
**UNIVERSIDADE DO ESDADO DA BAHIA**

**CAMPUS III - JUAZEIRO**

**DEPARTAMENTO DE TECNOLOGIA E CIÊNCIAS SOCIAIS**
**ENGENHARIA DE BIOPROCESSOS E BIOTECNOLOGIA**

# ANÁLISE COMPARATIVA DE FERRAMENTAS PARA CLASSIFICAÇÃO TAXONÔMICA DE METAGENOMAS

**GABRIEL AMORIM DE ALBUQUERQUE SILVA**

JUAZEIRO-BA

Julho, 2022

**GABRIEL AMORIM DE ALBUQUERQUE SILVA**

# ANÁLISE COMPARATIVA DE FERRAMENTAS PARA CLASSIFICAÇÃO TAXONÔMICA DE METAGENOMAS

Monografia apresentada ao Colegiado de Engenharia de Bioprocessos e Biotecnologia da Universidade do Estado da Bahia UNEB Campus III, como requisito parcial para avaliação do Trabalho de Conclusão do Curso de Engenharia de Bioprocessos e Biotecnologia.

Prof. Dr. João José de Simoni Gouveia

JUAZEIRO-BA
Julho, 2022

Dados Internacionais de Catalogação na Publicação
Regivaldo José da Silva/CRB-5-1169

S586a Silva, Gabriel Amorim de Albuquerque

Análise comparativa de ferramentas para classificação taxonômica de metagenomas / Gabriel Amorim de Albuquerque Silva. Juazeiro-BA, 2022.

19 fls.: il.

Orientador: Prof. Dr. João José de Simoni Gouveia.
Inclui Referências
TCC (Graduação – Engenharia de Bioprocessos e Biotecnologia) – Universidade do Estado da Bahia. Departamento de Tecnologia e Ciências Sociais. Campus III. 2022.

1. Metagenomas. 2. Microbioma. 3. Bioinformática. 4. Classificadores. I. Gouveia, João José de Simoni. II. Universidade do Estado da Bahia. Departamento de Tecnologia e Ciências Sociais. III. Título.

CDD: 636.20896

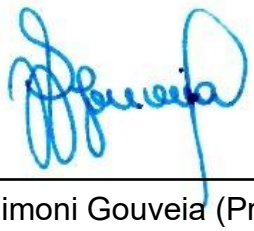
**GABRIEL AMORIM DE ALBUQUERQUE SILVA**

# ANÁLISE COMPARATIVA DE FERRAMENTAS PARA CLASSIFICAÇÃO TAXONÔMICA DE METAGENOMAS

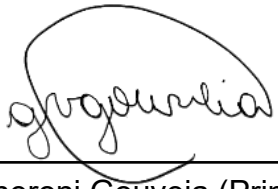
Trabalho de conclusão de curso aprovado como pré-requisito parcial à obtenção ao grau de Bacharel em Engenheira de Bioprocessos e Biotecnologia no curso de graduação em Engenharia Bioprocessos e Biotecnologia do Departamento de Tecnologia e Ciências Sociais da Universidade do Estado da Bahia.

Aprovado em: 16 de Julho de 2022

BANCA EXAMINADORA



Prof. Dr. João José de Simoni Gouveia (Presidente/Orientador)
Universidade Federal do Vale do São Francisco



Profa. Dra. Gisele Veneroni Gouveia (Primeiro examinador)
Universidade Federal do Vale do São Francisco

Me. Joel Fonseca Nogueira (Segundo examinador)
Universidade Federal do Vale do São Francisco

JUAZEIRO-BA
Julho, 2022

## AGRADECIMENTOS

À minha mãe Miriam e ao meu pai Luiz, que são a base incondicional da minha formação pessoal e acadêmica, que sempre acreditam e me ajudam a alcançar meus objetivos, por mais faustosos que possam ser, e estão sempre ao meu lado. Agradeço ao meu avô Tavares †, que sempre desejou ver seus netos formados e sempre me encorajou a seguir meus sonhos. Agradecimentos especiais à minha segunda mãe Gigi, que está ao meu lado há 23 anos com todo o amor e carinho possíveis, e a minha madrinha Lúcia e padrinho Marcelo †, que sempre me apoiaram.

Ao meu orientador Prof. Dr. João José, por ter me acolhido em seu laboratório e me introduzir ao mundo da bioinformática, bem como todas as correções e ensinamentos que me permitiram apresentar este trabalho. Assim como todo o pessoal do Laboratório de Genética e Biotecnologia da UNIVASF, que me auxiliou e com que pude desenvolver trabalhos nesses 3 últimos anos.

A Rita Nassur, Aurizângela de Sousa, Pricila de Jesus e Adailson Feitosa, que além de professores, foram mentores e amigos que acreditaram em mim e me permitiram ver e expandir meu potencial e me aperfeiçoar profissional e pessoalmente.

A todos o os meus amigos, sejam os antigos, cujos carrego desde o colégio, até os novos que UNEB me deu, em especial Maria Clara, Sophia e Bianca, os quais espero igualmente levá-los ao meu lado, assim como todos os meus colegas de turma com os quais dividi esses últimos anos.

Às pessoas com quem convivi ao longo desses anos de curso, que me incentivaram e que certamente tiveram impacto na minha formação acadêmica.

