UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
INSTITUTO DE BIOCIÊNCIAS - RIO CLARO

CIÊNCIAS BIOLÓGICAS

NICOLLE NAVES DE ARAUJO SANTOS

MINERAÇÃO DE DADOS EXPLORATÓRIA COM PYTHON: ANÁLISE DE PLATAFORMAS DE SEQUENCIAMENTO E FERRAMENTAS DE BIOINFORMÁTICA PARA PRÉ-PROCESSAMENTO DE DADOS TRANSCRIPTÔMICOS

Rio Claro - SP
2025

NICOLLE NAVES DE ARAUJO SANTOS

MINERAÇÃO DE DADOS EXPLORATÓRIA COM PYTHON: ANÁLISE DE PLATAFORMAS DE SEQUENCIAMENTO E FERRAMENTAS DE BIOINFORMÁTICA PARA PRÉ-PROCESSAMENTO DE DADOS TRANSCRIPTÔMICOS

Trabalho de Conclusão de Curso apresentado ao Instituto de Biociências – Câmpus de Rio Claro, da Universidade Estadual Paulista “Júlio de Mesquita Filho”, para obtenção do grau de Bacharela em Ciências Biológicas.

Orientador(a): Dra. Milene Ferro

Rio Claro - SP
2025

S237m

Santos, Nicolle Naves de Araujo

Mineração de dados exploratória com Python: análise de plataformas de sequenciamento e ferramentas de bioinformática para pré-processamento de dados transcriptômicos / Nicolle Naves de Araujo Santos. -- Rio Claro, 2025

38 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Ciências Biológicas) - Universidade Estadual Paulista (UNESP), Instituto de Biociências, Rio Claro

Orientador: Milene Ferro

1. Mineração de Dados. 2. Sequenciamento. 3. Pré-processamento. 4. Transcriptômica. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

NICOLLE NAVES DE ARAUJO SANTOS

MINERAÇÃO DE DADOS EXPLORATÓRIA COM PYTHON: ANÁLISE DE PLATAFORMAS DE SEQUENCIAMENTO E FERRAMENTAS DE BIOINFORMÁTICA PARA PRÉ-PROCESSAMENTO DE DADOS TRANSCRIPTÔMICOS

Trabalho de Conclusão de Curso apresentado ao Instituto de Biociências, Câmpus de Rio Claro, da Universidade Estadual Paulista "Júlio de Mesquita Filho", para obtenção do grau de Bacharela em Ciências Biológicas.

BANCA EXAMINADORA:

Profa. Dra. Milene Ferro

Dra. Marcela Paula Silva Zanão

Dr. Pepijn Wilhelmus Kooij

Aprovado em: 18 de Julho de 2025

Assinatura do discente

gov.br
Documento assinado digitalmente
NICOLLE NAVES DE ARAUJO SANTOS
Data: 21/07/2025 21:38:21-0300
Verifique em https://validar.iti.gov.br

Assinatura do(a) orientador(a)

gov.br
Documento assinado digitalmente
MILENE FERRO
Data: 25/07/2025 09:07:57-0300
Verifique em https://validar.iti.gov.br

AGRADECIMENTOS

Agradeço a minha base, à minha mãe, Joanadarque, e ao meu pai, Altamiro, por todo sacrifício, amor silencioso e por serem o exemplo de força que me inspira todos os dias. O que sou, e tudo que conquisto, começou com vocês. Vocês são a origem de tudo.

Às minhas irmãs Jaqueline, Evelyn e Thais, minhas eternas companheiras de jornada. Obrigada por estarem tão presentes mesmo com a distância física, me apoiando em todos os momentos. Por vibrarem com minhas vitórias como se fossem suas e por me oferecerem uma rede de apoio sempre que precisei. A jornada é muito mais leve e feliz porque tenho vocês.

Aos meus queridos sobrinhos, Guilherme, Gustavo, Elisa e Eva. Vocês nem imaginam o quanto estar com vocês, é como um combustível, que me lembra como a vida pode ser mais leve e divertida.

Agradeço à minha orientadora Milene, por ter me apresentado à bioinformática e, com isso, aberto um novo horizonte de possibilidades em minha jornada acadêmica e profissional. Sua orientação precisa, paciência infinita e dedicação, foram fundamentais para a realização deste trabalho.

Ao Lucca, uma das pessoas que tive o privilégio de conhecer. O pouco tempo que compartilhamos foi suficiente para me transformar para sempre. Levo sua memória e seu sorriso comigo.

À minha família e companheiras da República Mandala: Laiso, Ana, Ilaraê, Natt e Rafa. Obrigada por serem meu lar em Rio Claro e por dividirem comigo o caos e a delícia que foram esses anos. Todo companheirismo, amizade, risadas, rolês e tudo o que vivemos. Quero ter vocês sempre por perto.

Aos meus amigos que encontrei ao longo dessa jornada, Perdido, Prof, Joyce, Lu, João, Rudá e Isabelle vocês foram/são essenciais e como um presente em minha vida.

Agradeço à à Universidade Estadual Paulista "Júlio de Mesquita Filho" câmpus Rio Claro, por me proporcionar a formação em uma universidade de excelência, ensino de qualidade e gratuito. Ao corpo docente, funcionários, servidores e todos aqueles que colaboram para a qualidade da instituição. Aos programas de auxílio permanência que foram essenciais para a minha vivência em Rio Claro.

RESUMO

O crescente aumento no volume de dados gerados pelos sequenciamentos demonstra a importância e a evolução ao longo dos anos de ferramentas de Bioinformática e plataformas de sequenciamento. A transcriptômica é capaz de identificar transcritos (genes), principalmente RNAs que codificam proteínas, para um organismo e estes estudos estão envolvidos com a análise de expressão gênica em condições distintas. A primeira etapa a ser realizada na análise de transcriptomas é o pré-processamento das *reads* brutas, onde é feito o controle de qualidade e limpeza de dados e que incluem filtros e cortes de sequências para posterior montagem de transcritos visando obter os genes completos ou parciais. Considerando a importância desses estudos, foi realizado o levantamento descritivo observacional com abordagem qualitativa para montar o banco de dados, com as plataformas de sequenciamento de larga escala e ferramentas de pré-processamento mais utilizadas baseado nos artigos científicos publicados nos últimos 5 anos disponíveis no PubMed/NCBI. E também uma mineração de dados exploratória para categorizar e identificar os dados. Os resultados obtidos mostraram que a ferramentas de pré-processamento mais utilizadas é a Trimmomatic, com predominância em estudos humanos e em animais. E Cutadapt para estudos humanos. E para as plataformas de sequenciamento, a Illumina foi a predominante, com seus modelos HiSeq 2500 e NextSeq 500.

Palavras-chave: Pré-processamento; Sequenciamento; Transcriptômica; Mineração de Dados.

ABSTRACT

The growing increase in the volume of data generated by sequencing demonstrates the importance and evolution of Bioinformatics tools and sequencing platforms over the years. Transcriptomics is capable of identifying transcripts (genes), mainly protein-coding RNAs, for an organism, and these studies are involved with the analysis of gene expression under different conditions. The first step to be performed in the analysis of transcriptomes is the preprocessing of raw reads, where quality control and data cleaning are carried out, which include filters and trimming of sequences for subsequent transcript assembly aiming to obtain complete or partial genes. Considering the importance of these studies, an observational descriptive survey with a qualitative approach was carried out to assemble the database, with the most used large-scale sequencing platforms and preprocessing tools based on scientific articles published in the last 5 years available on PubMed/NCBI. And an exploratory data mining to categorize and identify the data. The results obtained showed that the most used preprocessing tool is Trimmomatic, with a predominance in human and animal studies. And Cutadapt for human studies. And for sequencing platforms, Illumina was the predominant one, with its HiSeq 2500 and NextSeq 500 models.

Keywords: Preprocessing; Sequencing; Transcriptomics; Data Mining.

SUMÁRIO

1 INTRODUÇÃO

1.1 Plataformas de sequenciamento

O sequenciamento é uma das etapas iniciais para a identificação de bases nitrogenadas das sequências nucleotídicas de DNA e RNA. Para descrição da estrutura tridimensional do DNA, o trabalho de Watson e Crick (1953) foi fundamental para compreender o processo de replicação e a codificação de proteínas e ácidos nucleicos (Christoff, 2017).

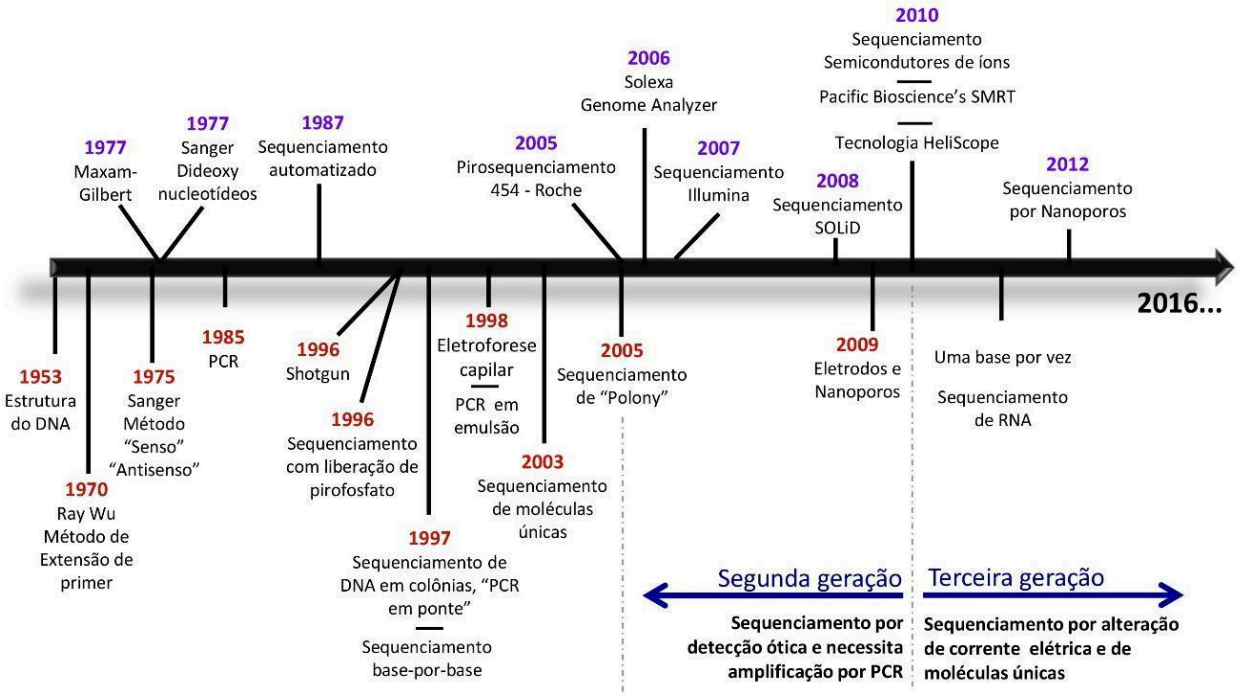
Visando identificar regiões de interesse e compreender a informação genética presente em um organismo. Em 1965, o pesquisador Robert Holley e seus colaboradores, realizaram a primeira montagem completa de uma sequência de ácido nucleico, aprimorando os estudos sobre genética molecular (Christoff, 2017). Entretanto, havia limitações para que o sequenciamento do DNA fosse possível devido as suas moléculas serem muito longas e geradas em unidades com similaridade elevada, o que tornou ainda mais difícil esse processo (Hutchison, 2007).

