

Scientific Electronic Archives

Issue ID: Sci. Elec. Arch. Vol. 17 (1)

January/February 2024

DOI: <http://dx.doi.org/10.36560/17120241872>

Article link: <https://sea.ufr.edu.br/SEA/article/view/1872>



O uso de ferramentas de bioinformática para análise de dados genéticos: uma revisão

The use of bioinformatics tools for genetic data analysis: a review

Corresponding author

Ruana Carolina Cabral da Silva

Universidade Federal de Ouro Preto

ruanacarolina08@gmail.com

Maria Cidinaria Silva Alves

Universidade Federal de Pernambuco

Resumo. A bioinformática vem revolucionando o modo como os cientistas analisam e interpretam os dados genéticos. Dessa forma, esta revisão destaca o papel fundamental das ferramentas de bioinformática na compreensão dos dados genômicos. O artigo explora a diversidade de softwares e algoritmos disponíveis para processar, analisar e interpretar dados genéticos. Aborda-se a relevância dessas ferramentas na identificação de genes, variações genéticas, predição de estruturas proteicas e estudos de evolução e filogenia. Além disso, são apresentados os desafios enfrentados na bioinformática, incluindo a integração de dados de diferentes fontes, padronização e interpretação dos resultados. É disponibilizado no artigo informações sobre alinhamento de sequências, limpeza de dados de sequenciamento, que são dados importantes quando se trabalha com conjunto de dados genéticos. Destaca-se ainda que discussões como estas são importantes, pois as ferramentas de bioinformática estão constantemente evoluindo, o que requer atualização constante de conhecimento e habilidades por parte dos pesquisadores.

Palavras-chave: Ferramentas genômicas, softwares e algoritmos, dados biológicos.

Abstract. Bioinformatics has been revolutionizing how scientists analyze and interpret genetic data. Thus, this review highlights the fundamental role of bioinformatics tools in understanding genomic data. The article explores the diversity of software and algorithms available for processing, analyzing, and interpreting genetic data. It addresses the relevance of these tools in identifying genes, genetic variations, predicting protein structures, and evolutionary and phylogenetic studies. Additionally, the challenges faced in bioinformatics, including integrating data from different sources, standardization, and interpreting results, are discussed. The article provides information on sequence alignment, sequencing data cleaning, which are crucial when working with genetic datasets. It is further emphasized that discussions like these are important because bioinformatics tools are constantly evolving, requiring researchers to continuously update their knowledge and skills.

Keywords: Genomic tools, software and algorithms, biological data

Contextualização e análise

A bioinformática é um campo interdisciplinar que combina biologia e ciência da computação para analisar e interpretar informações genéticas, utilizando dados ômicos (Iqbal; Kumar, 2023). Destaca-se que o genoma é o conjunto completo de genes e material genético de um organismo, e a bioinformática genômica se concentra em estudar essas informações em larga escala (Abdelkrim, 2023). Este campo de pesquisa desempenha um papel importante nos estudos de saúde, agricultura e outros, incluindo a medicina personalizada.

Atualmente, a partir da combinação da genética com a bioinformática tem-se diagnósticos específicos para diversas doenças, possibilitando previsões mais precisas sobre o desenvolvimento da doença. Estes avanços se devem em grande parte pela capacidade de examinar o conjunto completo de genes de um indivíduo em diferentes detalhes, desde mudanças nos cromossomos até modificações em níveis individuais de DNA (Pereira; Oliveira; Sousa, 2020).

Tem-se o uso de bioinformática para alinhar, comparar e anotar sequências de DNA, bem como

identificar genes, regiões regulatórias e variações genéticas. Estas ferramentas de bioinformática também trabalham com sequências de aminoácidos, as quais podem analisar estruturas proteicas, identificação de domínios funcionais e previsão de estruturas secundárias (Kanzi *et al.*, 2020).

A expressão gênica a partir de técnicas como microarranjos e RNA-Seq, podem ser analisados para identificar genes diferencialmente expressos ou padrões de expressão, por exemplo, entender como determinado gene é modulado de acordo com uma situação específica (Trapnell *et al.*, 2013; Rao *et al.*, 2019). Outro segmento é o estudo de metilação do DNA, sendo uma modificação epigenética importante que pode afetar a expressão gênica, nestas análises as ferramentas de bioinformática são usadas para analisar padrões de metilação do DNA em genomas e entender como essas modificações podem estar relacionadas a doenças e processos biológicos (Bock, 2012).