## RESUMO

Grande parte dos microrganismos não são cultiváveis, sendo a identificação de táxons em amostras complexas um desafio. A metagenômica permite analisar a comunidade microbiológica diretamente pelo material genético. Com isso, classificadores taxonômicos foram desenvolvidos para processar os dados do sequenciamento, utilizando diversos algorítimos e métodos. O trabalho objetivou a comparação de quatro classificadores, pela utilização do dataset simulado CAMI Toy Mouse Gut, sendo avaliados facilidade do uso, uso de RAM, tempo de execução, tamanho da base de dados, precisão e especificidade da classificação. Kaiju teve configuração e MetaCache uso mais fáceis. Centrifuge utilizou mais RAM e tempo na construção e MetaCache na classificação, enquanto Kraken2 teve os menores resultados gerais. Centrifuge teve o maior tamanho máximo e menor mínimo da base de dados, enquanto MetaCache teve o menor máximo e maior mínimo. Quanto a assertividade da classificação, Kraken2 e MetaCache se mostraram semelhantes entre si e com melhor precisão, Kaiju foi o mais específico em níveis a partir do gênero, porém especialmente inferior em espécie. Kraken2 foi mais rápido, com menor uso geral de RAM e com a segunda menor base de dados, tornando-o a principal recomendação.



Palavras-chave: Microbioma; Bioinformática; Classificadores

## ABSTRACT

Most microorganisms are not culturable, making taxon identification in complex samples a challenge. Metagenomics allows analyzing the microbial community directly from the genetic material. Thus, taxonomic classifiers have been developed to process the sequencing data, using various algorithms and methods. The aim was to compare four classifiers, through the use of the simulated CAMI Toy Mouse Gut dataset, and it was evaluated the ease of use, RAM usage, execution time, database size, classification accuracy and specificity. Kaiju had the easiest setup and MetaCache the usage. Centrifuge used the most RAM and time in building and MetaCache in classification, while Kraken2 was the smallest overall. Centrifuge had the largest maximum and smallest minimum database size, while MetaCache had the smallest maximum and largest minimum. As for classification assertiveness, Kraken2 and MetaCache were similar to each other with better accuracy, Kaiju was the most specific in levels from genus, but especially inferior in species. All in all, Kraken2 was faster, with lower overall RAM usage and the second smallest database, making it the top recommendation.



Keywords: Microiome; Bioinformatics; Classifiers

**SUMÁRIO**

# 1 INTRODUÇÃO

Uma vez que grande parte dos microrganismos ainda não são cultiváveis (STEEN *et al.*, 2019), um dos desafios mais antigos da microbiologia é a identificação de táxons em amostras biológicas complexas, como solo e rúmen. Com o advento do sequenciamento de nova geração e o surgimento da metagenômica, essa tarefa se tornou mais fácil, rápida e precisa, à medida que a composição da comunidade é verificada diretamente (BELLA *et al.*, 2013).

A metagenômica pode ser realizada através de diferentes metodologias, de acordo com seu alvo de estudo. O sequenciamento com a finalidade de demonstrar a relação evolutiva entre os microrganismos é tido como metataxonômica e a técnica de sequenciamento completo, conhecida também por *"shotgun"*, como metagenômica propriamente dita (MARCHESI; RAVEL, 2015). Ainda, a metagenômica é capaz de recuperar mais informação de táxons em baixa abundância e em maior resolução (DURAZZI *et al.*, 2021).

Em ambos os casos, há uma grande produção de dados genômicos de diversos microrganismos, acarretando necessidade de criação de métodos computacionais capazes de classificar e identificar as espécies presentes. Tais ferramentas são conhecidas como classificadores taxonômicos e diferem entre si pelos seus algoritmos, bases de dados e molécula alvo, produzindo, assim, resultados distintos. Alguns desses classificadores são: Centrifuge (KIM *et al.*, 2016), Kaiju (MENZEL; NG; KROGH, 2016) , Kraken2 (WOOD; LU; LANGMEAD, 2019; WOOD; SALZBERG, 2014) e MetaCache (MÜLLER *et al.*, 2017). Estes utilizam de métodos de anotação taxonômica direto das *reads* baseados na similaridade destas com sequências genômicas do banco de dados. Alguns dos desafios dessa técnica incluem a necessidade de realizar a comparação entre milhões de sequências, o que pode ter alto custo computacional; as bases de dados de referência devem ser mais completas possível para minimizar erros e não é possível a descoberta de novos táxons; além de que as *reads* podem não ter tamanho suficiente para produzir atribuições precisas (CARR; BORENSTEIN, 2014).

Centrifuge e MetaCache utilizam genomas completos do NCBI Reference Sequence Database (RefSeq) (O'LEARY *et al.*, 2016) como referência para construção do seu banco de dados. Kraken2 é flexível aceitando sequências de proteína ou DNA, genomas completos ou sequências específicas, incluindo todos os grupos mencionados anteriormente, além de bancos de proteínas e vetores, adaptadores e sequências de primers. Kaiju, utiliza somente proteínas, primariamente de genomas completos de RefSeq de Archaea, Bacteria e Viruses, porém aceita outros como fungos e outros bancos de proteínas. Além disso, Kaiju é o único que possui um servidor Web, não necessitando de instalação local, mas com a desvantagem de não

utilizar bancos de dados atualizados. Quanto ao algoritmo de busca das sequências, os classificadores dependem de alinhamentos das reads completas ou de k-mers. Centrifuge busca por alinhamentos exatos das reads (sem gaps ou mismatches) e Kaiju traduz a read para os seis quadros de leitura possíveis e busca pela correspondência no banco de dados proteíco, ambos utilizando a transformada de Burrows-Wheeler (BURROWS; WHEELER, 1994) e index de Ferragina-Manzini (FERRAGINA; MANZINI, 2000). Kraken2 e MetaCache utilizam o algoritmo de busca de k-mers, no qual o tamanho padrão para Kraken2 é de k=35 e MetaCache k=16.

Peabody *et al.* (2015) concluíram que não há objetivamente ferramenta melhor ou pior e que sua escolha depende das necessidades do estudo em questão, bem como condições específicas de infraestrutura disponível. Ye *et al.* (2019) corroboram apontando que um ponto crítico na velocidade de classificação de alguns programas é a velocidade de leitura do disco, que afeta diretamente a transferência da base de dados para memória.

Dessa forma, o presente estudo buscou a análise comparativa dos cinco classificadores citados, tanto em parâmetros quantitativos como qualitativos, para uso no Centro de Acesso Livre para Análises Genômicas (CALAnGO) em futuros estudos.

