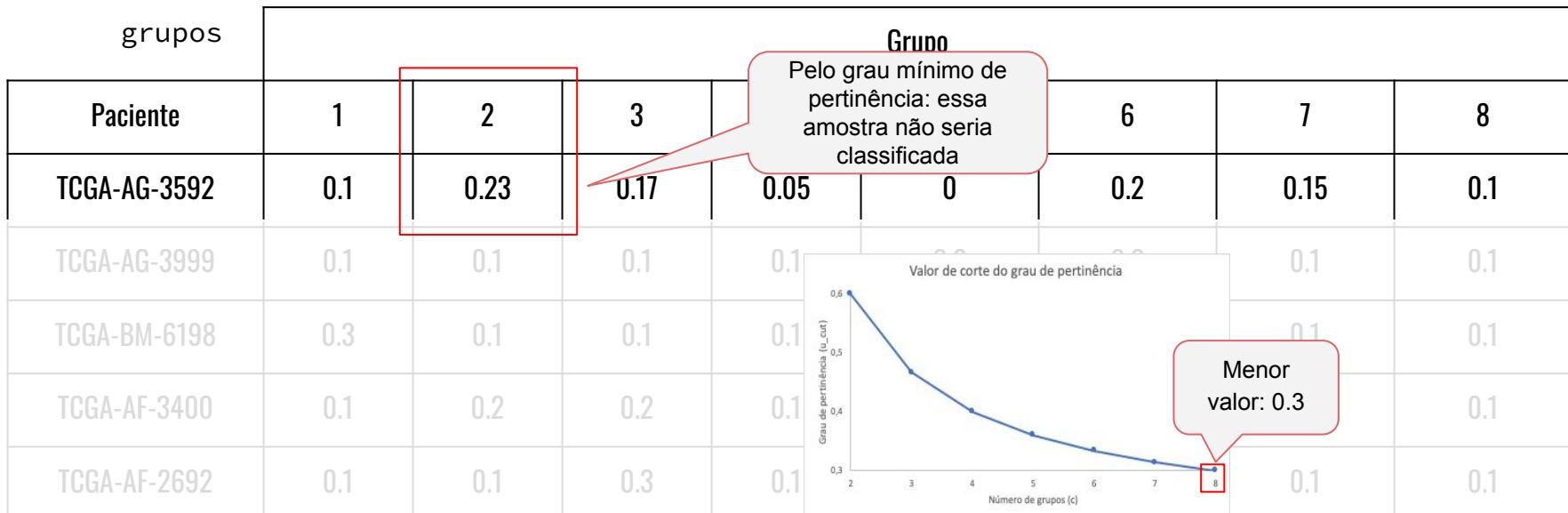
- Grau mínimo de pertinência ( $\mu_{fuzzy}$ ): que decresce com aumento do número de grupos

Paciente	Grupo							
	1	2	3	4	5	6	7	8
TCGA-AG-3592	0.1	0.23	0.17	0.05	0	0.2	0.15	0.1
TCGA-AG-3999	0.1	0.1	0.1	0.1	0.2	0.2	0.1	0.1
TCGA-BM-6198	0.3	0.1	0.1	0.1	0.1	0.1	0.1	0.1
TCGA-AF-3400	0.1	0.2	0.2	0.1	0.1	0.1	0.1	0.1
TCGA-AF-2692	0.1	0.1	0.3	0.1	0.1	0.1	0.1	0.1

# Metodologia - Grau de pertinência de maior valor

---

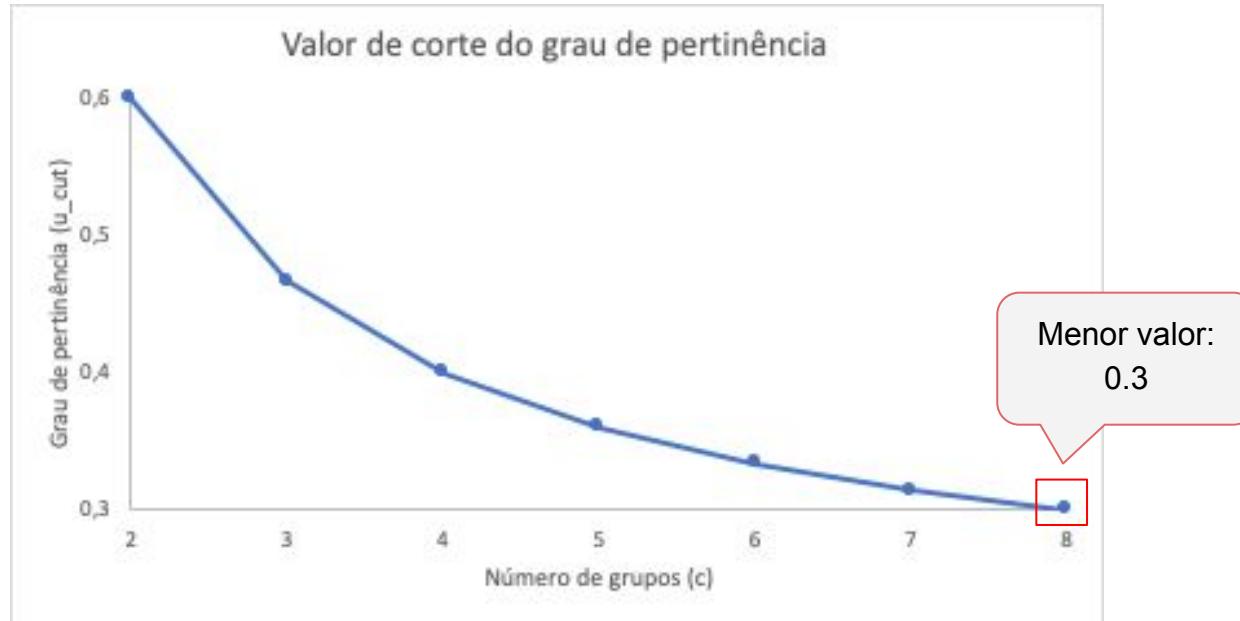
- Grau mínimo de pertinência ( $\mu_{fuzzy}$ ): que decresce com aumento do número de grupos



# Metodologia - Grau de pertinência de maior valor

---

- Grau mínimo de pertinência ( $u_{fuzzy}$ ): considerando grau mínimo de per



# Metodologia - Grau de pertinência de maior valor

---

- Grau mínimo de pertinência ( $\mu_{fuzzy}$ ): que decresce com aumento do número de grupos

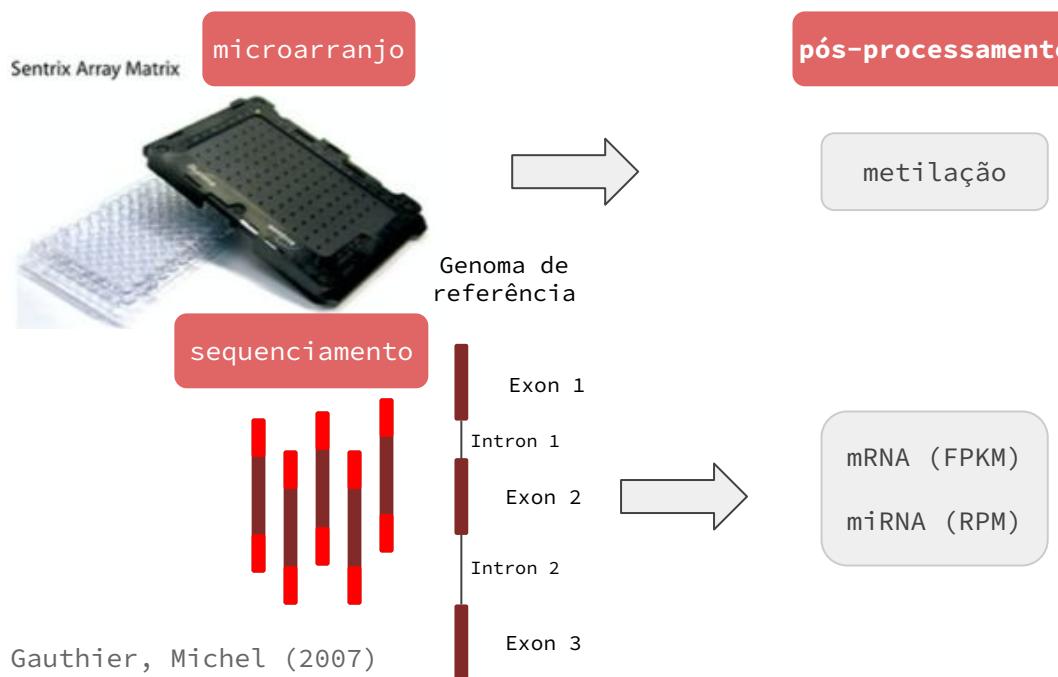
Paciente	Grupo							
	1	2	3	4	5	6	7	8
TCGA-AG-3592	0.1	0.23	0.17	0.05	0	0.2	0.15	0.1
TCGA-AG-3999	0.1	0.1	0.1	0.1	0.2	0.2	0.1	0.1
TCGA-BM-6198	0.3	0.1	0.1	0.1	0.1	0.1	0.1	0.1
TCGA-AF-3400	0.1	0.2	0.2	0.1	0.1	0.1	0.1	0.1
TCGA-AF-2692	0.1	0.1	0.3	0.1	0.1	0.1	0.1	0.1