A anotação funcional de genes/proteínas também vem sendo muito estudado, a partir das ferramentas de bioinformática é possível analisar

sobre as funções conhecidas ou previstas de genes e proteínas, bem como dados sobre vias metabólicas, interações proteína-proteína e redes biológicas (Mitra *et al.*, 2013). Tem-se pesquisas relacionadas à filogenia, buscando entender as relações evolutivas entre organismos, envolvendo a análise de sequências genéticas para construir árvores filogenéticas e entender a evolução de grupos gênicos ou organismos (Liu *et al.*, 2023).

Em contextos médicos, as ferramentas de bioinformática, ou pipelines, podem ser usados para analisar dados clínicos de pacientes, como informações genéticas, para personalizar tratamentos e entender melhor as bases genéticas de doenças hereditárias (Aronson; Rehm, 2015).

Destaca-se que o campo é altamente diversificado e utiliza abordagens computacionais para interpretar e extrair informações dos dados biológicos. Dessa forma, conhecer ferramentas que permitem a realização de estudos que busquem analisar dados genéticos pode favorecer não só pesquisas em saúde, mas em diferentes áreas do conhecimento (Figura 1).

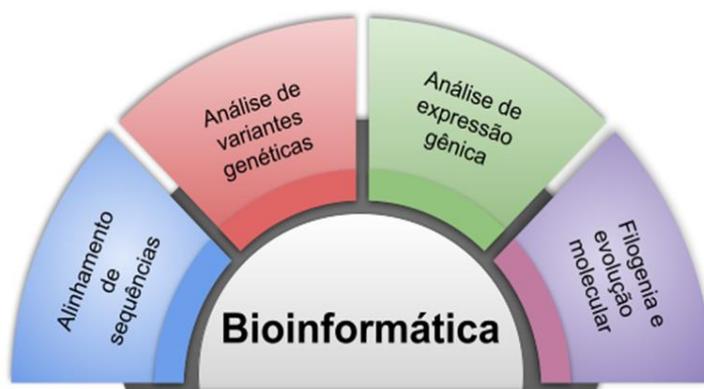


Figura 1. Aplicações de ferramentas de bioinformática. Fonte: Autoria própria.

Pré-processamento de dados e ferramentas de alinhamento

Quando se trabalha análise de dados, o pré-processamento é uma etapa fundamental, pois auxilia a preparar os dados brutos para análise, tornando-os mais limpos, organizados e adequados para a aplicação de técnicas. Em geral, tem-se algumas etapas que consistem em métodos de aprendizado de máquina para implementar limpeza e pré-processamento de dados, redução e transformação de dados (Figura 2).

busca de erros, valores ausentes, valores discrepantes (outliers) e outros problemas de qualidade. Os dados são corrigidos ou removidos conforme necessário.

A integração de dados acontece quando estes dados vêm de múltiplas fontes, podendo ser integrados em um único conjunto de dados. Isso pode envolver a combinação de tabelas, a resolução de discrepâncias de esquema e a criação de um

conjunto de dados coeso. Já na seleção de dados, ocorre porque nem todas as características (atributos) podem ser relevantes para o problema em questão, ou seja, a seleção de características envolve a escolha das características mais relevantes e informativas para a análise e modelagem subsequentes.

Na etapa de limpeza de dados, tem-se a transformação de dados pode incluir normalização para padronizar as escalas das características, discretizações de variáveis contínuas, e aplicação de técnicas matemáticas para melhorar a qualidade dos dados. Em conjuntos de dados grandes, a redução de dados pode ser necessária para diminuir a complexidade e o tempo de processamento. Técnicas como Análise de Componentes Principais (PCA) podem ser usadas para reduzir a dimensionalidade.

A discretização de dados, ocorre em alguns casos, sendo útil para transformar variáveis contínuas em variáveis discretas, o que pode simplificar a análise e modelagem. A etapa de engenharia de dados envolve a criação de novos atributos a partir dos existentes, com o objetivo de capturar informações importantes ou criar características mais significativas para a tarefa de aprendizado de máquina (Lan et al.2018).