Entre os anos de 1970 e 1980, as metodologias foram progressivamente aprimoradas para desenvolver o processo de determinação da ordem dos elementos que constituem o DNA (Christoff, 2017). Para a determinação, o método de degradação química utiliza produtos químicos que realizam a clivagem do DNA, agindo nas purinas ou pirimidinas específicas. Para isso, o DNA é marcado com o isótopo 32P, após tais alterações das bases, ocorre a clivagem a estrutura de açúcar-fosfato, resultando nos fragmentos que serão analisados por eletroforese (Maxam-Gilbert, 1977). Essa técnica permitiu que diversos pesquisadores adaptassem seus métodos para o sequenciamento de DNA. Em 1977, Sanger propôs o método de sequenciamento baseado na utilização da determinação de cadeia por dideoxinucleotídeos (ddNTPs). Esse método levou a avanços nos estudos de biologia molecular, automações, evolução das tecnologias e inovação nos protocolos de sequenciamento, impulsionando o surgimento das ciências ômicas. Desde então, novas frentes de estudo têm surgido, ampliando o alcance das pesquisas atuais (Heather, 2016).

Em meados de 2005, três grandes plataformas de sequenciamento de nova geração surgiram no mercado: o pirosequenciamento com detecção de pirofosfato; o sequenciamento por ligação da tecnologia Illumina e Solexa (Figura 1). Como algumas dessas tecnologias apresentaram problemas na geração de sequências para algumas ômicas, acabaram não sendo mais utilizadas nas pesquisas. Das três tecnologias de sequenciamento lançadas na época, a Illumina se manteve no mercado e hoje a maioria dos projetos transcriptômicos são realizados usando essa plataforma. Mais recentemente surgiram plataformas de sequenciamento como o Pacific Biosciences e Oxford Nanopore (Chistoff, 2017) que, diferentemente da Illumina, geram fragmentos longos de sequências.

O desenvolvimento de tecnologias de sequenciamento, permitiu o avanço de novas ciências ômicas e análises importantes de sistemas biológicos em larga escala (Wingender et al., 2007).

Figura 1 - Evolução das tecnologias de sequenciamento ao longo do tempo



Fonte: CHRISTOFF, 2017.

O sequenciamento em plataformas de alto rendimento foi um marco importante para a geração de dados moleculares de forma mais barata, rápida, acessível e em larga escala. As atuais tecnologias para a identificação de ácidos nucleicos (DNA e RNA), oferecem a vantagem de gerar informação explícita a partir de uma determinada região genômica, levando a milhares de dados gerados e adequados para estudos científicos atuais (Turchetto et al., 2017).

Na década de 1990, foi desenvolvida a técnica de pirosequenciamento, em que o sequenciamento ocorre através da detecção de liberação do pirofosfato (PPi) durante a síntese do DNA. Este método permite a transformação da reação em luz, na qual a sua intensidade refere-se à quantidade de nucleotídeos incorporados. A adição dessas bases determina a sequência através de sinais luminosos, sendo um método que não utiliza eletroforese em gel (Ronaghi et al., 1998).

Este método foi incorporado pela plataforma Roche 454 life sciences, na qual foi muito utilizada para o início de sequenciamento de genomas, pois os dados gerados são elevados se comparados ao método de Sanger. Porém, são leituras mais curtas e com menos precisão de leitura individual. Demonstrando assim que é possível montar genomas a partir de leituras curtas (Margulis et al., 2006). Após esse período, diversas técnicas foram desenvolvidas, mas a tecnologia Illumina obteve muito destaque.

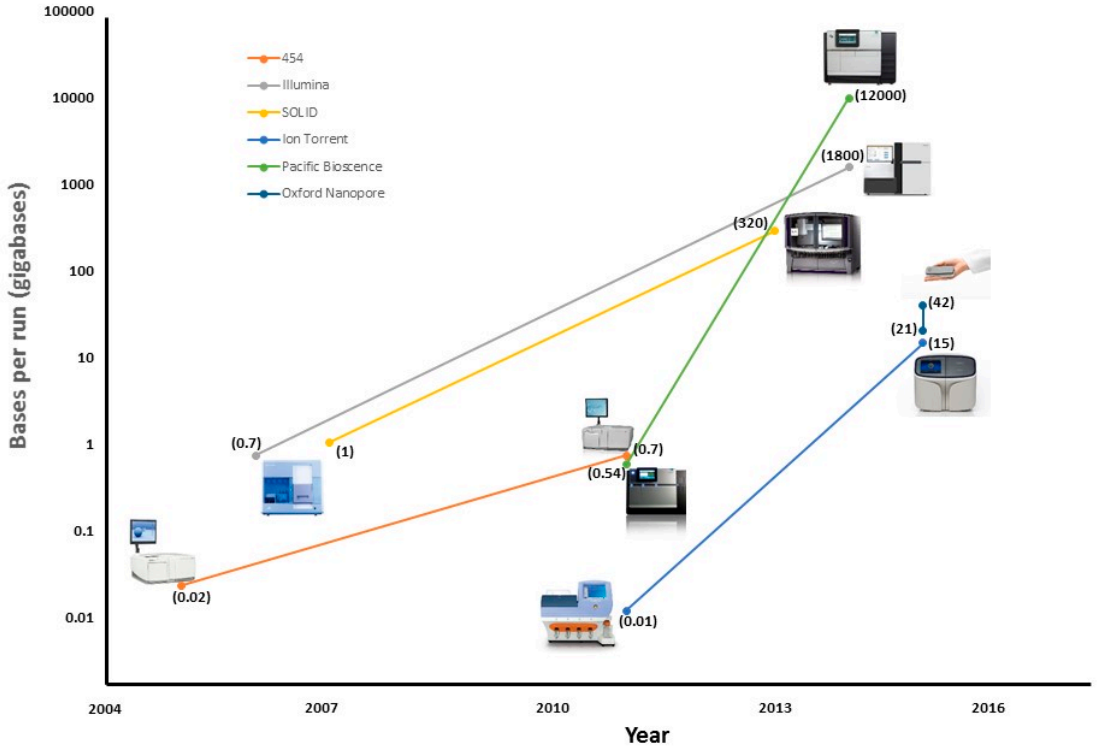
Uma das tecnologias mais utilizadas atualmente, é a plataforma Illumina, semelhante ao método estipulado por Sanger. Porém, a principal diferença é que o bloqueio de polimerização do DNA é reversível (Bentley, 2006). Esta utiliza o sequenciamento por síntese, em que há a fragmentação de ácidos nucleicos para elaborar as bases que serão adicionadas. Assim, são ligadas as sequências em formato de ferradura, flow cell, formando milhões de clusters de fragmentos clonais, em que se forma o primer que serão adicionados aos nucleotídeos que os reconhece emitindo um sinal de fluorescência e a sequência é formada (Liu et al., 2012). Os sequenciadores desta são os mais utilizados, pela sua alta precisão, baixo custo por Gb (*Gigabases*) obtidos e por apresentarem diferentes tipos de plataformas de sequenciamento que são adaptáveis e adequados a cada tipo de análise (Cardenas et al., 2017).

O sequenciador Ion Torrent também realiza o sequenciamento por síntese, no entanto é realizado por semicondutores. Essa tecnologia permite o sequenciamento de moléculas em tempo real, com alta velocidade e baixo custo. No entanto, o

tamanho dos fragmentos sequenciados e o volume de dados gerados ainda são limitados, embora esses aspectos tenham sido aprimorados ao longo dos anos. Estes possuem uma aplicação na identificação de patógenos microbianos (Liu et al., 2012).

O sequenciamento de terceira geração é marcado pela plataforma Pacific Bioscence (PacBio) pois tem a capacidade de sequenciar o DNA por uma síntese em tempo real de molécula única (SMRT). O sequenciamento ocorre através de nanosensores que faz a detecção de sinal na qual uma única DNA polimerase replica o DNA e produz leituras longas. Mesmo que o rendimento do PacBio RS seja menor do que o do sequenciador de segunda geração, as leituras mais longas são muito eficazes para pesquisas clínicas e biomédicas (Liu et al., 2012). Como demonstrado na figura 2:

Figura 2 - Número de gigabases de acordo com o tempo, demonstrando a alta capacidade das tecnologias conforme os avanços tecnológicos



Fonte: Cardenas, 2017.

O avanço da tecnologia de sequenciamento, permitiu com que organismos mais complexos pudessem ser sequenciados em um curto período de tempo. Embora a tecnologia NGS tenha se consolidado, o sequenciamento Sanger ainda é utilizado, porém para estudos mais pontuais como por exemplo identificação de

espécies por genes barcode como por exemplo pelo 16S (Heather, 2016).

Atualmente, o sequenciamento de alto rendimento vem sendo muito utilizado pela sua capacidade de produzir milhões/bilhões de sequências em um tempo e custo menor e com maior rendimento, se comparado com o sequenciamento Sanger (Reuter et al., 2015). Este método permite compreender, por exemplo, sobre variações genéticas, expressão gênica, estruturas de genomas e doenças. Para cada ciência ômica e objetivo de estudo existe uma plataforma de sequenciamento específica, seja está gerando leituras longas ou curtas.

1.2 Transcriptômica

A transcriptômica é uma ciência ômica que analisa transcritos, esta consiste em um conjunto de técnicas moleculares para análises de expressão gênica, com o auxílio de todos os tipos de RNA sob condições estabelecidas e específicas ou comparações entre espécies (Benedetto et al., 2021 & Han et al., 2013). Através delas, a análise da expressão gênica tem sido utilizada para caracterizar profundamente os transcriptomas, fornecendo medidas precisas, edição de RNA e expressão específica de alelos (Chambers et al., 2019).

O pré-processamento de dados permite um melhor desempenho, trazendo dados mais coesos, através de etapas desenvolvidas como o controle de qualidade através da limpeza que consiste em retirar dados inconsistentes (Binsheng et. al., 2020). Sendo uma etapa importante para o tratamento de sequências obtidas em bioinformática, o que consiste em filtragem com base em diferentes parâmetros, determinação do comprimento da sequência, número de bases, índice de qualidade e complexidade da sequência; corte de sequência em determinado comprimento e cortar as extremidades de acordo o índice de qualidade estabelecido, permitindo a geração de novas duplicatas de sequência (Schmieder && Edwards, 2011).

Com a quantidade crescente de dados de sequenciamento sendo gerados, se faz necessário o pré-processamento destas, pois dados brutos obtidos são suscetíveis a ter a presença de ruídos, valores ausentes e inconsistências, o que afeta diretamente a qualidade eficiência dos resultados (Chandrama et al., 2014).