# 2 METODOLOGIA

## 2.1 HARDWARE

Todos os programas foram rodados no servidor do Centro de Acesso Livre para Análises Genômicas (CALAnGO) do laboratório de Genética e Biotecnologia do Campus Ciências Agrárias (CCA) da Universidade Federal do Vale do São Francisco em Petrolina (PE). O servidor é do tipo Dell R44 com 2 processadores Intel Xeon Gold 6126 de 2,6 GHz (12 núcleos / 24 segmentos), 512 Gb RAM (32GB RDIMM, 2666MT/s, Dual Rank, BCC) e 8 HDs (1.8TB 10K RPM SAS 12Gbps 512e 2.5in Hot-plug Hard Drive).

## 2.2 DATASET

### 2.2.1 Simulado

Dez amostras (0 a 9) do dataset simulado Toy Mouse Gut provenientes do Critical Assessment of Metagenome Interpretation (CAMI) (<https://repository.publisso.de/resource/frl:6421672>) foram utilizadas. O CAMI (SCZYRBA *et al.*, 2017) é uma iniciativa criada com o intuíto avaliar métodos em metagenômica de forma isonômica, sem viéses e padronizada. Devido aos datasets não estarem descritos a nível de

espécie, os taxonomy ids corretos de cada read foram buscados utilizando os códigos accession presentes no arquivo ”readings_mapping.tsv.gz”.

### 2.2.2 Real

Foram utilizadas 12 amostras de metagenoma ruminal caprino provenientes de um estudo anterior de Silva de Sant’ana *et al.* (2022). As reads brutas do sequenciador foram submetidas à remoção dos adaptadores e sequencias de baixa qualidade utilizando o Trimmomatic v0.39 (BOLGER; LOHSE; USADEL, 2014) no modo paired end (ILLUMINACLIP:NexteraPE-PE:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:10:20 MINLEN:70). Em seguida, a biblioteca foi filtrada e as sequências do hospedeiro retiradas pelo programa Bowtie2 v2.4.1 (LANGMEAD; SALZBERG, 2012) utilizando o genoma de referência para Capra hircus (ARS1 GCF_001704415.1).

## 2.3 PROGRAMAS E PARÂMETROS

Foram empregadas ferramentas as quais os métodos de análise utilizam as sequências em fastq como entrada e possuem classificação por read. Dessa forma, os programas Centrifuge, Kaiju, Kraken2 e MetaCache foram escolhidos. De modo a padronizras análises, o mesmo conjunto de genomas foi utilizado na construção das bases de dados, sendo considerados apenas genomas completos de arqueias e bactérias. Os genomas de fungos, protozoários e outros eucariotos não foram considerados devido à relativa míngua quantidade disponível.

Todos os programas foram instalados em ambientes distintos através do sistema de gerenciamento de pacotes Conda (CONDA, 2021) e executados com 5 núcleos, o mínimo de configuração manual necessária e para o nível taxonômico máximo de espécie. Os comandos executados estão disponíveis e descritos em detalhe no repositório do GitHub.

A Tabela 1 fornece as características das ferramentas escolhidas para análise. O número de citações que cada uma recebeu é uma forma de avaliar a sua popularidade e influência. Ainda, alguns métodos foram lançados antes da publicação de seus artigos, portanto, são mais citados, enquanto novos métodos supostamente possuem abordagens mais modernas e melhoradas. Além disso, os métodos podem possuir desempenhos diferentes em diferentes situações e tipos de amostra.

Todos os programas possuem base de referência passível de customização, utilizam arquivos fastq pareados como entrada e nenhum possui formas de visualização dos resultados nativa.

Tabela 1 – Características das ferramentas analisadas.

| | **Centrifuge** | **Kaiju** | **Kraken2** | **MetaCache** |
|---|---|---|---|---|
| **Licença** | GNU GPL v3 | GNU GPL v3 | MIT License | GNU GPL v3 |
| **Implementação** | C++ | C/C++ | C++ e Perl | C++ |
| **Método** | Alinhamento exato da read | Alinhamento de proteínas das reads | Alinhamento exato de k-mer | Alinhamento de k-mer |
| **Versão** | 1.0.4 | 1.7.4 | 2.1.2 | 2.0.0 |
| **Método** | Alinhamento exato da read | Alinhamento de proteínas das reads | Alinhamento exato de k-mer | Alinhamento de k-mer |
| **Versão** | 1.0.4 | 1.7.4 | 2.1.2 | 2.0.0 |
| **Base de referência principal** | NCBI RefSeq | | | |
| **GZipped** | NÃO | SIM | SIM | NÃO |
| **Multiplos arquivos** | SIM | SIM | NÃO | SIM |
| **Citações (Google Scholar)** | 634 | 791 | 772 | 29 |

## 2.4 MÉTRICAS

Para comparação das diferentes ferramentas, o seguintes itens foram avaliados:

- Facilidade de instalação e uso;
- Uso máximo de RAM durante construção de database e classificação;
- Tempo de construção de database e classificação;
- Tamanho máximo e mínimo necessário da database no disco;
- Precisão e especificidade da classificação.
- Quantidade e concordância de táxons detectados.

Com exceção do último item, todos foram realizados apenas com o dataset simulado. A facilidade de uso dos programas foi avaliado pela quantidade de comandos necessários para execução. O uso de RAM e tempo de execução dos programas foi medido através do software GNU Time ”/usr/bin/time -v”e o tamanho das bases de dados através do comando ”du -h”.