Destaca-se que o alinhamento de sequências genéticas é uma etapa fundamental na

análise de dados genômicos, pois permite comparar e identificar semelhanças e diferenças entre sequências de DNA ou proteínas. Existem algumas ferramentas de alinhamentos de sequências muito usuais em pesquisas de bioinformática, seja de alinhamento local ou múltiplo, entre elas tem-se o *Basic Local Alignment Search Tool* (BLAST), *Multiple Alignment using Fast Fourier Transform* (MAFFT), *Sequence Alignment/Map tools* (SAMtools) e *Burrows-Wheeler Aligner* (BWA) (Figura 3).

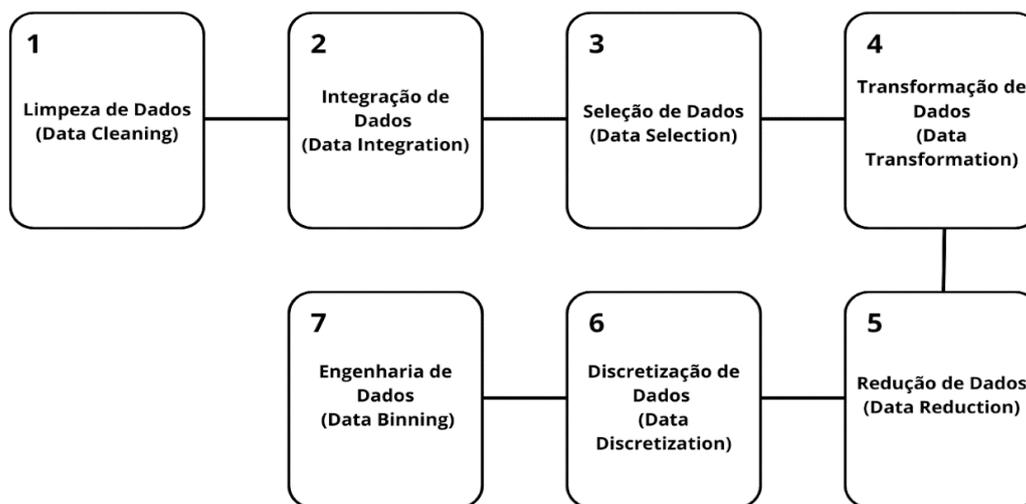


Figura 2. Etapas de pré-processamento, redução e transformação de dados.. Fonte: Autoria própria.

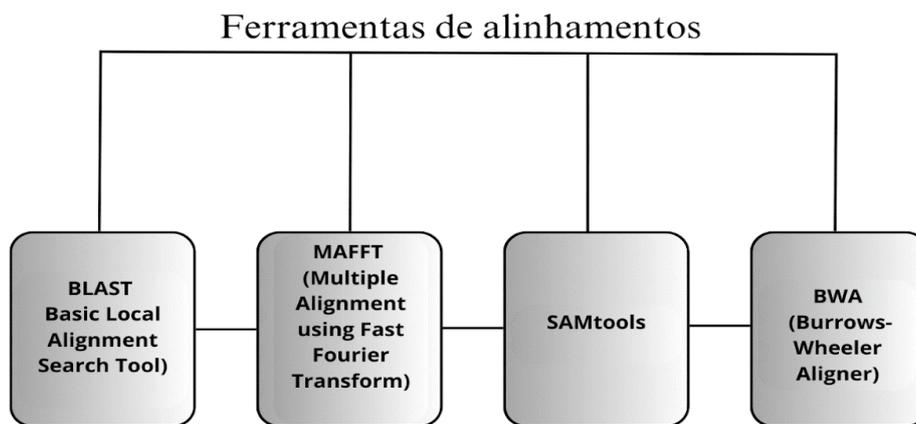


Figura 3. Ferramentas de alinhamentos mais comuns em análises de sequências genéticas. Fonte: Autoria própria.

O BLAST é um método de alinhamento local que permite a identificação de regiões homólogas em sequências biológicas, auxiliando na anotação de genes, na busca por similaridades funcionais e na inferência de evolução molecular, podendo ter uma ampla gama de aplicações, desde a pesquisa em genômica até a identificação de patógenos em amostras clínicas (Pearson, 2014).