# Estudo de caso

# Estudo de caso - Extração e normalização de dados biológicos

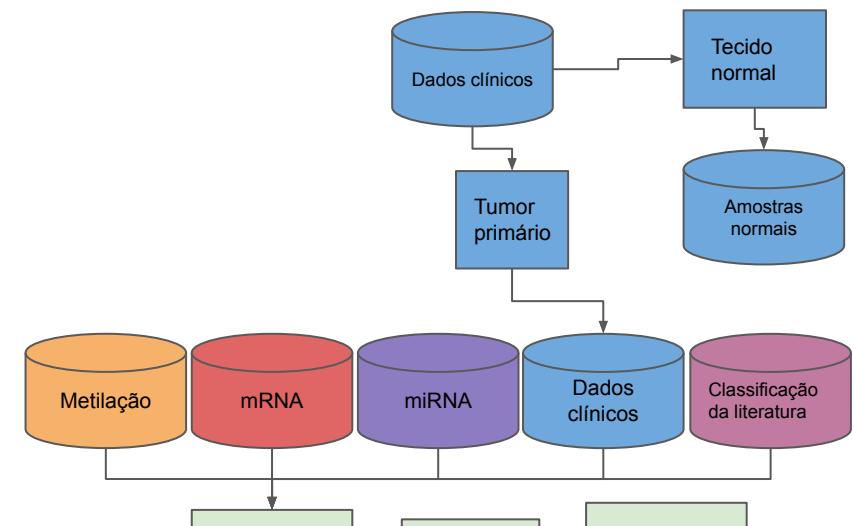
- Tecnologias de alto rendimento

THE CANCER GENOME ATLAS

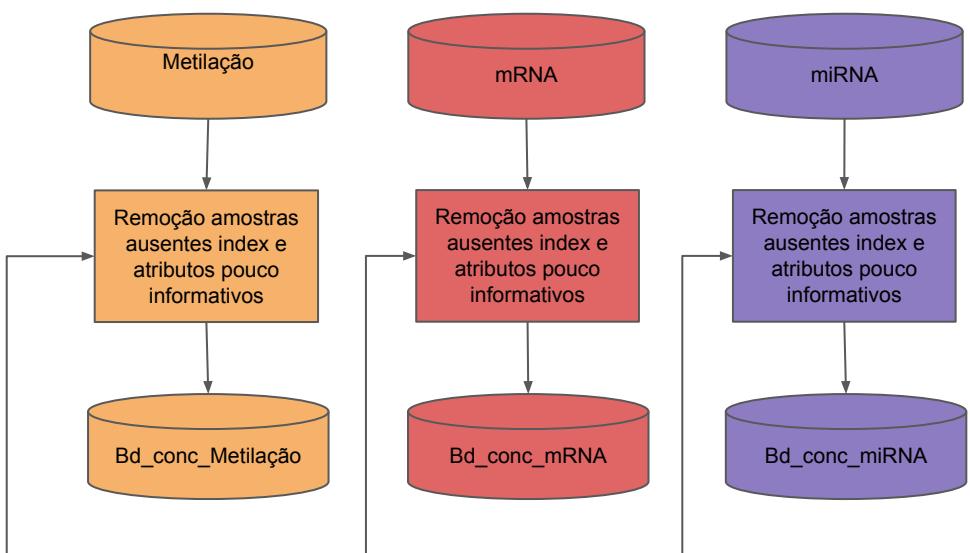


# Estudo de caso - Conciliação e limpeza de dados

## Conciliação entre bases de dados



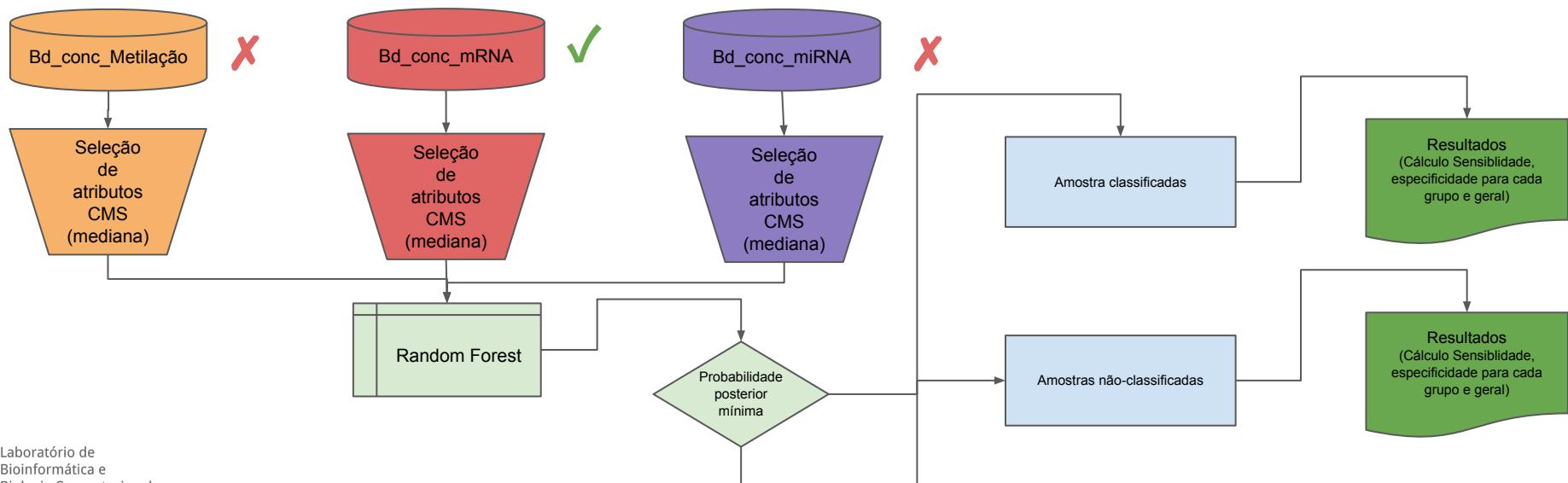
## Limpeza dados



# Estudo de caso - Escolha do conjuntos de dados

---

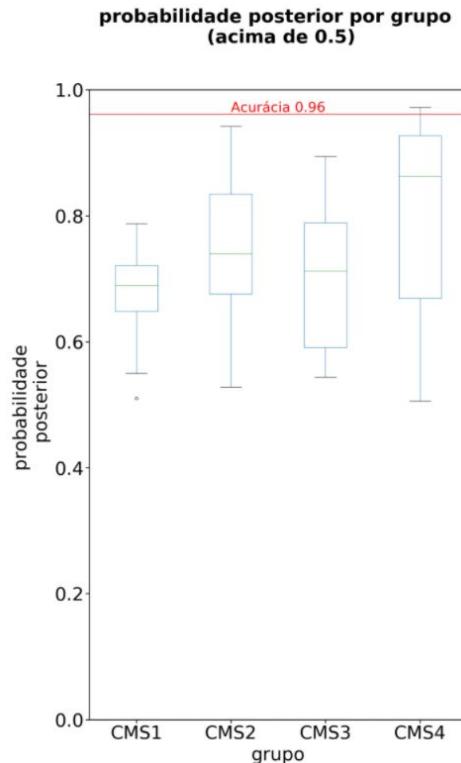
- Elaboração de classificador multiômico com base no classificador baseado em expressão de mRNA de Guinney, 2015 (CMS) e no trabalho de Liu, 2016.