A bioinformática desempenha um papel fundamental para a interpretação da linguagem dos genes por algoritmos oriundos da informática, leitura de reads,

criação de bancos de dados e desenvolvimento de plataformas e softwares capazes de realizar análises. O número de informações geradas e o agrupamento de resultados obtidos através de sequenciamentos, aumenta conforme o tempo passa, gerando um maior número de dados (Araújo et.al., 2008). A utilização de softwares e ferramentas são a base desta área para realizar análises e interpretações de dados provenientes das ciências ômicas, como a proteômica, metabolômica, genômica e transcriptômica (Hanussek et al., 2020).

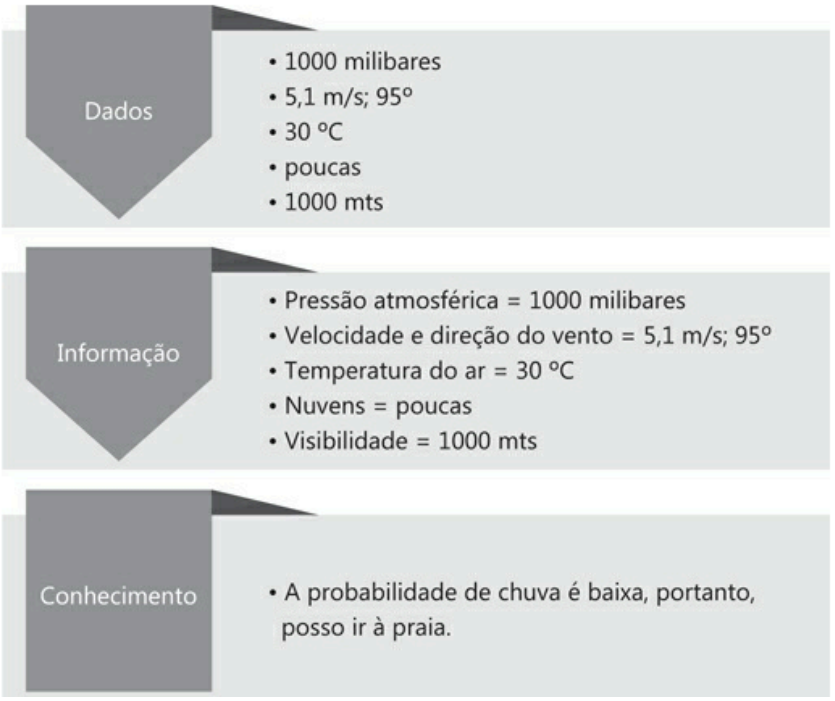
Atualmente, com a produção contínua de dados em grande escala, surge a necessidade de usar métodos capazes de transformar dados em informação à medida que estes são gerados.

1.3 Mineração de dados

Esse tratamento otimizado dos dados é chamado de mineração de dados ou data mining. Por meio desta é possível realizar análises, identificar padrões e afinidades entre os dados, previsões, classificações, agrupamentos e otimização para maximizar os resultados diante de um conjunto de informações (Cardoso & Machado, 2008). O uso de ferramentas adequadas são fundamentais para que esse processo seja realizado de forma efetiva, pois há etapas que devem ser realizadas como a limpeza de dados inconsistentes, integração dos dados a partir de fontes como banco de dados, seleção e escolha de dados relevantes ao objetivo e a transformação destes para formatos apropriados que permita que análises sejam realizadas. Com a extração de informações relevantes torna-se possível a tomada de decisão de forma estratégica (Castro, 2016).

Com o alto volume de dados no mundo atual, torna-se necessário metodologias que transformam dados em informações, para isso ferramentas para o processamento destes e tratamento que permitem melhor visualização e compreensão, são fundamentais (Junior & Rodrigues, 2021). Os dados, por si só, podem ser apenas símbolos, escritos ou termos não estruturados, sem um significado integrado. Já as informações, por outro lado, são resultados organizados e interpretados desses dados, apresentando significado e utilidade, como ilustrado na figura a seguir (Castro, 2016).

Figura 2 - Definição dos termos: Dados, informação e conhecimento.



Fonte:Castro, 2016.

No cenário atual da ciência de dados e da mineração de dados, a linguagem de programação Python emergiu e consolidou-se como uma das ferramentas predominantes e mais influentes. Sua popularidade deriva de sua sintaxe considerada clara e relativamente fácil de aprender, com bibliotecas integradas e especializadas que facilitam a execução de tarefas complexas (Mckinney, 2022).

Por apresentar um desempenho considerado de alto nível, a linguagem python vem como aliada para otimizar a manipulação de dados. Atualmente, vem sendo muito utilizada para análises de grande quantidade de dados, por ser uma linguagem de programação dinâmica e de fácil aprendizado. Esta pode ser utilizada de forma isolada, como também integrada a diversas bibliotecas e também outras linguagens . A biblioteca Pandas é fundamental, fornecendo estruturas de dados de alto desempenho e fáceis de usar (principalmente o DataFrame) e ferramentas para leitura, escrita, limpeza, transformação, agregação e análise de dados (Mckinney, 2022). Para visualização, a biblioteca Matplotlib é utilizada para uma visualização mais estabelecida, fornecendo um controle

detalhado sobre a criação de gráficos estáticos, animados e interativos. Para tarefas mais avançadas, como as que envolvem redes neurais e machine learning, bibliotecas como TensorFlow e PyTorch se tornaram referências no campo do aprendizado profundo. Além dessas, há ferramentas voltadas a áreas mais específicas, como o NLTK e o spaCy para o Processamento de Linguagem Natural (PLN), BeautifulSoup e Scrapy para extração de dados da web, e NetworkX, ideal para quem trabalha com análise de grafos e redes complexas. Na área de computação científica, destaca-se a biblioteca SciPy, que complementa o ecossistema com ferramentas robustas para cálculos técnicos e científicos (Mariano et al., 2015).

Em relação a Bioinformática, pode-se ter aplicações diversas como a classificação para compreender condições e características das ciências ômicas, tomadas de decisão e maximização do tempo nas vastas pesquisas em fontes de dados biológicos que continuam em uma rápida expansão (Lan et. al., 2018). Python possui bibliotecas voltadas ao tratamento de grandes volumes de dados biológicos. Dentre elas, destaca-se o Biopython, um projeto de código aberto mantido por uma comunidade internacional, que integra o termo “Bio”, ao lado de ferramentas como BioJava, BioPerl e BioRuby. Essa biblioteca permite desde a manipulação e anotação de sequências genéticas até análises filogenéticas, acesso a bancos de dados como o NCBI, e aplicações com aprendizado de máquina, contribuindo para pesquisas mais ágeis e eficientes no campo das ciências ômicas (Mariano et. al., 2015).

2 JUSTIFICATIVA

As etapas de pré-processamento de reads compreendem etapas cruciais para avaliar e melhorar a qualidade dos dados gerados pelas plataformas de sequenciamento. Todas as análises posteriores, como por exemplo montagem de transcritos e análise de expressão gênica, dependem da qualidade dos dados gerados na etapa de pré-processamento, que deve ser feita de forma criteriosa. Essa é a única fase da análise que deve ser realizada independente da ciência ômica que se esteja analisando.

Quando realizamos pesquisas na literatura encontramos trabalhos usando plataformas de sequenciamento distintas, bem como programas diferentes para realizar a etapa de pré-processamento. Muitos trabalhos comparam o uso de mais de uma ferramenta de Bioinformática usada na etapa de pré-processamento das reads, porém geralmente comparam a performance computacional dessas ferramentas. Não temos informações referentes às ferramentas predominantes utilizadas na etapa de pré-processamento atualmente.

3 OBJETIVO

O presente trabalho teve como objetivo realizar um levantamento baseado em artigos científicos publicados no banco de dados PubMed do NCBI, nos últimos 5 anos, das plataformas de sequenciamento em larga escala e ferramentas de Bioinformática para o pré-processamento de reads em dados transcriptômicos foram predominantemente utilizadas, através da mineração de dados (data mining) usando Python.

4 METODOLOGIA

A metodologia utilizada na pesquisa trata-se de um levantamento descritivo observacional com abordagem qualitativa, buscando revisar de forma sistemática artigos científicos disponíveis de forma online no banco de dados PubMed do NCBI (National Center for Biotechnology Information) publicados nos últimos 5 anos.

Para isso, foram definidas palavras-chave a partir de duas abordagens para busca dos termos, sendo que a primeira contou com os termos: Preprocessing tools “Bioinformatics” AND “Preprocessing tools” ; “Transcriptome” AND “Preprocessing tools” ; “RNA-Seq” AND “Preprocessing tools” e "Sequencing platform" AND "Preprocessing" AND "RNA-Seq". e para sequenciamento: “Sequencing platform” AND “RNA-seq” ; “Sequencing platform” AND “transcriptome” ; "Sequencing platform" AND transcriptomics; “Sequencing” AND “RNA-seq”; “Sequencing” AND “transcriptome” e "Sequencing" AND transcriptomics. Sendo a segunda demonstrada nas tabelas 2 e 4, o mais eficaz para a montagem do banco de dados.

Os artigos científicos identificados foram coletados e organizados inicialmente por meio do Google Sheets, o que permitiu estruturar um banco de dados com as informações extraídas de cada estudo. A categorização desses dados foi realizada com base na análise metodológica de cada artigo, identificando-se quais plataformas de sequenciamento e quais ferramentas bioinformáticas foram utilizadas na etapa de pré-processamento dos dados transcriptômicos. As variáveis categorizadas incluíram: título, autores, ano de publicação, identificação do artigo (PMID), tipo de estudo (humano, animal ou planta), fragmentos analisados, quantidade de células, plataforma utilizada e tipo de plataforma.

Ademais, foi realizada uma análise criteriosa para verificar se cada artigo se enquadra no escopo e objetivo do estudo, sendo desconsiderados aqueles que não atendiam aos critérios estabelecidos, como os que não apresentavam aplicação direta em estudos transcriptômicos ou não descreviam o uso de ferramentas de pré-processamento e plataformas de sequenciamento.

Baseado em técnicas de análise de dados (data mining) exploratória com enfoque na extração e análise descritiva das informações, a categorização do

conjunto de ferramentas e plataformas mais utilizadas atualmente foram relacionadas com as variáveis definidas como descrito acima.

Utilizou-se a linguagem de programação Python no ambiente interativo Jupyter Notebook, com o apoio das bibliotecas Pandas, para a manipulação e análise dos dados, e Matplotlib, para a geração de gráficos. Os dados tabulares foram importados em formato de DataFrame e, a partir disso, foram aplicadas técnicas como seleção de variáveis, filtragem, cruzamento de categorias para identificar possíveis relações entre as variáveis, sumarização de frequências e geração de gráficos descritivos.

5 RESULTADOS E DISCUSSÃO

5.1 Termos utilizados para a pesquisa

No presente trabalho, foi necessário refinar e analisar termos mais específicos para realizar a busca dos artigos. Os termos definidos para as buscas foram associados com RNA-seq. Inicialmente os termos definidos para a pesquisa utilizados estão descritos nas tabelas 1 e 2.