O MAFFT é muito utilizado para alinhar múltiplas sequências biológicas. Com sua eficiência computacional e precisão, permite analisar comparativamente sequências e identificar regiões conservadas ou variáveis. Sua aplicação é essencial em estudos de filogenia, estrutura de proteínas e análise funcional de genes, desempenhando um papel essencial na compreensão da biologia molecular (Kato; Asimeno; Toh, 2009).

O SAMtools é um conjunto de programas amplamente utilizado para processar, manipular e analisar dados de Sequenciamento de Nova Geração (NGS). Essa ferramenta desempenha um papel fundamental na extração de informações valiosas a partir de dados brutos de NGS, possibilitando a identificação de variantes genéticas, quantificação de expressão gênica e a análise de cobertura de sequenciamento. O SAMtools simplifica o gerenciamento de grandes volumes de dados genômicos (Spencer et al., 2014).

BWA é uma ferramenta de alinhamento de sequências genômicas altamente eficiente e amplamente utilizada, possuindo capacidade de mapear sequências de DNA rapidamente em um genoma de referência, utilizado em estudos de associação genética, identificação de mutações e análise de populações, a velocidade de análise e precisão o tornam uma escolha popular entre os pesquisadores que buscam alinhamento de sequências de alta qualidade (Jo; Koh, 2015).

Análise de variantes genéticas

A análise de variantes pode ser integrada para chamadas tanto dos Polimorfismos de Nucleotídeo Único (SNPs) quanto para INDELS (Inserções e Deleções) (Lam et al., 2012). As ferramentas de detecção identificam as posições onde as sequências diferem entre no conjunto de dados comparados. Algoritmos como GATK (Genome Analysis Toolkit), SAMtools e FreeBayes são frequentemente usados para essa finalidade (Yao et al., 2020).

Após a detecção de variantes genéticas, torna-se imperativo conduzir análises complementares e implementar procedimentos de filtragem rigorosa, visando a eliminação de variantes falsas positivas e a seleção das variantes de maior relevância. Estes processos de seleção podem envolver critérios diversificados, abrangendo

aspectos relacionados à qualidade dos dados, frequência dos alelos, bem como os potenciais efeitos funcionais das variantes, entre outros parâmetros relevantes (Belkadi et al., 2015). Para conduzir tais análises e filtragens, diversas ferramentas computacionais estão disponíveis, com destaque para softwares como VCFtools, ANNOVAR e VariantAnnotation, que desempenham um papel crucial na identificação e avaliação dessas variações genéticas.

Tem-se também a avaliação do impacto biológico das variantes que frequentemente inclui análises funcionais, tais como a previsão dos efeitos das variantes em proteínas ou em regiões regulatórias, proporcionando uma compreensão mais profunda de seu potencial efeito biológico (Pabinger et al., 2014). Além disso, uma análise de associação pode ser conduzida para identificar variantes que possam estar correlacionadas com características clínicas ou fenotípicas específicas, abrindo caminho para a compreensão de suas implicações em contextos médicos e biológicos mais amplos.

Os pipelines de chamadas de variantes podem ser aplicados para diferentes finalidades, a depender do contexto estudado, mas em geral seguem algumas etapas padrão: (1) Geração dos dados, que em geração são os FASTQs oriundos do sequenciamento; (2) Alinhamento, essa etapa irá depender do método escolhido, se alinhadas sequências com genomas de referência ou outros métodos; (3) Chamada de variantes, para essa etapa são utilizadas as ferramentas que realizam essas chamadas; (4) Anotação das variantes, tem-se aqui um filtro de acordo com o interesse do estudo; (5) Seleção das variantes, nesta etapa são selecionadas aquelas de interesse de acordo com o objetivo proposto na pesquisa de interesse ou análise de clínica (Figura 4).

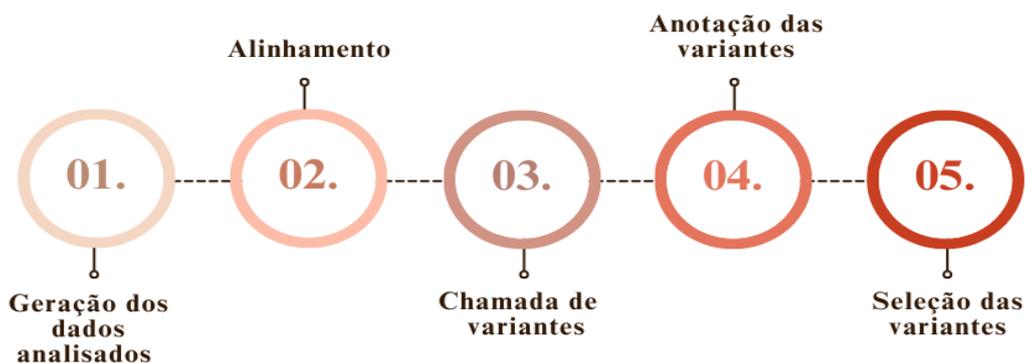


Figura 4. Exemplo de pipeline utilizado para realização de chamada de variantes. Fonte: Autoria própria

Dessa forma, a chamada de variantes na genômica auxilia na compreensão de doenças genéticas, evolução populacional e outras aplicações biológicas. Através da análise de dados NGS são geradas informações que ao aplicados métodos de bioinformática ajudam na identificação das

alterações no genoma, fornecendo informações relevantes para a pesquisa em biologia, medicina e genética populacional.

Análise de expressão gênica

Para identificar genes diferencialmente expressos entre diferentes grupos de amostras são utilizadas técnicas estatísticas, estes dados podem ser analisados através de diferentes métodos, alguns dos mais populares incluem DESeq2 e edgeR para análise de RNA-seq (Varet *et al.*, 2016; Liu *et al.*, 2021) e SAM (*Significance Analysis of Microarrays*) para microarranjos (Larsson; Wahlestedt; Timmons, 2005).

A visualização de dados oriundos de expressão gênica pode ser analisada através de ferramentas como RStudio e bibliotecas R (ggplot2, pheatmap) são frequentemente usadas para criar gráficos e mapas de calor que representam os dados de maneira informativa. Além disso, são usualmente realizadas análises de enriquecimento de vias através do Kyoto Encyclopedia of Genes and Genomes (KEGG) (<https://www.genome.jp/kegg/>) ou Gene Ontology (GO) (<https://geneontology.org/>) (White; Adams; Bu, 2000; Kanehisa, 2002), que podem auxiliar a entender as funções biológicas das redes de regulação. Estas análises permitem identificar genes centrais nas redes que são altamente regulados ou que desempenham um papel crítico em processos biológicos específicos (Thompson; Regev; Roy, 2015).

A análise da expressão gênica é um procedimento essencial na busca por uma compreensão aprofundada do papel de grupos de genes envolvidos na resposta de plantas à defesa contra ameaças. Nesse contexto, diferentes pesquisas têm sido conduzidas, empregando técnicas de sequenciamento de RNA em diversas espécies vegetais, a fim de desvendar os intrincados padrões de expressão gênica e prospectar ou caracterizar conjuntos de dados relevantes (Cervantes-Pérez *et al.*, 2022).

Um exemplo de investigação na área da expressão gênica é o uso da técnica de RNA-Seq para desvendar como os genes são regulados em distintos contextos. Nessa abordagem, amostras experimentais são confrontadas com amostras de controle (grupo controle vs. grupo tratamento), permitindo a identificação de vias de sinalização, genes cuja expressão é induzida ou constitutiva, bem como diversas outras análises relevantes. Essas abordagens são empregadas em pesquisas de aprimoramento genético, aplicações biotecnológicas ou na saúde.

Filogenia e Evolução Molecular

A reconstrução da história evolutiva de um organismo utilizando a filogenética é um método fundamental para representar evidências sobre as relações evolutivas entre ou dentro dos organismos e para estudar a evolução e a inovação funcional dos genes (Lam *et al.*, 2010; Lee *et al.*, 2014).

A filogenia é frequentemente baseada em dados genéticos, como sequências de DNA ou proteínas. A ideia subjacente é que as espécies que compartilham sequências genéticas mais semelhantes têm um ancestral comum mais recente. Na filogenia e evolução molecular, a análise de SNPs

têm se mostrado a forma dominante de mudança evolutiva, tornando-se uma estratégia indispensável para análises filogenéticas (Lee *et al.*, 2014; Shakya *et al.*, 2020).

Para construção de árvores filogenéticas existem diversas ferramentas disponíveis. A PhyML é uma ferramenta de máxima verossimilhança usada para construir árvores filogenéticas a partir de sequências moleculares (Guindon *et al.*, 2009). Para análises baseadas em modelos bayesianos o BEAST (Bayesian Evolutionary Analysis by Sampling Trees) destaca-se estimando datas de divergência (Bouckaert *et al.*, 2014). Análises de genes e famílias gênicas são realizadas através do PhyloBayes (Lartillot, 2020). O OrthoMCL é usado para identificar grupos de genes ortólogos e parálogos a partir de várias sequências genômicas (Li; Stoeckert; Roos, 2003).

Na evolução molecular o objetivo é analisar as mudanças nas sequências de DNA e proteínas ao longo do tempo, fornecendo informações sobre como as moléculas biológicas evoluem e como as mutações genéticas levam à diversidade entre as espécies. Para Análise de Evolução Molecular, o software PAML (Phylogenetic Analysis by Maximum Likelihood) é utilizado para identificar genes sob seleção positiva ou negativa (Yang, 2007), enquanto a ferramenta HyPhy utiliza o modelo bayesiano para detecção de sinais de seleção e outros processos evolutivos (Pond; Frost; Muse, 2005).

Contudo, a filogenia e a evolução molecular são áreas interligadas que desempenham um papel crucial na descoberta das relações evolutivas e na análise das mudanças genéticas ao longo do tempo, fornecendo insights valiosos para a biologia evolutiva, a taxonomia e a compreensão da diversidade da vida.

Considerações finais

Ao longo da revisão, observou-se que o uso de ferramentas de bioinformática na análise de dados genéticos é fundamental para avanços significativos na pesquisa genômica. Entretanto, alguns pontos precisam ser analisados, entre os principais desafios está relacionado à complexidade dos dados genéticos, que podem ser vastos e heterogêneos. Isso demanda o desenvolvimento de algoritmos e métodos computacionais sofisticados para o processamento e interpretação dessas informações. Além disso, a qualidade dos dados é importante, e erros na coleta ou sequenciamento podem afetar a precisão das análises. Dessa forma, as ferramentas de bioinformática estão constantemente evoluindo, o que requer atualização constante de conhecimento e habilidades por parte dos pesquisadores.

Nesse expansivo crescimento e progresso da bioinformática têm inaugurado cada vez mais oportunidades para desenvolvimento de diferentes pesquisas que se baseiam em análise de dados, descoberta de conhecimento e aplicações inovadoras. Em geral, as tendências emergentes na bioinformática genômica apontam para uma empolgante evolução no campo. O crescente uso de

NGS está gerando enormes volumes de dados genômicos, tornando a análise de dados de larga escala mais acessível. O desenvolvimento de abordagens de aprendizado de máquina e inteligência artificial está revolucionando a interpretação de dados genômicos, permitindo previsões mais precisas e a identificação de associações complexas. No futuro, podemos esperar uma maior automação, aprimoramento da interoperabilidade entre diferentes ferramentas e uma ênfase crescente na análise de dados em tempo real.

Referências

ABDELKRIM, R. Bioinformatics: An Exciting Field of Science-Importance and Applications. *Journal of Concepts in Structural Biology & Bioinformatics (JSBB)*, v. 1, n. 4, 2023.

ARONSON, S. J.; REHM, H. L. Building the foundation for genomics in precision medicine. *Nature*, v. 526, n. 7573, p. 336-342, 2015.

BELKADI, A. *et al.* Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences*, v. 112, n. 17, p. 5473-5478, 2015.

BOCK, C. Analysing and interpreting DNA methylation data. *Nature Reviews Genetics*, v. 13, n. 10, p. 705-719, 2012.

BOUCKAERT, R. *et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLOS Computational Biology*, v. 10, n. 4, p. e1003537, 2014.

CERVANTES-PÉREZ, S. A. *et al.* Challenges and perspectives in applying single nuclei RNA-seq technology in plant biology. *Plant Science*, v. 325, p. 111486, 2022.

GUINDON, S. *et al.* Estimating Maximum Likelihood Phylogenies with PhyML. Em: POSADA, D. (Ed.). *Bioinformatics for DNA Sequence Analysis. Methods in Molecular Biology*. Totowa, NJ: Humana Press, p. 113-137, 2009.

IQBAL, N.; KUMAR, P. From Data Science to Bioscience: Emerging era of bioinformatics applications, tools and challenges. *Procedia Computer Science*, v. 218, p. 1516-1528, 2023.

JO, H.; KOH, G. Faster single-end alignment generation utilizing multi-thread for BWA. *Bio-medical materials and engineering*, v. 26, n. s1, p. S1791-S1796, 2015.

KANEHISA, M. The KEGG database. In: 'In silico'simulation of biological processes: Novartis Foundation Symposium 247. Chichester, UK: John Wiley & Sons, Ltd, 2002. p. 91-103.

KANZI, A. M. *et al.* Next generation sequencing and bioinformatics analysis of family genetic inheritance. *Frontiers in Genetics*, v. 11, p. 544162, 2020.

KATOH, K.; ASIMENOS, G.; TOH, H. Multiple alignment of DNA sequences with MAFFT. *Bioinformatics for DNA sequence analysis*, p. 39-64, 2009.

LAM, H. Y. K. *et al.* Detecting and annotating genetic variations using the HugeSeq pipeline. *Nature biotechnology*, v. 30, n. 3, p. 226-229, 2012.

LAM, H-M. *et al.* Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet* 42, 1053-1059, 2010.

LAN, K. *et al.* A survey of data mining and deep learning in bioinformatics. *Journal of medical systems*, v. 42, p. 1-20, 2018.

LARSSON, O.; WAHLESTEDT, C.; TIMMONS, J. A. Considerations when using the significance analysis of microarrays (SAM) algorithm. *BMC bioinformatics*, v. 6, n. 1, p. 1-6, 2005.

LARTILLOT, N. PhyloBayes: Bayesian Phylogenetics Using Site-heterogeneous Models. IN: SCORNAVACCA, C.; DELSUC, F.; GALTIER, N. (Eds.). *Phylogenetics in the Genomic Era*. p. 1.5:1-1.5:16, 2020.

LEE, T.-H. *et al.* SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics* 15, 162, 2014.

LI, L.; STOECKERT, C. J.; ROOS, D. S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, v. 13, n. 9, p. 2178-2189, 2003.

LIU, C. *et al.* The community coevolution model with application to the study of evolutionary relationships between genes based on phylogenetic profiles. *Systematic Biology*, v. 72, n. 3, p. 559-574, 2023.

LIU, S. *et al.* Three differential expression analysis methods for RNA sequencing: limma, EdgeR, DESeq2. *JoVE (Journal of Visualized Experiments)*, n. 175, p. e62528, 2021.

MITRA, K. *et al.* Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, v. 14, n. 10, p. 719-732, 2013.

PABINGER, S. *et al.* A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*, v. 15, n. 2, p. 256-278, 2014.

PEARSON, W. R. BLAST and FASTA similarity searching for multiple sequence alignment. Multiple sequence alignment methods, p. 75-101, 2014.

PEREIRA, R.; OLIVEIRA, J.; SOUSA, M. Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. Journal of clinical medicine, v. 9, n. 1, p. 132, 2020.

POND, S. L. K.; FROST, S. D. W.; MUSE, S. V. HyPhy: hypothesis testing using phylogenies. Bioinformatics, v. 21, n. 5, p. 676–679, 2005.

RAO, M. S. *et al.* Comparison of RNA-Seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat toxicity studies. Frontiers in genetics, v. 9, p. 636, 2019.

SHAKYA, M. *et al.* Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life. Sci Rep 10, 1723, 2020.

SPENCER, D. H. *et al.* Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data. The Journal of Molecular Diagnostics, v. 16, n. 1, p. 75-88, 2014.

THOMPSON, D.; REGEV, A.; ROY, S. Comparative analysis of gene regulatory networks: from network reconstruction to evolution. Annual review of cell and developmental biology, v. 31, p. 399-428, 2015.

TRAPNELL, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature biotechnology, v. 31, n. 1, p. 46-53, 2013.

VARET, H. *et al.* SARTools: a DESeq2-and EdgeR-based R pipeline for comprehensive differential analysis of RNA-Seq data. PloS one, v. 11, n. 6, p. e0157022, 2016.

WHITE, M. H.; ADAMS, D. A.; BU, J. On the go with SONOS. IEEE Circuits and Devices Magazine, v. 16, n. 4, p. 22-31, 2000.

YANG, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution, v. 24, n. 8, p. 1586–1591, 2007.

YAO, Z. *et al.* Evaluation of variant calling tools for large plant genome re-sequencing. BMC bioinformatics, v. 21, n. 1, p. 1-16, 2020.