# Estudo de caso - Escolha do conjuntos de dados

---

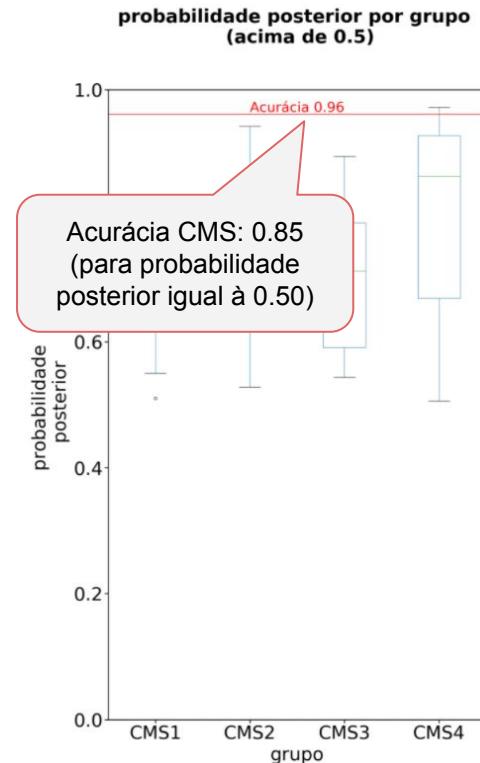
- Parâmetros:
  - Pré-seleção de features: 80 para miRNA e 5000 para metilação e expressão
  - Número de árvores: 500
  - Número de nós por árvore: 70
  - Divisão amostras teste/treino:  $\frac{1}{3}$  (150) e  $\frac{2}{3}$  (300)
  - Probabilidade posterior mínima: varia de 0.2 à 0.80, com intervalos de 0.1



# Estudo de caso - Escolha do conjuntos de dados

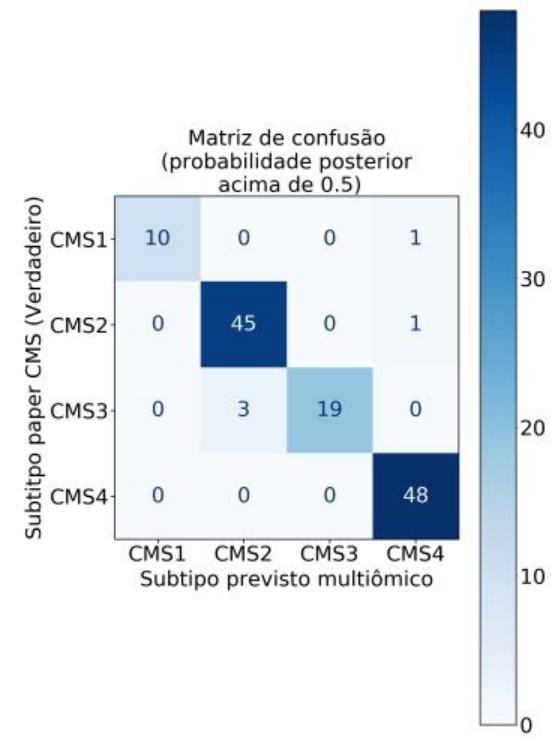
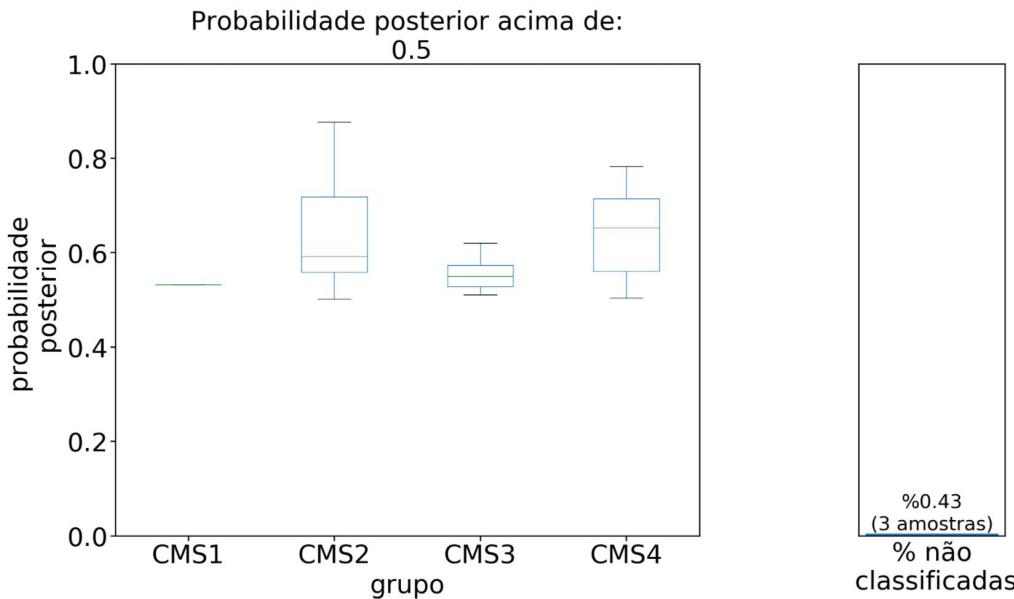
---

- Parâmetros:
  - Pré-seleção de features: 80 para miRNA e 5000 para metilação e expressão
  - Número de árvores: 500
  - Número de nós por árvore: 70
  - Divisão amostras teste/treino:  $\frac{1}{3}$  (150) e  $\frac{2}{3}$  (300)
  - Probabilidade posterior mínima: varia de 0.2 à 0.80, com intervalos de 0.1



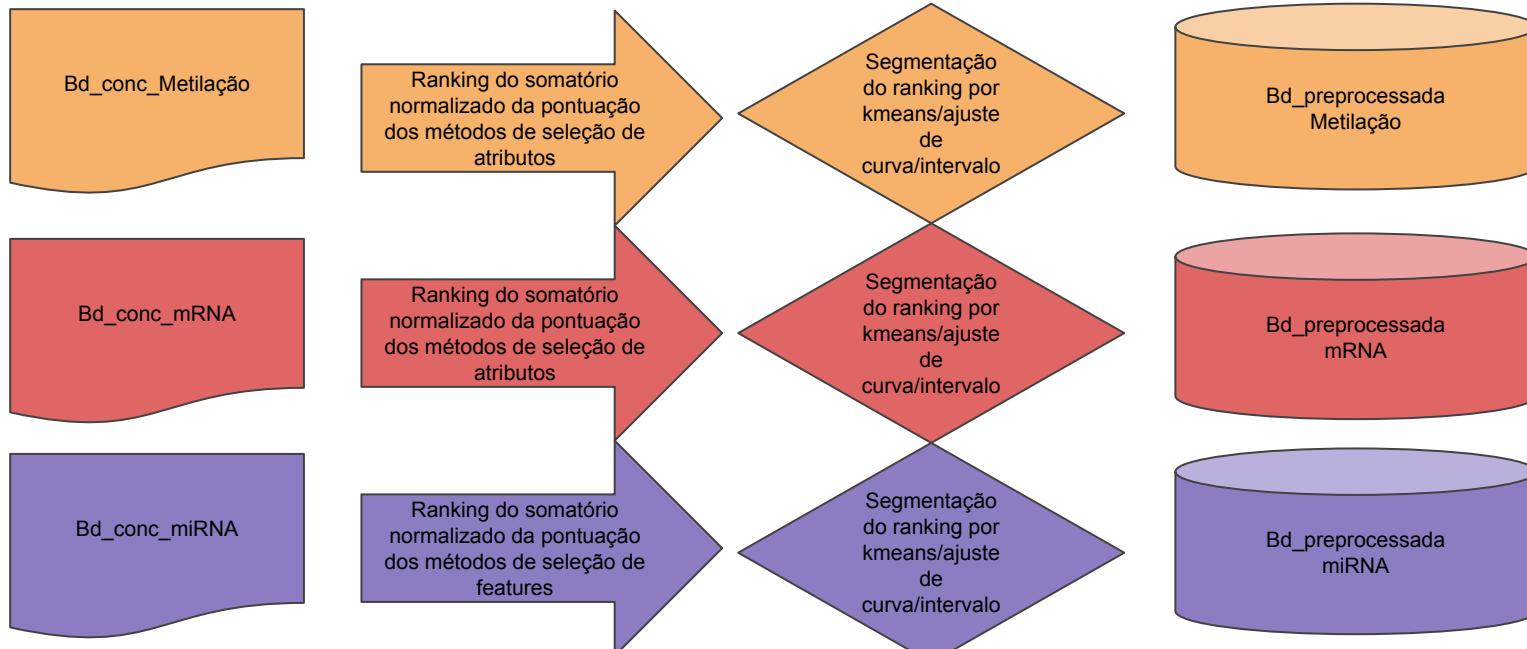
# Estudo de caso - Escolha do conjuntos de dados