Tabela 1 - Termos utilizados para pesquisa referentes a etapa de pré-processamento

Termos definidos a priori	Termos utilizados
Preprocessing tools	“Bioinformatics” AND “Preprocessing tools” ; “Transcriptome” AND “Preprocessing tools” ; “RNA-Seq” AND “Preprocessing tools”
Preprocessing	"Sequencing platform" AND "Preprocessing" AND "RNA-Seq"

Fonte: Elaborado pela Autora

Tabela 2 - Termos utilizados para pesquisa em relação ao sequenciamento

Termos definidos a priori	Termos utilizados
Sequencing platform	“Sequencing platform” AND “RNA-seq” ; “Sequencing platform” AND “transcriptome” ; "Sequencing platform" AND transcriptomics.
Sequencing	“Sequencing” AND “RNA-seq”; “Sequencing” AND “transcriptome”; "Sequencing" AND transcriptomics.

Fonte: Elaborado pela Autora

No entanto, ao realizar a busca tais termos não estavam direcionando para os artigos específicos e de interesse. Nesse sentido, foi necessário refinar a busca e assim, optou-se por utilizar na etapa de pré-processamento as ferramentas mais utilizadas para a limpeza das reads geradas através do sequenciamento. Este ajuste metodológico, destacou e priorizou a busca por nomes de ferramentas e também das plataformas de sequenciamento, permitiu uma seleção mais direcionada e eficiente dos artigos. Esta dificuldade e necessidade de refinamento

considera-se como um resultado relevante, pois é possível compreender como a terminologia e a forma como as informações são descritas nos artigos podem impactar significativamente a recuperação de dados em estudos de caráter sistemático e exploratório, especialmente em áreas como a bioinformática, em que há uma grande diversidade de nomenclaturas e fluxos de trabalho.

Dessa forma, novos termos foram estabelecidos, baseados nos nomes de plataformas e ferramentas comumente utilizadas em estudos de bioinformática, como demonstrado nas tabelas 3 e 4.

Tabela 3 - Termos utilizados para a montagem do banco de dados para busca das ferramentas de pré-processamento

Ferramenta de Bioinformática	Termos utilizados	N° de artigos	Referência:
FASTQC	"FASTQC" AND "RNA-SEQ"	27	ANDREWS, S.(2010)
MUltiQC	"MUltiQC" AND "RNA-SEQ"	3	EWELS, P. et al. (2016)
Prinseq	"Prinseq" AND "RNA-SEQ"	0	SCHMIEDER, R.; EDWARDS, R. (2011).
Trim galore	"Trim galore" AND "RNA-SEQ"	1	KRUEGER, F. (2015).
FASTX Toolkit	"FASTX Toolkit" AND "RNA-SEQ"	0	GORDON, A. (2010)
FASTP	"FASTP" AND "RNA-SEQ"	4	CHEN, S. et al. (2018).
Trimmomatic	"Trimmomatic" AND "RNA-SEQ"	5	BOLGER, A. M.; LOHSE, M.; USADEL, B. (2014).
SeqyClean	"SeqyClean" AND "RNA-SEQ"	0	ZHBANNIKOV, et al. (2017).
SeqTrim	"SeqTrim" AND "RNA-SEQ"	0	FALGUERAS, A. et al. (2010).
Cutadapt	“Cutadapt” AND “RNA-SEQ”	3	MARTIN, M. (2011).

Fonte: Elaborado pela Autora

Tabela 4 - Termos utilizados e considerados para os resultados presentes neste estudo

Termos definidos para a busca	Termos utilizados	Quantidade de artigos
Illumina	"Illumina" AND "RNA-SEQ"	900
PacBio	"PacBio" AND "RNA-SEQ"	241
miniON	"miniON" AND "RNA-SEQ"	27

Fonte: Elaborado pela autora

Os novos termos permitiram uma busca mais direcionada, para além da seleção de palavras chaves, alguns critérios relevantes para o levantamento foram estabelecidos, como desconsiderar artigos que não apresentem estudos de caso, duplicatas, revisões sistemáticas e bibliográficas. Os resultados obtidos em algumas pesquisas tiveram os mesmos artigos encontrados, nos quais os que estavam duplicados foram filtrados e retirados os de repetição, permanecendo na pesquisa apenas um, para que não houvesse resultados duplicados.

Selecionando assim, aqueles que se enquadram com o objetivo do presente trabalho. No levantamento de ferramentas de pré-processamento, 27 artigos foram selecionados, e para as plataformas de sequenciamento, 719 artigos foram considerados para compor o banco de dados. O que possibilitou as análises posteriores de forma mais precisa sobre as ferramentas de pré-processamento e plataformas de sequenciamento atualmente utilizadas em estudos transcriptômicos.

Foi-se analisado primeiramente os resumos para compreender se a pesquisa realizada no artigo seria utilizada e depois os materiais e métodos, a fim de identificar como foi realizada e os pipelines bioinformática utilizados. Para compor o presente trabalho e as tabelas ficarem adaptadas, optou-se por utilizar o PMID como identificador dos artigos selecionados.

5.2 Análise dos dados obtidos - Pré-processamento

Com o auxílio da aplicação Jupyter notebook, juntamente, com as bibliotecas Pandas e Matplotlib, foram filtrados e analisados os dados do dataframe. Os dados obtidos de acordo com o número de artigos encontrados, revelam um predomínio significativo do uso da ferramenta FastQC para controle de qualidade, com um total de 15 artigos, representando a maior concentração entre todas as ferramentas avaliadas. Em seguida, aparece a combinação entre FastQC e MultiQC, mencionada em 6 estudos. Pelo levantamento essa associação foi a mais comum, se comparado com FastQC e Trimmomatic (3 artigos) demonstra que alguns autores optaram por empregar diferentes ferramentas em conjunto. A ferramenta Trimmomatic utilizada de forma isolada, foi observada em apenas 2 artigos, o que indica menor prevalência desta nos artigos levantados para controle de qualidade, como demonstrado na tabela 5 e na figura 3.

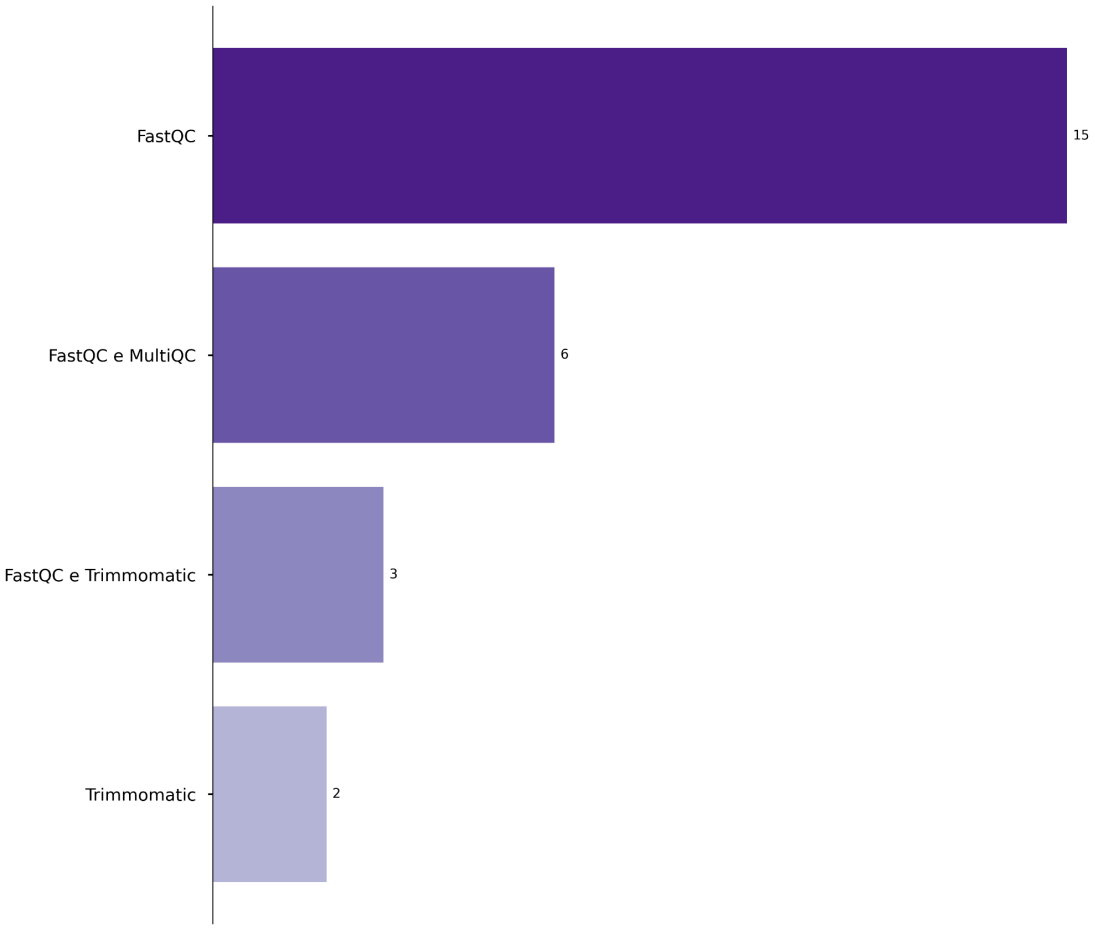
Tabela 5 - Número de ferramentas de pré-processamento utilizadas nos anos de 2020 a 2024, representando o tipo de estudo.

Ano	PMID	Ferramenta - controle de qualidade	Ferramenta - Trimming	Tipo de estudo
2020	32547551	FastQC	Não Menciona	Humano
2020	32642521	FastQC	Trimmomatic	Animal
2020	32986356	FastQC	Não Menciona	Humano
2020	33318985	FastQC e MultiQC	Cutadapt	Humano
2020	33318985	FastQC e MultiQC	Cutadapt	Humano
2021	34360925	FastQC	BBDUK2	Humano
2021	33978066	FastQC e Trimmomatic	Trimmomatic	Animal
2022	35099752	FastQC	Cutadapt	Humano
2022	35576023	FastQC	Não Menciona	Animal
2022	36381277	FastQC e Trimmomatic	Trimmomatic	Humano
2023	36598628	FastQC	Não Menciona	Animal
2023	37683796	FastQC	Não Menciona	Animal
2023	37733230	FastQC	Não Menciona	Animal
2023	36724030	Trimmomatic	Trimmomatic	Planta
2023	38115564	FastQC	Não Menciona	Planta
2023	37727589	FastQC	Não Menciona	Animal
2023	37089209	FastQC	Não Menciona	Animal
2023	36969977	FastQC e MultiQC	Não Menciona	Humano

2023	37006386	FastQC e MultiQC	Não Menciona	Planta
2024	38406249	FastQC	Não Menciona	Humano
2024	38734729	FastQC e MultiQC	Não Menciona	Animal
2024	38990974	FastQC	Cutadapt	Humano
2024	38097858	FastQC e MultiQC	Cutadapt	Humano
2024	38998009	Trimmomatic	Trimmomatic	Animal
2024	38406249	FastQC	Fastp	Humano

Fonte: Elaborado pela autora.

Figura 3 - Relação dos artigos científicos de acordo com a ferramenta de pré-processamento para controle de qualidade.



Fonte: Elaborado pela autora.