Para avaliar o desempenho dos programas, foi utilizada uma abordagem similar a Lindgreen, Adair e Gardner (2016) e Cárdenas, Neuenschwander e Malaspinas (2022), na qual cada read foi analisada a nível taxonômico de espécie, gênero, família e ordem

e categorizada como Corretamente classificada (RC), Incorretamente classificada (RI) ou Não classificada (RN). Dessa forma, a especificidade foi dada pela divisão do número de reads Corretamente classificadas (RC) pelo total de reads na amostra (RC + RI + RN) (Equação 2.1) e a precisão pela divisão do número de reads corretamente classificadas (RC) pelo total de reads classificadas (RC + RI) (Equação 2.2). A precisão e especificidade para os programas foram então apresentadas como a média das precisões e especificidades de cada amostra.

$$Especificidade = \frac{R_C}{R_C + R_I + R_N} \tag{2.1}$$

$$Precisão = \frac{R_C}{R_C + R_I} \tag{2.2}$$

A quantidade na detecção de táxons foi avaliada considerando a contagem de táxons individuais em todas as amostras e a concordância considerando os táxons em comum entre as ferramentas.

Os dados foram processados em Python 3.9 e os gráficos construídos em R 4.1.1 utilizando scripts próprios.

## 2.5 DISPONIBILIDADE DOS DADOS

Todos os comandos utilizados na execução dos programas e scripts construídos para as análises e visualização estão disponíveis em <https://github.com/bielasilva/metagenomic-tools-benchmark>.

# 3 RESULTADOS E DISCUSSÃO

Dos programas analisados (Figura 1), destaca-se o Kaiju como o mais simples de todos de ser configurado, necessitando de um comando único para criação da database. Em contrapartida, Centrifuge foi o de maior complexidade, necessitando de quatro comandos: dois para download dos dados necessários, um para agregar os genomas em arquivo único e por fim mais um para construção de fato da database. A execução não foi problemática e as instruções são bem claras, entretanto, comparado aos demais, nota-se sua complexidade.

Quanto aos comandos para classificação, destaca-se MetaCache por possuir apenas uma linha de comando. Kraken2 não permite o processamento de múltiplas amostras nativamente, fazendo necessário o uso de um loop para o processamento sequencial das 10 amostras. O classificador Kaiju define que os caminhos para as reads sejam separados por vírgula, assim, para evitar a construção de um comando

gigante com todos os caminhos, um script simples teve de ser construído. Centrifuge possui duas maneiras de adicionar múltiplas amostras para processamento sequencial. A primeira é similar à do Kaiju e a segunda é através de um arquivo TSV com os caminhos para as reads e os arquivos que serão salvos com o resultado das análises. Essa última foi a escolhida por ter maior escalabilidade e apesar de necessitar da criação de um arquivo extra, é a melhor opção para se ter maior controle das amostras e principalmente caso seus nomes não sejam padronizados.

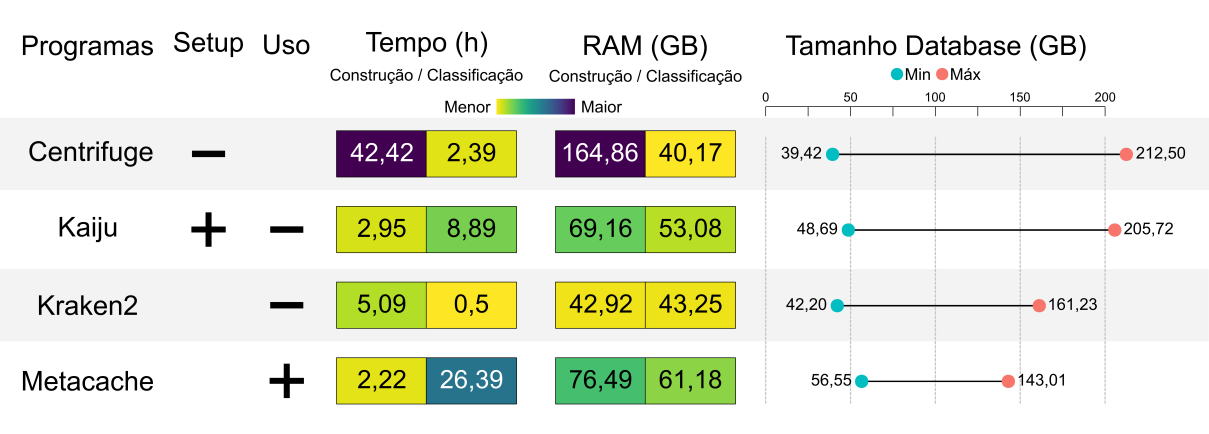



Figura 1 – Facilidade de setup e uso, tempo de execução, uso de RAM e tamanho da database

Quanto ao tempo de execução, Centrifuge foi consideravelmente o mais demorado, com 42,42 h apenas na construção da database, muito superior ao menor, MetaCache com 2,22 h, mas também ao segundo maior, Kraken2, com 5,09 h. Já no tempo para classificação, destaca-se o MetaCache com 26,39 h, também bem superior a Kaiju com 8,89 h. O mais rápido foi kraken2 com apenas 0,5 h para o processamento do dataset. Curiosamente, Centrifuge foi o programa que obteve os dois extremos de uso de RAM, com o maior na construção (164,86 GB) e o menor na classificação (40,17 GB). Kaiju e MetaCache obtiveram valores intermediários abaixo de 80 GB e Kraken2 teve a menor discrepância entre as duas etapas, se mantendo abaixo de 44 GB.

A respeito do tamanho ocupado pelas bases de dados, Centrifuge novamente se mostrou em extremos, tendo pico de 212,50 GB e após limpeza dos arquivos dispensáveis restou apenas 39,42 GB. Esse valor máximo provavelmente se dá pelo fato de que é necessário concatenar todas as sequências baixadas em um único arquivo para construção da base, o que, na prática, duplica toda a biblioteca de genomas. Já a maior base mínima pertence a MetaCache (56,55 GB), assim como a menor máxima (143,01 GB). É importante notar que os bancos de sequências dos quais os programas constroem suas bases estão em constante expansão, sendo razoável esperar que no futuro o tamanho ocupado em disco aumente, assim como ocorra aumento no número de reads classificadas em metagenomas reais, uma vez que novos táxons estão sendo

incluídos, como demonstrado por Nasko *et al.* (2018). Além disso, a inclusão de grupos como vírus, fungos, ou outros disponíveis, certamente impacta nesses valores.