---



# Estudo de caso - Seleção de atributos preliminar

---



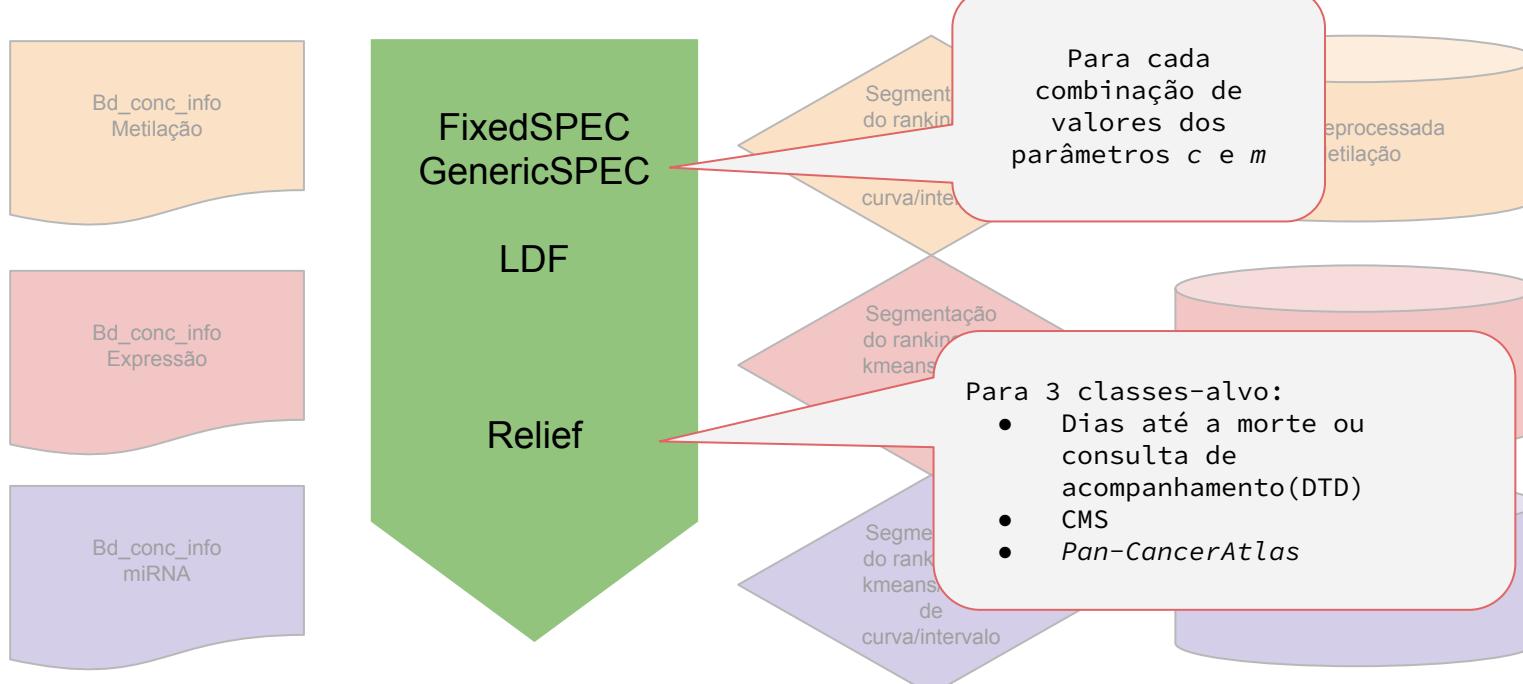
# Estudo de caso - Seleção de atributos preliminar

---



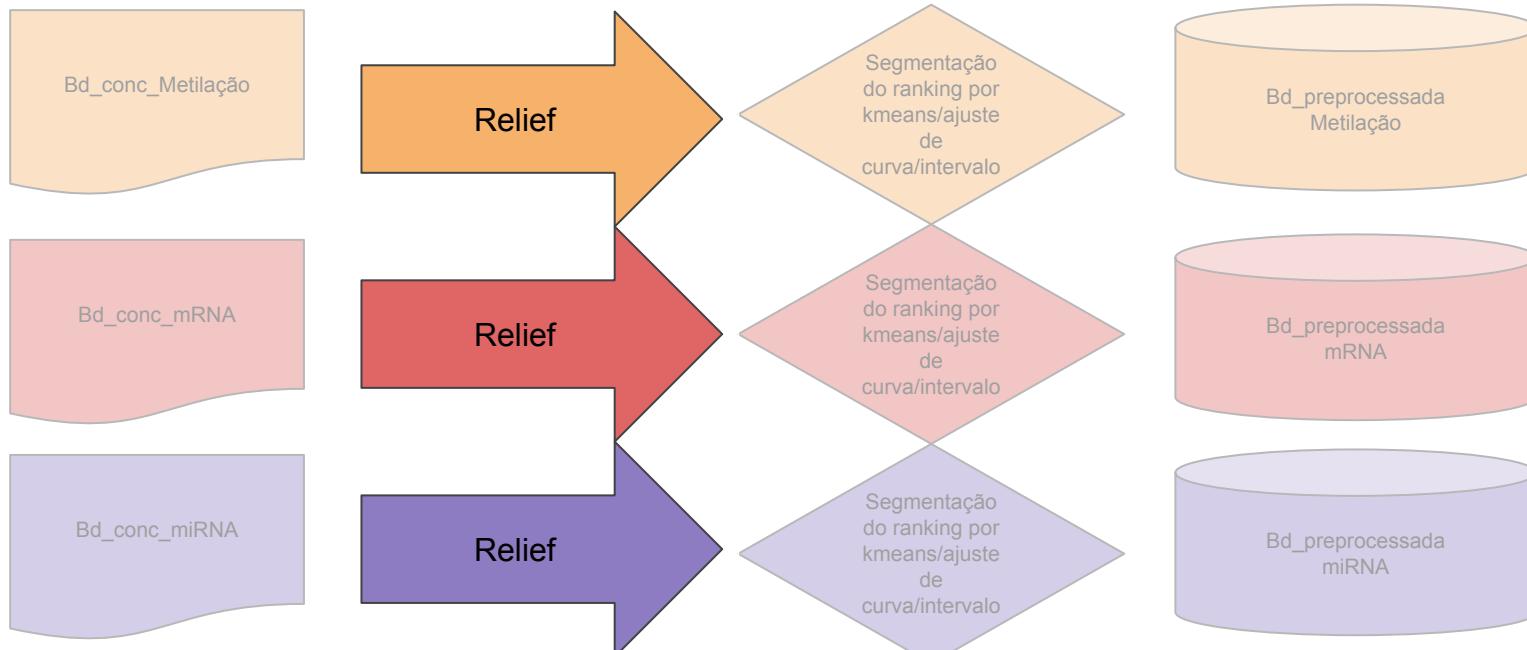
# Estudo de caso - Seleção de atributos preliminar