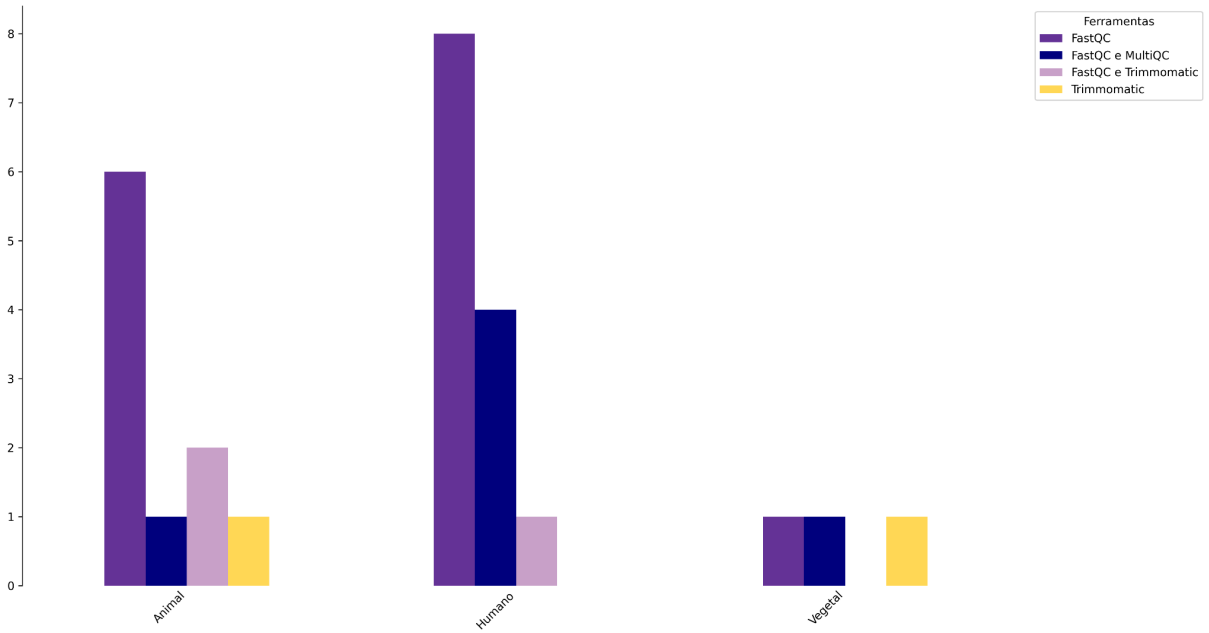
O uso da ferramenta FastQC na maioria dos estudos, pode ser explicado pela sua facilidade de uso e a geração de relatórios precisos e confiáveis. É possível integrá-la às demais ferramentas para potencializar determinado estudo, sendo uma ferramenta de código aberto, compatível com diferentes pipelines bioinformáticos. O formato FastQ é o padrão para leituras das sequências geradas a partir de tecnologias de sequenciamento de nova

geração. Além disso, fornece análises automatizadas de múltiplos parâmetros essenciais, como a qualidade por base, distribuição do conteúdo GC e detecção de sequências adaptadoras (Andrews, 2010). O seu uso é eficaz para o pré-processamento de dados em larga escala.

As ferramentas utilizadas para o controle de qualidade podem ser associadas a outras dependendo da qualidade das leituras obtidas. Nesse sentido, estas foram levantadas também, o que resultou com a prevalência das ferramentas Cutadapt e Trimmomatic, cada uma presentes em 5 dos artigos levantados.

A partir da relação da distribuição das ferramentas de pré-processamento em relação ao tipo de estudo permitiu a categorização dos artigos conforme o organismo: humano, animal e planta. É possível observar que a ferramenta FastQC é predominante em estudos humanos e em animais, sendo 8 e 7 ocorrências. A combinação das ferramentas FastQC e MultiQC é menos frequente, em sua maioria em estudos humanos, 4 artigos, animais e plantas e apenas 1 artigo para cada organismo.

Figura 4 - Frequência dos estudos em relação ao tipo de estudo com a ferramenta de controle de qualidade utilizada.



Fonte: Elaborado pela autora.

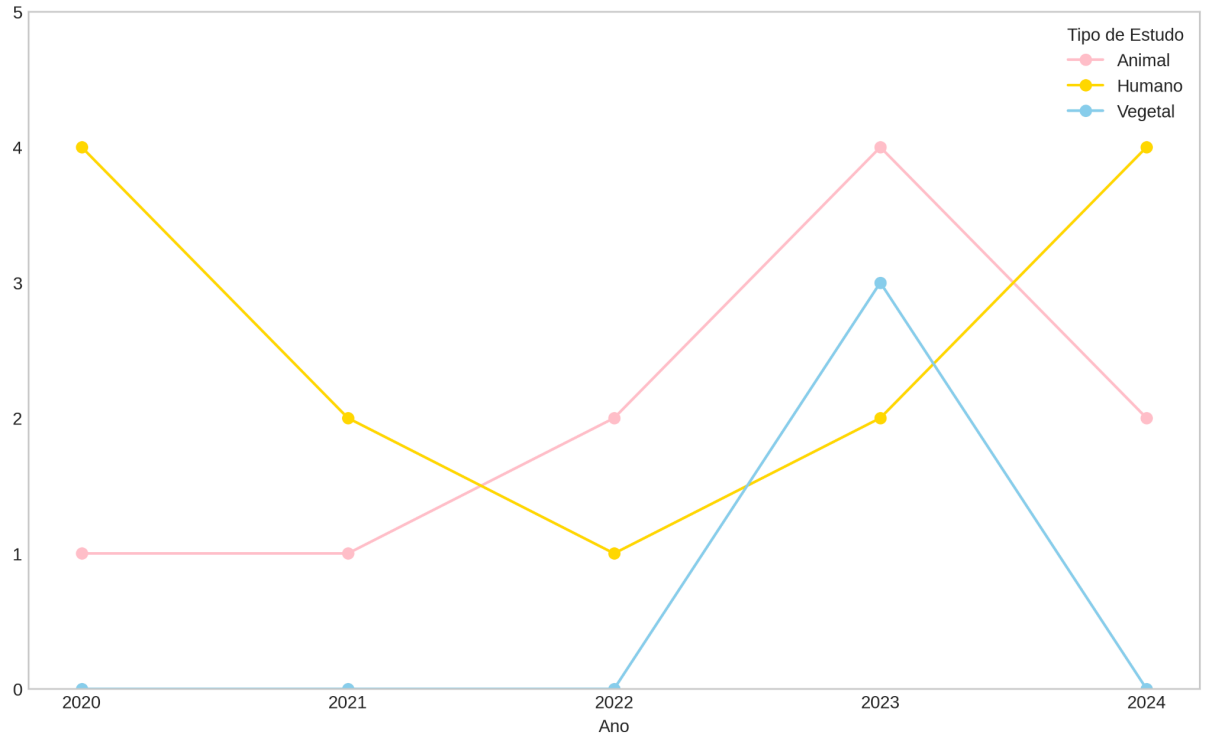
Os estudos humanos se destacam em relação ao total de análises com

as ferramentas de controle de qualidade, nos quais também predominam as ferramentas utilizadas para a etapa de trimming (processo de remoção de sequências indesejadas ou de baixa qualidade das leituras).

O tipo de estudo ao longo dos último 5 anos (2020 a 2024), pode-se verificar que os estudos humanos no período todo analisado, é o que mais se destaca em relação às quantidades, tendo predominância em 2020 e 2024, o que indica como tais pesquisas estão em destaque para estudos transcriptômicos. No caso de estudos com animais, o crescimento significativo é a partir de 2022, com pico em 2023, assim como estudos envolvendo plantas.

De forma geral, o tipo de estudo planta apresentou a menor frequência e diversidade de ferramentas, o que pode refletir tanto limitações em iniciativas científicas voltadas a esse grupo quanto possíveis diferenças nos requisitos metodológicos.

Figura 5 - Frequência de cada tipo de estudo ao longo dos anos



Fonte: Elaborado pela autora.

A predominância do uso de FastQC nos artigos científicos levantados,

pode indicar que este é fundamental em pipelines de bioinformática, para a verificação do controle de qualidade de leituras obtidas. Indicando que não necessariamente seja uma concorrente com as demais, mas sim ferramentas que se complementam no controle de qualidade (QC) e pré-processamento. Esta ferramenta demonstra-se como ponto de partida para o diagnóstico (Andrews, 2010), o que pode direcionar o uso das demais ferramentas que estão mais associadas com o tratamento e agregação de resultados. Devido a isso há a distinção nas colunas da planilha como “controle de qualidade” e “trimming”.

Em estudos humanos e animais acompanha com maior frequência esses estudos ao longo dos anos. A constância dos estudos humanos justifica o uso diversificado de ferramentas, como combinações com MultiQC e Trimmomatic. Já o pico de estudos com animais em 2023 coincide com a diversificação das ferramentas. Os estudos com plantas, por sua vez, são os menos frequentes e apresentam pouca diversidade de ferramentas. Em síntese, a crescente relevância dos estudos com modelos animais e a subrepresentação das pesquisas com plantas revelam tendências e lacunas que podem orientar futuras investigações e estratégias de padronização metodológica.

5.3 Análise dos dados obtidos para Plataformas de Sequenciamento

A partir da análise dos artigos, foi realizado um mapeamento das principais plataformas de sequenciamento de dados transcriptômicos para identificar a predominância de uso de cada uma. Verificou-se a relação completa dos artigos analisados, com as respectivas plataformas utilizadas, está apresentada na Tabela 6 que pode ser acessada através do link: {https://drive.google.com/file/d/11QqLB00MCrxB0Gy7gCkVTHZGph6zQLRh/view?usp=drive_link}.

Com base nesses dados, foi realizada uma análise da distribuição temporal, quantificando o uso das plataformas por ano de publicação. A Tabela 7 permite observar que a plataforma Illumina é a mais frequente, embora muitos estudos combinem o uso de mais de uma tecnologia. É notável que a dominância desta plataforma não apenas se manteve, mas se intensificou proporcionalmente ao

longo do período. Em 2021, a Illumina correspondia a 80.6% das publicações, fração que aumentou para 84.2% em 2024.

Tabela 7 - Relação entre as plataformas mais utilizadas de acordo com a quantidade de artigos publicados no ano período de 2020 a 2024.

Plataforma(s) utilizadas(s)	2020	2021	2022	2023	2024	Total por plataforma
Illumina	82	150	130	105	107	574
Illumina e Pacbio	25	27	21	20	14	107
Illumina e miniON	3	2	4	2	-	11
Pacbio	1	1	1	2	1	6
Illumina e Ion Torrent	-	1	-	-	-	1
miniON	-	3	3	-	3	9
miniON e Sanger	-	-	-	-	1	1
Illumina e ONT	2	2	2	-	1	7
Pacbio e Sanger	-	-	-	1	-	1
MGI FC e Illumina	1	-	-	-	-	1
Pacbio e ONT	1	-	-	-	-	1
Total por ano	115	186	161	130	127	

Elaborado pela autora.

Uma tendência secundária relevante é o uso combinado da tecnologia de leituras curtas da Illumina com a de leituras longas da PacBio. Essa abordagem se manteve como a segunda mais frequente em todos os anos analisados, correspondendo a aproximadamente 14-21% dos estudos, o que sugere uma estratégia para a obtenção de transcriptomas mais completos.