De maneira geral, Kraken2 e MetaCache obtiveram as menores médias na especificidade e as maiores na precisão (Figura 2). Isso significa que os programas conseguiram classificar menos reads no total, diminuindo sua especificidade, porém sua taxa de acerto foi maior, o que eleva a precisão dado que se desconsidera reads não classificadas. É importante destacar que ambos os programas utilizam a mesma abordagem de classificação utilizando k-mers

Centrifuge mostrou-se o melhor na especificidade aos níveis de espécie (46,73%) e gênero (53,69%), sendo este, entretanto, bastante similar ao Kaiju (53,44%). Interessantemente, Kaiju possui especial menor desempenho na classificação de espécies, obtendo distintivamente menor média de precisão (37,17%) e especificidade (29,77%) que os outros programas. Essa diferença tende a diminuir à medida que níveis taxonômicos mais altos são considerados, chegando às maiores médias de especificidade para família (62,26%) e ordem (70,87%). Kraken2 foi o programa que melhor obteve resultados de precisão em todos os níveis taxonômicos analisados com médias de 67,28%, 81,75%, 87,63%, 94,53% para espécie, gênero, família e ordem, respectivamente.

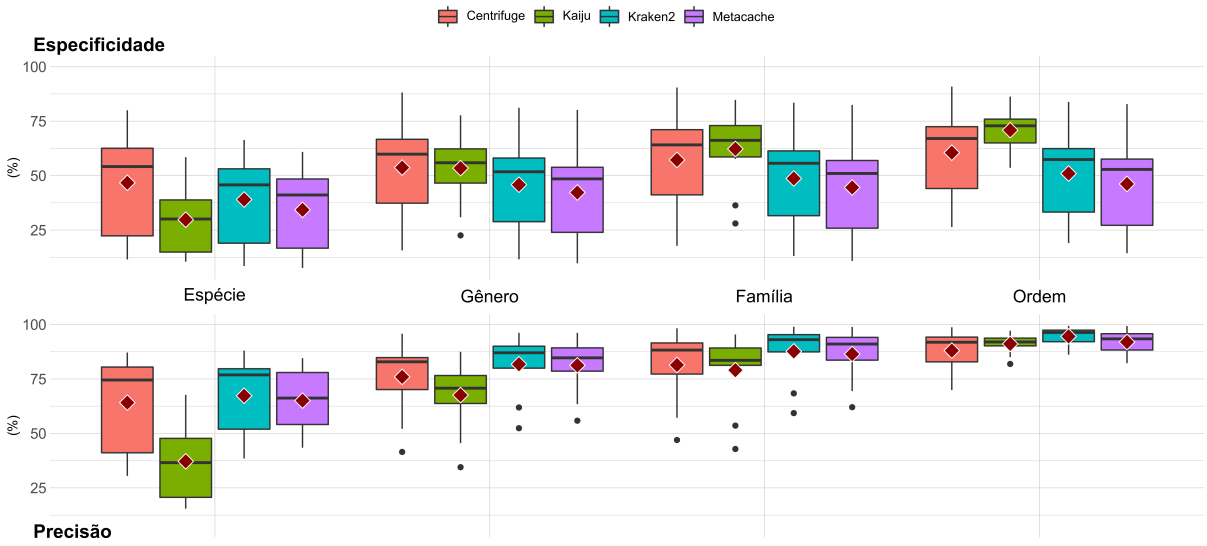



Figura 2 – Boxplots Precisão e Especificidade

A baixa precisão do Kaiju em espécie e gênero talvez possa ser explicada por seu método de classificação ser baseado em aminoácidos e não DNA, como nos outros, uma vez que a probabilidade de substituição de um aminoácido é menor que a de um nucleotídeo e o acumulo de substituições nucleotídicas necessárias para causar mudanças significativas em uma sequência polipeptídica pode levar também à mudança na classificação taxonômica.

Os Boxplots ainda revelam que Centrifuge teve, em geral, maior amplitude. Isso sinaliza que a classificação dele pode ser inconsistente a depender da amostra. Kraken2 e Kaiju, por outro lado, possuem menor variação nos resultados, o que indica melhor uniformidade do método na análise das diferentes amostras.

A Figura 3 demonstra que o desempenho dos programas tende a cair em amostras com maior complexidade. A especificidade foi em especial prejudicada à medida que mais reads permanecem sem elucidação quanto à sua taxonomia. Esse problema é reduzido em níveis taxonômicos mais altos, podendo talvez ser desprezível acima de ordem, fato que deve ser considerado à medida que a escolha de qual nível taxonômico utilizar é bastante particular ao design e objetivo do estudo em questão.

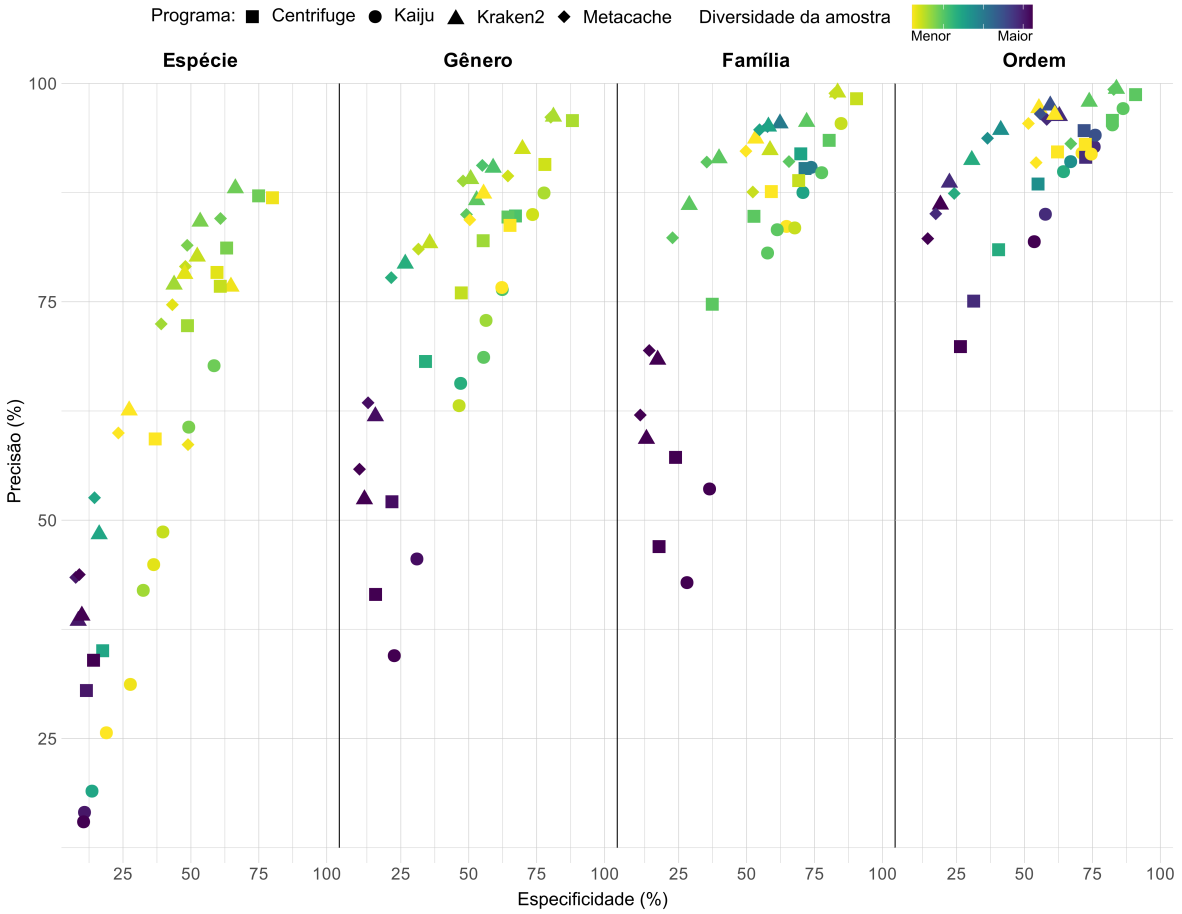



Figura 3 – Quantidade de espécies por amostra com Precisão e Especificidade

Ainda, Centrifuge, Kaiju e Kraken2 apresentaram quantidades de táxons detectados maiores que MetaCache (Figura 4A). A nível de espécie, os três classificaram 6445, 6272 e 6176, respectivamente, e Metacache apenas 3690. Apesar disso, os dados simulados tinham apenas 367, sinalizando a quantidade de falsos positivos dos programas.

Isso pode ser verificado também com a concordância entre os programas

(Figura 4B). Os quatro formam o grupo com maior intersecção, ainda que Centrifuge, Kaiju e Kraken2 formem um segundo distinto com os táxons não classificados por MetaCache. Além disso, é possível verificar os táxons exclusivos do dataset e, portanto, não classificados por nenhuma ferramenta. Esses táxons não elucidados podem ser explicados pelo fato de que o dataset do CAMI II inclui genomas não disponíveis nas bancos públicos e, portanto, não presentes nas bases usadas para classificação.

Assim, conforme o nível taxonômico aumenta, os programas tendem a convergir para maior concordância e com menos táxons sem elucidação. Já a nível de família, apenas 3 não foram classificados por nenhuma ferramenta e nenhum a nível de ordem.

Esses dados demonstram que as quatro ferramentas conseguem detectar os mesmos táxons verdadeiros, porém acusam falsos positivos e corroboram com a avaliação do baixo desempenho em níveis de espécie e gênero, especialmente em amostras mais complexas. Isso auxilia a explicar as classificações erradas, as quais reduziram a precisão das ferramentas, vistas nas Figuras 2 e 3.

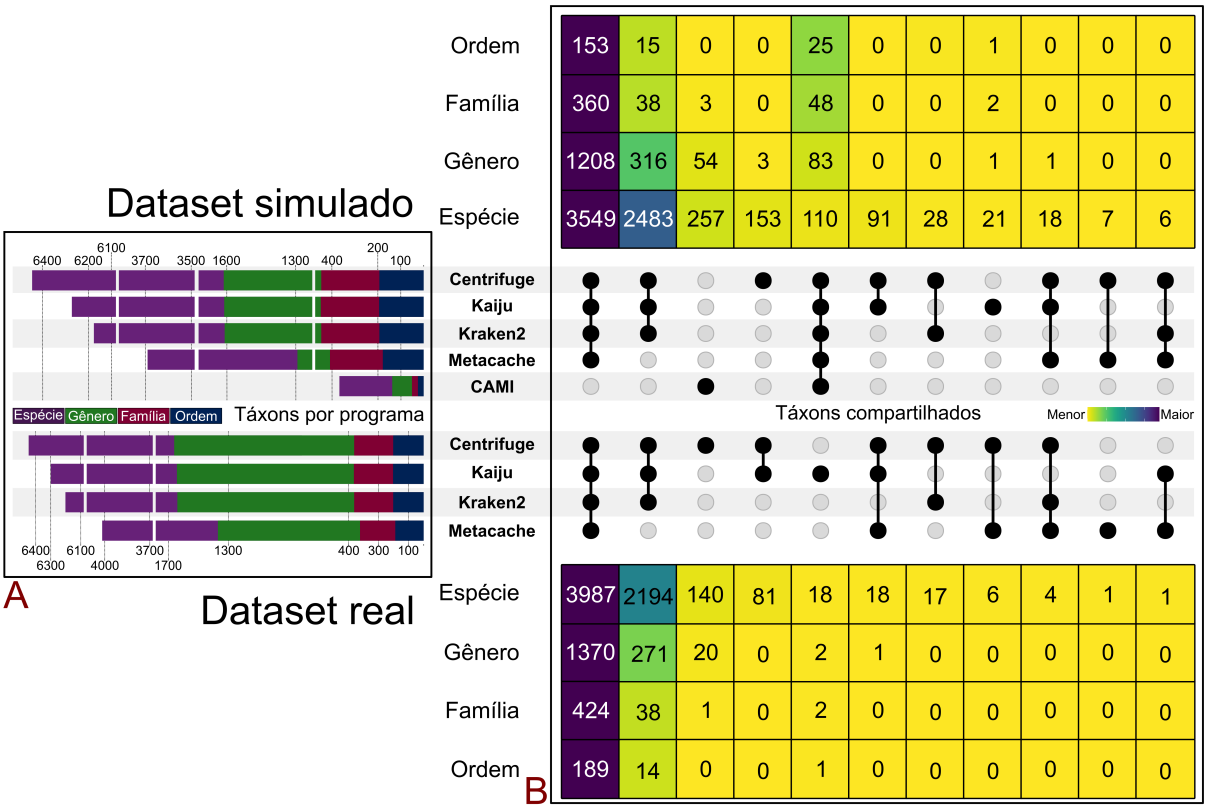



Figura 4 – Quantidade (A) e Concordância (B) táxons detectados pelos programas

De forma geral, o perfil dos dados reais assemelha-se ao dataset simulado, entretanto não há como verificar a real composição da microbiota. O que se pode dizer, é que este tenha sofrido com problemas similares quanto à presença de falsos positivos. Como os programas foram executados de maneira simples, um recurso a ser utilizado

é aumentar a exigência dos programas para que resultados menos confiáveis sejam excluídos e o usuário aplicar um threshold no número de reads presentes em um táxon para que ele possa ser considerado positivo. Essa estratégia pode diminuir os falsos positivos com o tradeoff de detectar menos táxons raros.

## 4 CONCLUSÃO

Nota-se que a assertividade dos programas tende a cair com a maior complexidade da amostra, sendo o desempenho dos programas melhor em níveis taxonômicos mais altos como família e ordem. Os resultados de Kraken2 e MetaCache se mostraram os mais semelhantes entre si, e Kaiju o mais variável entre os níveis taxonômicos. A nível de espécie, Centrifuge teve a melhor especificidade e Kraken2 a precisão. Quando considerados os níveis mais altos Kaiju se mostrou mais específico, o que indica classificação de maior número de reads e apresentou menor amplitude entre os valores máximos e mínimos encontrados, sugerindo melhor uniformidade na análise das diferentes amostras. Em termos de precisão, Kraken2 continuou sendo o melhor em todos os níveis.

Todos os programas detectaram os mesmos táxons verdadeiros e acusaram diversos falsos positivos, a maior parte compartilhados entre si. MetaCache, porém, classificou consideravelmente menos falsos positivos.

De forma geral, Kraken2 ainda que tenha a menor facilidade de uso, foi o mais rápido, com menor uso geral de RAM e com a segunda menor base de dados, além de ter a melhor precisão em todos os níveis taxonômicos, sendo, portanto, a principal recomendação desse estudo, especialmente para sistema menos robustos e com limitação de hardware. MetaCache segue em seguida como boa opção, porém peca especialmente no tempo de classificação das amostras, um tradeoff por ter o menor número de falsos positivos.

No estabelecimento de protocolos específicos, uma alternativa para aumentar o desempenho e reduzir o número de falsos positivos é restringir os parâmetros de classificação, ainda que provavelmente haja perdas no que se refere a táxons raros.

## REFERÊNCIAS

BELLA, J. M. D.; BAO, Y.; GLOOR, G. B.; BURTON, J. P.; REID, G. High throughput sequencing methods and analysis for microbiome research. **Journal of Microbiological Methods**, v. 95, n. 3, p. 401–414, dez. 2013.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: A flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, ago. 2014.

BURROWS, M.; WHEELER, D. J. A block-sorting lossless data compression algorithm. In: . [S.l.: s.n.], 1994.

CÁRDENAS, Y. O. A.; NEUENSCHWANDER, S.; MALASPINAS, A.-S. Benchmarking metagenomics classifiers on ancient viral DNA: A simulation study. **PeerJ**, v. 10, p. e12784, mar. 2022.

CARR, R.; BORENSTEIN, E. Comparative Analysis of Functional Metagenomic Annotation and the Mappability of Short Reads. **PLoS ONE**, v. 9, n. 8, p. e105776, ago. 2014.

CONDA. 2021. Anaconda Software Distribution.

DURAZZI, F.; SALA, C.; CASTELLANI, G.; MANFREDA, G.; REMONDINI, D.; CESARE, A. D. Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. **Scientific Reports**, v. 11, n. 1, p. 3030, dez. 2021.

FERRAGINA, P.; MANZINI, G. Opportunistic data structures with applications. In: **Proceedings 41st Annual Symposium on Foundations of Computer Science**. [S.l.: s.n.], 2000. p. 390–398.

KIM, D.; SONG, L.; BREITWIESER, F. P.; SALZBERG, S. L. Centrifuge: Rapid and sensitive classification of metagenomic sequences. **Genome Research**, v. 26, n. 12, p. 1721–1729, dez. 2016.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, v. 9, n. 4, p. 357–359, abr. 2012.

LINDGREEN, S.; ADAIR, K. L.; GARDNER, P. P. An evaluation of the accuracy and speed of metagenome analysis tools. **Scientific Reports**, v. 6, n. 1, p. 19233, maio 2016.

MARCHESI, J. R.; RAVEL, J. The vocabulary of microbiome research: A proposal. **Microbiome**, v. 3, n. 1, p. 31, s40168–015–0094–5, dez. 2015.

MENZEL, P.; NG, K. L.; KROGH, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. **Nature Communications**, v. 7, n. 1, p. 11257, set. 2016.

MÜLLER, A.; HUNDT, C.; HILDEBRANDT, A.; HANKELN, T.; SCHMIDT, B. MetaCache: Context-aware classification of metagenomic reads using minhashing. **Bioinformatics**, v. 33, n. 23, p. 3740–3748, dez. 2017.

NASKO, D. J.; KOREN, S.; PHILLIPPY, A. M.; TREANGEN, T. J. RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. **Genome Biology**, v. 19, n. 1, p. 165, dez. 2018.

O'LEARY, N. A.; WRIGHT, M. W.; BRISTER, J. R.; CIUFO, S.; HADDAD, D.; MCVEIGH, R.; RAJPUT, B.; ROBBERTSE, B.; SMITH-WHITE, B.; AKO-ADJEI, D.; ASTASHYN, A.; BADRETDIN, A.; BAO, Y.; BLINKOVA, O.; BROVER, V.; CHETVERNIN, V.; CHOI, J.; COX, E.; ERMOLAEVA, O.; FARRELL, C. M.; GOLDFARB, T.; GUPTA, T.; HAFT, D.; HATCHER, E.; HLAVINA, W.; JOARDAR, V. S.; KODALI, V. K.; LI, W.; MAGLOTT, D.; MASTERSON, P.; MCGARVEY, K. M.; MURPHY, M. R.; O'NEILL, K.; PUJAR, S.; RANGWALA, S. H.; RAUSCH, D.; RIDDICK, L. D.; SCHOCH, C.; SHKEDA, A.; STORZ, S. S.; SUN, H.; THIBAUD-NISSEN, F.; TOLSTOY, I.; TULLY, R. E.; VATSAN, A. R.; WALLIN, C.; WEBB, D.; WU, W.; LANDRUM, M. J.; KIMCHI, A.; TATUSOVA, T.; DICUCCIO, M.; KITTS, P.; MURPHY, T. D.; PRUITT, K. D. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. **Nucleic Acids Research**, v. 44, n. D1, p. D733–D745, jan. 2016.

PEABODY, M. A.; ROSSUM, T. V.; LO, R.; BRINKMAN, F. S. L. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. **BMC Bioinformatics**, v. 16, n. 1, p. 362, dez. 2015.

SCZYRBA, A.; HOFMANN, P.; BELMANN, P.; KOSLICKI, D.; JANSSEN, S.; DRÖGE, J.; GREGOR, I.; MAJDA, S.; FIEDLER, J.; DAHMS, E.; BREMGES, A.; FRITZ, A.; Garrido-Oter, R.; JØRGENSEN, T. S.; SHAPIRO, N.; BLOOD, P. D.; GUREVICH, A.; BAI, Y.; TURAEV, D.; DEMAERE, M. Z.; CHIKHI, R.; NAGARAJAN, N.; QUINCE, C.; MEYER, F.; BALVOČIŪTĖ, M.; HANSEN, L. H.; SØRENSEN, S. J.; CHIA, B. K. H.; DENIS, B.; FROULA, J. L.; WANG, Z.; EGAN, R.; KANG, D. D.; COOK, J. J.; DELTEL, C.; BECKSTETTE, M.; LEMAITRE, C.; PETERLONGO, P.; RIZK, G.; LAVENIER, D.; WU, Y.-W.; SINGER, S. W.; JAIN, C.; STROUS, M.; KLINGENBERG, H.; MEINICKE, P.; BARTON, M. D.; LINGNER, T.; LIN, H.-H.; LIAO, Y.-C.; SILVA, G. G. Z.; CUEVAS, D. A.; EDWARDS, R. A.; SAHA, S.; PIRO, V. C.; RENARD, B. Y.; POP, M.; KLENK, H.-P.; GÖKER, M.; KYRPIDES, N. C.; WOYKE, T.; VORHOLT, J. A.; Schulze-Lefert, P.; RUBIN, E. M.; DARLING, A. E.; RATTEI, T.; MCHARDY, A. C. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. **Nature Methods**, v. 14, n. 11, p. 1063–1071, nov. 2017.

SILVA DE SANT'ANA, A.; SILVA, A. P. R.; DO NASCIMENTO, S. P. O.; MORAES, A. A.; NOGUEIRA, J. F.; BEZERRA, F. C. M.; COSTA, C. F. da; GOUVEIA, J. J. d. S.; GOUVEIA, G. V.; RODRIGUES, R. T. d. S.; BONFA, H. C.; MENEZES, D. R. Tannin as a modulator of rumen microbial profile, apparent digestibility and ingestive behavior of lactating goats: A preliminary metagenomic view of goats adaptability to tannin. **Research in Veterinary Science**, v. 145, p. 159–168, jul. 2022.

STEEN, A. D.; CRITS-CHRISTOPH, A.; CARINI, P.; DEANGELIS, K. M.; FIERER, N.; LLOYD, K. G.; THRASH, J. C. High proportions of bacteria and archaea across most biomes remain uncultured. **The ISME Journal**, v. 13, n. 12, p. 3126–3130, dez. 2019.

WOOD, D. E.; LU, J.; LANGMEAD, B. Improved metagenomic analysis with Kraken 2. **Genome Biology**, v. 20, n. 1, p. 257, dez. 2019.

WOOD, D. E.; SALZBERG, S. L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. **Genome Biology**, v. 15, n. 3, p. R46, 2014.

YE, S. H.; SIDDLE, K. J.; PARK, D. J.; SABETI, P. C. Benchmarking Metagenomics Tools for Taxonomic Classification. **Cell**, v. 178, n. 4, p. 779–794, ago. 2019.