---



# Estudo de caso - Seleção de atributos preliminar

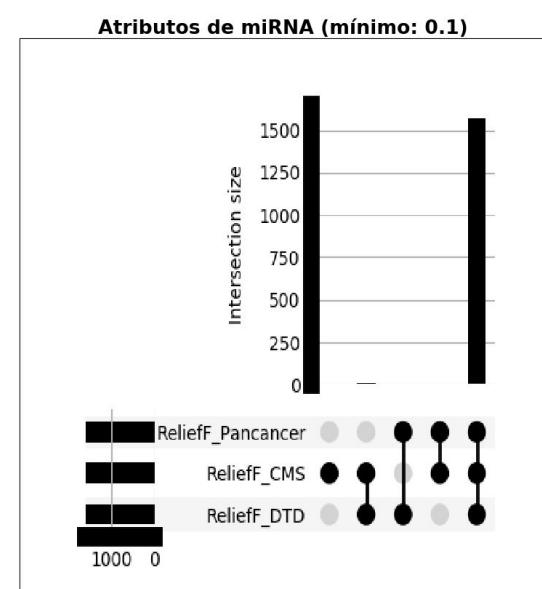
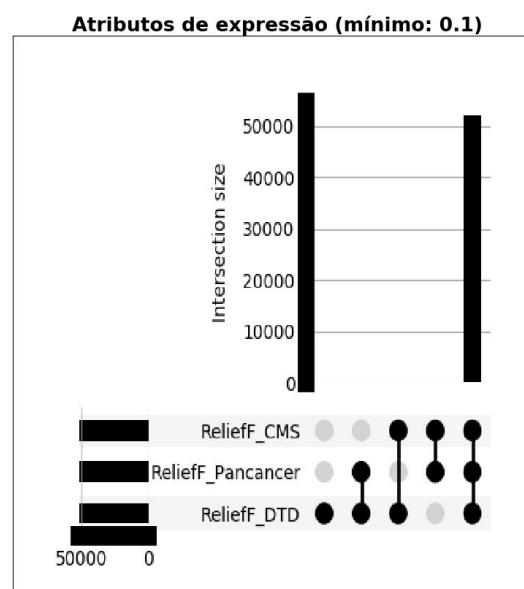
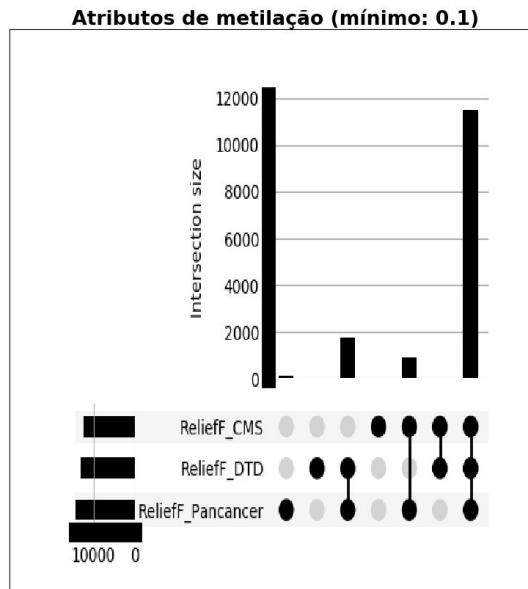
---