O ano de 2021 destacou-se com o pico de publicações (n=186) na área durante o período analisado. Embora em menor número, a presença de tecnologias como a MiniON (Oxford Nanopore) e Pacbio de forma isolada ou combinada sugere uma exploração contínua de alternativas para sequenciamentos de leituras longas.

Com os avanços das tecnologias de sequenciamento, organismos com maior complexidade podem ser sequenciados em período curto de tempo com diferentes comprimentos de leitura e rendimento, o que justifica a plataforma illumina como a mais utilizada. Pois esta promove um custo menor por gigabase, tem capacidade de produzir milhões/bilhões de sequências em um tempo e diversos tipos de plataforma eficientes para o objetivo dos estudos. No geral, devido aos seus sistemas de alto rendimento e custo-benefício, a Illumina dominou o cenário comercial da NGS Technologies nas últimas décadas (Akaçin et al., 2022). À medida que as tecnologias de sequenciamento continuam a avançar na década de 2020, elas prometem aprimorar ainda mais nossa compreensão de sistemas biológicos complexos. (Kazim et al., 2024).

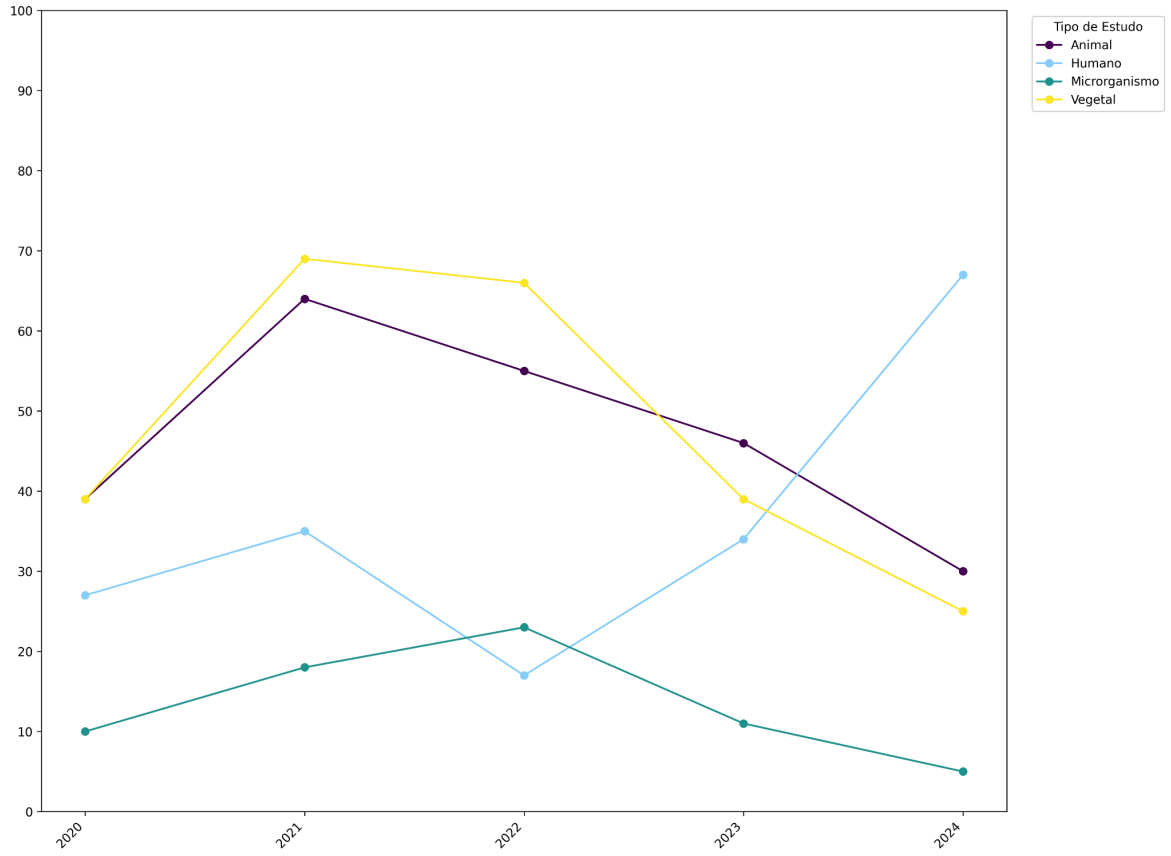
Adicionalmente, o uso da plataforma Pacbio associado a plataforma illumina, permite obter dados de alta qualidade com longo comprimento de leitura e precisão, o que pode superar as limitações de uma única tecnologia. O sequenciamento de leitura curta geralmente pode fornecer alta cobertura (e, portanto, sensibilidade), enquanto o sequenciamento de leitura longa pode fornecer mais informações sobre isoformas, permitir o faseamento de variantes e o splicing de variantes (Gong et al., 2024).

A fim de identificar quais os tipos de estudo mais frequentes, para todos os tipos de plataforma, foi realizada a relação do número de artigos levantados com o tipo de estudo e ano. Os dados demonstraram que há um certo dinamismo a respeito dos focos de pesquisa nos anos de 2020 e 2024.

No ano de 2020, houve a predominância nos estudos que envolvem organismos plantas (n=69) e animais (n=64), ambas tiveram o ponto de declínio contínuo até 2024. No entanto, os estudos humanos apresentaram uma expressão acentuada em relação ao crescimento, o que praticamente duplicou (n=68) em relação ao início do período analisado

Houve uma crescente geral em todos os tipos de estudo no ano de 2021, o que coincide com o ano com mais publicações, como apresentado anteriormente. Ademais, estudos com microrganismos, têm uma diferença expressiva se comparado aos demais, sendo a categoria com menor número de publicações, em que teve um aumento em 2022. Todos os aspectos aqui evidenciados, podem ser analisados de forma gráfica na figura 6.

Figura 6 - Frequência do tipo de estudo no decorrer do período de 2020 a 2024



Elaborado pela autora

O crescimento generalizado de todos os tipos de estudos em 2021, sugere como a comunidade acadêmica produziu muitos artigos científicos nesse período. Possivelmente, esse resultado está relacionado com a pandemia de COVID-19, em que houve uma intensificação de investigações que demandou esforços para diferentes tipos de estudo, como com microrganismos e plantas até a aplicação em modelos animais e estudos preliminares em humanos. Os estudos bioinformáticos foram fundamentais na pandemia em diversos aspectos com o desenvolvimento de ferramentas, metodologias e a integração de dados.

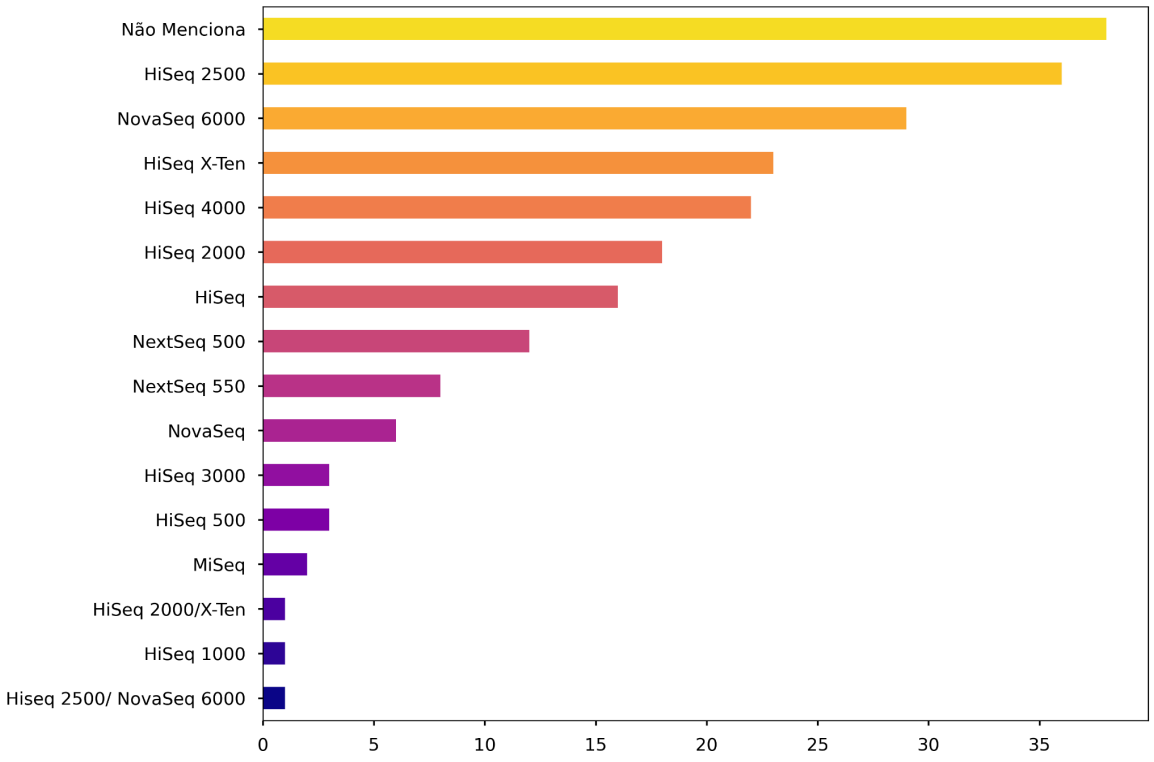
Nos estudos humanos, o expressivo aumento a partir do ano de 2022, deve ter correlação com diretrizes e novos protocolos, a OMS estabeleceu como prioridade o fortalecimento da capacidade dos países para sequenciar patógenos e, crucialmente, para utilizar os dados genômicos gerados na tomada de decisões em saúde pública. O avanço abrangeu todo o campo das ciências ômicas, como a importância dos estudos transcriptômicos, que foram cruciais para potencializar o combate à pandemia (Organização Mundial da Saúde, 2023).

Para a realização do sequenciamento, há tipos de plataformas específicas

variando significativamente conforme o tipo de organismo estudado. Um aspecto metodológico importante, é a alta frequência de artigos que não especificam o modelo da plataforma utilizada nos tipos de estudo planta e humano, em que esta é uma etapa técnica fundamental, baseada em critérios como custo, rendimento de dados e aplicação.

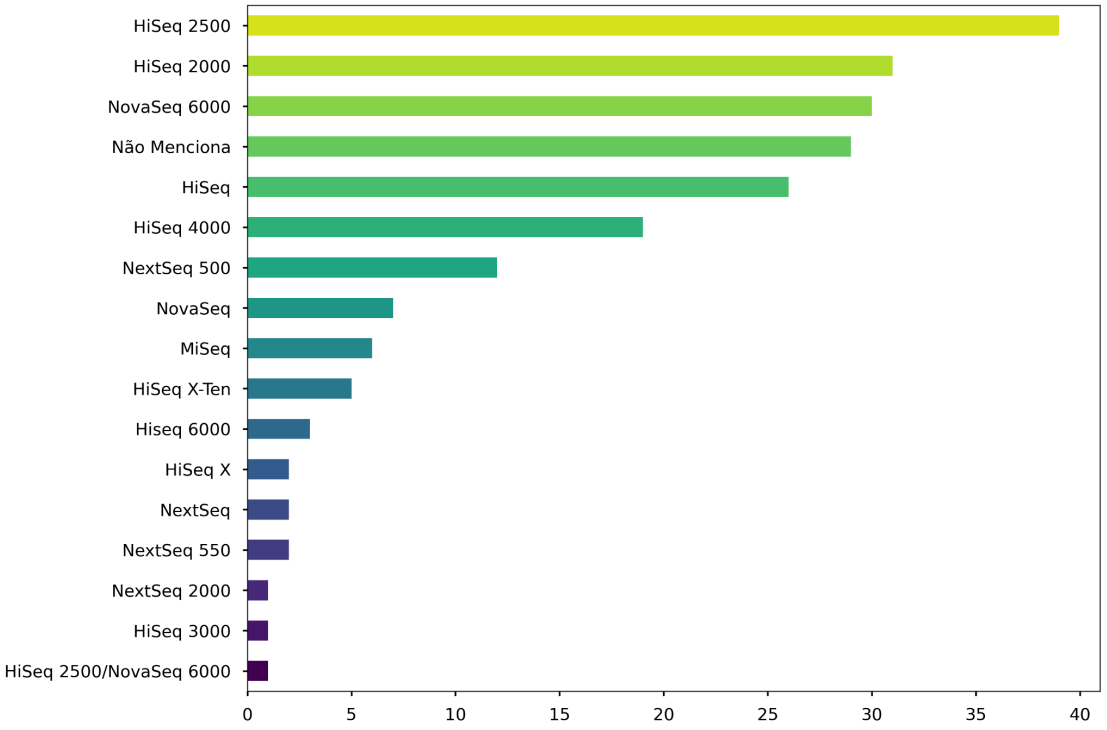
A illumina possui múltiplas plataformas e pela sua predominância e relevância na quantidade de artigos, aqui serão descritas as plataformas mais utilizadas por tipo de estudo, como demonstrado nas figuras 7, 8, 9 e 10 respectivamente.

Figura 7 - Relação dos tipos de plataforma Illumina em estudos plantas.



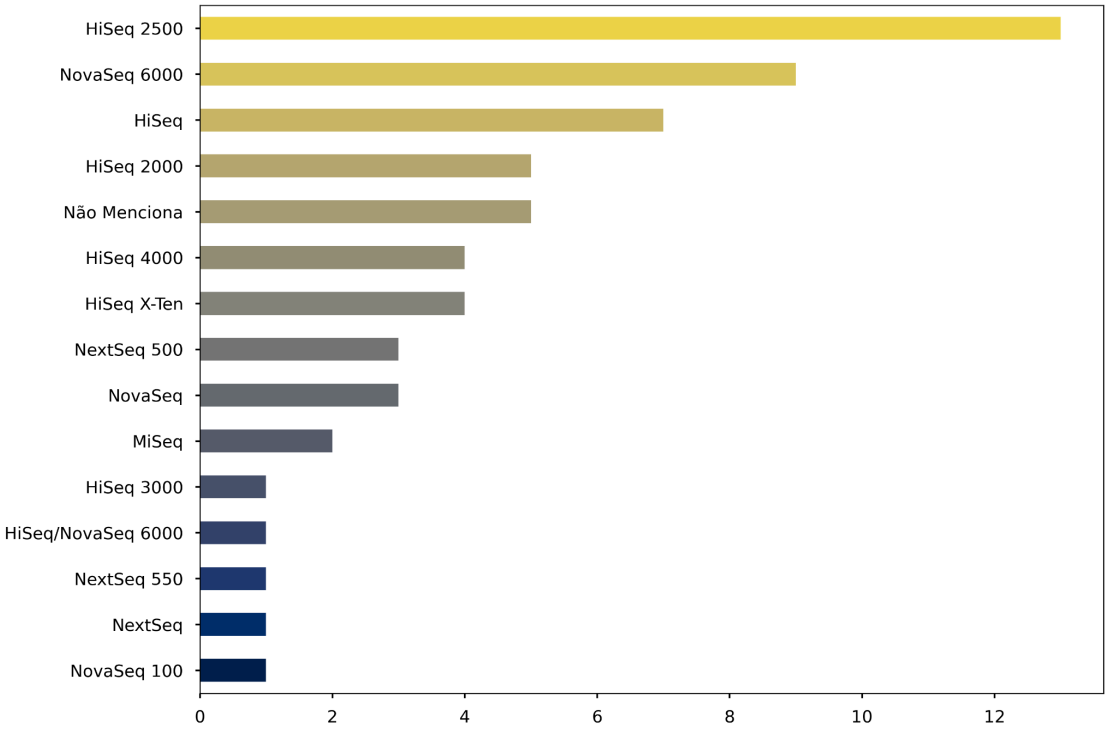
Elaborado pela autora.

Figura 8 - Relação dos tipos de plataforma Illumina em estudos animais.



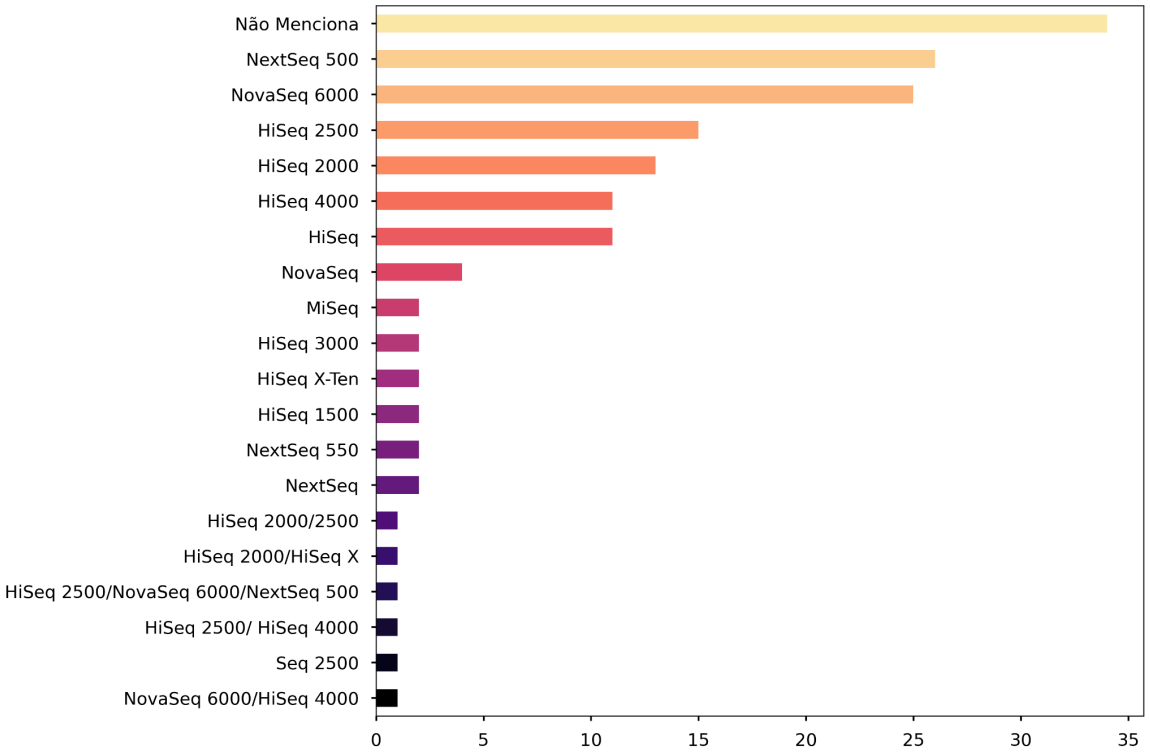
Elaborado pela autora.

Figura 9 - Relação dos tipos de plataforma Illumina em estudos com microrganismos.



Elaborado pela autora.

Figura 10 - Relação dos tipos de plataforma Illumina em estudos humanos.



Elaborado pela autora.

Em um panorama geral, há uma diversidade considerável de plataformas utilizadas, sendo a principal a “HiSeq 2500” para os tipos de estudo de microorganismo, planta e animal e para estudos humanos a plataforma que se destaca é “NextSeq 500”.

Foram relacionadas estas da forma que estava descrito nos artigos científicos analisados. Foi encontrada a menção "Seq 2500" , o que não foi possível identificar de forma concreta qual era, mas com base nos modelos comercializados e na popularidade observada nos gráficos, presume-se que o termo se refira à plataforma “HiSeq 2500”.

Segundo a documentação Illumina 2015, HiSeq 2500 é apresentada como um único sistema que combina saída escalável de 10 Gb a 1 Tb com modos de execução flexíveis. A possibilidade de escolher entre o "modo de execução rápida e alta saída" permite que os pesquisadores adaptem o percurso de seu projeto específico, seja ele pequeno ou grande, garantindo versatilidade e otimização de custos e tempo. A NextSeq 500 atendeu a uma demanda crescente por agilidade e flexibilidade, especialmente no campo da pesquisa clínica e humana. Com um tempo de resposta inferior a 30 horas e kits de saída variável, ela se tornou a

ferramenta ideal para aplicações como o sequenciamento de exomas, transcriptomas e paineis de genes, onde a capacidade de analisar lotes menores de forma ágil e com resultados em poucas horas é mais valiosa do que o volume bruto de dados.

A diversidade e coexistência das plataformas descritas nas figuras, mostram um cenário científico que representa uma importante evolução no processo de sequenciamento.

6 CONCLUSÕES

A tecnologia de sequenciamento de alto rendimento (high-throughput sequencing) permite obter milhões de sequências (reads) de DNA ou RNA de forma rápida e com baixo custo. As plataformas de sequenciamento se desenvolvem rapidamente e novas plataformas são lançadas periodicamente, inclusive gerando um volume de dados cada vez maior.

No que se refere às ferramentas utilizadas para a aplicação dos métodos de mineração de dados e demais procedimentos adotados, destaca-se a eficiência do ambiente Jupyter Notebook. Sua versatilidade foi fundamental para a realização de filtros e a análise de arquivos A biblioteca Pandas desempenhou um papel essencial ao possibilitar a leitura e manipulação desses arquivos na forma de DataFrames, facilitando a organização e interpretação dos dados.

Observando os resultados para plataformas de sequenciamento, notou-se que a maioria dos trabalhos dos últimos 5 anos utilizaram na obtenção de RNA-Seq/transcriptômica a plataforma Illumina HiSeq 2500 para os estudos que envolvem animais, microrganismos e plantas. Enquanto que para estudos humanos a plataforma de predominância foi a NextSeq 500.

Já para a análise de ferramentas de pré-processamento de dados transcriptômicos as ferramentas Trimmomatic e Cutadapt foram as mais utilizadas como o padrão da área, refletindo a necessidade por soluções eficientes e bem estabelecidas para garantir a qualidade dos dados.

O presente estudo evidenciou a importância de pipelines de bioinformática bem descritos e com informações necessárias para que um estudo seja replicado, a quantidade de artigos que não mencionaram as ferramentas e plataformas que utilizou, impactou os resultados obtidos.

REFERÊNCIAS

AKAÇIN, I.; ERSOY, S.; DOLUCA, O.; GUNGORMUSLER, M. Comparing the significance of the utilization of next generation and third generation sequencing technologies in microbial metagenomics, **Microbiological Research**, v. 264, 2022. Disponível em: https://doi.org/10.1016/j.micres.2022.127154. Acesso em 15 jan. 2025.

ANDREWS, S. A Quality Control Tool for High Throughput Sequence, 2010. Disponível em: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ . Acesso em: 08 out. 24

ARAÚJO, N. et al. A era da bioinformática: seu potencial e suas implicações para as ciências da saúde. **Estudos de Biologia**, v. 30, n. 70/72, 2008. Disponível em: https://www.researchgate.net/publication/267748258_A_ERA_DA_BIOINFORMATICA_SEU_POTENCIAL_E_SUAS_IMPLICACOES_PARA_AS_CIENCIAS_DA_SAUDE . Acesso em: 27 fev. 2024.

BENEDETTO, A.; PEZZOLATO, M.; BIASIBETTI, E.; BOZZETTA, E. Omics applications in the fight against abuse of anabolic substances in cattle: challenges, perspectives and opportunities. **Current Opinion in Food Science**, v. 40, p. 112-120, 2021. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S2214799321000424. Acesso em: 27 fev. 2024.

BENTLEY, D. R. Whole-genome re-sequencing. **Current Opinion in Genetics & Development**, v. 16, n. 6, p. 545–552, 2006. Disponível em: https://pubmed.ncbi.nlm.nih.gov/17055251/. Acesso em: 25 fev. 2024.

BINSHENG, H. et al. Assessing the impact of data pre-processing on analysing next generation sequencing data. **Frontiers in Bioengineering and Biotechnology**, v. 8, jul. 2020. Disponível em: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00817/full. Acesso em: 1 mar. 2024.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: A flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114-2120, 2014. Disponível em: https://pubmed.ncbi.nlm.nih.gov/24695404/. Acesso em: 09 out. 2024.

BUTLER, J. M. Massively parallel sequencing techniques for forensics: a review. **Forensic Science International: Genetics**, v. 37, p. 190–196, 2018. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC6282972/. Acesso em: 25 fev. 2024.

CARDOSO, O.; MACHADO, R. Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras. **Revista de Administração Pública**, v. 42, n. 3, p. 495-528, maio/jun. 2008. Disponível em: https://www.scielo.br/j/rap/a/4ScBD9DkFprnH7MFyKC3ydv/?lang=pt. Acesso em: 1 mar. 2024.

CASTRO, L.; FERRARI, D. **Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações.** 1. ed. São Paulo: **Saraiva,** 2016.

CHAMBERS, D. C. et al. Transcriptomics and single-cell RNA-sequencing. **Respirology**, 2019. Disponível em: https://onlinelibrary.wiley.com/doi/10.1111/resp.13412. Acesso em: 27 fev. 2024.

CHANDRAMA, W.; DEVALE, P. R.; RAVINDRA, M. Survey on data preprocessing method of web usage mining. **International Journal of Computer Science and Information Technologies**, v. 5, 2014. Disponível em: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=fabab5698365dba510dbc4e0da878a802b08db0d. Acesso em: 27 fev. 2024.

CHEN, S. et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, v. 34, n. 17, p. i884–i890, 2018. Disponível em: https://pubmed.ncbi.nlm.nih.gov/30423086/ . Acesso em: 25 ago. 2024.

CHRISTOFF, A. P. Genômica e sequenciamento de nova geração. Marcadores Moleculares na Era Genômica: Metodologias e Aplicações, Cap. 2, p. 21-50, 2017.

EWELS, P. et al. MultiQC: summarize analysis results for multiple tools and samples in a single report. **Bioinformatics**, v. 32, n. 19, p. 3047-3048, 2016. Disponível em: https://pubmed.ncbi.nlm.nih.gov/27312411/. Acesso em 25 ago. 2024

FALGUERAS, J. et al. SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. **BMC Bioinformatics**, v. 11, p. 70, 2010. Disponível em: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-38. Acesso em 25 ago. 2024

GARRIDO-CARDENAS, J. A.; GARCIA-MAROTO, F. DNA sequencing sensors: an overview. **Sensors**, v. 17, n. 3, p. 588, 2017. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC5375874/. Acesso em: 25 fev. 2024.

GONGO, B.; LI, D.; LABAJ, PP.; NOVORADOVSKAYA. et al. Targeted DNA-seq and RNA-seq of Reference Samples with Short-read and Long-read Sequencing. **Sci Data.** v. 11(1), p. 892. Disponível em: https://pubmed.ncbi.nlm.nih.gov/39152166/ . Acesso em: 02 fev. 2025.

GORDON, A. FASTX-Toolkit: FASTQ/A short-reads pre-processing tools. Cold Spring Harbor: **Hannon Lab.** 2010. Disponível em: http://hannonlab.cshl.edu/fastx_toolkit/. Acesso em: 09 out. 2024.

HAN, X. J. et al. Transcriptome sequencing and expression analysis of terpenoid biosynthesis genes in Litsea cubeba. **PLoS ONE**, 2013. Disponível em: https://pubmed.ncbi.nlm.nih.gov/24130803/. Acesso em: 26 fev. 2024.

HANUSSEK, M.; BARTUSCH, F.; KRÜGER, J. BOOTABLE: bioinformatics benchmark tool suite for applications and hardware. **Future Generation Computer Systems**, v. 102, p. 1016-1026, 2020. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0167739X1931533X. Acesso em: 26 fev. 2024.

HEATHER, James M.; CHAIN, Benjamin. The sequence of sequencers: The history of sequencing DNA. Genomics, v. 107, n. 1, p. 1-8, 2016. DOI: 10.1016/j.ygeno.2015.11.003. Disponível em:

https://pmc.ncbi.nlm.nih.gov/articles/PMC4727787/. Acesso em: 27 fev. 2024.

HUTCHISON, Clyde A., III. DNA sequencing: bench to bedside and beyond. Nucleic Acids Research, v. 35, n. 18, p. 6227–6237, 2007. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2094077. Acesso em: 01 mar. 2024

ILLUMINA. HiSeq 2500 Sequencing System. **Illumina**, 2020. Disponível em:https://www.illumina.com/documents/products/datasheets/datasheet_hiseq2500.pdf. Acesso em: 16 fev. 2025.

ILLUMINA. NextSeq 500 System WGS Solution. **Illumina**, 2021. Disponível em: https://www.illumina.com/documents/products/appnotes/appnote-nextseq-500-wgs.pdf. Acesso em: 16 fev. 2025.

KAZIM, I.; GRANDE, T.; REYHER, E.; GYATSHO, K.; BHUTIA.; DHINGRA, K. Advancements in sequencing technologies: From genomic revolution to single-cell insights in precision medicine. **Journal of Knowledge Learning and Science Technology.** v. 3, n. 4, p. 108-124, 2024. Disponível em: https://jklst.org/index.php/home/article/view/242 . Acesso em: 16 jan. 2025.

KRUEGER, F. Trim Galore!. **Babraham Institute**, 2019. Disponível em: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. Acesso em: 16 ago. 2024.

LAN, K.; WANG, D. T.; FONG, S. et al. A survey of data mining and deep learning in bioinformatics. **Journal of Medical Systems**, v. 42, p. 139, 2018. Disponível em: https://pubmed.ncbi.nlm.nih.gov/29956014/. Acesso em: 26 fev. 2024.

LIU, L. et al. Comparison of next-generation sequencing systems. **Journal of Biomedicine and Biotechnology**, v. 2012, Artigo ID 251364, 2012. Disponível em: https://pubmed.ncbi.nlm.nih.gov/22829749/. Acesso em: 25 fev. 2024.

MARGULIES, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. **Nature**, v. 437, p. 376–380, 2005. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC1464427/. Acesso em: 25 fev. 2024.

MARIANO, D.; BARROSO, J. et al. Introdução à Programação para Bioninformática com Biopython. 3. ed. Minas Gerais, 2016. Disponível em: https://www.researchgate.net/profile/Diego-Mariano/publication/344756626_Introducao_a_Programacao_para_Bioinformatica_com_Biopython/links/5f8e4b83458515b7cf8dc262/Introducao-a-Programacao-para-Bioinformatica-com-Biopython.pdf
Acesso em: 9 fev. 2024.

MARTIN, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. **EMBnet**. v. 17, p. 10–12. Disponível em: https://journal.embnet.org/index.php/embnetjournal/article/view/200 . Acesso em: 08 out 2024

MCKINNEY, Wes. Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter. 3. ed. **Sebastopol**: O'Reilly Media, 2022. Acesso em 24 fev. 2024.

REUTER, J. A.; SPACEK, D. V.; SNYDER, M. P. High-throughput sequencing

technologies. **Molecular Cell**, v. 58, 2015. Disponível em: https://pubmed.ncbi.nlm.nih.gov/26000844/. Acesso em: 25 fev. 2024.

RONAGHI, M.; UHLÉN, M.; NYRÉN, P. A sequencing method based on real-time pyrophosphate. **Science**, v. 281, n. 5375, p. 363–365, 1998. Disponível em: https://pubmed.ncbi.nlm.nih.gov/9705713/. Acesso em: 25 fev. 2024.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences of the United States of America**, v. 74, n. 12, p. 5463–5467, 1977. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC431765/. Acesso em: 25 fev. 2024.

SATAM, H. et al. Next-generation sequencing technology: current trends and advancements. **Biology**, v. 12, n. 7, p. 997, 2023. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC10376292/. Acesso em: 25 fev. 2024.

SCHMIEDER, R.; EDWARDS, R. Quality control and preprocessing of metagenomic datasets. **Bioinformatics**, v. 27, p. 863-864, mar. 2011. Disponível em: https://academic.oup.com/bioinformatics/article/27/6/863/236283. Acesso em: 25 fev. 2024.

TURCHETTO, Z. , et al. Marcadores Moleculares na Era Genômica: Metodologias e Aplicações. Ribeirão Preto: **Sociedade Brasileira de Genética**, 2017. 181 p. ISBN 978-85-89265-26-3

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-seq: a revolutionary tool for transcriptomics. **Nature Reviews Genetics**, v. 10, n. 1, p. 57-63, 2009. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2949280/. Acesso em: 25 fev. 2024.

WINGENDER, E.; CRASS, T. et al. Integrative content-driven concepts for bioinformatics "beyond the cell". **Journal of Biosciences**, v. 32, p. 80-169, 2007. Disponível em: https://doi.org/10.1007/BF02779119. Acesso em: 25 fev. 2024.

WORLD HEALTH ORGANIZATION. Global genomic surveillance strategy for pathogens with pandemic and epidemic potential 2022–2032. Geneva: **World Health Organization**, 2022. Disponível em: https://www.who.int/initiatives/genomic-surveillance-strategy . Acesso em: 15 jan. 2025.

ZHBANNIKOV, IL; HUNTER, S.; FOSTER, J.; SETTLES, M.SeqyClean: A Pipeline for High-throughput Sequence Data Preprocessing. **ACM-BCB**. P. 20-23, 2017. Disponível em: https://dl.acm.org/doi/epdf/10.1145/3107411.3107446. Acesso em: 09 out. 2024