# Estudo de caso - Seleção de atributos preliminar

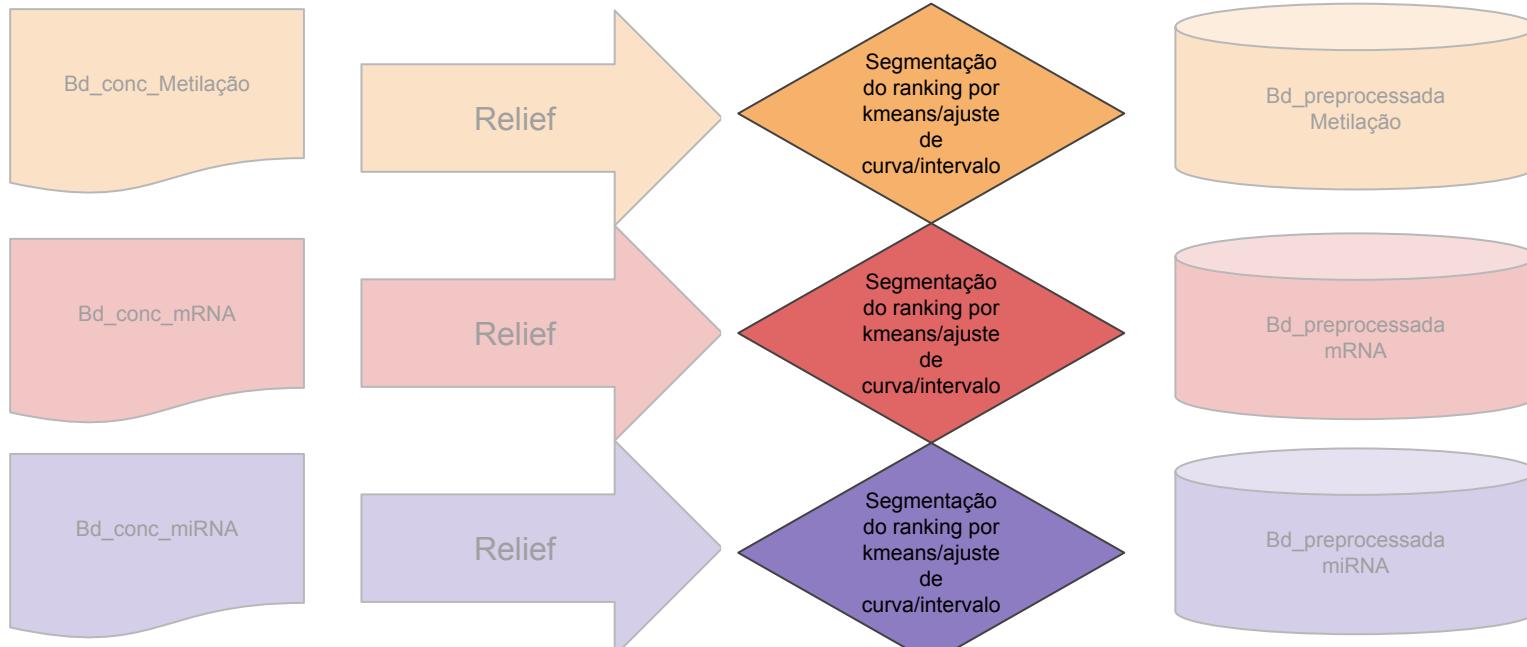
---

- Interseção entre métodos de seleção de atributos

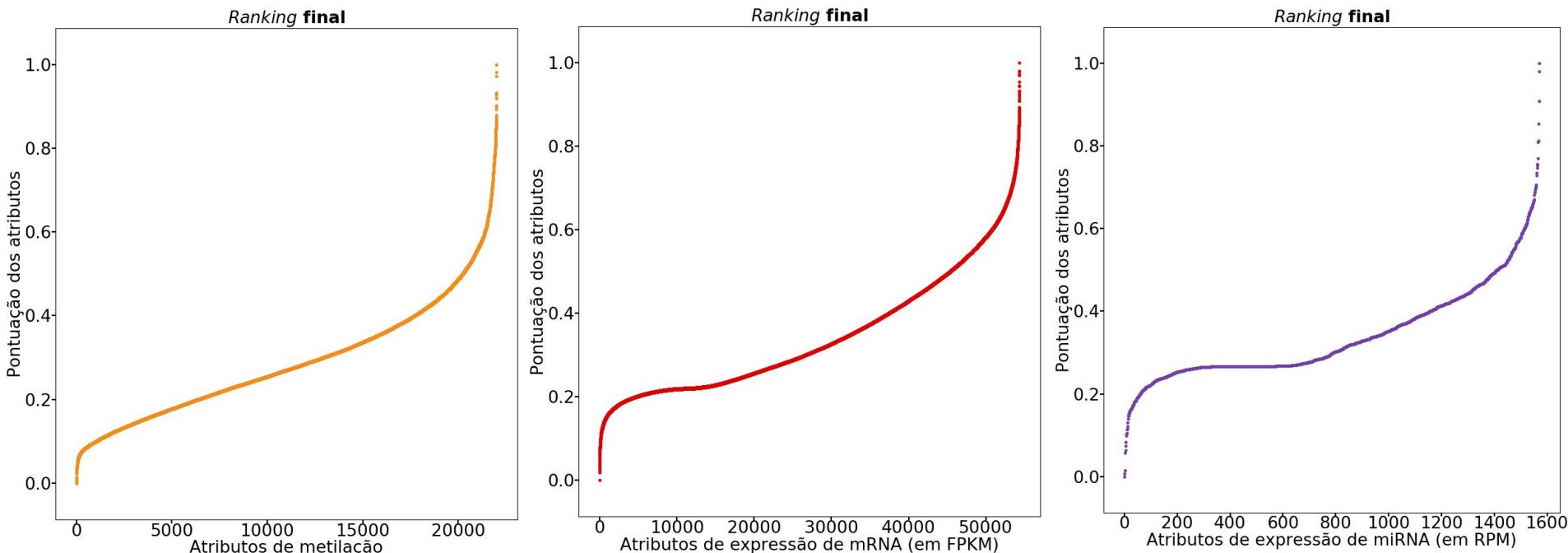


# Estudo de caso - Segmentação do ranking de atributos

---

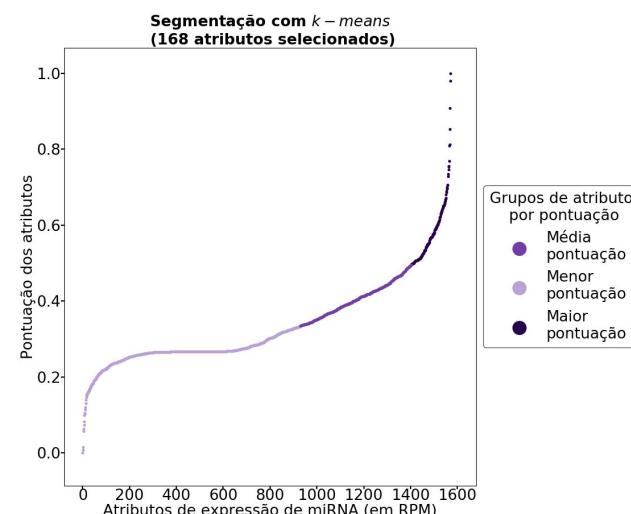
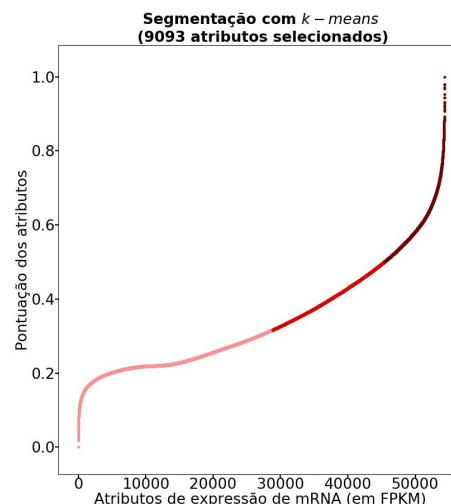
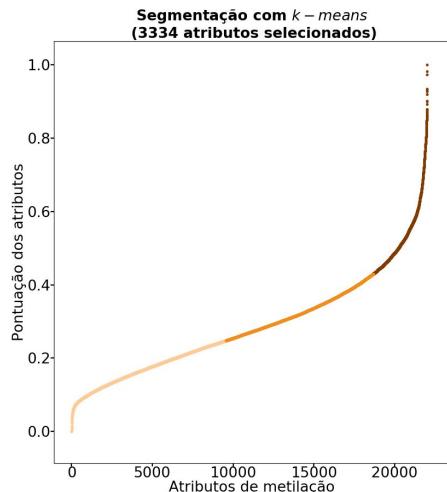


# Estudo de caso - Segmentação do ranking de atributos



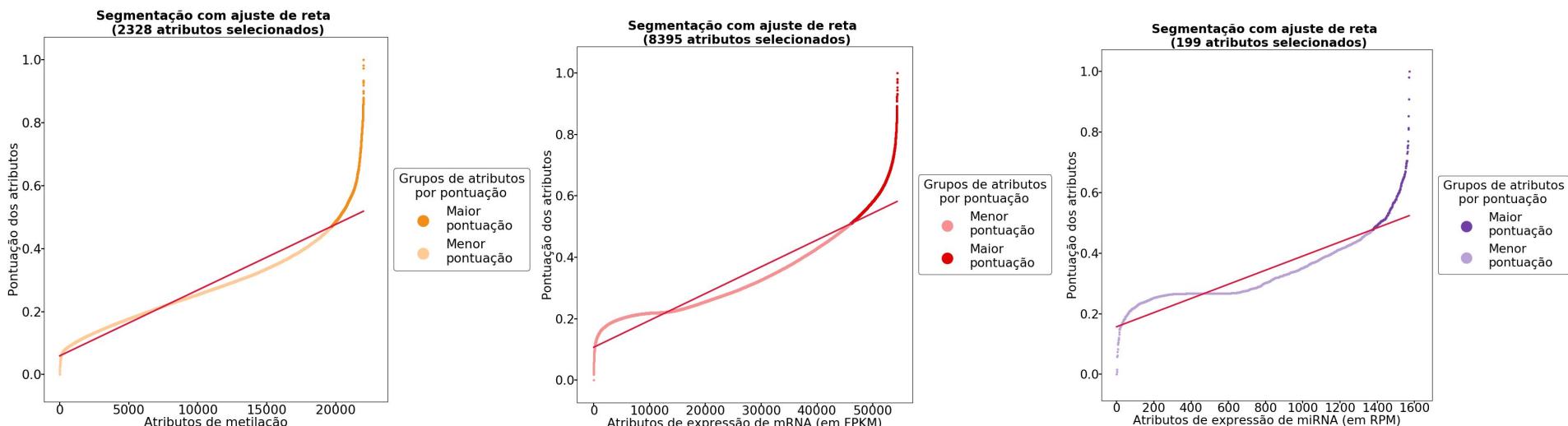
# Estudo de caso - Segmentação do ranking de atributos

---



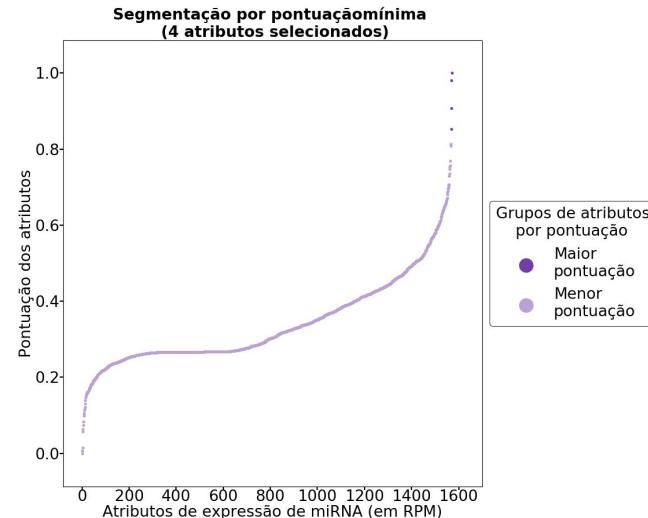
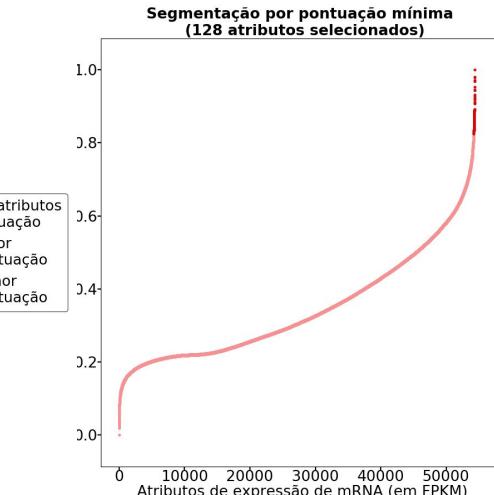
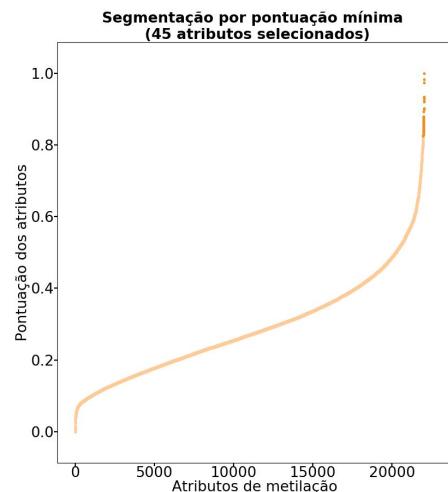
# Estudo de caso - Segmentação do ranking de atributos

---



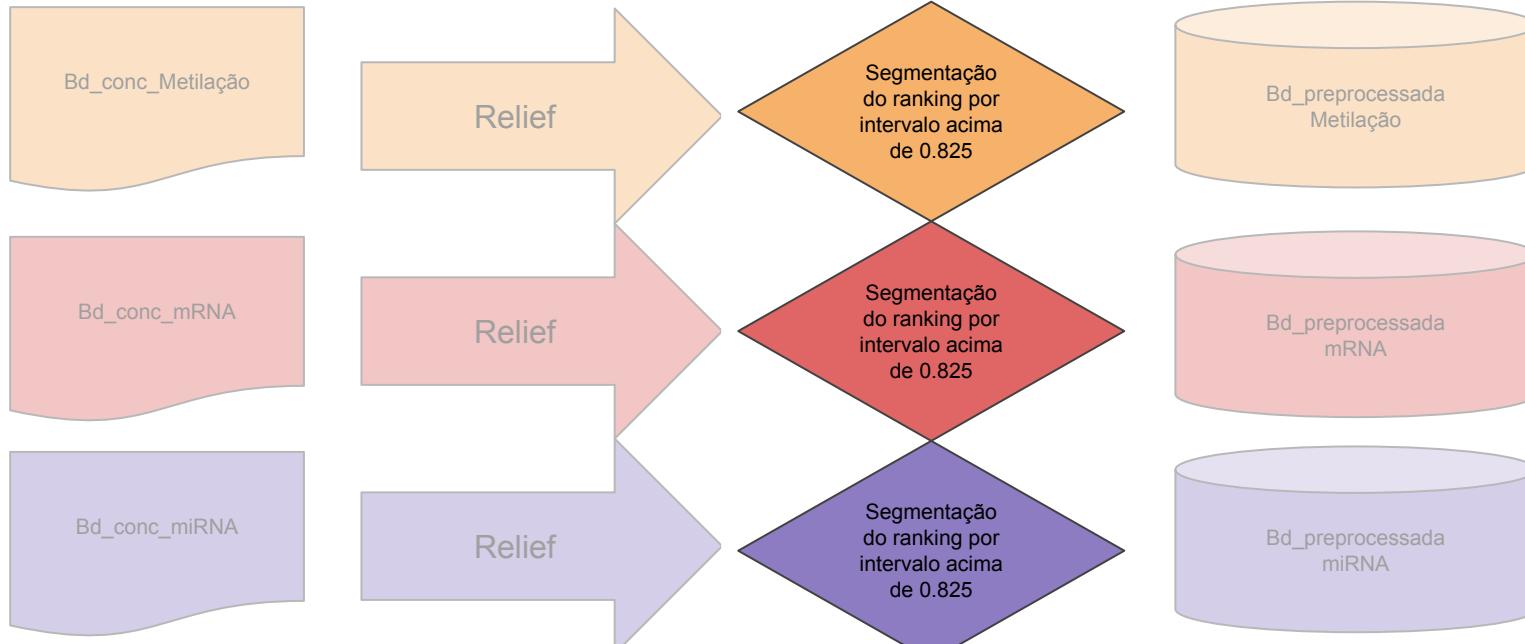
# Estudo de caso - Segmentação do ranking de atributos

---



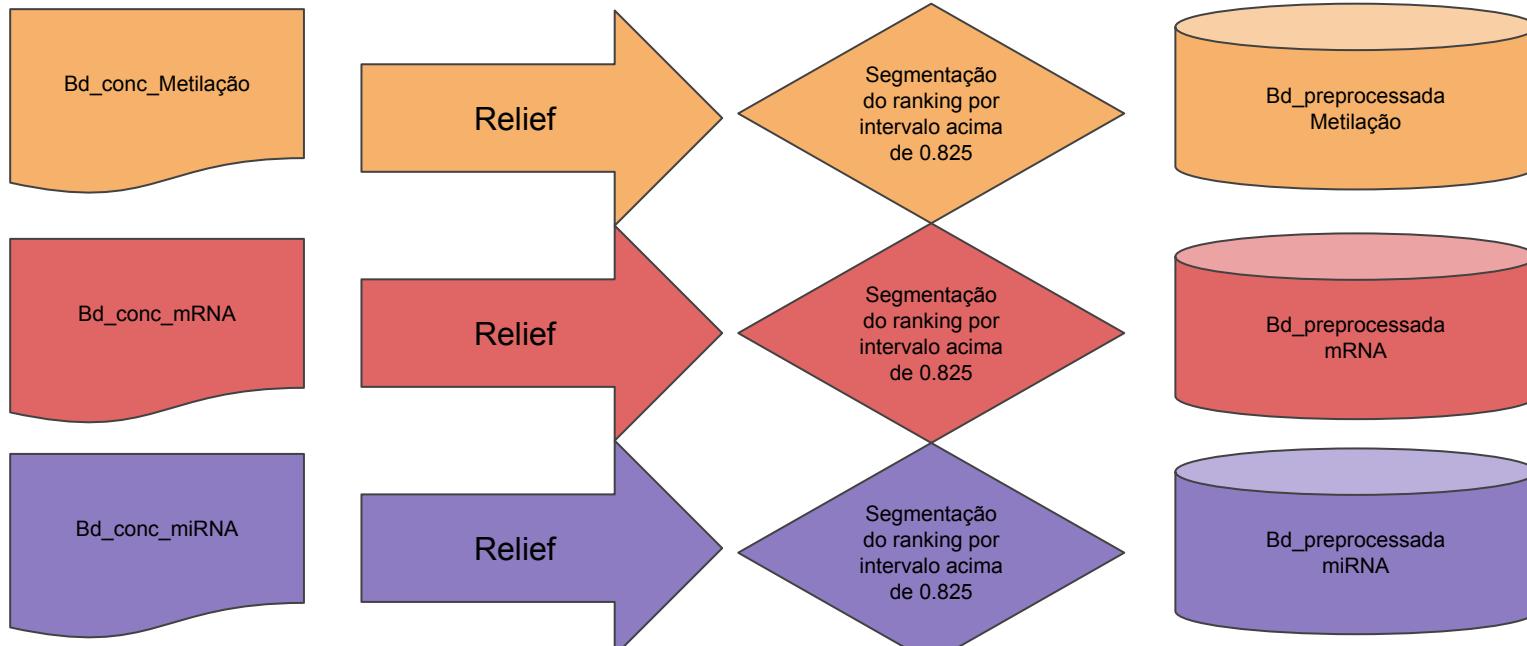
# Estudo de caso - Segmentação do ranking de atributos

---



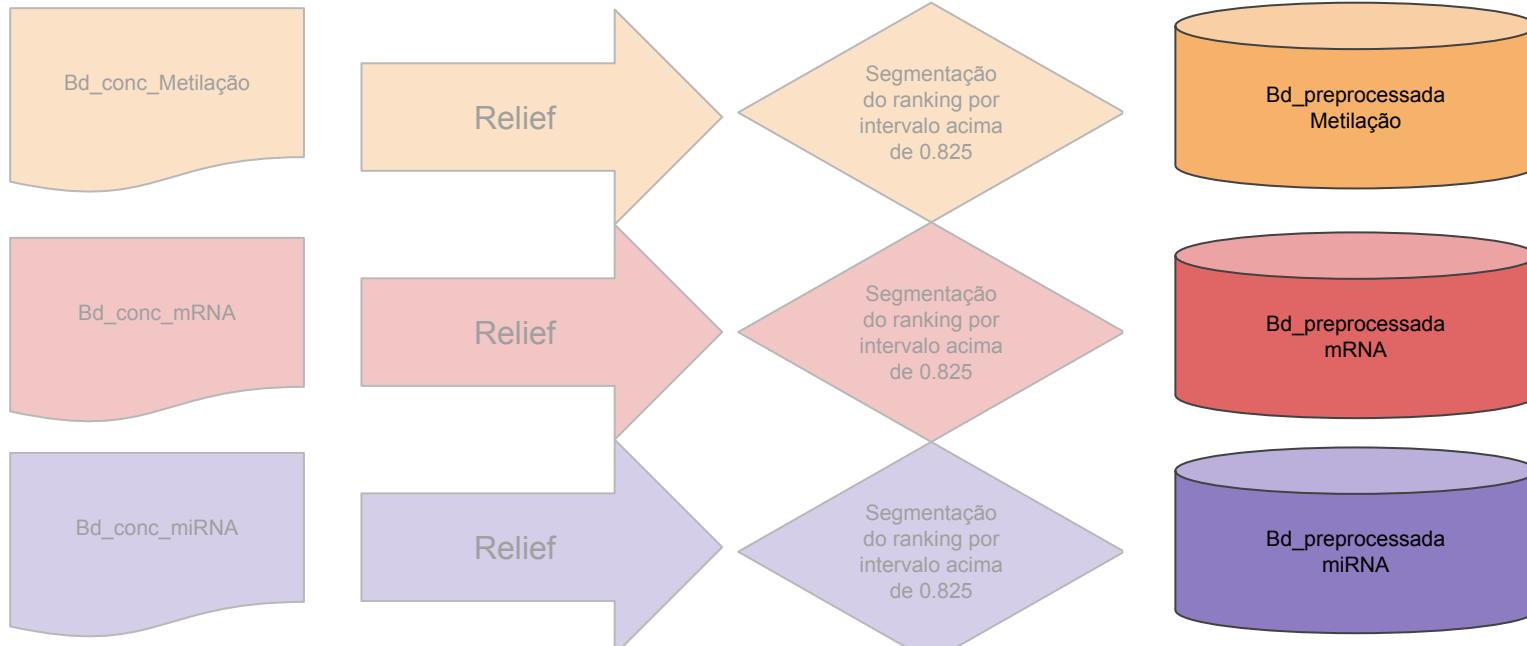
# Estudo de caso - Pré-processamento

---



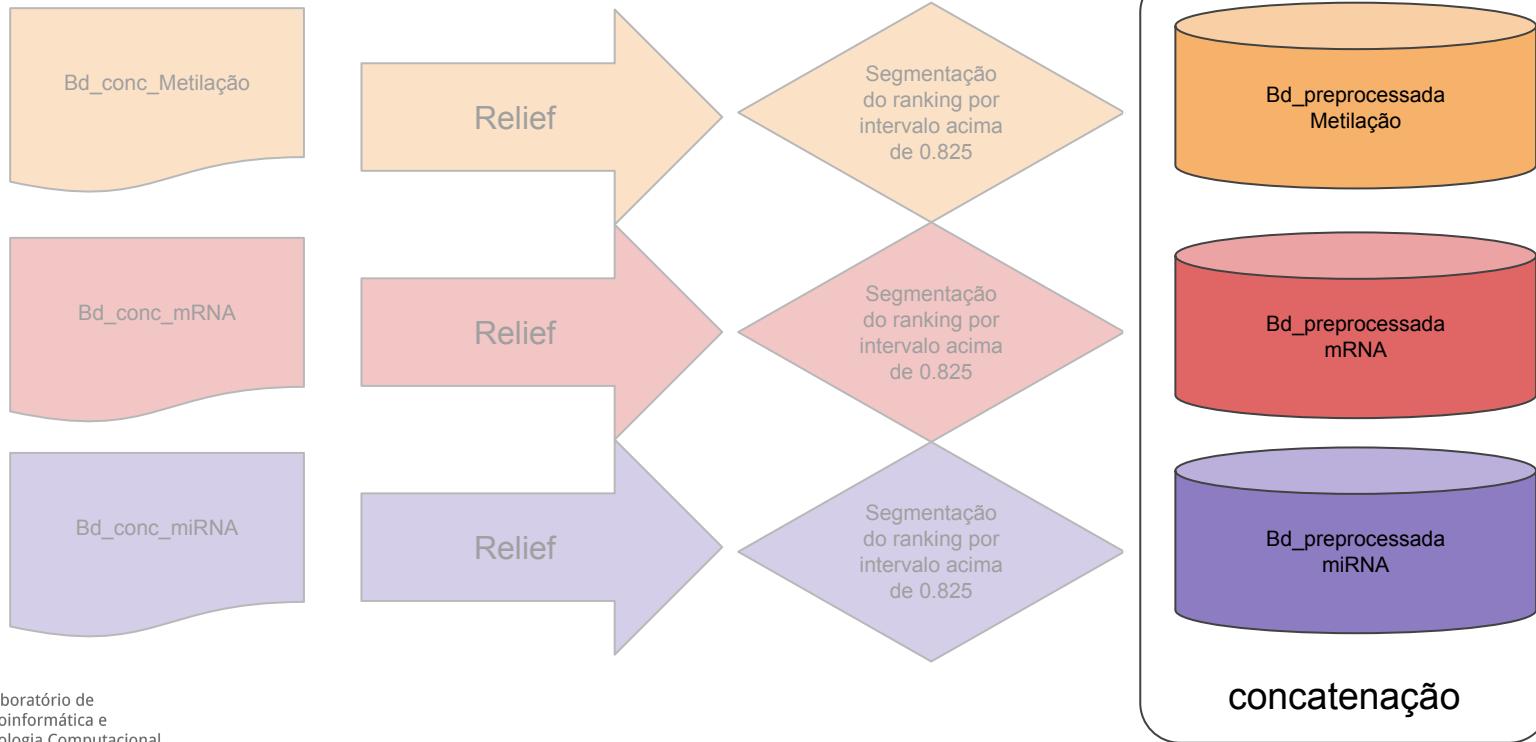
# Estudo de caso - Pré-processamento

---



# Estudo de caso - Pré-processamento

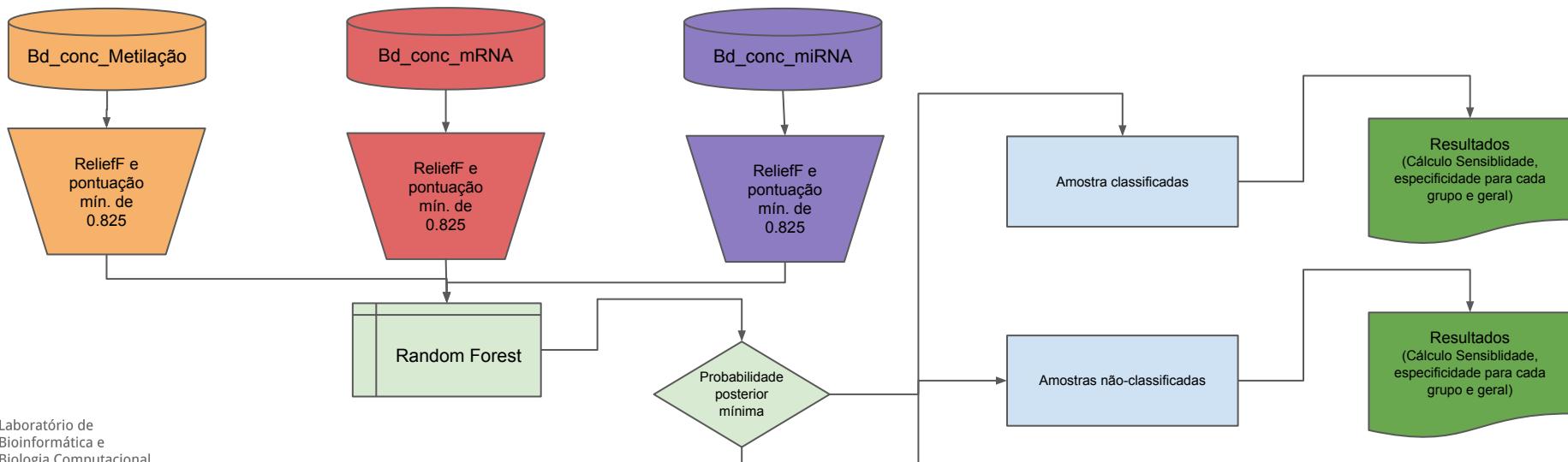
---



# Estudo de caso - Pré-processamento

---

- Elaboração de classificador multiômico com base no classificador baseado em expressão de mRNA de Guinney, 2015 (CMS), no trabalho de Liu, 2016 com método de seleção de atributos ReliefF e importância mínima do ranking de atributos de 0.825.



# Estudo de caso - Agrupamento dados concatenados

---

- Para cada par de valores dos parâmetros número de grupos (c), grau de fuzzificação (m) e intervalo do ranking de atributos:

