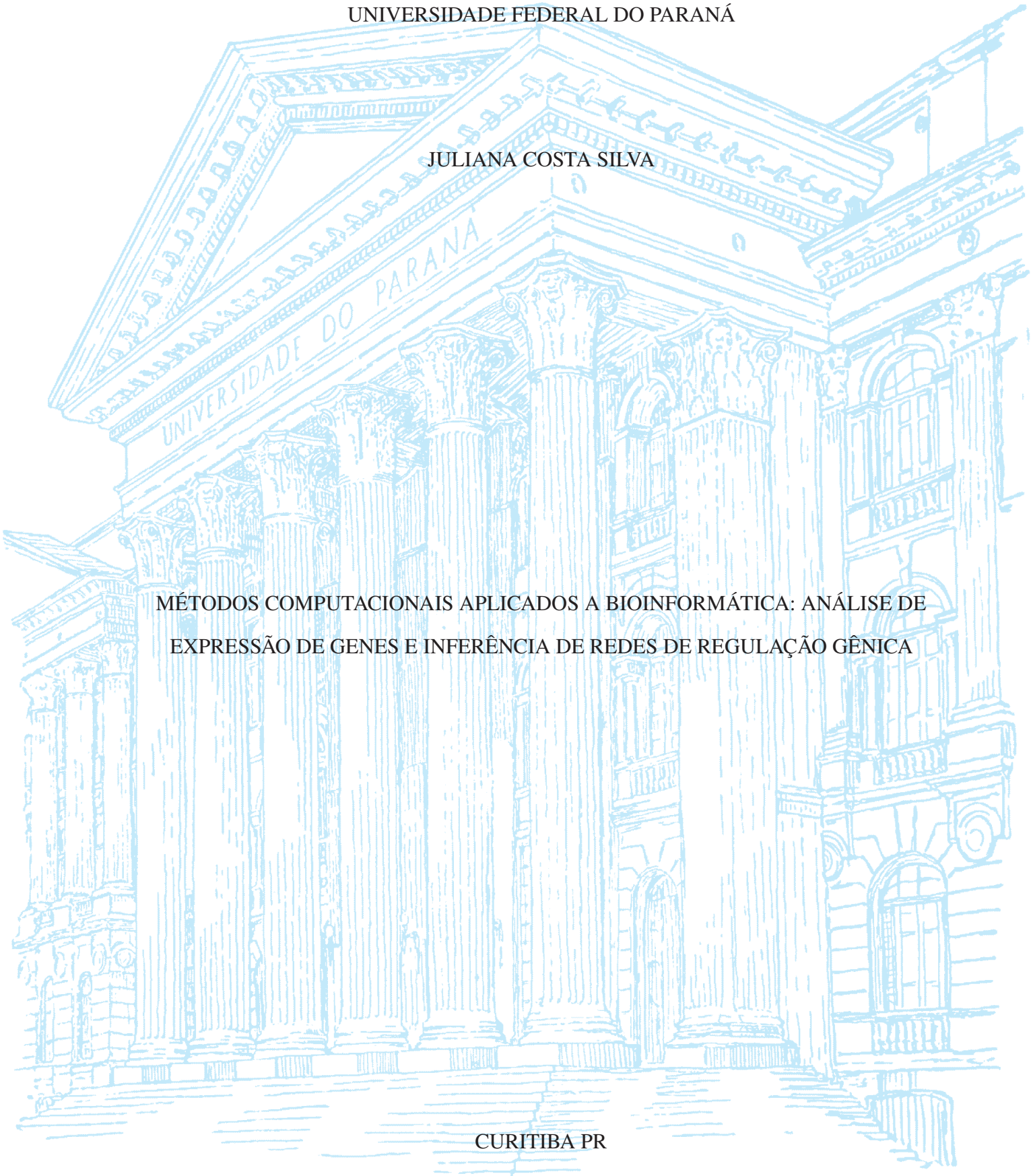
UNIVERSIDADE FEDERAL DO PARANÁ

JULIANA COSTA SILVA



MÉTODOS COMPUTACIONAIS APLICADOS A BIOINFORMÁTICA: ANÁLISE DE EXPRESSÃO DE GENES E INFERÊNCIA DE REDES DE REGULAÇÃO GÊNICA

CURITIBA PR

2025

JULIANA COSTA SILVA

MÉTODOS COMPUTACIONAIS APLICADOS A BIOINFORMÁTICA: ANÁLISE DE EXPRESSÃO DE GENES E INFERÊNCIA DE REDES DE REGULAÇÃO GÊNICA

Tese apresentada como requisito parcial à obtenção do grau de Doutora em Computação no Programa de Pós-Graduação em Informática, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Computação*.

Orientador: Profº. Drº. David Menotti.

Coorientador: Profº. Drº. Fabricio M. Lopes.

CURITIBA PR

2025

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA CIÊNCIA E TECNOLOGIA

Silva, Juliana Costa
Métodos computacionais aplicados a bioinformática: análise de expressão de genes e inferência de redes de regulação gênica. / Juliana Costa Silva. – Curitiba, 2025.
1 recurso on-line : PDF.

Tese (Doutorado) - Universidade Federal do Paraná, Setor de Ciências Exatas. Programa de Pós-graduação em Informática.

Orientador: Profº. Drº. David Menotti.
Coorientador: Profº. Drº. Fabricio Martins Lopes.

1. Redes reguladoras de genes. 2. Genes (Expressão diferencial de). 3. Biologia computacional – Métodos. I. Universidade Federal do Paraná. II. Programa de Pós-Graduação em Informática. III. Menotti, David. IV. Lopes, Fabricio Martins. V. Título.

Bibliotecária: Roseny Rivelini Morciani CRB-9/1585

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIÊNCIAS EXATAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO INFORMÁTICA - 40001016034P5

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação INFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **JULIANA COSTA SILVA**, intitulada: **Métodos Computacionais Aplicados a Bioinformática: Análise de Expressão de Genes e Inferência de Redes de Regulação Genica**, sob orientação do Prof. Dr. DAVID MENOTTI GOMES, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de doutora está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 30 de Maio de 2025.

Assinatura Eletrônica
11/06/2025 09:46:03.0
DAVID MENOTTI GOMES
Presidente da Banca Examinadora

Assinatura Eletrônica
17/06/2025 09:29:05.0
LUCAS FERRARI DE OLIVEIRA
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
13/06/2025 19:50:58.0
MÁRCIO DORN
Avaliador Externo (UNIVER. FEDERAL DO RIO GRANDE DO SUL)

Assinatura Eletrônica
11/06/2025 09:17:05.0
MAURO ANTONIO ALVES CASTRO
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ - UFPR)

Assinatura Eletrônica
11/06/2025 08:17:42.0
FABRICIO MARTINS LOPES
Coorientador(a) (UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ - CAMPUS CORNÉLIO PROCÓPIO)

A todos que, assim como eu, acreditam que a educação e a ciência tem o poder de transformar vidas e mudar o mundo.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por todas as pessoas que colocou em meu caminho antes e durante o doutorado. Ter boas pessoas ao meu redor foi essencial na minha vida, e imprescindível para o desenvolvimento desta tese.

Agradeço aos meus pais, **Terezinha de Jesus Costa Silva** e **Osvaldo Ferreira da Silva**, por toda a dedicação que tiveram durante a minha vida, e em especial durante o doutorado, o apoio e alegria de vocês a cada conquista é um dos meus maiores motivadores na vida acadêmica.

Agradeço ao meu esposo **Eliel Fernandes Batista**, por todo o apoio, incentivo, abnegações, cuidado e por ser meu porto seguro durante as dificuldades da pós-graduação. Também deixo aqui registrada a minha gratidão a todos os que de alguma forma me escutaram e apoiaram nesse período, pessoal ou profissionalmente.

Ao **Profº. Drº David Menotti**, meu orientador, pelo apoio, compreensão, aconselhamentos e estímulo constante, atitudes que tanto contribuíram para o desenvolvimento desta tese.

Ao **Profº. Drº Fabrício M. Lopes**, meu co-orientador, professor da Universidade Tecnológica Federal do Paraná - Campus Cornélio Procópio, pelo apoio, orientações assertivas, confiança depositada e tempo empenhado na orientação desta tese. Sem o olhar atento do Profº. Drº Fabrício, ao potencial de seus alunos, minha vida acadêmica sequer teria começado.

Agradeço a **Universidade Federal do Paraná** especialmente a equipe da secretaria do Programa de Pós-Graduação em Informática, pela presteza nos atendimentos e orientações. Agradeço a Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa parcial, concedida pelo processo SCBA 88887.516979/2020-00.

RESUMO

Compreender a complexa rede de interações que forma e mantém um organismo é uma tarefa desafiadora, composta por múltiplos passos, muitos dos quais envolvem análises computacionais. A identificação de genes envolvidos em um processo biológico representa um desses passos essenciais para o entendimento dessa rede de interações, podendo incluir o sequenciamento de nova geração (RNA-Seq), a identificação de genes diferencialmente expressos e a inferência de redes de regulação gênica. Contudo, a vasta quantidade de metodologias computacionais existentes comumente gera dúvidas quanto à definição de um *pipeline* e à seleção do método mais adequado para cada tipo de análise no contexto biológico. Além da análise de genes diferencialmente expressos, pode ser necessário identificar a rede de regulação desses genes, a qual indica quais genes são possíveis agentes de aumento da expressão ou de silenciamento de outros. Esta indicação é feita utilizando grafos, nos quais as arestas indicam a possível influência da expressão de um gene sobre a expressão de outro. Para esta tarefa, igualmente computacional, existe uma ampla gama de métodos disponíveis. Considerando as ferramentas computacionais desenvolvidas para a análise de expressão diferencial de genes, mesmo que apenas as especificamente desenvolvidas para dados de RNA-Seq, identifica-se um grande volume de metodologias, porém não é trivial encontrar uma categorização e/ou detalhamento das estratégias utilizadas em cada uma delas. Ainda neste contexto, a identificação da rede de relações entre genes é geralmente responsável por definir características ou respostas em diferentes organismos, entretanto a inferência de redes de regulação gênica ainda é uma atividade desafiadora, uma vez que as metodologias atuais apresentam baixas taxas de recuperação dessas relações. A identificação de lacunas das metodologias atualmente disponíveis para inferência de redes de regulação gênica pode indicar caminhos de melhorias para novos métodos. Esta tese selecionou e classificou as metodologias computacionais para análise de expressão gênica com dados de RNA-Seq mais relevantes desde a popularização do RNA-Seq até os dias atuais. Como resultado desta classificação, foi observado que, dentre as metodologias para análise de expressão, mais de 30% são desenvolvidas com dependência, total ou parcial de outras metodologias. Além disso, foi desenvolvido um pacote R para a análise de expressão que indica genes diferencialmente expressos com base no consenso entre várias metodologias. Considerando a análise das metodologias para a inferência de redes de regulação gênica, foram avaliadas 10 metodologias, sendo identificado que algumas das interações entre fatores de transcrição e genes não são detectadas, mesmo quando várias metodologias são empregadas conjuntamente para essa finalidade. Esta tese também caracterizou as relações encontradas, as não encontradas e as exclusivamente identificadas pelas metodologias utilizadas, para tanto utilizamos a entropia do sinal de expressão de cada gene da relação. Além disso foi possível definir um intervalo de valores de entropia para as arestas não detectadas, o que pode apoiar trabalhos futuros.

Palavras-chave: Bioinformática; Expressão diferencial de genes; Métodos computacionais; Redes regulatórias.

ABSTRACT

Understanding the complex network of interactions that form and sustain an organism is a challenging task involving multiple steps, many of which rely on computational analyses. Identifying genes involved in a specific biological process is one of these essential steps, often requiring next-generation sequencing (RNA-Seq), differential gene expression analysis, and the inference of gene regulatory networks (GRNs). However, the wide array of available computational methodologies frequently raises questions regarding the construction of an appropriate analysis pipeline and the selection of the most suitable tools for a given biological context. Besides identifying differentially expressed genes (DEGs), it may be necessary to infer the regulatory network underlying their expression—i.e., to determine which genes may upregulate or suppress others. This inference is commonly represented as a graph, where edges suggest potential regulatory influence between genes. Similar to DEG analysis, diverse methods have been proposed to handle the computationally intensive task of GRNs inference. Despite the abundance of tools developed specifically for RNA-Seq-based differential expression analysis, categorizing and understanding the methodological strategies behind them remains non-trivial. Furthermore, the reconstruction of gene regulatory networks is critical for elucidating the mechanisms driving phenotypic traits or responses across different organisms, yet it remains a challenging endeavor. Current GRN inference approaches often yield low recall rates, failing to recover many true regulatory interactions. Identifying the limitations of these existing methods may help guide the development of more effective approaches. This thesis systematically selected and categorized the most relevant computational methods for RNA-Seq-based gene expression analysis since the widespread adoption of RNA-Seq technology. The analysis revealed that over 30% of DEG analysis tools rely, either partially or entirely, on pre-existing methods. Additionally, an R package was developed to identify differentially expressed genes based on consensus across multiple methodologies. For the GRN inference component, ten methods were evaluated, and the results showed that certain transcription factor–gene interactions were not recovered, even when combining multiple inference strategies. This thesis also characterized the recovered, not recovered, and uniquely inferred regulatory interactions using the entropy of gene expression signals. Furthermore, this study identified a specific entropy range for undetected edges, which may support future investigations into the properties of missing regulatory links.

Keywords: Bioinformatics; Differetnial expression gene; Computational methods; Regulatory networks.

LISTA DE SIGLAS

ACC	Acurácia, do inglês *Accuracy*.
BS	Biologia de sistemas
CLR	Probabilidade de relação entre contextos, do inglês *Context Likelihood of Relatedness*
DBN	distribuição binomial negativa
DEGs	Differentrial Expressed Genes
ENCODE	Enciclopédia de elementos do DNA, do inglês Encyclopedia of DNA Elements
FN	Falso Negativo, do inglês *False Negative*.
FP	Falso Positivo, do inglês *False Positive*.
GEO	Gene Expression Omnibus
GFF	Formato Geral de Características, do inglês *General Feature Format*
GO	Consórcio de ontologia de genes, do inglês *Gene Ontology Consortium*
GRNs	Redes Regulatórias de Genes, do inglês *Gene Regulatory Networks*
GTF	Formato Geral de Transferência, do inglês *General Transfer Format*
MI	Informação Mútua, do inglês *Mutual Information*
MIM	Matriz de Informação Mútua, do inglês *Mutual Information Matrix*
NCBI	Centro Nacional para informação em Biotecnologia, do inglês *National Center for Biotechnology Information*
NGS	Sequenciadores de nova geração, do inglês: *Next-Generation Sequencing*
RNA	Ácido Ribonucleico, do inglês *Ribonucleic Acid*
SAGE	Análise em Série de Expressão de Genes, do inglês Serial Analysis of Gene Expression
SARS-CoV-2	Coronavirus 2 da síndrome respiratória aguda grave, do inglês *Severe Acute Respiratory Syndrome — Related Coronavirus 2*

TFs	Fatores de transcrição, do inglês *Transcriptor Factors*
TMM	Média ajustada de valores M, do inglês *Trimmed Mean of M-values*.
TN	Verdadeiro Negativo, do inglês *True Negative*.
TP	Verdadeiro Positivo, do inglês *True Positive*.
cDNA	DNA complementar, do inglês *complementary DNA*, DNA sintético que é transcrito a partir de uma molécula de mRNA
mRNAs	RNAs mensageiros
minet	Redes de Informação Mútua, do inglês *Mutual Information NETworks*
qPCR	Reação em Cadeia da Polimerase Quantitativa, do inglês *Quantitative Polimerase Chain Reaction*

SUMÁRIO

1 INTRODUÇÃO

A intersecção entre a biologia molecular e a ciência da computação tem proporcionado avanços notáveis na compreensão da complexidade biológica de organismos (Gahlawat et al., 2023). No entanto, essa intersecção também apresenta desafios significativos. Um dos desafios que veem sendo superado é o sequenciamento do genoma de várias espécies como commodities agrícolas: milho (espécie *Zea mays L.*) (Hansey et al., 2012), café (espécie *Coffea*) (Salojärvi et al., 2024), patógenos de doenças infecciosas em humanos: chikungunya (Sahadeo et al., 2017), SARS-Cov-2 (Wu et al., 2020). Estes sequenciamentos produzem um grande volume de dados biológicos em formato digital, no entanto, entender ou prever o comportamento fisiológico dos seres vivos a partir desses dados, pode ser uma tarefa complexa (Alam et al., 2024), e geralmente utiliza processamento computacional.

Neste contexto, o desenvolvimento de algoritmos que atendam a grande diversidade de análises que a genética demanda é fundamental. Esses algoritmos são utilizados para analisar o grande volume de dados digitais gerados por sequenciadores de DNA/RNA, como exemplo podemos utilizar o conjunto de dados gerados em um experimento com humanos, envolvendo células cerebrais e um mix de tecidos (Bullard et al., 2010), que gerou mais de 6 GB de dados de sequenciamento. Devido à variedade de possibilidades nas reações de um organismo a diferentes estímulos do meio em que vive, a computação também tem esforçado-se no sentido de viabilizar a implementação de modelos que ajudam na identificação da dinâmica de sistemas biológicos. Essa abordagem é conhecida como biologia de sistemas (Klipp et al., 2016; Hillmer, 2015).

Organismos multicelulares podem ser compreendidos como conjuntos de sistemas biológicos que atuam de forma coordenada e interdependente para garantir a manutenção da vida. Esses sistemas são compostos por diferentes tecidos, os quais, por sua vez, são formados por células especializadas. Em organismos eucariotos, todas as células compartilham o mesmo material genético — o DNA (ácido desoxirribonucleico) presente no núcleo —, apesar de exercerem funções diversas e apresentarem características morfológicas distintas. Essa diversidade funcional, a partir de uma mesma informação genética, é possível graças a mecanismos complexos de regulação da expressão gênica, os quais controlam quais genes são ativados ou silenciados em cada tipo celular, em resposta a sinais internos e externos. Porém, apesar de todas as células possuírem o mesmo DNA, cada uma pode ter uma função específica (ação/influência) no organismo. Isso se deve aos genes expressos em cada uma das células. A expressão ocorre pelo processo de transcrição, no qual o DNA é transcrito em RNA (Ácido Ribonucleico) (Snustad et al., 2000). O RNA transporta as informações contidas nos genes para a maquinaria de síntese presente nos ribossomos, que estão fora do núcleo celular em organismos eucarióticos. Esta cadeia de reações representa o dogma central da biologia molecular, apresentado de forma ilustrada na Figura 1.1.

É conhecido que a maneira como um organismo se adapta a fatores ambientais é controlada por meio de processos que modulam a atividade de muitos de seus genes (Seki et al., 2002). Esses processos podem ser investigados por meio de técnicas como microarrays e RNA-Seq, que permitem identificar e quantificar os genes expressos em determinadas condições. Por exemplo, sob estresse hídrico ou salino, plantas ativam vias de sinalização que envolvem fatores de transcrição da família DREB/CBF, os quais regulam a expressão de genes como RD29A e COR15A, associados à tolerância ao estresse (Yamaguchi-Shinozaki e Shinozaki, 1994, 2006). Além disso, em situações de alta salinidade, a via SOS (*Salt Overly Sensitive*) é ativada para manter o equilíbrio iônico celular, regulando genes como SOS1. Esses mecanismos são

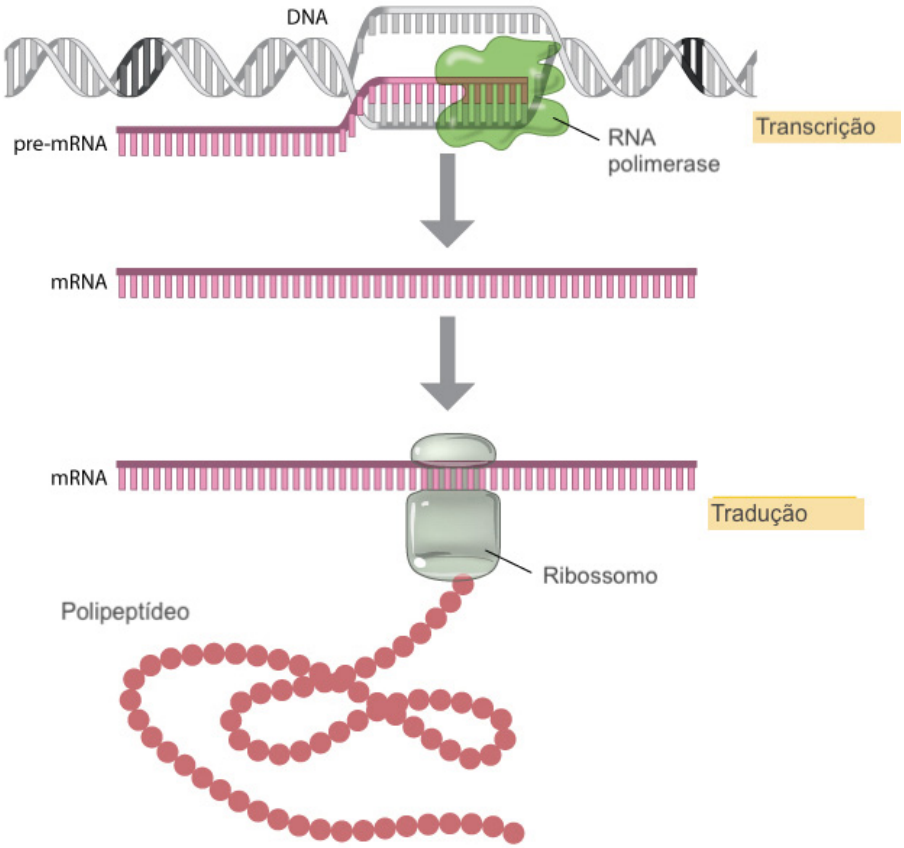


Figura 1.1: Dogma Central da Biologia Molecular. Esquema representando o processo de transcrição e tradução que ocorre durante a replicação do DNA.
Fonte: Traduzido de (Clancy e Brown, 2008).

geralmente mediados por hormônios como o ácido abscísico (ABA), que atua como sinalizador interno. Quando o organismo percebe alterações ambientais, seja por fatores externos (como frio ou seca) ou internos (como níveis hormonais), ocorre uma reprogramação da expressão gênica com o objetivo de proteger ou adaptar as células às novas condições (Alberts et al., 2009).

A "expressão" de um gene é o processo pelo qual um organismo produz uma proteína a partir da transcrição do DNA contido em um gene, este processo é representado na Figura 1.1. Isso envolve a produção de moléculas de RNA, que são a transcrição da informação do gene, e servem como um "molde" para a produção da proteína. Os genes que mostram uma grande mudança na quantidade de RNA produzido em resposta a diferentes situações são considerados "diferencialmente expressos" (Snustad et al., 2000). Portanto, para entender a funcionalidade de um organismo em nível molecular, precisamos entender a sua expressão.

Muitas técnicas foram desenvolvidas para quantificar a expressão de genes (Schena et al., 1995; Bustin, 2000; Bainbridge et al., 2006). De modo geral, essas técnicas tem como principais objetivos: i) estimar a quantidade de RNA produzida em uma determinada condição; ii) Quais genes são responsáveis por esta produção de RNA; e iii) Como é a produção do mesmo RNA em outras condições.

As primeiras metodologias desenvolvidas para quantificar expressão gênica (Velculescu et al., 1995; Bustin, 2000), dependem do conhecimento prévio dos genes de interesse (Wang et al., 2009). Os métodos iniciais também apresentam algumas particularidades. Uma delas é a necessidade de definir previamente quais genes devem ser observados no experimento. Nesse contexto, os experimentos conseguem verificar se alguns genes, selecionados antecipadamente,

estão envolvidos na produção de RNA em uma determinada condição. No entanto, todos os outros genes do organismo estudado não são avaliados, o que limita a possibilidade de novas descobertas.

Apesar dessa particularidade e de outras, esses métodos são considerados sensíveis e robustos, um exemplo é o qPCR (reação em cadeia da polimerase quantitativa) (Morgante et al., 2016), recentemente utilizado em testes rápidos para identificação da presença de cópia do material genético do vírus SARS-CoV-2 em pacientes sintomáticos de forma rápida e precisa (de Ciências UNICAMP, 2020). O qPCR também é amplamente utilizado como padrão ouro para estudos de análise de expressão (Corchete et al., 2020b).

Para superar as limitações das técnicas iniciais, foram desenvolvidos métodos de sequenciamento que não exigem a seleção de um gene específico para análise. Os dispositivos que executam essa técnica de sequenciamento são chamados de sequenciadores de nova geração (NGS) . Os dados do NGS tiveram um impacto positivo nos avanços nas análises de expressão gênica, permitindo até mesmo a identificação de genes diferencialmente expressos em organismos sobre os quais não haviam muitas informações genéticas disponíveis, i.e., sem o seu genoma sequenciado e anotado.

Essa capacidade de identificar genes diferencialmente expressos sem a necessidade de um genoma sequenciado e anotado abriu novas possibilidades para a pesquisa em biologia molecular e ciência da computação. Através da aplicação de técnicas de sequenciamento NGS, os pesquisadores agora podem sequenciar os RNAs mensageiros (mRNAs) que são produzidos por um organismo em um determinado tecido, ampliando o escopo de suas investigações.

A técnica de sequenciamento NGS, conhecida como RNA-Seq, pode ser aplicada a vários tipos de estudos genéticos, dentre eles, na visualização da atividade transcricional de um organismo em determinada condição ou tratamento (Zhang et al., 2014). Em estudos para identificação de genes diferencialmente expressos (DEGs) , as técnicas de sequenciamento NGS permitem que sejam sequenciados os mRNAs produzidos pelo organismo em um determinado tecido. Em um protocolo típico, os mRNAs são convertidos em DNA complementar (cDNA) (Wang et al., 2009), para que seja possível a aplicação da técnica RNA-Seq.

Devido ao grande volume de dados gerados pelos sequenciadores NGS, as análises de expressão são realizadas através de *software* e/ou pacotes. Essas análises envolvem no mínimo três passos que demandam muitos recursos computacionais: mapeamento, contagem e identificação de DEGs.

Para compreender o mapeamento, é necessário considerar que: no sequenciamento RNA-Seq, não há necessidade de conhecimento prévio do genoma ou seleção de genes de interesse, a técnica de RNA-Seq gera uma saída com todos os cDNAs transcritos pelo organismo na condição ou tratamento estudado. Nesse ponto, e em muitos outros da análise, a computação é uma grande aliada, visto que um sequenciador NGS pode gerar milhões de sequências por condição ou tratamento (Bullard et al., 2010). As sequências geradas (chamadas de *reads*) são fragmentos de cDNA, e não existe uma marcação da origem desses fragmentos. Portanto, é necessário identificar qual gene produziu cada fragmento para, então, contar o volume de transcrição (e/ou sua expressão) em determinada condição para cada gene.

Nesse contexto, o mapeamento de transcritos pode ser realizado de duas maneiras principais. i) Em organismos com o genoma sequenciado e anotado, os fragmentos de cDNA gerados pelo sequenciador são alinhados às regiões do genoma que apresentam alta similaridade, permitindo a identificação dos loci transcritos. ii) Em organismos sem genoma de referência disponível, os fragmentos de cDNA são utilizados para reconstruir os transcritos por meio de técnicas de montagem *de novo*. Essas técnicas consistem em métodos computacionais que permitem a reconstrução das sequências de RNA a partir dos fragmentos de cDNA, que podem

ser comparados a peças de um quebra-cabeça. A montagem *de novo* busca encaixar essas peças de forma a reconstituir os transcritos originais. Mais detalhes sobre essas técnicas podem ser encontrados em (Li e Dewey, 2011; Trapnell et al., 2012; Frazee et al., 2015).

A partir do mapeamento, realiza-se a contagem dos cDNAs alinhados às regiões genômicas anotadas como genes, gerando-se uma tabela de contagem de transcritos para cada gene em cada condição experimental. Em experimentos com organismos sem genoma de referência, após a reconstrução dos transcritos, os fragmentos de cDNA são mapeados de volta às sequências geradas na montagem *de novo*. Esse mapeamento permite identificar quais transcritos estão presentes nas amostras e estimar sua respectiva abundância. Abordagens e ferramentas voltadas à contagem de *reads* em contextos de mapeamento podem ser encontradas em (Anders et al., 2015; Liao et al., 2019, 2014).

Com os dados de contagem em mãos, diversas análises podem ser conduzidas, sendo uma das mais comuns a identificação de genes diferencialmente expressos (DEGs) entre diferentes condições. Um exemplo é o estudo de Cai et al. (2020), no qual foi analisada a expressão gênica em tecidos pulmonares de indivíduos classificados como não fumantes, ex-fumantes e fumantes atuais, com o objetivo de investigar alterações na expressão de genes associadas ao tabagismo. Os resultados mostraram que o gene ACE2, responsável por codificar o receptor do vírus SARS-CoV-2, apresentou regulação positiva (i.e., aumento de expressão) em fumantes, evidenciando uma possível ligação entre o hábito de fumar e maior susceptibilidade à infecção viral.

Entre a contagem de *reads* mapeados e a inferência de expressão diferencial, é necessário considerar fatores que podem introduzir viés nas análises. Um dos principais aspectos é o tamanho dos genes. Por exemplo, suponha que o gene A possua 1500 pares de base (pb), enquanto o gene B possua apenas 300 pb. Considerando que os sequenciadores podem gerar *reads* com tamanhos entre 35 e 150 pb, uma contagem de 100 *reads* para o gene A e 20 *reads* para o gene B não implica necessariamente maior expressão de A, pois o número de *reads* pode ser proporcional ao tamanho do gene. Para evitar interpretações incorretas, é fundamental aplicar técnicas de normalização dos dados de contagem. Diversas abordagens têm sido propostas para esse fim, conforme discutido por Wagner et al. (2012); Bullard et al. (2010).

Diferentes maneiras de normalizar dados de contagem foram propostas, porém a decisão a respeito de qual metodologia de normalização utilizar, esta diretamente atrelada aos dados e modelo do experimento. O artigo dos autores Zhao et al. (2021b), intitulado: "*TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository*", faz uma comparação entre as metodologias de normalização, dentre elas: TPM (do inglês, *transcript per million*), RPKM (do inglês, *reads per kilobase of transcript per million reads mapped*), e FPKM (do inglês, *fragments per kilobase of transcript per million reads mapped*).

Com os dados de contagem normalizados, a análise chega ao seu último passo, identificar se a contagem de *reads* de um gene, quando observada em diferentes condições, possui variação relevante, que possa indicar um comportamento diferencial, ou seja, uma variação de queda ou aumento acima do que é comumente observado nos dados. Para tanto as metodologias mais utilizadas, aplicam uma distribuição aos dados de contagem de genes vs. condições. Como resultado, os genes observados fora/ distantes desta distribuição, são considerados diferencialmente expressos (de modo geral o limiar dessas considerações pode ser parametrizado).

Desde a popularização do RNA-Seq como técnica de análise transcriptômica a partir de 2008, uma ampla gama de métodos computacionais foi desenvolvida para a identificação de genes diferencialmente expressos (DEGs) (Overbey et al., 2021). Embora esse avanço tenha

ampliado significativamente as possibilidades analíticas, ele também introduziu novos desafios, especialmente no que se refere à definição de *pipeline* de análise robustos e reprodutíveis. Em cada etapa do processo — do mapeamento e quantificação à normalização e análise estatística — existe uma multiplicidade de ferramentas disponíveis, cada qual baseada em pressupostos estatísticos e estratégias distintas. Assim, a escolha dos métodos deve ser orientada não apenas pela popularidade ou facilidade de uso, mas principalmente pela adequação ao desenho experimental, ao tipo de dado e à hipótese biológica em estudo.

No contexto da identificação de DEGs, os métodos disponíveis podem ser classificados, de forma geral, em três grandes categorias: paramétricos, não-paramétricos e híbridos. Os métodos paramétricos, como os implementados em ferramentas como edgeR e DESeq2, assumem que os dados seguem distribuições estatísticas específicas — frequentemente a distribuição binomial negativa — e modelam a variabilidade dos dados com base em estimativas de dispersão. Já os métodos não-paramétricos, como SAMseq e abordagens baseadas em permutação, não fazem suposições explícitas sobre a distribuição dos dados, sendo mais apropriados em cenários com poucos replicatas ou com distribuição assimétrica dos dados. Por fim, os métodos híbridos combinam características de ambos os paradigmas, buscando maior flexibilidade analítica; eles geralmente integram modelagens paramétricas com ajustes não-paramétricos, ou utilizam heurísticas para estimar significância de forma mais robusta em situações de alta variabilidade biológica.

Essa categorização não apenas contribui para uma melhor compreensão das estratégias existentes, como também orienta escolhas metodológicas mais informadas, especialmente em estudos que envolvem múltiplas condições experimentais, baixo número de replicatas biológicas ou dados com alto ruído técnico. Assim, conhecer as bases teóricas e computacionais que sustentam cada tipo de abordagem é fundamental para garantir inferências estatísticas confiáveis e biologicamente relevantes.

Uma ferramenta computacional que integre todas as etapas do *pipeline* para identificação de *DEGs* pode oferecer diversas vantagens, como maior reprodutibilidade, automação do fluxo de trabalho e padronização das análises. No entanto, a complexidade envolvida na integração dessas etapas — que vão desde o pré-processamento das leituras até a análise estatística dos resultados — implica a necessidade de manipulação de múltiplos formatos de arquivos e a coordenação de diferentes módulos computacionais. Consequentemente, o número elevado de dependências pode tornar essas ferramentas mais suscetíveis a problemas de manutenção, como conflitos entre versões, descontinuidade de pacotes ou incompatibilidades em atualizações de bibliotecas. Esses fatores podem comprometer a estabilidade da ferramenta ao longo do tempo e demandar esforços adicionais de configuração por parte do usuário.

A primeira parte desta tese consistiu no desenvolvimento de uma "revisão das metodologias computacionais para identificação de genes diferencialmente expressos (*DEGs*) (Costa-Silva et al., 2023)". Essas metodologias foram classificadas com base em sua abordagem estatística predominante — *paramétrica*, *não paramétrica*, *híbrida* — além de incluir um grupo de métodos considerados *seminais*, por seu impacto histórico ou inovação conceitual. A revisão também explorou as relações de dependência e reutilização entre diferentes métodos, evidenciando uma rede interligada de desenvolvimento de ferramentas na área.

Os resultados dessa análise indicaram que, além dos métodos amplamente adotados baseados em modelos *paramétricos*, há uma variedade significativa de abordagens alternativas, incluindo aquelas que adotam estratégias *não paramétricas* ou combinações híbridas. Um aspecto notável identificado foi a elevada taxa de reutilização de componentes metodológicos entre diferentes ferramentas, o que aponta para um ecossistema altamente modular e interdependente. Também foi observada uma diversidade nos formatos de disponibilização dessas metodologias,

que podem ser distribuídas como pacotes, bibliotecas, *software* para instalação local ou como plataformas *web*. Entre essas opções, destacam-se as ferramentas disponíveis em repositórios especializados em análises biológicas, como o *Bioconductor* (Huber et al., 2015), que concentram grande parte das metodologias mais utilizadas na área.

Dentre todas as metodologias analisadas, apenas uma foi classificada como *híbrida*: "a metodologia anteriormente proposta por esta autora, denominada *consexpression* (Costa-Silva et al., 2017a)", que integra os resultados de sete métodos distintos para identificação de *DEGs*. Essa estratégia visa combinar diferentes abordagens para obter maior robustez nos resultados, ao reduzir vieses associados ao uso de uma única metodologia.

Na segunda parte desta tese, avaliou-se o desempenho da estratégia de uso combinado de métodos de análise de expressão diferencial, com foco em seu impacto sobre métricas como acurácia e precisão. Essa avaliação teve como objetivo verificar se a integração de múltiplas abordagens melhora a confiabilidade dos resultados obtidos.

Com a finalidade de tornar essa abordagem mais acessível a usuários da linguagem R e ampliar seu escopo de aplicação, foi desenvolvida uma nova versão da metodologia *consexpression*, agora implementada como um pacote R. Esta nova versão, denominada *consexpressionR*, é compatível com análises realizadas com ou sem genoma de referência, possui interface gráfica para o usuário (*GUI*) e executa sete métodos distintos de análise de expressão. O pacote está disponível para instalação via *GitHub*, e sua documentação completa, incluindo manual de instalação, pode ser acessada em: https://costasilvati.github.io/consexpressionR/.

Indicar com assertividade quais são os genes diferencialmente expressos tem grande relevância quando se quer identificar quais genes estão associados a uma característica fenotípica de interesse. Apesar de apresentar resultados relevantes, métodos que analisam dados biológicos como componentes isolados não conseguiram desvendar os mecanismos que justificam seus resultados (Otero e Nielsen, 2010).

A análise de *DEGs* é fundamental para estudar os componentes de cada organismo de forma isolada. Porém, é necessário relacionar esse comportamento a uma cadeia de reações que pode ser gerada, a linha de expressão desses elementos geralmente reflete a função reguladora dos genes (Zhao et al., 2021a). A partir desta premissa, emergem na década de 1950 as primeiras concepções sobre a Biologia de Sistemas (BS) . Essa área, pautada nos conceitos de sistema e de complexidade, envolve um estudo sistemático de interações em um sistema biológico (Morin et al., 2014).

Uma definição para o termo biologia de sistemas foi apresentada por Edgar Morin e Jean-Louis Le Moigne em seu estudo *Biologia de Sistemas* (Morin et al., 2014), que faz parte do livro *Bioinformática: da Biologia à Flexibilidade*:

> A Biologia de Sistemas é um campo que investiga as interações entre os componentes de um sistema biológico, buscando contribuir para o entendimento de como essas interações influenciam a função e o comportamento do sistema.

A medição das interações entre os genes é uma tarefa desafiadora. Por outro lado, medir a abundância dos componentes (por exemplo, os níveis de mRNA) é consideravelmente mais direta. Os avanços das tecnologias de sequenciamento permitiram medições cada vez maiores da expressão gênica a custos cada vez menores. Essa tendência proporcionou uma motivação com o intuito de tentar reconstruir computacionalmente as estruturas de interação que fundamentam os padrões de expressão gênica: essas interações são coletivamente denominadas *Gene Regulatory Networks* (GRNs - Redes Regulatórias de Genes) (Lopes, 2011; Hashimoto et al., 2004a). A reconstrução dessas redes tem sido um esforço central do campo interdisciplinar da Biologia de Sistemas (Huynh-Thu e Sanguinetti, 2019).

Utilizando como base as tecnologias de sequenciamento citadas anteriormente e muitas outras, foram criados vários bancos de dados para registrar e categorizar interações biológicas, dentre eles:

- GEO - *Gene Expression Omnibus* : é um banco de dados aberto de registros de expressão gênica desenvolvido pelo NCBI (do inglês, *National Center for Biotechnology Information*) em 2000 (Barrett et al., 2011);
- RegulonDB: registra de forma ampla e detalhada informações sobre as relações regulatórias da *Escherichia coli* K-12 (Gama-Castro et al., 2011);
- KEGG: registra redes de interações moleculares e vias metabólicas (Kanehisa et al., 2016);
- GO - *Gene Ontology Consortium* : categoriza e cria ontologias para funções de genes e produtos gênicos (Blake et al., 2015);
- ENCODE - *Encyclopedia of DNA Elements* : registra elementos funcionais no genoma humano (de Souza, 2012).

Apesar de um número crescente de conexões regulatórias já ter sido registrado em diversas bases de dados, essas conexões ainda representam apenas uma fração das inúmeras interações e relações complexas que ocorrem nos sistemas biológicos (Maetschke et al., 2014). Esse cenário evidencia uma lacuna significativa entre o conhecimento atualmente consolidado e a totalidade das redes de regulação gênica que, de fato, operam nos organismos. Com o avanço contínuo das tecnologias de sequenciamento de alto desempenho (*high-throughput sequencing*) e de métodos experimentais, a quantidade de dados disponíveis sobre expressão gênica tem crescido exponencialmente, além de apresentar ampla diversidade de fontes, condições e organismos.

Diante desse grande volume de dados, torna-se impraticável validar experimentalmente, de forma individual, todas as possíveis conexões regulatórias utilizando exclusivamente recursos humanos (Zhao et al., 2021a). Esse desafio ressalta a importância das abordagens computacionais, que desempenham um papel essencial na triagem, inferência e priorização de interações gênicas com potencial relevância biológica, contribuindo para o avanço do conhecimento sobre redes de regulação gênica de maneira escalável e sistemática.

Diante das limitações mencionadas anteriormente — como a baixa recuperação de interações validadas e o volume crescente de dados — e considerando que as redes de regulação gênica (*GRNs*) são fundamentais para a representação e compreensão de sistemas biológicos complexos (Huynh-Thu e Sanguinetti, 2019; da Rocha Vicente e Lopes, 2014), diversas abordagens computacionais têm sido propostas com o objetivo de inferir tais redes (Lopes, 2011). Essas metodologias baseiam-se em pressupostos e estratégias distintas, refletindo a diversidade de perspectivas adotadas na área de inferência de redes de regulação gênica (*GRNs*) (Margolin et al., 2006a; Lopes et al., 2008; Sławek e Arodź, 2013; Mercatelli et al., 2020; Kuang et al., 2023).

A ferramenta *ARACNE* (*Algorithm for the Reconstruction of Accurate Cellular Networks*) propõe uma abordagem baseada em teoria da informação, utilizando *mutual information* para medir a dependência estatística entre pares de genes e aplicando o princípio de exclusão de informação (*Data Processing Inequality*) para remover interações indiretas, o que favorece a obtenção de redes mais parcimoniosas (Margolin et al., 2006a).

Em outra perspectiva, o método proposto por (Lopes et al., 2008) adota uma abordagem orientada à seleção de atributos (*feature selection*), combinando algoritmos genéticos e métodos

estatísticos para identificar subconjuntos de genes reguladores mais relevantes, com ênfase na redução da dimensionalidade e na explicação local das interações.

A ferramenta *ENNET* utiliza técnicas de aprendizado de máquina, mais especificamente *gradient boosting*, para inferir redes a partir de dados de expressão gênica em larga escala. Essa abordagem visa capturar relações complexas e não lineares entre genes, sendo adequada para cenários com alto volume de dados e múltiplas variáveis preditoras (Sławek e Arodź, 2013).

O pacote *corto*, por sua vez, apresenta uma implementação leve e eficiente de inferência de redes baseada em correlação e análise de reguladores mestres (*master regulators*). Ele utiliza uma versão simplificada de algoritmos de inferência baseados em teoria da informação, com foco em acessibilidade, performance e integração com o ecossistema *Bioconductor* (Mercatelli et al., 2020).

Por fim, a ferramenta *GeCoNet-Tool* propõe uma abordagem centrada na construção e análise de redes de coexpressão gênica, explorando métricas topológicas e propriedades estruturais para identificar padrões de regulação gênica. Seu foco está na análise global da rede e na identificação de módulos funcionalmente relevantes (Kuang et al., 2023).

Essas distinções entre as abordagens demonstram como diferentes pressupostos — desde a dependência estatística e seleção de variáveis até técnicas de aprendizado de máquina e análise topológica — moldam o modo como as *GRNs* são inferidas a partir de dados de expressão gênica. A diversidade metodológica reflete não apenas a complexidade dos sistemas biológicos, mas também a variedade de estratégias computacionais empregadas para enfrentá-la.

Enquanto algumas abordagens exploram características estruturais das *GRNs*, como sua topologia e conectividade (Lopes et al., 2010; da Rocha Vicente e Lopes, 2014; Lopes et al., 2014; Martins Jr et al., 2016), outras propõem modelos estatísticos e algoritmos mais robustos, capazes de oferecer inferências com maior confiabilidade (Mendoza et al., 2012).

Apesar desses avanços, a capacidade dos métodos disponíveis de recuperar interações biologicamente validadas ainda é limitada (Marbach et al., 2012; Hashimoto et al., 2004b). Ampliar o poder de identificação dessas interações permanece, portanto, como um desafio central e de grande relevância para o campo. Revisões recentes, como as realizadas por Martins Jr et al., 2016 e Marku e Pancaldi, 2023, reforçam essa limitação e destacam que, embora novos algoritmos venham sendo desenvolvidos, a acurácia geral das inferências ainda está aquém do desejado.

Nesse cenário, a terceira parte desta tese dedica-se à comparação entre diferentes métodos de inferência de *GRNs*, abrangendo tanto algoritmos amplamente utilizados quanto propostas mais recentes. Essa comparação busca avaliar o desempenho relativo de cada abordagem, assim como identificar interações consistentemente inferidas entre os métodos, aquelas nunca recuperadas e aquelas exclusivamente identificadas por uma ou poucas metodologias.

Conforme discutido ao longo deste trabalho, persistem lacunas importantes nas análises computacionais de expressão gênica, tanto no que diz respeito à integração de múltiplos métodos para a identificação de *DEGs*, quanto à acessibilidade e usabilidade das ferramentas disponíveis. A inferência de redes de regulação gênica surge, assim, como uma etapa subsequente e complementar aos estudos de expressão diferencial, oferecendo uma perspectiva mais abrangente sobre os mecanismos moleculares envolvidos.

Diante disso, esta tese tem como objetivo investigar tais lacunas, propondo soluções metodológicas e computacionais que contribuam para a robustez analítica, a reprodutibilidade dos resultados e a expansão das possibilidades de inferência em contextos biológicos diversos.

1.1 OBJETIVOS

Conforme apresentado, a análise de expressão gênica e a inferência de redes regulatórias constituem etapas essenciais para a compreensão de sistemas biológicos. Essas etapas são geralmente complementares. Os estudos, por meio da análise de expressão, buscam identificar genes envolvidos em um determinado processo biológico. A partir dos genes identificados como diferencialmente expressos, procura-se entender como a expressão desses genes influencia o sistema biológico em questão.

No contexto de análise de expressão, esta tese tem como principal objetivo fornecer um panorama histórico e temporal das principais metodologias computacionais para análise de expressão diferencial desenvolvidas e implementadas até o ano de 2022, além de categorizá-las, facilitando a escolha dos usuários.

Outro objetivo desta tese é implementar a segunda versão da metodologia para análise de *DEGs* híbrida "*consexpression* (Costa-Silva et al., 2017a)", inicialmente desenvolvida para análise de dados de expressão desde a etapa de mapeamento, contagem, análise de *DEGs*. O pacote R *consexpressionR* desenvolvido nesta tese indica genes diferencialmente expressos através do consenso de pelo menos cinco métodos, com base na teoria de conhecimento de multidão (Marbach et al., 2012), e em resultados prévios. O pacote R *consexpressionR*, conta com distribuição em repositórios oficiais de pacotes R, com uso facilitado através de vasta documentação e interface gráfica. Esta metodologia gera resultados com base em no mínimo cinco metodologias e permite a identificação de genes diferencialmente expressos de forma mais robusta, além de ser acessível a experimentos que possuam mapeamento e contagem, com ou sem genoma.

A indicação de *DEGs* com mais precisão permite avaliar o perfil transcricional de um organismo em diferentes condições. É possível avaliar a relação entre os níveis de expressão dos genes e detalhar como a atividade transcricional de um gene, ou um grupo de genes, influencia na expressão dos demais.

Esta tese também apresenta como objetivo a inferência de redes regulatórias. Mais especificamente se dedica a avaliação atualizada de métodos de inferência de *GRNs* e avaliação de eficiência. Além disso, propõe a identificação e a caracterização de sinais de expressão dos genes que os métodos de inferência não identificaram os relacionamentos. Logo, contribuindo na caracterização das relações e indicando possibilidade de melhorias nos métodos, bem como o desenvolvimento de novas abordagens.

1.2 CONTRIBUIÇÕES

Esta tese apresenta contribuições relevantes para o campo da bioinformática e, de forma indireta, para áreas como saúde, agronomia e biologia molecular. Isso se deve ao fato de que a identificação de genes diferencialmente expressos (*DEGs*) é uma etapa essencial em diversos tipos de análises genéticas, como a caracterização de perfis de expressão, a identificação de vias metabólicas e a investigação de mecanismos de defesa de organismos. Apesar de sua importância, a identificação de *DEGs*, por si só, raramente é suficiente para subsidiar intervenções práticas, como terapias ou manipulações genéticas direcionadas.

Nesse contexto, parte-se do princípio de que os genes operam em redes interconectadas de regulação. Portanto, uma intervenção sobre um gene específico — por exemplo, um *knockout*[1] — pode desencadear efeitos em cascata sobre outros genes. Antecipar essas consequências é

[1]Técnica da genética que consiste em bloquear a expressão de um gene específico, substituindo-o em seu *locus* original por uma versão modificada.

fundamental para compreender os impactos sistêmicos da manipulação gênica e orientar decisões experimentais ou terapêuticas com maior segurança.

Entretanto, prever os efeitos de uma intervenção exige, muitas vezes, a realização de uma grande quantidade de testes empíricos, o que implica altos custos e longos prazos. Esta tese propõe uma alternativa computacional a esse processo, por meio da aplicação de métodos de inferência de redes de regulação gênica (*GRNs*) capazes de estimar, com alto grau de confiabilidade, os padrões de interação entre genes. Ao empregar estratégias de análise *in silico*, torna-se possível simular cenários de intervenção e prever seus efeitos potenciais de forma mais rápida, econômica e escalável.

Com isso, esta tese contribui para a consolidação de ferramentas computacionais que apoiam a compreensão de sistemas biológicos complexos, auxiliando na priorização de alvos experimentais e no planejamento de intervenções mais assertivas, com implicações práticas para pesquisas biomédicas, agrícolas e ambientais.

1.3 ORGANIZAÇÃO DO DOCUMENTO

Esta tese apresenta uma revisão detalhada dos passos para a análise computacional de genes diferencialmente expressos, conforme descrito no Capítulo 2. Ainda no Capítulo 2, introduzimos os conceitos fundamentais para a compreensão dos métodos de inferência de redes regulatórias na Seção 4.3, bem como suas propriedades e possíveis aplicações biológicas. No Capítulo 3, descrevemos os dados e as metodologias empregadas para a obtenção dos resultados.

Os resultados alcançados são discutidos no Capítulo 4. No Capítulo 5, apresentamos uma breve discussão e por fim, no Capítulo 6, apresentamos as conclusões e as direções futuras do projeto.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os conceitos fundamentais que sustentam as análises desenvolvidas nesta tese, oferecendo uma contextualização sobre os métodos computacionais aplicados à análise de expressão diferencial de genes e à inferência de redes regulatórias de genes (*GRNs*). Para compreender adequadamente esses métodos, é indispensável conhecer a origem e as características dos dados utilizados como entrada nas análises.

A forma como os dados de expressão gênica são gerados — ou seja, a tecnologia empregada para sua obtenção — pode influenciar significativamente as etapas subsequentes da análise, assim como a interpretação dos resultados obtidos. Diferentes plataformas de sequenciamento, protocolos experimentais e estratégias de preparo de amostras impactam diretamente a estrutura, a complexidade e a qualidade dos dados. Por essa razão, torna-se necessário apresentar uma visão geral sobre os processos biológicos envolvidos e sobre os principais métodos de obtenção de dados de expressão.

A Seção 2.1 introduz o conceito de expressão gênica, com base no dogma central da biologia molecular, e descreve as principais técnicas empregadas para a quantificação dos níveis de expressão de genes ou transcritos. Em seguida, a Seção 2.2 detalha as etapas computacionais típicas dos fluxos de análise utilizados na identificação de genes diferencialmente expressos (*DEGs*), desde o pré-processamento até a inferência estatística. Por fim, a Seção 2.4 aborda os conceitos centrais relacionados às *GRNs* e descreve as metodologias computacionais avaliadas ao longo deste trabalho para a reconstrução dessas redes a partir de dados de expressão gênica.

2.1 DADOS DE EXPRESSÃO GÊNICA

O conjunto de moléculas de *DNA* que caracteriza um organismo constitui seu genoma. A informação contida nesse genoma está organizada em unidades estruturais denominadas cromossomos, os quais, por sua vez, são compostos por sequências menores chamadas genes. Em organismos procarióticos (como bactérias), os genes estão concentrados em um único cromossomo. Já em organismos eucarióticos (como humanos, plantas e outros), os genes encontram-se distribuídos entre múltiplos cromossomos localizados no núcleo celular.

Segundo (Zaha et al., 2014), do ponto de vista molecular, um gene — seja de um procarioto ou de um eucarioto — pode ser definido como toda a sequência nucleotídica necessária e suficiente para a síntese de um polipeptídeo ou de uma molécula de *RNA* estável. De acordo com essa definição, cada gene inclui uma região codificadora, responsável pela determinação da sequência de aminoácidos de uma proteína ou da sequência de um *RNA* funcional (como *rRNA* ou *tRNA*), bem como sequências regulatórias associadas à sua transcrição.

O conjunto completo de moléculas de *mRNA* expressas em uma célula ou em uma população celular é denominado *transcriptoma* (McGettigan, 2013). O termo foi inicialmente proposto por Charles Auffray em 1996 (Pietu et al., 1999) e utilizado pela primeira vez na literatura científica em 1997 por (Velculescu et al., 1997).

Embora a maior parte do *DNA* esteja localizada nos cromossomos presentes no núcleo celular, a síntese de proteínas ocorre majoritariamente no citoplasma. Nesse processo, o *RNA* atua como molécula intermediária, transferindo a informação genética do *DNA* nuclear para os ribossomos citoplasmáticos. A relação entre *DNA*, *RNA* e proteínas é tradicionalmente descrita pelo *dogma central da biologia molecular*, conforme ilustrado na Figura 1.1. Segundo esse modelo, o *DNA* é transcrito em *RNA*, e este, por sua vez, é traduzido em proteína.

Inicialmente, considerava-se que a relação entre gene e proteína era linear e sem ambiguidade, ou seja, um gene seria responsável pela codificação de uma única proteína. No entanto, com os avanços da genômica, verificou-se que essa relação é substancialmente mais complexa (Almeida et al., 2022). Os genes podem apresentar múltiplos sítios de início de transcrição, gerando diferentes variantes de transcritos. Adicionalmente, mecanismos como o *splicing* alternativo e a edição do pré-*mRNA* permitem a geração de múltimas isoformas proteicas a partir de um único gene (Najjar e Mustelin, 2023). Estima-se que mais de 50% dos genes humanos sejam capazes de produzir mais de uma proteína (Tao et al., 2024), o que contribui para que o *proteoma* humano seja significativamente mais complexo que seu *genoma*. Dados atuais indicam que cerca de 25.000 genes humanos codificam mais de 100.000 proteínas distintas (Zhao, 2012).

Alguns estudos têm como objetivo identificar os genes cuja atividade é aumentada ou silenciada em determinadas condições fisiológicas, ambientais ou experimentais. Esse processo é denominado análise de expressão diferencial de genes (*differential gene expression analysis*). Trata-se de uma tarefa essencial para compreender como organismos respondem a estímulos ou alterações em seu ambiente, sendo frequentemente aplicada em estudos sobre doenças, desenvolvimento celular e resistência a estresses.

Por envolver a quantificação da transcrição de *mRNA* em diferentes condições, essa análise demanda o processamento de grandes volumes de dados, o que torna indispensável o uso de métodos computacionais. Nesse contexto, os algoritmos desenvolvidos para análise de dados de expressão têm como objetivo estimar o nível de transcrição gênica e identificar variações estatisticamente significativas — ou seja, diferenças que não se devem ao acaso, mas que indicam uma possível associação funcional entre a expressão do gene e a condição biológica investigada. Tais variações são consideradas *relevantes* por poderem estar associadas a processos regulatórios, ativação de vias metabólicas ou respostas a estímulos internos e externos, e são frequentemente utilizadas para gerar hipóteses em estudos funcionais posteriores.

Nas próximas seções são apresentadas as técnicas de coleta de dados de expressão utilizadas por esta tese.

2.1.1 *PCR*

Com o avanço na compreensão dos conceitos de gene e genoma, diversas técnicas foram desenvolvidas para viabilizar o estudo da expressão diferencial de genes. As primeiras tentativas sistemáticas de identificação de perfis transcricionais em mamíferos remontam ao início da década de 1990 (Adams et al., 1991), com o uso da tecnologia de sequenciamento de Sanger, que possibilitou o desenvolvimento de métodos como o *SAGE* (*Serial Analysis of Gene Expression*) (Velculescu et al., 1995) . De forma paralela, estudos baseados na técnica de *microarray* também ganharam destaque (Schena et al., 1995), estabelecendo-se como a principal abordagem para análise de perfis de transcrição por vários anos.

Entre as técnicas mais amplamente empregadas, destaca-se a reação em cadeia da polimerase (do inglês *polymerase chain reaction*, *PCR*), desenvolvida por Kary Mullis na década de 1980 (Mullis e Faloona, 1987; Mullis, 1993). Trata-se de uma técnica amplamente utilizada para amplificar sequências específicas de *DNA in vitro*[1]. Dentre suas múltiplas aplicações, a *PCR* tem papel relevante na análise de expressão gênica, sendo também empregada como ferramenta de validação de resultados obtidos por outras metodologias computacionais ou experimentais.

Uma variação amplamente difundida da técnica original é a *qPCR* (*quantitative real-time PCR*), que incorpora a quantificação do *DNA* amplificado em tempo real ao longo dos ciclos

[1]Dispensando a necessidade de um organismo vivo para multiplicação do material genético

da reação (Ladeira et al., 2011). Segundo Morgante et al. (2016), a *qPCR* pode ser definida como: "A *PCR* combina a amplificação exponencial de um fragmento de *DNA* alvo específico com métodos de quantificação, por meio de medidas da fluorescência associada à síntese de um amplicon, ao longo dos ciclos da *PCR*".

A *qPCR* simula o processo de transcrição do *DNA*, promovendo a amplificação da sequência-alvo por meio de ciclos térmicos que ativam reações enzimáticas específicas. A cada ciclo, uma nova cópia da sequência de interesse é gerada, resultando em um aumento exponencial da quantidade total de *DNA* produzido. Esse crescimento é proporcional à abundância inicial da sequência, o que permite inferir a quantidade relativa de expressão de um gene.

Para ilustrar esse processo, considere um experimento hipotético com cinco genes, entre eles o *gene1*, responsável pela produção da *Proteína X*. Se o *gene1* originar três cópias de uma determinada sequência codificadora, a *qPCR* amplificará essas cópias a cada ciclo, dobrando sua quantidade: após o primeiro ciclo haverá seis cópias, doze no segundo, vinte e quatro no terceiro, e assim sucessivamente. Esse processo de amplificação continua até atingir um ponto de saturação — conhecido como fase de platô — no qual a produção de *DNA* se estabiliza e não aumenta significativamente com ciclos adicionais. Na Figura 2.1, essa fase é representada pela linha azul.

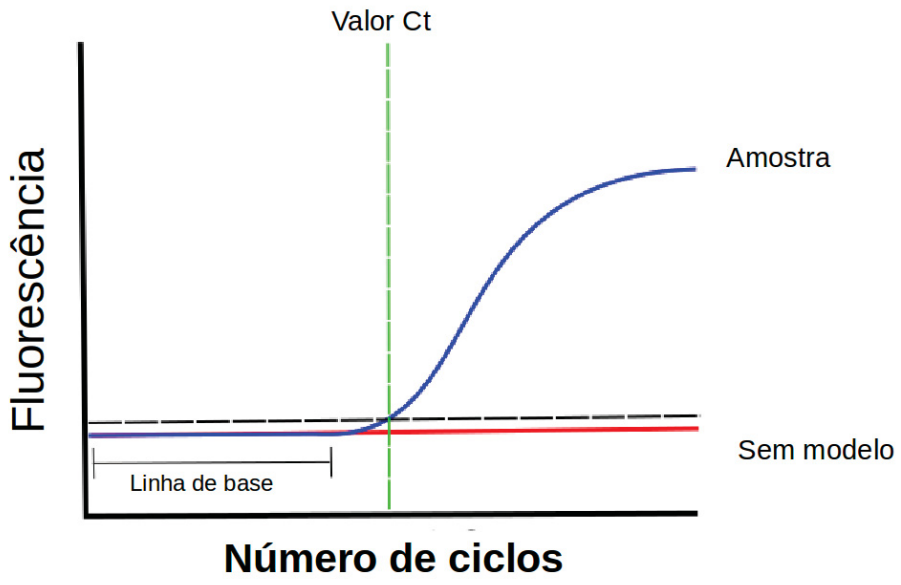


Figura 2.1: Gráfico de Limiar Curva *qPCR*

Gráfico de Limiar Curva *qPCR*: Nível de limiar em uma curva de amplificação *qPCR*. A linha vermelha indica o nível de fluorescência do controle utilizado (*Sem modelo*), a linha azul indica a curva de amplificação da sequência de interesse (*Amostra*), a linha tracejada em verde é o limiar dos ciclos da *qPCR* (C_t) **Fonte:** Adaptado de (Bustin et al., 2009).

A *PCR* emprega a técnica de transcrição reversa (*RT*) para converter moléculas de *RNA* mensageiro (*mRNA*) em *DNA* fita dupla, denominado *DNA* complementar (*cDNA*). Esse processo é viabilizado por meio de uma enzima transcriptase reversa, capaz de sintetizar *DNA* a partir de uma fita molde de *RNA*.

Na reação em cadeia da polimerase, o *cDNA* gerado é amplificado e monitorado em tempo real, o que caracteriza a técnica como *real-time PCR*. Para viabilizar esse monitoramento, a mistura reacional contém fluoróforos que emitem fluorescência proporcional à quantidade de *DNA* sintetizado, permitindo acompanhar o progresso da reação ciclo a ciclo (Morgante et al.,

2016). Dessa forma, a *qPCR* possibilita não apenas a detecção da presença de sequências gênicas específicas, mas também a estimativa da sua quantidade relativa na amostra biológica.

A análise da *qPCR* é realizada por equipamentos especializados, que capturam a emissão de fluorescência ao longo dos ciclos e geram representações gráficas da amplificação. A Figura 2.1 apresenta um exemplo simplificado desse gráfico, em que o eixo y representa a intensidade de fluorescência e o eixo x indica o número de ciclos. A curva azul representa a amplificação de uma amostra, enquanto a curva vermelha corresponde ao sinal basal de uma sequência controle (como uma reação sem *DNA*, uma sequência alvo conhecida ou uma referência interna). A linha verde pontilhada indica o limiar de detecção, também chamado de *threshold cycle* (C_t), definido como o ponto em que a fluorescência da amostra ultrapassa o sinal basal (Heid et al., 1996). A partir desse limiar, são realizados os cálculos de normalização e quantificação, permitindo estimar a abundância relativa dos transcritos. Em termos práticos, quanto menor o valor de C_t, maior a expressão da sequência alvo.

A técnica de *qPCR* é amplamente reconhecida por sua alta sensibilidade e especificidade na quantificação de transcritos (Wang e Brown, 1999). No entanto, algumas limitações devem ser consideradas, como a dependência de sequências-alvo previamente conhecidas para o desenho dos iniciadores, o número restrito de genes que podem ser analisados simultaneamente, além de questões associadas à reprodutibilidade dos experimentos, conforme apontado em (Rieu e Powers, 2009).

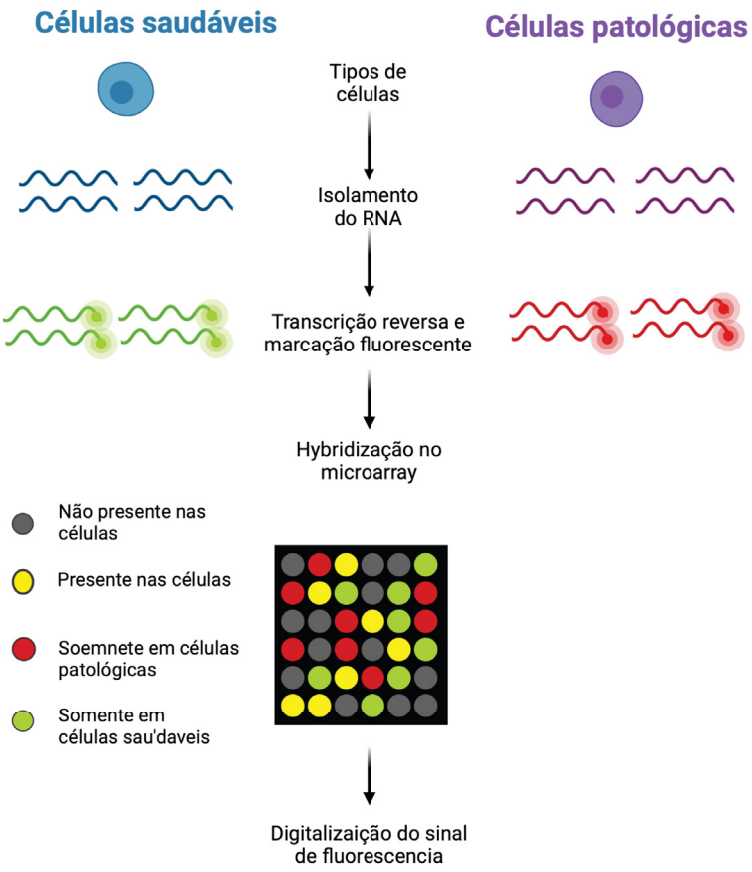
Nesta tese, dados obtidos por *qPCR* foram utilizados como padrão-ouro para validar metodologias computacionais de análise de *DEGs* com base em dados de *RNA-Seq*. Além disso, foi empregado um conjunto de dados obtido por *microarray* (Schena et al., 1995) para avaliar a capacidade preditiva dos métodos de inferência de *GRNs*. Para esse fim, são apresentadas nas seções seguintes descrições complementares sobre as abordagens de *microarray* e de *RNA-Seq*, a fim de contextualizar sua aplicação e impacto na análise computacional da expressão gênica.

2.1.2 *Microarray*

A tecnologia de *microarray* representou um marco na biotecnologia ao possibilitar a análise simultânea da expressão de milhares de segmentos de *DNA* gênico. Os *microarrays*, também denominados chips de *DNA*, consistem em lâminas sólidas sobre as quais segmentos de fita simples, denominados sondas, são fixados de maneira ordenada em regiões denominadas células de sonda. Cada célula contém múltiplas cópias de um transcrito ou segmento gênico específico, possibilitando sua posterior identificação (Guindalini e Tufik, 2007). Atualmente, chips tradicionais incluem representações de praticamente todos os genes do genoma de organismos modelo, como humanos, ratos, camundongos e *Drosophila melanogaster*, conforme descrito em (ThermoFisher, 2024).

A geração de dados de expressão por meio de *microarrays* tem início com a seleção de segmentos de *DNA* correspondentes a genes previamente identificados. Esses segmentos são amplificados por meio da técnica de *PCR* e fixados em uma superfície sólida. Ao término desse processo, obtêm-se múltiplas cópias dos segmentos de interesse, dispostas em locais específicos (os chamados *spots*) na lâmina do chip.

Nos experimentos com *microarrays*, a molécula de interesse é o *RNA* mensageiro, uma vez que se objetiva mensurar a expressão gênica em diferentes condições biológicas, como em situações fisiológicas normais ou em presença de patologias. Os *mRNAs* extraídos das amostras são convertidos em *cDNAs* por meio de transcrição reversa. Posteriormente, esses *cDNAs* são marcados com fluoróforos que emitem fluorescência quando excitados por luz em comprimento de onda específico. Essa marcação pode ser feita com fluoróforos distintos para diferentes



	Controle1	Controle2	Tratamento1	Tratamento2
Gene 1	0.00	1.988	0.7888	0.444
Gene 2	0.755	2.888	2.956	3.998
Gene 3	0.555	1.000	4.777	3.787

Figura 2.2: Esquema geral de um experimento de *microarray*.
Esquema geral de um experimento de *microarray*. O *mRNA* é extraído tanto das células saudáveis (controle) quanto das células patológicas (caso de interesse). Utilizando a transcriptase reversa, o *mRNA* é transformado em cDNA, que é marcado com fluoróforos com cores diferentes (uma cor para o controle e outra cor para o caso de interesse). O *cDNA* é então exposto ao *microarray* , no qual estão os genes de interesse (*spots*). O *microarray* é digitalizado e os sinais fluorescentes dos *spots* são convertidos para uma escala numérica de intensidade.
Fonte: Adaptado de Sagar Aryal utilizando biorender.com.

amostras, como amostras controle e tratadas, permitindo comparações diretas entre os perfis de expressão.

Os *cDNAs* marcados são hibridizados com as sondas presentes no *microarray*. A intensidade da fluorescência observada em cada *spot* é proporcional à quantidade de *mRNA* presente na amostra original, fornecendo, assim, uma estimativa do nível de expressão de cada gene em diferentes condições experimentais (Lopes, 2011). A Figura 2.2 ilustra o processo descrito.

Embora técnicas como a *qPCR* apresentem alta acurácia na quantificação de transcritos, sua capacidade de análise simultânea é limitada a poucos genes por experimento. O *microarray*, por sua vez, superou essa limitação ao permitir a avaliação de milhares de genes em uma única

análise. No entanto, sua aplicabilidade permanece restrita ao conhecimento prévio das sequências de *DNA* dos genes de interesse, uma vez que depende da hibridização com sondas previamente desenhadas.

A crescente demanda por abordagens mais abrangentes e menos dependentes de informações prévias impulsionou o desenvolvimento de metodologias mais modernas, como o sequenciamento de nova geração (NGS) para análise de expressão gênica. Dentre essas, destaca-se a tecnologia de *RNA-Seq*, que será abordada na seção seguinte.

2.1.3 *RNA-Seq*

Com a demanda crescente na geração de dados de expressão gênica em larga escala, houve um avanço com as técnicas de sequenciamento "de nova geração", conhecidas como *NGS* (*Next-Generation Sequencing*), dentre elas o *RNA-Seq* (Mortazavi et al., 2008), o qual apresenta a característica de não requerer o conhecimento prévio da sequência de *DNA* dos genes de interesse. As primeiras técnicas de sequenciamento *NGS* passaram a gerar sequências curtas de *mRNA* (chamados de *reads*, ou leituras em tradução literal), em grande quantidade. Os aparelhos sequenciadores que executam este tipo de sequenciamento também são chamados sequenciadores de nova geração ou de alto rendimento.

A metodologia *RNA-Seq* representou um avanço significativo dentre as abordagens de análise de expressão gênica, o primeiro trabalho utilizando dados de *RNA-Seq* foi publicado em 2006 (Bainbridge et al., 2006), utilizando a tecnologia 454/Roche (Margulies et al., 2005), os dados gerados foram 200.000 pequenas sequências, com 110 pares de base (pb) de tamanho. Alguns anos depois, a tecnologia de *RNA-Seq* começou a se popularizar, em 2008 um trio de estudos científicos demonstraram o início da popularização (Mortazavi et al., 2008; Sultan et al., 2008; Wilhelm et al., 2008).

Atualmente, existem várias tecnologias de sequenciamento *RNA-Seq* (Hong et al., 2020), permitindo a produção de *reads* longos *single* ou *paired-end*. Dessa forma, *RNA-Seq* possibilita mapeamentos de qualidade, identificação precisa de *splicing* alternativo, reconstrução de transcritos, entre outros estudos.

A metodologia de sequenciamento *RNA-Seq* consiste em uma população de *RNA* (inteira ou fracionada) convertida em uma biblioteca de fragmentos de *cDNA* com adaptadores (sequências conhecidas de 6 a 12 nucleotídeos) ligados as extremidades dos fragmentos, como apresentado na Figura 2.3, cada fragmento (amplificado ou não) é sequenciado, obtendo-se pequenas sequências de uma extremidade (sequenciamento *single-end*) ou das duas extremidades (sequenciamento *paired-end*). As sequências geradas (*reads*) possuem tipicamente entre 30 e 400 pb (pares de base, nucleotídeos identificados).

Em estudos que avaliam a expressão diferencial de genes, os dados gerados após o sequenciamento passam essencialmente por análises computacionais, que incluem limpeza, mapeamento e contagem, conforme apresentado na Figura 2.4. Cada uma dessas etapas possui diversas metodologias e formatos de arquivos específicos. Portanto, a escolha da metodologia a ser utilizada em cada fase deve ser feita com base no estudo realizado. Na seção seguinte, apresentamos métodos e ferramentas de limpeza, mapeamento e contagem que são comumente aplicados a estudos de expressão com dados de *RNA-Seq*. Também apresentamos as especificidades de cada método, a fim de apoiar essa escolha.

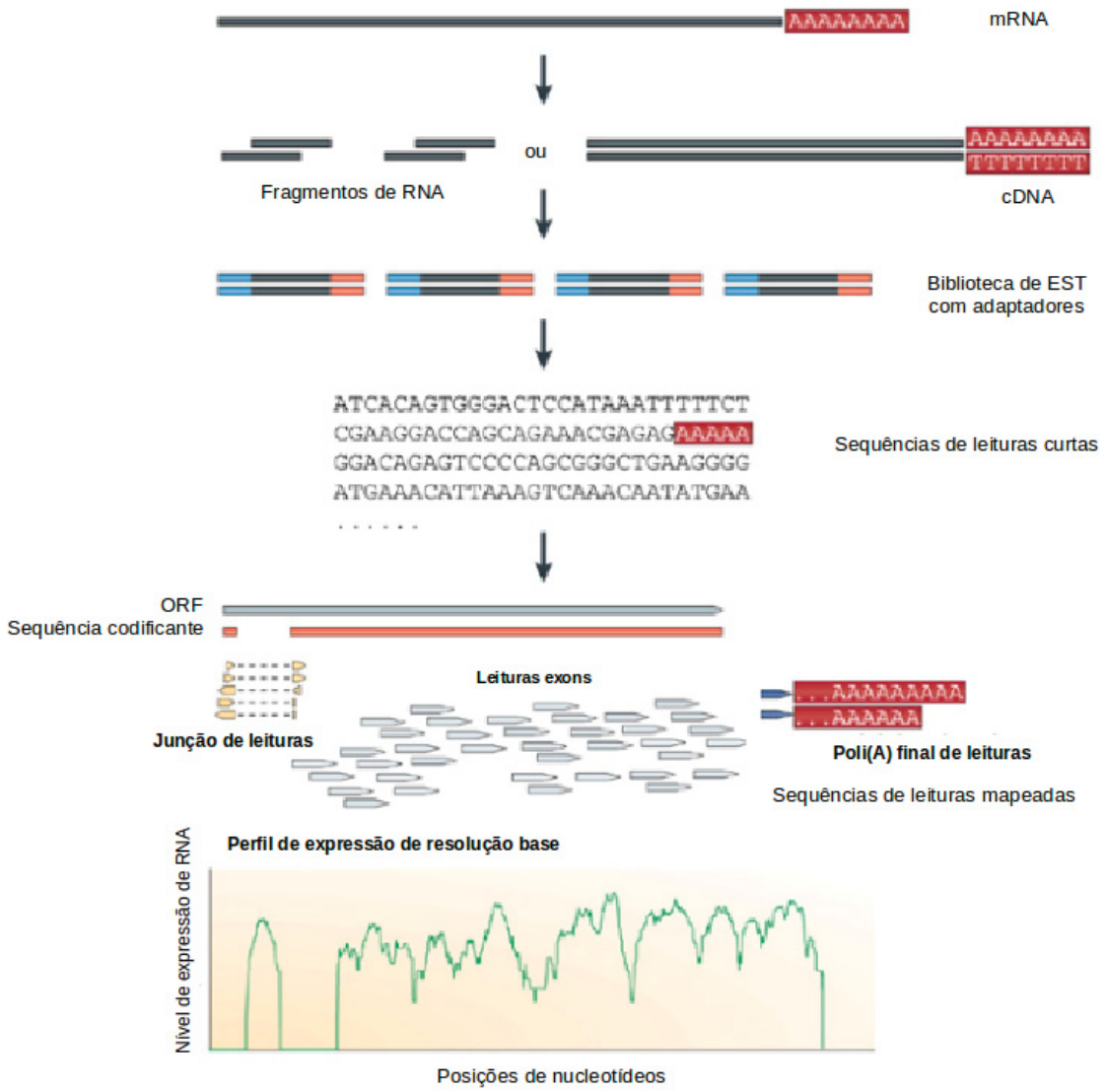


Figura 2.3: Exemplo de um experimento típico de *RNA-Seq* .

Exemplo de um experimento típico de *RNA-Seq* : Adaptadores (azul) são adicionados a cada fragmento de *cDNA* e sequências curtas são obtidas por meio do sequenciamento de cada fragmento de *cDNA*, utilizando tecnologias de alto rendimento. Os *reads* resultantes do sequenciamento são alinhados com o genoma de referência ou transcriptoma e classificados em três tipos: *reads* exônicos, *reads* de junção e *reads* poli-A. Esses três tipos de *reads* são utilizados para gerar um perfil de expressão para cada gene. **Fonte:** Adaptado de (Wang et al., 2009).

2.2 ETAPAS DA ANÁLISE DE GENES DIFERENCIALMENTE EXPRESSOS COM DADOS DE *RNA-SEQ*

Os sequenciadores *NGS* geram, como saída, arquivos em formato *FASTQ* (Cock et al., 2010), que contêm as sequências identificadas automaticamente e uma nota de qualidade para cada nucleotídeo dessas sequências. As notas de qualidade estão representadas por um caractere e podem ser convertidas em valores numéricos utilizando um intervalo da tabela ASCII, dependendo da origem dos dados. Mais detalhes sobre as notas de qualidade podem ser encontrados no artigo (Cock et al., 2010).

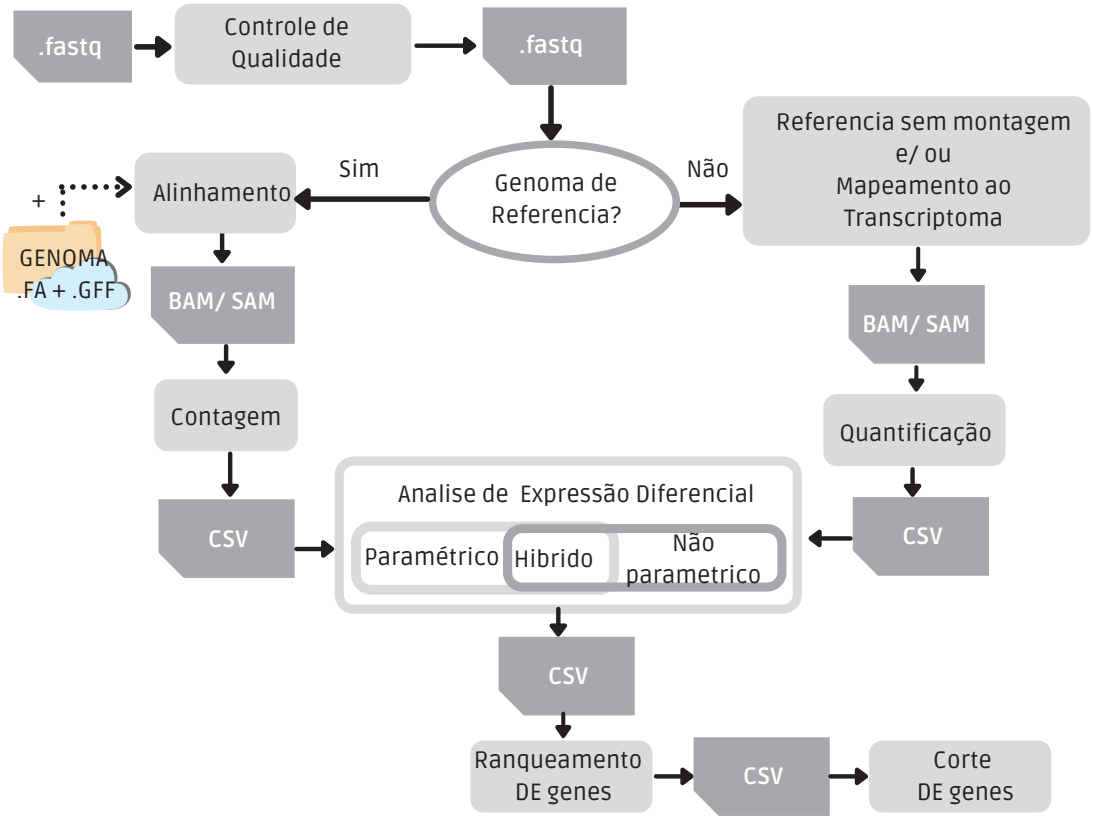


Figura 2.4: Experimento comum para análise de expressão com dados de *RNA-Seq* .
Experimento comum para análise de expressão com dados de *RNA-Seq* : Esquema om os principais passos da análise de expressão em massa com dados de *RNA-Seq* .
Fonte: De autoria própria.

2.2.1 Limpeza - Análise de Qualidade

A nota de qualidade associada aos nucleotídeos das sequências indica a confiabilidade dos nucleotídeos indicados pelo sequenciador, o formato de arquivo *FASTQ* é como apresentado na Figura 2.5. Sequências com baixa qualidade de leitura podem ser excluídas, dependendo do experimento. A verificação e exclusão de sequências de baixa qualidade é chamada de limpeza (Li et al., 2015b).

Os fragmentos gerados pelo sequenciador são alinhados ao genoma de referência do organismo que se deseja investigar após a limpeza. O alinhamento busca definir qual região do genoma gerou aquele fragmento (Canzar e Salzberg, 2017). Geralmente, esses fragmentos são alinhados a regiões anotadas como regiões exônicas do *DNA* (Wang et al., 2009). As regiões exônicas são segmentos de um gene que são transcritos em *RNA* mensageiro (mRNA) e, posteriormente, traduzidos em proteínas. Elas contêm a informação genética essencial para a síntese de proteínas específicas.

2.2.2 Mapeamento

Alguns organismos não possuem genoma sequenciado. Nestes casos, em experimentos de *RNA-Seq* para análise de expressão, os *reads* gerados são utilizados para a reconstrução da sequência dos transcritos, processo este que é chamado de montagem *de novo* dos transcritos. Na montagem *de novo*, os *reads* são utilizados para reconstruir a sequência de *mRNA* que deu

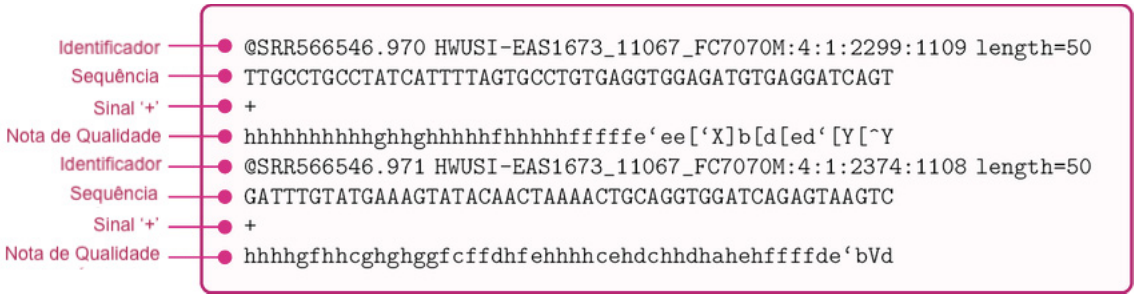


Figura 2.5: Exemplo de um arquivo *FASTQ*.
Exemplo de um arquivo *FASTQ*, que apresenta duas leituras (*reads*). Os caracteres apresentados na linha indicada como 'pontuação de qualidade' são convertidos com base em seu valor na tabela ASCII para o cálculo da qualidade do *read*. Para os *reads* com qualidade maior ou igual ao limiar estabelecido, serão utilizados apenas o identificador e a sequência desses arquivos nas próximas etapas da análise.
Fonte: Adaptado de (Hosseini et al., 2016).

origem a eles. Ao término da montagem, obtemos uma lista de transcritos (sequências maiores montadas através da sobreposição dos pequenos fragmentos gerados pelo sequenciador, também conhecidas como *contigs*). Após a montagem, os *reads* são mapeados aos transcritos que foram montados a partir deles mesmos (Simoneau et al., 2021).

Para a execução do mapeamento, em organismos com referência (genoma ou transcriptoma) sequenciada e anotada, é necessário fornecer um arquivo em formato *FASTA* contendo a sequência de referência e um arquivo de anotação. Este arquivo de anotação, geralmente é fornecido em formato *GFF* (Formato Geral de Características, do inglês *General Feature Format*) ou *GTF* (Formato Geral de Transferência, do inglês *General Transfer Format*) , o qual indica a região do genoma corresponde a qual característica (Zhang, 2016). As características podem ser: gene, éxon, entre outros. O arquivo também indica a posição inicial e final de cada característica. Mais detalhes sobre os formatos de arquivos de sequenciamento podem ser encontrados no estudo *"Overview of Sequence Data"* (Zhang, 2016).

Com os arquivos *FASTQ*, que contêm os *reads* que passaram pelo filtro de qualidade (limpeza), e com o arquivo de anotação *GFF* ou *GTF*, as ferramentas de mapeamento buscam os locais de alinhamento dos *reads* na referência (genoma ou transcriptoma). Elas geram arquivos de mapeamento com registros de cada alinhamento (característica, posição inicial e final). No entanto, um *read* pode ter alinhamento com mais de uma característica do genoma ou alinhar-se parcialmente a uma característica.

O alinhamento é um problema clássico da Bioinformática, com muitas propostas de soluções (Mount, 2007; Kent, 2002; Wu et al., 2005) as quais se aplicam especialmente a marcadores de sequência expressa (ESTs, do inglês *Expressed Sequence Tag*). No entanto, para o mapeamento de dados de *RNA-Seq* , é necessária uma abordagem diferente de alinhamento, pois, em alguns casos, os *reads* são pequenos (~30-125 pares de base), os índices de erro são consideráveis e alguns *reads* podem ser oriundos de junções de éxons (Garber et al., 2011). Outro fator a ser considerado é a quantidade de dados, pois alguns experimentos podem chegar a centenas de milhões de *reads*.

Existem dois algoritmos principais e, respectivos métodos computacionais, para mapear *reads* a uma referência. Uma das abordagens é denominada alinhador de *reads* com emendas (do inglês *spliced read aligner*), o qual leva em consideração a possibilidade de um *read* ser oriundo de junções exônicas. A outra abordagem, chamada de alinhador de *reads* sem emendas (do inglês *unspliced read aligner*), não considera junções exônicas.

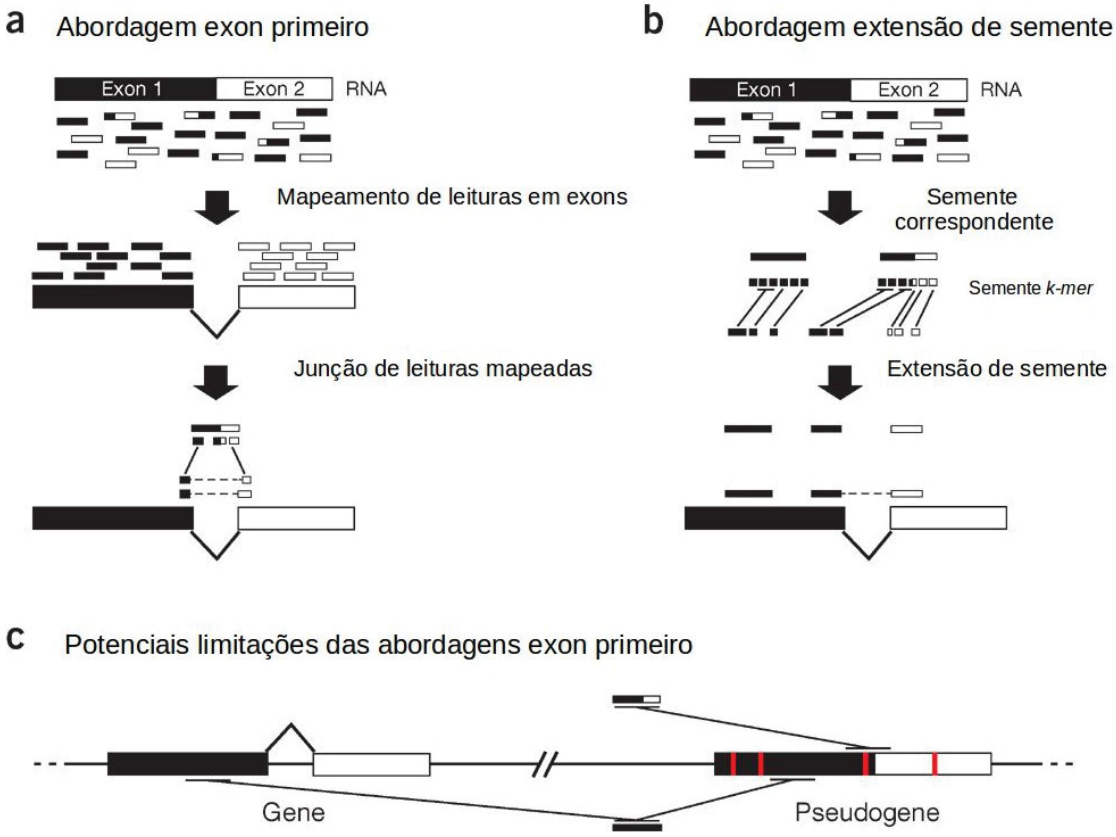


Figura 2.6: Estratégias de alinhamento.
Estratégias de alinhamento com intervalos (*gaps*) para *reads* de *RNA-Seq* contra o genoma. Ilustração de *reads* obtidos de uma região com dois éxons. **(a)** Método *éxon-first*: realiza o mapeamento completo em regiões exônicas (sem junções). Os *reads* que não obtiveram mapeamento são fragmentados em pequenas sequências e mapeados no genoma. Em seguida, estende-se as sequências mapeadas, permitindo *gaps*, para encontrar regiões candidatas a junção. **(b)** Registra pequenas sequências (*k-mers*) de tamanho similar no genoma em uma estrutura de dados altamente eficiente para pesquisa. Os *reads* são fragmentados em pequenas sequências, que são mapeadas na estrutura de dados e estendidas em grandes alinhamentos, que podem conter *gaps*. **(c)** Uma potencial limitação da metodologia *éxon-first* ocorre em casos de pseudogenes, com a associação de retrotransposons. *Reads* exônicos podem mapear tanto em um gene quanto em um pseudogene, podendo ser associados erroneamente.
Fonte: Adaptado de (Garber et al., 2011).

A abordagem *spliced read aligner* pode ser dividida em duas grandes metodologias: a do éxon primeiro (do inglês *éxon-first*) e a de extensão de semente (do inglês *seed and extend*).

- ***Éxon-first***: Esta abordagem executa o processo de alinhamento em dois passos. Primeiramente, os *reads* são alinhados a uma referência utilizando a metodologia *unspliced*, conforme exibido na Figura 2.6. (a). No segundo passo, após o alinhamento inicial, os *reads* que não obtiveram mapeamento são fragmentados em segmentos curtos e alinhados de forma independente. As regiões adjacentes aos segmentos de *reads* mapeados são analisadas para possíveis junções de éxons. Esta metodologia é muito eficiente quando apenas uma pequena porção de *reads* precisa ser processada no segundo passo (Garber et al., 2011). Ferramentas como TopHat (Trapnell et al., 2009) implementam a metodologia *éxon-first*.

- ***Seed-extend***: Inicialmente, os *reads* são fragmentados em pequenos pedaços (sementes), que são confrontados ao genoma para localizar alinhamentos, conforme ilustrado na Figura 2.6. (b). As regiões candidatas, que possuem alinhamento, são verificadas com métodos mais sensíveis, como *Smith-Waterman* (De Bona et al., 2008), e unidas às sementes iniciais. Com o alinhamento inicial de sementes e a extensão com métodos sensíveis, é possível determinar com mais exatidão o local da junção para o *read*. A ferramenta *GSNAP* (do inglês *Genomic Short Read Alignment Program*) (Wu e Nacu, 2010) implementa a abordagem *Seed-extend*.

As metodologias *unspliced read aligner* alinham *reads* sem permitir grandes extensões de discordância (*gaps*) e geralmente se enquadram em pelo menos uma de duas categorias: método semente e transformada de *Burrows-Wheeler*.

- **Método semente**: Este método encontra subsequências que se alinham perfeitamente com a referência, chamadas 'sementes'. Cada semente é utilizada para restringir as regiões onde métodos mais sensíveis, como *Smith-Waterman*, podem tentar estender as sementes para um alinhamento completo. Ferramentas como *HISAT* e *HISAT2* implementam esta metodologia (Kim et al., 2015, 2019).

- **Transformada de *Burrows-Wheeler***: Este método compacta a referência em um índice (de *Burrows-Wheeler*), que é uma estrutura muito eficiente para busca de concordâncias (*match*) perfeitas. No entanto, ao permitir discordâncias (*mismatches*), a performance da técnica diminui de forma exponencial em relação ao número de *mismatches* (Li e Durbin, 2009; Langmead et al., 2009). Ferramentas como *BWA* (*Burrows-Wheeler Alignment*) (Li e Durbin, 2009) e *Bowtie* e *Bowtie2* (Langmead et al., 2009; Langmead e Salzberg, 2012) implementam esta metodologia.

A abordagem *unspliced* é ideal para o mapeamento de *reads* contra bancos de cDNA, especialmente em casos de quantificação (Mortazavi et al., 2008; Griffith et al., 2010). Após o alinhamento, a quantidade de *reads* alinhados a cada região da referência é contabilizada. Essa contagem permite algumas inferências, dentre elas, a verificação de regiões do genoma/transcriptoma com maior atividade na situação analisada.

2.2.3 Contagem de *reads* mapeados

O mapeamento e a contagem de leituras constituem, portanto, um fluxo de trabalho comum pelo qual as leituras brutas são resumidas em uma matriz de frequências que pode ser usada para análises posteriores. Essas duas etapas geralmente representam a parte mais dispendiosa em termos computacionais de uma análise de *RNA-Seq* , com o mapeamento e a contagem contribuindo substancialmente para o custo total (Liao et al., 2019).

As metodologias de mapeamento influenciam de forma muito sutil os resultados finais de experimentos que avaliam a expressão gênica (Schaarschmidt et al., 2020; Costa-Silva et al., 2017a). No entanto, os resultados do mapeamento é uma das bases para a análise de expressão diferencial de genes, por isso a forma como os arquivos de mapeamento são considerados é importante para o resultado final da análise.

Utilizando os arquivos de resultados do mapeamento, são extraídas as contagens de *reads* mapeados para gerar a tabela de frequências. De modo geral, relaciona-se os genes ou transcritos do genoma ou transcriptoma de referência à condição analisada . As linhas da tabela de contagem são genes ou transcritos e as colunas são as condições ou perturbações analisadas.

Em algumas ferramentas, esta tabela de contagem pode ser gerada de forma transposta ao descrito, com condições nas linhas e genes ou transcritos nas colunas.

A contagem é a etapa na qual é possível estimar quais regiões genômicas estavam mais ativas (em termos de transcrição) na amostra. Esta etapa não define quais genes são diferencialmente expressos, mas representa a base para os passos seguintes da análise. Isso ocorre porque, para cada arquivo de sequenciamento apresentado ao mapeador, será obtida uma contagem de *reads* mapeados aos genes ou transcritos da referência.

Consequentemente, para que este processo seja executado, existe a necessidade de um arquivo de anotação do genoma de referência. O arquivo de anotação é geralmente no formato *General Feature Format* (GFF), que consiste em uma linha por característica, a característica pode ser gene, *mRNA* ou outro tipo de sequência. GFF é um formato de arquivo texto amplamente utilizado para armazenar anotações do genoma, descrevendo anotações baseadas em sequências. Além disso, os arquivos *GFF* apresentam características do genoma em uma tabela delimitada por tabulação, com uma característica única por linha, tornando-o ideal para uso com vários pipelines de análise de dados (Rastogi e Gupta, 2014). O arquivo *GFF* é utilizado para traduzir a informação de alinhamento, que, por exemplo, apresenta apenas a seguinte informação: os *reads* A e B foram mapeados no cromossomo X, entre os pares de base i e j. Com a informação de região de alinhamento associada ao arquivo de anotação, é possível indicar a qual gene de um cromossomo X os pares de base entre i e j se referem. Por outro lado, o arquivo de anotação indicará que o gene Z possui parte ou toda a sua sequência entre as posições i e j do cromossomo X, portanto os *reads* A e B serão considerados na contagem de mapeamento do gene Z.

Antes de realizar a contagem, é necessário analisar os tipos de alinhamentos que serão considerados. Essas opções definem quais *reads* serão contabilizados como alinhados, dentre as seguintes opções:

- *Read* totalmente alinhado a um gene;
- *Read* parcialmente alinhado a um gene;
- *Read* alinhado a uma junção (intron e éxon);
- *Read* alinhado a junção de éxons (sem alinhamento com intron);
- *Read* parcialmente alinhado a dois genes;
- *Read* alinhado a dois genes.

Para esta tarefa, existem algumas metodologias que podem ser utilizadas associadas a pipelines de análise ou de forma isolada, tais como *HTSeq-count* (Anders et al., 2015), que faz parte do framework *HTSeq*, o conjunto de ferramentas *BEDTools* (Quinlan e Hall, 2010) e os *software* featureCounts (Liao et al., 2014) e *RSubread* (Liao et al., 2019).

A escolha da ferramenta e de como considerar os mapeamentos na contagem deve ser realizada com base no conjunto de dados e suas propriedades. Para eventuais situações em que se tem pouco conhecimento prévio, recomenda-se a comparação da contagem no modo mais restritivo e no modo mais abrangente de cada ferramenta para a definição de uma parametrização adequada.

Após a etapa de contagem, é necessário considerar a metodologia de normalização dos dados. As principais metodologias são apresentadas na seção 2.2.4.

2.2.4 Normalização

RPKM

A metodologia *RPKM* (do inglês *Reads per Kilobase per Million*), foi a primeira metodologia proposta para uma quantificação precisa de expressão de genes com dados de *RNA-Seq* . Publicada em 2008 (Mortazavi et al., 2008), esta metodologia quantifica a expressão de dados de *RNA-Seq* através da normalização do tamanho total do transcrito e do número de *reads* sequenciados.

Utilizando a quantidade de nucleotídeos do genoma ou gene de referência e, a quantidade de *reads* mapeados para obter um valor de expressão, o que permite que genes ou transcritos pequenos não sejam penalizados, se comparados a sequências maiores. *RPKM* pode ser definido pela equação 2.1:

$$RPKM = \frac{10^9 \, r_g}{R \, fl_g} \tag{2.1}$$

onde, g representa um gene, ou uma região específica da referência r_g é a quantidade de *reads* mapeados em uma região particular (gene), R é o número total de *reads* do experimento, e fl_g é o total de nucleotídeos contidos na referência (gene, ou região) em pb (pares de base).

O *RPKM* é uma das metodologias mais utilizadas para quantificação de expressão em dados de *RNA-Seq* (Li et al., 2015a), e foi inicialmente introduzida para facilitar a comparação entre genes em uma amostra e, entre amostras, pois reescala a contagem de genes para corrigir diferenças de tamanho da biblioteca e do gene.

FPKM

A metodologia *FPKM* é análoga ao *RPKM*, mas suporta uma, duas ou mais (se necessário em futuras tecnologias) sequências da mesma fonte molecular (Trapnell et al., 2010). Quando a técnica de sequenciamento utilizada é *paired-end* a metodologia para identificação de genes diferencialmente expressos é um pouco diferente, e utiliza a palavra fragmento ao invés de *reads*, pois nesse contexto é possível encontrar *reads foward* e *reverse* (nos dois sentidos da dupla fita de *DNA* 5'- 3' ou 3' - 5') mapeados em uma mesma região.

O *FPKM* pode ser definido como na equação 2.2.

$$FPKM = \frac{N}{(L/1000)/(R/10^6)} \tag{2.2}$$

onde N representa o total de fragmentos do experimento, L o tamanho total do transcrito (referência) em Kilobase (/1000) e R representa o total de *reads* mapeados em milhões ($/10^6$).

A abordagem *FPKM* considera que fragmentos nem sempre são representados por um *read*, mas se referem a fragmentos gerados por um experimento de *RNA-Seq* , partindo da afirmação que, um sequenciamento paired-end gera mais *reads* do que fragmentos, ou seja, se existe a a ocorrência de dois *reads* mapeados em um mesmo local, mas um *read* é reverso complementar do outro (representa o outro lado da fita de *DNA*). FPKM propõe para que aquele *read* não seja interpretado como mais uma sequência expressa e que, dois *reads* nessa condição sejam considerados como um fragmento.

TPM

Outra abordagem de normalização foi proposta por (Wagner et al., 2012), onde os autores propõem a normalização TPM (transcritos por milhão). *TPM* é uma modificação da abordagem RPKM e, busca a remoção de tendencias do *RPKM* (Wagner et al., 2012). O valor de *TPM* é calculado como na equação 2.3, onde Y_{gk} é o total de leituras mapeadas para o gene g na biblioteca k, rl é a média de tamanho das leituras mapeadas, fl_g é o número nucleotídeos do transcrito mapeável e, N_k é a quantidade de leituras da biblioteca k:

$$TPM = \frac{Y_{gk} \times rl \times 10^6}{fl_g \times N_k} \tag{2.3}$$

TMM

Ao buscar uma métrica apropriada de expressão de genes, que possa ser utilizada para comparação entre amostras, foi desenvolvida a metodologia Média Aparada por Valores M (do inglês *Trimmed Mean of M-values* (*TMM*)) (Robinson e Oshlack, 2010).

Uma metodologia para estimativa de expressão de genes, deve garantir que um gene com nível de expressão igual em duas amostras não seja detectado como diferencialmente expresso. Para uma estimativa precisa de níveis de expressão é necessário quantificar a produção total de *RNA*, S_k, o que não pode ser estimado diretamente. Entretanto a produção relativa de *RNA* entre duas amostras pode ser mais facilmente determinada calculando a mudança global entre amostras (*global fold-change*) $f_k = S_k/Sk'$.

O método *TMM* foi proposto como um caminho simples e robusto de estimar a produção de *RNA*. A contagem observada para o gene g na biblioteca k é definida por Y_{gk} e, o total de leituras da biblioteca k é definido por N_k. O cálculo de mudança para um gene conhecido é definido por:

$$M_g = \log_2 \frac{Y_{gk}/N_k}{Y_{gk'}/N_{k'}} \tag{2.4}$$

O nível absoluto de expressão é definido, considerando a conjunção ($\bullet$) das amostras, ou seja, considera o cálculo de mudança para o gene g na amostra (condição) k e k', para todo gene que possui contagem em ambas as amostras $Yg\bullet$ diferente de 0, como definido por (Robinson e Oshlack, 2010) na equação 2.5:

$$A_g = \frac{1}{2} \log_2(Y_{gk}/N_k \bullet Y_{gk'}/N_{k'}) \; para \; Y_{g\bullet} \neq 0 \tag{2.5}$$

O cálculo é considerado apenas para genes com contagem de *reads* diferente de 0. O método *TMM* é duplamente cortado, pelo cálculo de mudança (log-fold-change) M^r_{gk} (amostra k em relação a amostra r do gene g) e pela intensidade absoluta A_g.

Especificamente a normalização para a amostra k usando como referência a amostra r é calculado como:

$$\log_2(TMM_k^{(r)}) = \frac{\sum\limits_{g \in G^*} w^r_{gk} M^r_{gk}}{\sum\limits_{g \in G^*} w^r_{gk}} \tag{2.6}$$

onde:

$$M^r_{gk} = \frac{\log_2(Y_{gk}/N_k)}{\log_2(Y_{gr}/N_r)} \; e \; w^r_{gk} = \frac{Nk - Y_{gk}}{NkY_{gk}} + \frac{N_r - Y_{gr}}{N_r Y_{gr}}; \; Y_{gk}, Y_{gr} > 0 \tag{2.7}$$

Os casos em que Y_{gk} ou $Y_{gr} = 0$, são excluídos anteriormente aos cálculos apresentados nas equações 2.6 e 2.4, visto que a variação não pode ser calculada. G^* representa o conjunto de genes com valores de M_g e A_g válidos.

A metodologia *TMM* é utilizada pelo pacote *edgeR* (Robinson et al., 2010), na prática muito semelhante a metodologia utilizada pelo pacote *DESeq* (Anders e Huber, 2010), os resultados também são semelhantes em alguns pontos (Li et al., 2015a; Soneson e Delorenzi, 2013).

2.3 IDENTIFICAÇÃO DE GENES DIFERENCIALMENTE EXPRESSOS

A análise de expressão diferencial de genes tem como objetivo principal identificar genes que aumentam ou diminuem sua transcrição em condições específicas. Essas condições, de modo geral, estão relacionadas a características fenotípicas. Em plantas, podemos utilizar como exemplo o tamanho, formato ou sabor de um fruto.

Como apresentado anteriormente e ilustrado na Figura 2.4, a análise de expressão com dados de *RNA-Seq* é dividida em várias fases, sendo a identificação e ranqueamento dos genes diferencialmente expressos a etapa final desta análise. Portanto, é importante a compreensão das propriedades fundamentais dos métodos que analisam a distribuição dos dados de expressão, bem como as principais diferenças dessas análises.

A metodologia de sequenciamento *RNA-Seq* começou a se popularizar entre 2006 e 2009. Desde então, muitas metodologias para análise de expressão de genes com dados de *RNA-Seq* foram propostas. Em um estudo recente (Costa-Silva et al., 2023), avaliamos quais são as características comuns a essas metodologias e em que elas se diferem. Observou-se que muitos métodos avaliam a expressão gênica com base em uma distribuição de dados. Desse modo, nesta tese, os métodos foram classificados em três grupos principais: Métodos Paramétricos, Não paramétricos e Híbridos.

2.3.1 Métodos paramétricos

Os métodos paramétricos de *DEGs* são aqueles que aplicam uma distribuição estatística específica para identificar genes diferencialmente expressos. Os métodos paramétricos são definidos como aqueles que inferem que um dado segue uma determinada distribuição. Para a análise de *DEGs* com dados de *RNA-Seq* , algumas distribuições tendem a apresentar melhores resultados devido às suas características específicas. Os dados da revisão desenvolvida nesta tese indicam que a distribuição binomial negativa (DBN) foi registrada como a mais utilizada, correspondendo a aproximadamente 67% dos métodos classificados como paramétricos. A segunda distribuição mais utilizada é a de *Poisson*, com cerca de 10% dos métodos paramétricos.

Neste contexto, é primordial compreender as principais características dessas duas distribuições e como elas se aplicam à análise de *DEGs*. Portanto, abordaremos a DBN e a Distribuição de *Poisson* de maneira mais detalhada.

A DBN é amplamente utilizada por métodos paramétricos para a identificação de *DEGs*. Ela pode ser definida como uma distribuição que nos permite identificar a probabilidade de necessitarmos de X tentativas de *Bernoulli* (experimentos onde o resultado pode ser apenas fracasso ou sucesso) para obtermos r sucessos. Esta distribuição foi definida pelo biólogo e estatístico Ronald A. Fisher (Fisher, 1941).

A DBN possui algumas propriedades que a descrevem: i) O experimento consiste em X eventos (tentativas/testes) repetidos; ii) Cada evento pode resultar em apenas dois resultados (sucesso ou fracasso); iii) A probabilidade de sucesso (notação: p) é a mesma para cada evento;

iv) Os eventos são independentes, ou seja, um evento não influencia o resultado de outro; v) O experimento termina apenas quando são observados r sucessos, sendo r definido previamente.

A DBN modela a contagem de *reads* mapeados em um gene como uma variável aleatória discreta, assumindo que cada *read* é mapeado de forma independente e com probabilidade constante ao longo do gene. A probabilidade de sucesso (mapeamento de um *read*) é representada por p, e o número de falhas (*reads* que não mapeiam) antes do k-ésimo sucesso (k *reads* mapeados) é dado por $r = k - 1$.

A Distribuição Binomial Negativa é definida na equação 2.8, na qual r é o número de sucessos pretendidos, p é a probabilidade de sucesso na realização de um evento e k é o número de repetições necessárias para obter r sucessos.

$$Pr(X = k) = \begin{pmatrix} k-1 \\ r-1 \end{pmatrix} = p^r(1-p)^{k-r} \tag{2.8}$$

Dessa forma, a última repetição necessariamente deve ser bem sucedida, e os eventos anteriores devem contabilizar $r - 1$ sucessos e $k - 1$ tentativas (MORETTIN, 2009).

A Distribuição Binomial Negativa é aplicada à identificação de *DEGs*, modelando a contagem de *reads* para cada gene. Em dados de contagem, a quantidade de *reads* mapeados para um gene específico pode ser considerada como a contagem de eventos de sucesso. As contagens podem variar devido a variações na eficiência do sequenciamento, extensão do gene (quantidade de pares de base) e variação biológica.

Como exemplo de uso da DBN para análise de *DEGs*, existe o pacote R chamado *edgeR* (Robinson et al., 2010), disponível no Bioconductor (Huber et al., 2015). O pacote *edgeR* inicia o processo de análise de expressão diferencial de genes com a normalização dos dados. Essa normalização é realizada com base no tamanho das bibliotecas, que é definido pela soma dos valores de cada coluna (contagem de *reads* mapeados em cada condição).

Durante a criação de um objeto da classe *DGEList*, o tamanho das bibliotecas é definido automaticamente e o valor do fator de normalização é definido como 1 para todas as bibliotecas. O fator de normalização é calculado no passo seguinte, onde os valores de contagem são normalizados, utilizando o método *TMM* (média ajustada de valores M) , apresentado por Robinson e Oshlack, 2010.

É sabido que os dados de expressão podem ser resumidos em uma tabela de contagens, com linhas correspondentes a genes (éxons ou transcritos) e colunas correspondentes a amostras ou condições. O pacote *edgeR* modela os dados como uma distribuição binomial negativa (BN),

$$Y_{gi} \sim BN(M_i p_{gj}, \phi_g), \tag{2.9}$$

para o gene g e a amostra (replicata) i. Na equação 2.9, M_i é o tamanho da amostra, ϕ_g é a dispersão e p_{gj} é a abundância relativa do gene g no grupo experimental j, ao qual pertence a amostra i. O *edgeR* utiliza a parametrização BN, onde a média é $\mu_{gi} = M_i p_{gi}$ e a variância é $\mu_{gi}(1 + \mu_{gi}\phi_g)$ (Robinson et al., 2010). No contexto de análise de expressão diferencial, o parâmetro de interesse é p_{gi}. Mais detalhes sobre a execução da análise de expressão diferencial de genes com o pacote *edgeR* podem ser encontrados no artigo Chen et al., 2024.

Conforme mencionado anteriormente, nosso estudo identificou que a segunda distribuição mais utilizada em métodos paramétricos é a distribuição de *Poisson* (utilizada por 10% dos métodos analisados). Portanto é importante avaliar as características dessa distribuição para compreendermos como se aplica a análise de genes diferencialmente expressos.

A distribuição de *Poisson* é definida como uma distribuição estatística utilizada em estatística e em simulações de resultados. Ela é utilizada para representar a probabilidade de que

um determinado evento (valor) aconteça, quando a média de probabilidade é conhecida. Esta distribuição foi descrita por Siméon-Denis *Poisson* no estudo Poisson e Schnuse, 1841.

A distribuição de *Poisson* possui algumas propriedades que a tornam adequada a análise de expressão diferencial de genes. **i)** A probabilidade de uma ocorrência é a mesma para qualquer um dos intervalos de tempo, ou seja, a distribuição não possui nenhuma região com maior probabilidade do que outras regiões; **ii)** A probabilidade de acontecer mais de uma ocorrência no mesmo ponto é quase nula, ou se aproxima de zero; **iii)** Os eventos são independentes, de modo que a ocorrência de um evento não interfere na ocorrência de outro.

A definição da distribuição de *Poisson* é definida na equação 2.10, na qual X é o número de ocorrências em um intervalo, λ é a taxa (média) de ocorrências do evento X, e é uma constante natural ($e \approx 2,71828$):

$$P(x) = \frac{\lambda^X \cdot e^{-\lambda}}{X!}. \tag{2.10}$$

A distribuição de *Poisson* é frequentemente indicada como adequada para descrever dados de expressão gênica, pois representa a probabilidade de um evento ocorrer em um intervalo de tempo esperado. Estudos anteriores identificaram que a contagem de dados de *RNA-Seq* é bem representada por uma distribuição de *Poisson* (Marioni et al., 2008).

Existem algumas aplicações que utilizam a distribuição de *Poisson* para a análise de expressão diferencial de genes. Como exemplo, temos o pacote R DEGSeq (Wang et al., 2010), que permite ao usuário escolher entre a distribuição de *Poisson* ou a binomial negativa. No entanto, a distribuição utilizada por padrão é a *Poisson*.

Os métodos *edgeR* (Robinson et al., 2010), *baySeq* (Hardcastle e Kelly, 2010), *EBSeq* (Leng et al., 2013), *DESeq2* (Love et al., 2014) e *limma* (Law et al., 2014), supõem que as contagens de leitura em réplicas biológicas sigam uma distribuição binomial negativa. Tanto o *DESeq2* quanto o *edgeR* aplicam um modelo linear generalizado, mas estimam a variabilidade de uma maneira diferente, enquanto uma abordagem bayesiana é usada pelo *baySeq* (Tarazona et al., 2011). Cada um desses métodos possuem características particulares e, apesar de serem todos paramétricos, buscam formas diferentes de identificar *DEGs*, conforme resumidamente apresentado a seguir:

- *edgeR* (Robinson et al., 2010): utiliza um modelo de super dispersão de *Poisson* para levar em conta a variação técnica e biológica. Aplica o método empírico bayesiano para moderar o grau de super dispersão em relação às transcrições;

- *baySeq* (Hardcastle e Kelly, 2010): utiliza a abordagem empírica bayesiana para estimar a probabilidade *a posteriori* de cada conjunto de modelos. Cada conjunto define padrões de expressão diferencial para cada tupla;

- *EBSeq* (Leng et al., 2013): foi desenvolvido com o objetivo principal de identificar isoformas diferencialmente expressas. Além disso, é robusto na identificação de *DEGs*. É semelhante ao *baySeq* (Hardcastle e Kelly, 2010) que também aplica a abordagem empírica bayesiana;

- *DESeq2* (Love et al., 2014): baseado em uma distribuição binomial negativa. A variância e a média são limitadas por regressão local. Primeiramente, cria um modelo com contagens observadas. Então, ele se ajusta usando o mesmo método do *DESeq* original, ou se ajusta em duas etapas: encontra o valor do parâmetro que torna a verossimilhança maior, o que é chamado de estimativa de verossimilhança máxima. Em seguida, busca todos os valores dos genes e move esses valores em direção a um valor médio. O

DESeq2 usa o teorema de Bayes (Bayes et al., 1763) para orientar a quantidade de movimento de cada gene: se a informação do gene for baixa, seu valor será movido para perto da média; se a informação do gene for alta, seu valor será movido muito pouco. Assim, os valores deslocados são úteis para avaliar diferentes conjuntos de genes, bem como para aplicar um limite.

- *limma* (Law et al., 2014): baseado no modelo linear e foi originalmente desenvolvido para analisar dados de *microarray* . Atualmente, foi estendido para análise de *RNA-Seq* . O guia do usuário do *limma* recomenda o uso da normalização *TMM* (Robinson e Oshlack, 2010) do pacote *edgeR* associado ao uso da conversão voom. Essencialmente, este método transforma as contagens normalizadas em logaritmos de base 2 e estima a relação média-variância para determinar o peso de cada observação feita inicialmente por um modelo linear;

Outros dois métodos também utilizados nesta tese são não paramétricos e serão descritos na Seção 2.3.2.

2.3.2 Métodos não paramétricos

Os métodos não paramétricos para análise de *DEGs* incluem métodos de inferências, estatísticas descritivas não paramétricas, modelos estatísticos e testes estatísticos. Esses métodos não estabelecem, *a priori*, um modelo de distribuição de dados. A estrutura dos modelos é definida com base na distribuição dos dados, processo comumente conhecido como *data-driven*. A expressão "não paramétrica", quando associada a uma ferramenta para análise de *DEG*, indica que o número e a natureza dos parâmetros são ajustados de acordo com a distribuição dos dados(Penfold et al., 2012).

Não existe uma definição única para esses métodos, ou seja, cada um utiliza uma abordagem distinta para identificar a expressão diferencial. A seguir, são descritas brevemente algumas metodologias utilizadas por duas ferramentas de análise de *DEGs* avaliadas nesta tese, como: *SAMSeq* (Li e Tibshirani, 2013) e *NOISeq* (Tarazona et al., 2011), que são métodos não paramétricos e, portanto, não fazem suposições sobre a distribuição de dados.

O *NOISeq* compara a alteração na expressão entre as condições com uma distribuição de ruído gerada pela comparação de pares de réplicas dentro da mesma condição. O *NOISeq* é um método que avalia a expressão gênica diferencial entre grupos por meio da relação de variação de expressão e diferenças de expressão absoluta (Tarazona et al., 2011, 2015). Este método utiliza dados de contagem de *reads* mapeados, corrigidos e normalizados, modela a distribuição do ruído através do contraste do logaritmo da mudança de expressão e das diferenças de expressão absoluta entre os grupos. Define-se um gene como diferencialmente expresso entre grupos se o logaritmo correspondente da mudança de expressão e os valores de diferença absoluta de expressão tiverem uma alta probabilidade de serem superiores aos valores de ruído (Li, 2019).

Outra ferramenta para análise de *DEG* não paramétrica é o *SAMSeq* (Li e Tibshirani, 2013). O *SAMseq*, baseado na estatística de classificação de Wilcoxon (Wilcoxon, 1945), utiliza uma abordagem de permutação para corrigir o viés de profundidade do sequenciamento. Para comparações entre grupos, o *SAMSeq* emprega as estatísticas de classificação Wilcoxon de duas amostras. Além disso, o *SAMSeq* considera as diferentes profundidades por meio de um processo de reamostragem na análise de dados diferenciais. Na estatística de classificação Wilcoxon e *FDR*, a distribuição nula é estimada usando o método de permutação (Li, 2019).

Outro método para análise de *DEGs* é o *RSEM*. A ferramenta *RSEM* (Li e Dewey, 2011), que utiliza a premissa de pseudo-alinhamento, não é voltada exclusivamente para a análise de

expressão. Essa ferramenta ganhou popularidade devido à possibilidade de realizar análise de expressão diferencial de genes para dados sem genoma de referência, uma necessidade latente na época.

Desse modo, é notável que a variedade de métodos não paramétricos é menor que a dos paramétricos, o que indica fortemente que existem áreas a serem exploradas na análise de expressão diferencial de genes.

2.3.3 Métodos híbridos

Nesta tese nos referimos a uma metodologia paramétrica associada a uma não paramétrica através do termo "Método Híbrido". Esse termo identificará metodologias que empregam análises paramétricas em conjunto com análises não paramétricas para inferir *DEGs*. Foi identificada apenas uma ferramenta classificada como híbrida nesta tese: *consexpression* (Costa-Silva et al., 2017a).

Embora os métodos híbridos não possuam uma definição formal, nesta tese, podemos afirmar que os métodos de análise de expressão diferencial de genes, classificados como híbridos, são aqueles que associam resultados de métodos não paramétricos a paramétricos para gerar a lista de genes considerados diferencialmente expressos.

A ferramenta *consexpression* é um pipeline para a análise de expressão, desenvolvido em linguagem de programação Python no ano de 2017, que adota a identificação dos *DEGs* a partir da indicação de *DEGs* de nove ferramentas. Dentre elas, duas são não paramétricas: *NOISeq* (Tarazona et al., 2011) e *SAMSeq* (Li e Tibshirani, 2013), e sete são paramétricas: *edgeR* (Robinson et al., 2010), *baySeq* (Hardcastle e Kelly, 2010), *EBSeq* (Leng et al., 2013), *DESeq* (Anders e Huber, 2010), *DESeq2* (Love et al., 2014), *limma* (Law et al., 2014; Ritchie et al., 2015) e sleuth (Pimentel et al., 2017). No método consexpression, são considerados diferencialmente expressos os genes indicados pelo consenso de pelo menos cinco ou mais metodologias.

O *consexpression* demonstrou reduzir a quantidade de falsos positivos na indicação de *DEGs*, visto que as indicações são feitas com base em um conjunto de métodos. Dessa forma, reduz-se a probabilidade de um estudo prosseguir com testes laboratoriais em genes erroneamente indicados como DE.

Também foram identificadas as ferramentas que possuem a opção de análise paramétrica ou não paramétrica na Figura 4.1. Essas ferramentas utilizam uma das metodologias conforme a escolha do usuário, diferentemente do critério adotado nesta tese.

Conforme apresentado, a análise de expressão diferencial de genes é uma abordagem que permite identificar genes cujos níveis de expressão variam significativamente entre diferentes condições experimentais. Entretanto, a expressão gênica não ocorre de maneira isolada, mas sim em um contexto de interações complexas e regulatórias entre genes (del Val et al., 2024).

As redes regulatórias de genes (*GRNs* do inglês *Gene regulatory networks*) são modelos utilizados para descrever essas interações, representando os genes como nós e as interações entre eles como arestas. Essas redes podem revelar padrões de co-expressão, identificar genes-chave (ou *hubs*) que regulam muitos outros genes e ajudar a elucidar os mecanismos subjacentes a fenômenos biológicos complexos.

Os dados de expressão gênica são frequentemente utilizados para inferir essas redes regulatórias porque fornecem uma visão abrangente da atividade gênica em uma célula ou tecido. Ao comparar os níveis de expressão de diferentes genes entre várias condições, é possível identificar padrões de co-expressão que sugerem uma interação regulatória. Por exemplo, se dois genes são consistentemente super expressos (*up-regulated*) ou sub expressos (*down-regulated*) juntos em resposta a uma variedade de condições, isso pode sugerir que eles estão envolvidos em uma via comum ou que um gene regula o outro.

Portanto, a análise de *DEGs* e a inferência de *GRNs* são complementares e, quando usadas em conjunto, podem fornecer observações valiosas sobre a biologia molecular de uma célula ou organismo e a descoberta de conhecimento.

2.4 MODELAGEM

Uma interação gênica pode ser vista como uma relação na qual o efeito de um gene alvo é modificado pelo efeito de um ou mais outros genes preditores (Madhukar et al., 2015).

Os componentes fundamentais de uma rede de regulação gênica são delineados por Mira et al., 2012, conforme segue:

> A expressão gênica é o processo de conversão das informações armazenadas nos genes em produtos gênicos funcionais que podem ser *mRNA* ou proteínas por meio de transcrição e tradução. Os fatores de transcrição (TFs) que modulam a transcrição, atuando como ativadores ou inibidores, também são codificados por genes e também são regulados, formando assim uma rede regulatória complexa.

Essas interações são importantes para delinear relações funcionais entre genes e suas proteínas correspondentes, bem como para elucidar processos biológicos complexos e doenças (Boucher e Jenna, 2013). As interações genicas são bem representadas por redes, que por sua vez são representadas por grafos. A seguir, apresentamos os conceitos essenciais para a compreensão da estrutura de grafos.

Um grafo pode ser definido como um par (V, E), sendo V um conjunto de vértices que representam os genes, e E um conjunto de arestas que representam a conexão entre os genes. Define-se como $E = \{(i, j)|i, j \in V\}$ uma conexão entre os vértices i e j (Pavlopoulos et al., 2011). Neste caso, pode-se dizer que i e j são vizinhos ou adjacentes. Conexões com várias arestas também podem ocorrer e devem ser consideradas, especialmente em redes de interação proteína-proteína.

Um grafo dirigido (ou dígrafo) é definido como $G = (V, E, f)$, onde f é uma função que mapeia cada elemento em E para um par ordenado de vértices em V. Os pares ordenados de vértices são chamados de arestas direcionadas. Uma aresta $E = (i, j)$ é considerada como tendo direção de i a j. Os grafos dirigidos são mais adequados para a representação de esquemas que descrevem caminhos ou procedimentos biológicos, mostrando a interação sequencial de elementos em um ou vários pontos de tempo e o fluxo de informações através da rede. Estes são, principalmente, redes metabólicas, de transmissão de sinal ou reguladoras (Pavlopoulos et al., 2011).

As redes regulatórias transcricionais podem ser modeladas como grafos dirigidos ou não dirigidos. Como exemplo, em uma rede regulatória transcricional, os nós representam genes, com arestas indicando as interações entre eles. Como cada uma dessas interações tem uma direção natural associada, tais redes são modeladas como grafos dirigidos (Mason e Verwoerd, 2007), quando se deseja identificar apenas a relação entre os genes em uma determinada condição, e verificar se os dados de expressão de um gene influenciam a mudança de outro, podemos utilizar o grafo não direcionados.

Outra especificação de grafo é o ponderado. Um grafo ponderado é definido como um grafo $G = (V, E)$, onde V é um conjunto de vértices e E é um conjunto de arestas entre os vértices $E = \{(u, v)|u, v \in V\}$ associado a ele uma função de peso $w : E \rightarrow R$, onde R denota o conjunto de todos os números reais. Na maioria das vezes, o peso w_{ij} da borda entre os vértices i e j representa a relevância da conexão. Normalmente, um peso maior corresponde a uma maior confiabilidade de uma conexão. Os grafos ponderados são atualmente as redes mais utilizadas na

Bioinformática (Pavlopoulos et al., 2011). Como exemplo, as relações cuja importância é variável são frequentemente atribuídas a dados biológicos para capturar a relevância de co-ocorrências identificadas por mineração de texto, sequência ou semelhanças estruturais entre proteínas ou co-expressão de genes (Jensen et al., 2009).

Um grafo bipartido é um grafo não direcionado $G = (V, E)$, onde V pode ser dividido em 2 conjuntos V_1 e V_2, onde $(u, v) \in E$ indica que $u \in V_1$ e $v \in V_2$ OU $v \in V_1$ e $u \in V_2$. Este tipo de grafo pode ser aplicado na visualização ou modelagem de redes biológicas, desde a representação de ligações em reações enzimáticas ou vias metabólicas, até ontologias ou conexões orgânicas (Pavlopoulos et al., 2011).

Neste contexto, se $G = (V, E)$ é um grafo, então $G_1 = (V_1, E_1)$ é nomeado subgrafo se $V_1 \subseteq V$ e $E_1 \subseteq E$, onde cada vértice em E_1 é afetado por vértices em V_1 (Mason e Verwoerd, 2007).

O grau de um vértice em um grafo não direcionado é o número de arestas que conectam um vértice a outro. Em um dígrafo, cada vértice i pode possuir duas definições de grau: grau de entrada (do inglês *in-degree*, $deg_{in}(i)$), que define a quantidade de arestas que saem de algum vértice e se conectam ao vértice i; grau de saída (do inglês *out-degree*, $deg_{out}(i)$) que define a quantidade de arestas que saem do vértice i e se conectam a outros vértices (Alcalá-Corona et al., 2021).

A conectividade total de uma rede é definida como na equação 2.11, na qual E é o número de arestas e N o total de vértices.

$$C = \frac{E}{N(N-1)} \tag{2.11}$$

A estrutura de conectividade de redes biológicas é geralmente informativa, a respeito de interação e reversibilidade de reações, compostos que estruturam a rede, como no metabolismo, ou nas relações tróficas, como nas redes de alimentação (Huynh-Thu e Sanguinetti, 2019).

Os relacionamentos descritos nas redes geralmente são representados através da estrutura de dados chamada matriz de adjacência. Uma matriz de adjacência pode ser definida como $M = (e_{i,j})$, correspondente ao grafo $G = (V, E)$, no qual onde, $e_{i,j} = 1 \quad se \quad \in E$ ou $e_{i,j} = 0 \quad se \quad \notin E$. Um exemplo de matriz de adjacência e dígrafo correspondente pode ser observado na Figura 2.7.

Os grafos podem ser representados por listas de adjacência (Figura 2.7. (b), estas listas também podem registrar os pesos das conexões, adaptando-se à estrutura de dados utilizada para representá-las. A representação de redes biológicas por meio de grafos pode trazer à luz informações relevantes para o conhecimento estrutural e organizacional dessas redes.

A estrutura da rede permite o cálculo de várias medidas que capturam diferentes recursos da topologia da rede que podem revelar informações importantes sobre a biologia subjacente do sistema (Marku e Pancaldi, 2023). Portanto, a observação de diferentes propriedades de redes biológicas pode proporcionar perspectivas valiosas a respeito de diversos processos, evolutivos, regulatórios de expressão, entre outros. Algumas das principais propriedades serão apresentadas a seguir.

A densidade de um grafo mostra o quão esparso ou denso o grafo é de acordo com o número de conexões por conjunto de vértices, a densidade de um grafo é definida como:

$$density = \frac{2|E|}{|V|(|V|-1)}$$

Já a densidade de um dígrafo é definida como:

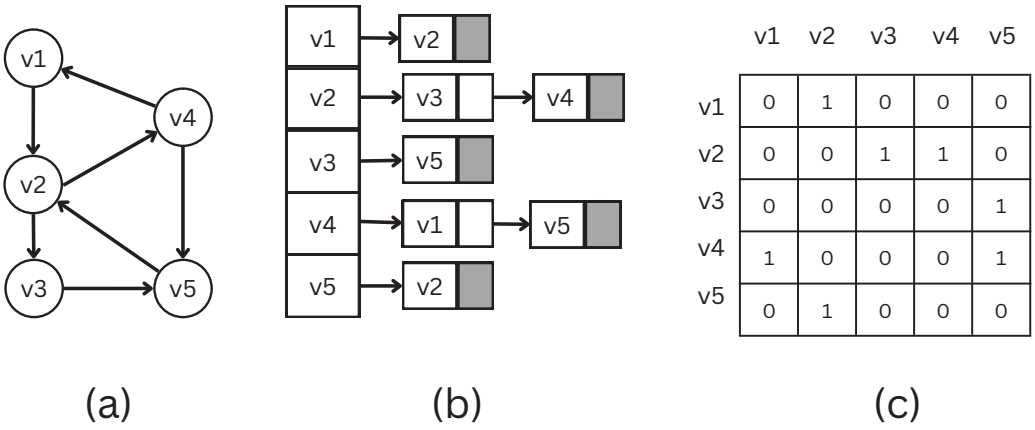


Figura 2.7: Exemplo de um dígrafo com 5 vértices.
Exemplo de um dígrafo com 5 vértices. (a) Dígrafo (b) lista e (c) matriz de adjacência. Cada elemento relacionado a um vértice na lista representa um relacionamento do vértice, assim como cada elemento igual a 1 na matriz de adjacências representa um relacionamento entre dois vértices do dígrafo.
Fonte: Adaptado de (Pavlopoulos et al., 2011).

$$density = \frac{E}{V(V-1)}$$

Um grafo é considerado de alta densidade quando sua densidade se aproxima de 1, e de baixa densidade quando a densidade se aproxima de 0. No contexto de grafos, um caminho é definido como uma sequência específica de vértices. Este caminho pode ser simples, sem repetição de vértices, ou um ciclo, com repetição de vértices, de tal forma que o último vértice é igual ao primeiro (Redhu e Thakur, 2022).

Um grafo é denominado cíclico se contém pelo menos um ciclo. Um ciclo é uma sequência de vértices na qual o primeiro e o último vértices são idênticos e todos os demais são distintos. Em contrapartida, um grafo é denominado acíclico se não contém nenhum ciclo. Um grafo completo é aquele no qual cada par de vértices é adjacente. Um grafo é considerado conectado se, em um grafo não direcionado, for possível chegar de qualquer vértice a qualquer outro, seguindo uma sequência de arestas. Um grafo direcionado é fortemente conectado se existir um caminho direcionado de qualquer vértice para qualquer outro vértice .

A propriedade de distância é amplamente utilizada nas análises de grafos. A distância $\delta(i, j)$ de i até j é o comprimento do caminho mais curto de i a j em G. Se tal caminho não existir, então definimos $\delta(i, j) = \infty$, assumindo que os nós estão tão distantes uns dos outros que não estão conectados.

O comprimento médio do caminho e o diâmetro de um grafo G são definidos como sendo o valor médio e máximo de $\delta(i, j)$ tomado sobre todos os pares de nós distintos, $i, j \in V(G)$ que estão conectados por pelo menos um caminho (Redhu e Thakur, 2022).

Com base no conhecimento acerca das propriedades dos grafos e considerando que estes são excelentes representações de redes regulatórias, é possível integrar a teoria dos grafos aos

conceitos biológicos. Dessa forma, conseguimos descrever de maneira mais precisa o sistema biológico de regulação gênica.

Vários métodos experimentais, matemáticos e *in silico* foram introduzidos para a detecção, visualização e análise de interações entre genes e TFs (Meyer et al., 2008; Kuang et al., 2023). As técnicas experimentais para a detecção da interação gênica de TFs são a imunoprecipitação de cromatina (ChIP), DamID e o sistema híbrido de levedura (Y1H), enquanto as relações regulatórias são detectadas pela correlação de perfis transcricionais de genes e reguladores putativos (MacNeil e Walhout, 2011). A combinação de relações físicas e regulatórias é necessária para gerar *GRNs* completas e precisos

As características e propriedades dos grafos são utilizadas na análise de redes. No contexto das redes de regulação gênica, existe uma ampla variedade de soluções computacionais para inferência (Redhu e Thakur, 2022). Nesta tese, iremos abordar as metodologias para a inferência de redes regulatórias de genes. As análises selecionadas para esta tese são exclusivamente aquelas que resultam em grafos não direcionados.

2.4.1 Métodos de inferência de redes regulatórias

Conforme descrito nas seções anteriores, a inferência de redes regulatórias de genes (*GRNs*) tem como base a modelagem matemática para representar as interações biológicas. As estratégias, portanto, buscam modelar dependências estatísticas entre genes, com base nas variações de perfis de expressão.

Os métodos computacionais para a reconstrução de *GRNs* podem ser divididos em três grandes categorias, conforme proposto em (Marbach et al., 2012). A primeira inclui métodos de teoria da informação (Steuer et al., 2002; Butte e Kohane, 2000), os quais identificam dependências estatísticas entre duas moléculas. O termo “dependência”, nesse contexto, refere-se a qualquer situação em que as variáveis aleatórias não satisfazem uma condição matemática de independência probabilística. Esses métodos se baseiam em duas medidas básicas da teoria da informação: informação mútua (e coeficiente de correlação (Pham et al., 2012). A informação mútua é a medida mais utilizada, pois leva em conta qualquer tipo de dependência, enquanto coeficiente de correlação responde apenas à dependência linear (veja uma comparação em Camacho et al., 2007). A segunda categoria consiste em redes bayesianas e gráficas, que maximizam uma função de pontuação em alguns modelos de rede alternativos para encontrar o que melhor se ajusta aos dados (Liu et al., 2016). A terceira inclui equações diferenciais e de diferença (Gardner et al., 2003).

Uma técnica comumente utilizada na inferência de *GRNs* é a seleção de características (Lopes et al., 2008), incluindo estratégias adicionais como a proposta de algoritmo para o crescimento de pequenas sub-redes a partir de um conjunto de genes (Hashimoto et al., 2004b). Algoritmos de busca em conjunto com a seleção de características também são explorados por outros métodos de inferência de *GRNs* (Lopes et al., 2010; da Rocha Vicente e Lopes, 2014; Lopes et al., 2014).

A literatura sobre inferência de *GRNs* é extensa, e algumas revisões e comparações corroboram com avanços metodológicos. Delgado e Gómez-Vela, 2019 aborda os principais conceitos sobre *GRNs*, como tipo de dados, métodos de aprendizado de máquina e ferramentas utilizadas na inferência de *GRNs*. Huynh-Thu e Sanguinetti, 2019 apresenta uma introdução aos conceitos básicos das ferramentas de inferência de *GRNs* e destaca os pontos em comum e os pontos fortes dos métodos existentes da literatura. Pušnik et al., 2022 revisita os métodos de inferência de redes booleanas e apresenta uma nova metodologia. Redhu e Thakur, 2022 apresenta uma revisão detalhada com as características gerais e tipos de redes biológicas, além de metodologias experimentais, bancos de dados biológicos e perspectivas para o futuro. As

aplicações para a inferência de *GRNs* com dados de sequenciamento de célula única (*Single-cell*) também são discutidas em Wang et al., 2023.

Dentre as metodologias desenvolvidas para inferir *GRNs*, algumas se destacam pela facilidade de uso, como as implementadas na linguagem R. Estas geralmente são distribuídas como pacotes e podem ser utilizadas com poucos comandos.

O pacote R, denominado *minet* (Redes de Informação Mútua, do inglês *Mutual Information NETworks*) (Meyer et al., 2008), oferece diversas metodologias para realizar essa inferência. Nesse contexto, destacam-se as análises *CLR*, *MRNET*, *MRNETB* e *ARACNE*. Apresentamos a seguir o conceito de informação mútua e aplicações em inferência de redes regulatórias.

Duas formas populares de calcular a informação mútua são: i) discretizar as variáveis com intervalo de frequência equivalente; ii) assumir uma variável distribuída normalmente (Meyer et al., 2010a).

A discretização é um processo usado para transformar variáveis contínuas em um formato discreto. Para isso, pode-se utilizar uma categoria, ou *bin*. O *binning* de frequência equivalente é um método específico de discretização que busca garantir que cada *bin* contenha um número igual de observações. Esse método é aplicado em casos nos quais temos variáveis contínuas com valores distribuídos de forma desigual. Ao aplicar a discretização, espera-se que as distribuições marginais (a distribuição de cada variável individualmente) se tornem uniformes. Isso significa que cada valor possível da variável tem a mesma probabilidade de ocorrer. O pacote R *infotheo* implementa a discretização por meio do método *discretize* (Meyer, 2010).

Após a discretização, as análises do pacote *minet* seguem um processo que começa com a geração de uma matriz de informação mútua. A informação mútua é uma medida da dependência mútua entre duas variáveis. Essa medida é calculada para cada par de nós (neste caso, genes). Mais especificamente, a informação mútua quantifica a informação obtida sobre uma variável aleatória pela observação de outra variável aleatória. A matriz de informação mútua é composta pelos elementos i e j

$$MIM_{ij} = I(X_i; X_j) = \sum_{x_i \in x} \sum_{X_j \in X} p(x_i, x_j) \log \left(\frac{p(x_i, x_j)}{p(x_i)p(x_j)} \right) \tag{2.12}$$

onde informação mutua entre X_i e X_y dado que $X_i \in \chi, i = 1, ..., n$, é uma variável randômica discreta que denota o nível de expressão do i-ésimo gene, conforme apresentado no equação 2.12.

Utilizando a *MIM* (Matriz de Informação Mútua, do inglês *Mutual Information Matrix*) , descrita na equação 2.12, foram desenvolvidos alguns métodos.

O método RELNET (Rede de Relevância, do inglês *Relevance Network*) (Butte e Kohane, 2000), foi aplicado com sucesso na inferência de relações em dados de expressão gênica (Butte et al., 2000). A abordagem consiste em inferir uma GRN, na qual um par de genes $\{X_i; X_j\}$ está relacionado a um gene se a informação mútua $I(X_i; X_j)$ for maior que um dado limiar I_0. Um ponto importante sobre esse método é o fato de ser propenso a inferir falsos positivos no caso de interações indiretas entre genes. Por exemplo, se o gene $X1$ regula tanto o gene $X2$ quanto o gene $X3$, uma alta informação mútua entre os pares $\{X1, X2\}$, $\{X1, X3\}$ e $\{X2, X3\}$ estaria presente. Como consequência, o algoritmo infere uma aresta entre $X2$ e $X3$, embora esses dois genes interajam somente por meio do gene $X1$ (Meyer et al., 2007).

Existem outros pacotes R que implementam algoritmos para a inferência de *GRNs*. Nesta tese, os métodos listados abaixo foram considerados para uma avaliação de desempenho:

1. *ARACNE* (Algoritmo para a Reconstrução de Redes Celulares Precisas, do inglês *Algorithm for the Reconstruction of Accurate Cellular Networks*): O *ARACNE* adota uma abordagem baseada na não conformidade, que utiliza informação mútua. A não conformidade afirma que, se o gene X_1 interage com o gene X_3 por meio do gene X_2, então:

$$I(X_1, X_3) \leq \min(I(X_1, X_2), I(X_2, X_3)) \tag{2.13}$$

   O *ARACNE* inicia associando a cada par de genes um peso equivalente à informação mútua. Assim como em uma rede de relevância, todos os nós com $I(X_1; X_2) < I_0$ são removidos, sendo I_0 um limiar definido. Eventualmente, um gene mais fraco em um trio pode ser interpretado como uma interação indireta e é removido, se a diferença entre os dois menores pesos estiver acima de um limiar W_0. É importante observar que ao aumentar I_0, o número de arestas inferidas tende a ser menor, enquanto o efeito oposto é obtido ao aumentar W_0.

   Essa metodologia é particularmente eficiente na detecção de relações não lineares e na eliminação de interações espúrias. O algoritmo realiza uma filtragem rigorosa, removendo as relações que não apresentam informações mútuas significativas. O *ARACNE* se destaca por sua capacidade de identificar interações específicas e confiáveis, contribuindo para redes regulatórias mais precisas (Margolin et al., 2006b);

2. *MRNET* (Rede de Redundância Mínima, do inglês *Minimum Redundancy Network*): A metodologia *MRNET* enfatiza a minimização da redundância nas redes inferidas. Esse método adota uma abordagem de seleção de características para identificar as interações mais informativas e, ao mesmo tempo, evitar sobreposições significativas. Para tanto *MRNET* infere uma GRN utilizando o método *MRMR* (Máxima Relevância/ Miníma Redundância) de seleção de características. A ideia consiste em executar uma série de seleções supervisionadas de genes com *MRMR*, onde cada gene por sua vez desempenha o papel de saída esperada.

   Considere um classificador supervisionado no qual a saída é denominada Y e o conjunto de variáveis de entrada denominado V. O método *MRNET* classifica o conjunto de dados de entrada V de acordo com uma pontuação que é a diferença entre a *MI* com a variável de saída Y (máxima relevância) e a média da informação mútua com todas as variáveis previamente classificadas (redundância mínima). As interações diretas devem ser "bem"classificadas, entretanto as interações indiretas "mal"classificadas pelo método.

   Inicia-se então a busca gulosa, que começa selecionando a variável X_i com a maior informação mútua para o alvo Y. A segunda variável selecionada X_j será aquela com uma alta informação $I(X_j; Y)$ para o alvo e, ao mesmo tempo, uma baixa informação $I(X_j; X_i)$ para a variável selecionada anteriormente. Nas etapas a seguir, dado um conjunto S de variáveis selecionadas, o critério atualiza S escolhendo a variável

$$X_j^{MRMR} = \arg \max_{X_j \in VS} (u_j - r_j) \tag{2.14}$$

   que maximiza a pontuação

$$s_j = u_j - r_{j'} \tag{2.15}$$

Onde u_j é um termo de relevância e r_j é um termo de relevância. Mais precisamente,

$$u_j = I(X_i; Y) \tag{2.16}$$

é a *MI* onde X_j com a variável alvo Y, e

$$r_j = \frac{1}{|S|} \sum_{x_k \in S} I(X_j; X_k) \tag{2.17}$$

mede a redundância média de X_j para cada uma das variáveis já selecionadas $X_k \in S$. Em cada etapa do algoritmo, espera-se que a variável selecionada permita uma compensação eficiente entre relevância e redundância.

A análise *MRNET* é particularmente eficaz na identificação de conexões não redundantes, contribuindo para uma representação simplificada e informativa da GRN (Meyer et al., 2007);

3. *CLR* (Probabilidade de relação entre contextos, do inglês *Context Likelihood of Relatedness*) : A abordagem *CLR* busca identificar bordas entre genes com base na probabilidade condicional dada a expressão dos genes vizinhos. O algoritmo *CLR* (Faith et al., 2007) é uma extensão do RELNET (Butte e Kohane, 2000). O *CLR* calcula a informação mútua (*MI*) para cada par de genes e obtém uma pontuação relacionada à distribuição empírica desses valores de *MI*. Ao invés de considerar a informação mútua $I(X_i; X_j)$ entre o gene X_i e X_j e considera o score $z_{i,j} = \sqrt{z_i^2 + z_j^2}$

$$z_j = \max\left(0, \frac{I(X_i; X_j) - \mu_i}{\sigma_i}\right) \tag{2.18}$$

e μ_i e σ_i são, respectivamente, a média e o desvio padrão da distribuição empírica dos valores de informação mútua $I(X_i; X_k), k = 1, ..., n)$. O método *CLR* se destaca por sua capacidade de capturar bordas não lineares, sendo especialmente útil em contextos em que as interações não seguem padrões lineares previsíveis (Meyer et al., 2008). A interpretação dos resultados da *CLR* é baseada na medida de probabilidade condicional, fornecendo informações sobre as bordas de dependência entre os genes (Faith et al., 2007);

4. *GENIE3* é um algoritmo que usa conjuntos de árvores de regressão para inferir *GRNs* a partir de dados de expressão. A estrutura da árvore de decisão (DT) para resolver problemas de regressão ou classificação. A rede é representada como uma matriz de adjacência ponderada.

   A abordagem de *GENIE3* é apresentada na Figura 2.8, na qual, para cada gene, uma amostra de aprendizado é gerada com níveis de expressão de j como valores de saída e níveis de expressão de todos os outros genes como valores de entrada. Uma função é aprendida e uma classificação local de todos os genes, exceto j, é calculada. As p classificações locais são então agregadas para obter uma classificação global de todas as ligações regulatórias. O *GENIE3* foi identificado como um dos principais métodos em termos de precisão na inferência de *GRNs* de genes a partir de dados transcriptômicos (Huynh-Thu et al., 2010);

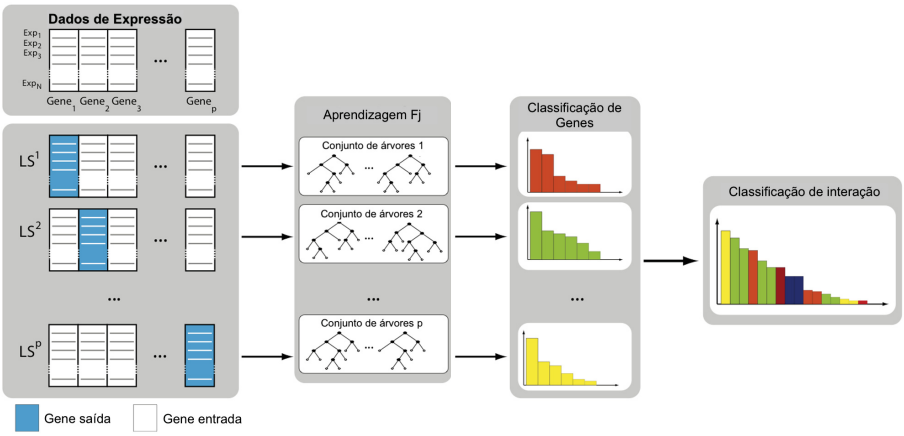


Figura 2.8: Metodologia do pacote R *GENIE3*.
Metodologia do pacote R *GENIE3*. Cada gene é considerado uma amostra de aprendizagem, e os outros são considerados dados de entrada, gerando uma árvore de decisão, a junção dessas árvores gera um ranking de interações.
Fonte: Adaptado de (Huynh-Thu et al., 2010).

5. *C3NET*: adota aprendizado híbrido de rede Bayesiana dinâmica para inferir *GRNs* com atraso de tempo. O *C3NET* é um algoritmo não supervisionado que usa dados de expressão de *microarray* para inferir *GRNs* com interações físicas diretas.

   A ideia básica da *C3NET* consiste na identificação de uma rede de informações mútuas de máximas significâncias, o núcleo causal conservador, de modo que dois genes só estejam conectados entre si se o valor de informações mútuas significativas compartilhadas for, pelo menos para um desses dois genes, máximo em relação a todos os outros genes (Altay e Emmert-Streib, 2010a);

6. *MRNETB* (Bootstrapped MRNET): Uma versão do *MRNET* (Meyer et al., 2007) com melhorias incrementais. O *MRNETB* incorpora técnicas de bootstrap para avaliar a robustez das interações inferidas. A versão anterior *MRNET* infere interações entre TFs e genes utilizando o método "*foward selection*"(busca progressiva), estratégia para identificar um conjunto de vizinhos com independência máxima para cada variável. Entretanto métodos baseados em *foward selection*, dependem da escolha correta do primeiro vizinho, e são afetados em performance se a escolha do primeiro vizinho estiver incorreta. *MRNETB* supera esta limitação implementado uma combinação de "*backward elimination*"(eliminação regressiva) e um procedimento de substituição sequencial (Sutter e Kalivas, 1993). Esse método usa amostragem de substituição para gerar vários conjuntos de dados, permitindo uma análise estatística da estabilidade das bordas identificadas. A inclusão do bootstrap no *MRNETB* melhora a confiabilidade dos resultados, fornecendo informações sobre a variabilidade das interações na GRN (Meyer et al., 2010b);

7. *BC3NET*: A *GRN G* é inferida de um conjunto de *bootstrap* gerado a partir de um único conjunto de dados de expressão gênica D. Para cada conjunto de dados gerado no conjunto, D_k^b, uma *GRN* G_k^b é inferida usando o *C3NET*. A partir de $G_{k\,k=1}^{b\,B}$, é obtida uma rede agregada G_w^b, cujas bordas são usadas como estatísticas de teste para obter a rede final G (de Matos Simoes e Emmert-Streib, 2012).

8. *ENNET*: emprega a técnica de aumento de gradiente para discernir associações regulatórias entre genes. Aumento de gradiente é um método de otimização em aprendizado de máquina que busca minimizar uma função de perda. O método realiza essa tarefa de modo iterativo, movendo-se na direção do gradiente negativo da função. Em cada passo, um novo modelo é adicionado que é treinado para corrigir os erros cometidos pelo modelo existente. Isso é feito ajustando os pesos do modelo na direção que minimiza o erro, daí o nome "aumento de gradiente". Esta técnica é amplamente utilizada em problemas de classificação e regressão. Inicialmente, o *ENNET* padroniza os dados de expressão gênica e elimina os *outliers*. Em seguida, ele utiliza um conjunto de dados de treinamento que inclui genes rotulados como "reguladores" e "alvos". O modelo é treinado para reconhecer padrões de expressão gênica que diferenciam essas duas categorias. Para inferir a *GRN*, o modelo treinado é usado para estimar a probabilidade de um gene regular outro. Essas previsões são posteriormente empregadas para construir uma rede de genes reguladores (Sławek e Arodź, 2013);

9. *Corto*: O *Corto* combina duas técnicas fundamentais: A correlação de *Pearson* e a desigualdade de processamento de dados (DPI). Ele requer uma lista de centroides (fatores de transcrição) e um conjunto de genes-alvo. Para cada par centroide-alvo, é calculado o coeficiente de correlação de *Pearson*, que quantifica a correlação linear entre os perfis de expressão do centroide e do alvo. As bordas com coeficientes de correlação acima de um determinado limite de significância são consideradas significativas. O limite pode ser definido pelo usuário ou determinado com base em critérios estatísticos (por exemplo, valor de p) (Mercatelli et al., 2020);

10. *GeCoNet-Tool*: calcula o Coeficiente de Correlação de *Pearson* (*PCC*) entre cada par de genes com base nos dados processados. A matriz *PCC* é salva como uma matriz triangular superior se o usuário optar por salvar a matriz *PCC*. O *GeCoNet-Tool* classifica os *PCCs* em diferentes intervalos (tamanho do compartimento) com base no número de condições emparelhadas de pares de genes, o que pode ser especificado pelo usuário. Para selecionar bordas com base no valor de *PCC*, usa o valor de corte expresso como a porcentagem superior escolhida de todos os *PCCs* em um determinado intervalo (por exemplo, 0,005, 0,01 ou 0,02), que é usado para determinar o limiar deslizante, ajustando a curva a seguir:

$$f^{thres}(x) = \alpha - \frac{1}{\eta + \lambda_e^{-\frac{x}{\beta}}} \tag{2.19}$$

Onde α, η, λ e β, são quatro parâmetros que foram ajustados, e x é o número de elementos emparelhados (Kuang et al., 2023).

Estudos anteriores têm comparado metodologias de inferência de redes regulatórias gênicas (Lopes et al., 2009; Delgado e Gómez-Vela, 2019; Huynh-Thu e Sanguinetti, 2019; Maetschke et al., 2014; Marku e Pancaldi, 2023), destacando desafios importantes e deixando em aberto algumas questões, como a validação de *GRNs* e a inferência orientada por dados, especialmente no contexto de dados de *scRNA-Seq* (Aibar et al., 2017; Kharchenko et al., 2014). Pesquisas mais recentes vêm explorando abordagens específicas para inferência de *GRNs* a partir de dados de célula única, ampliando o leque de possibilidades analíticas (Chen e Mar, 2018; Kim et al., 2023; Todorov et al., 2019).

A literatura sugere que a combinação de múltiplas metodologias pode gerar resultados mais robustos do que o uso isolado de um único método (Marbach et al., 2012). De fato, já

foi demonstrado que estratégias baseadas em consenso podem melhorar significativamente a acurácia em tarefas como a análise de expressão diferencial de genes (Costa-Silva et al., 2017a).

Alguns trabalhos propõem o uso de conjuntos de algoritmos de aprendizado de máquina para inferência de *GRNs*, demonstrando desempenho superior quando comparado a métodos tradicionais (Alawad et al., 2023; Musilova et al., 2024). A maioria dessas avaliações tem utilizado como referência os conjuntos de dados dos desafios DREAM 4 (Marbach et al., 2009) e DREAM 5 (Marbach et al., 2012), amplamente reconhecidos na área por fornecerem dados simulados e reais com redes de referência conhecidas.

Esta tese tem como objetivo atualizar e expandir as avaliações de metodologias clássicas e recentes de inferência de *GRNs*, comparando seus desempenhos individuais e em combinações consensuais. Adicionalmente, realiza-se uma análise dos sinais de expressão utilizados como entrada para esses métodos, com foco na caracterização das arestas inferidas. São analisadas as conexões identificadas pela maioria dos métodos, aquelas detectadas apenas por subconjuntos e também as arestas que não foram inferidas por nenhum dos métodos avaliados.

Essa caracterização é conduzida com base nas propriedades dos sinais de expressão, em especial sua entropia. O objetivo é compreender como essas propriedades se associam ao sucesso ou fracasso dos métodos de inferência em identificar determinadas interações. Como contribuição, esta abordagem propõe uma nova perspectiva para avaliar e melhorar a eficácia dos algoritmos de inferência, fornecendo subsídios para o aprimoramento de metodologias existentes e o desenvolvimento de novas estratégias para reconstrução de *GRNs*.

3 MATERIAIS E MÉTODOS

Este capítulo apresenta os conjuntos de dados e as metodologias utilizadas para obter os resultados desta tese. Ele é composto por subseções que detalham os materiais e métodos de cada fase do projeto, incluindo: a revisão de métodos computacionais para análise de expressão diferencial de genes (Costa-Silva et al., 2023); a implementação de um pacote R que aplica a metodologia de consenso atualizada e revisada (Costa-Silva et al., 2017a); e a análise de métodos computacionais para inferência de redes gênicas regulatórias.

Como apresentado no Capítulo 1, a identificação de *DEGs* é uma tarefa que requer vários passos, independentemente da origem dos dados de expressão. No entanto, identificar *DEGs* é apenas desvendar a primeira parte de uma cadeia de processos biológicos inter-relacionados. De modo geral, se faz necessário identificar quais mecanismos ativam ou silenciam a expressão dos genes envolvidos no processo biológico de interesse.

Conforme apresentado no Capítulo 2, os dados de expressão podem ser gerados de várias maneiras. Nesta tese, utilizamos dados de *RNA-Seq* e Microarray. Os dados de *RNA-Seq* são gerados a partir do sequenciamento, que tem formato de saída FASTQ, contendo milhares de pequenas sequências, os *reads*, que são mapeados ao genoma ou transcriptoma de referência. A saída de mapeamento em formato de arquivos BAM ou SAM é utilizada para a contagem de *reads* mapeados, gerando assim a tabela de contagem contendo genes nas linhas e amostras nas colunas (várias colunas podem pertencer a um mesmo grupo de amostras). As análises partem dos valores inteiros da tabela de contagem, com os quais se executa a análise de expressão diferencial de genes. Os métodos que realizam essas análises, em geral, normalizam os dados e, com base em uma distribuição (métodos paramétricos) ou em uma hipótese (métodos não paramétricos), calculam o log Fold-Change, P-Value ou outra métrica para cada gene (dependendo do método). Esse cálculo é realizado com base nos valores de contagem do gene comparados aos diferentes grupos do experimento.

Com base nos resultados de um estudo anterior, que apresentou uma metodologia de consenso para análise de expressão diferencial de genes aplicada a dados de *RNA-Seq* (Costa-Silva et al., 2017a), esta tese buscou identificar o estado da arte nesse contexto e compreender a evolução dos métodos de análise de expressão diferencial de genes com dados de *RNA-Seq*. A metodologia da revisão desenvolvida sobre esse tema é apresentada na Seção 3.1.

Neste contexto, foi implementada uma nova versão da metodologia de consenso em formato de pacote R, o *consexpressionR* (Costa-Silva et al., 2024). Os passos dessa implementação, bem como os dados utilizados nos testes, são apresentados na Seção 3.1.

Considerando a sequência natural das análises de expressão gênica, esta tese realizou uma análise sobre métodos de inferência de redes de regulação gênica. Os dados utilizados nesta análise e a metodologia empregada são apresentados na Seção 3.3.

3.1 REVISÃO DE MÉTODOS COMPUTACIONAIS PARA ANÁLISE DE EXPRESSÃO

Para avaliar a literatura recente sobre as metodologias para análise de *DEGs* com dados de *RNA-Seq*, foram investigados os métodos computacionais desenvolvidos com essa finalidade. Devido ao grande volume de estudos sobre essas análises (*DEGs*), especificamente para dados de *RNA-Seq*, definimos um termo de busca. Esse termo tem como objetivo ser abrangente, mas restringir o tema central dos trabalhos encontrados. O termo de busca "*RNA-Seq differential*

expression analysis" foi utilizado na ferramenta de busca acadêmica do *Google Scholar* (Schoolar, 2024).

Conforme apresentado na Figura 3.1, os resultados da busca foram refinados por alguns critérios. Consideramos apenas métodos implementados em forma de software ou pacote (com base na definição de (Pressman e Maxim, 2021) para os termos *software*, pacote ou biblioteca) e que disponibilizam a opção de análise de *DEGs*, exceto os estudos seminais. Aqui consideramos como seminais os estudos que serviram como base para as análises de *DEGs* com dados de *RNA-Seq* e que foram os primeiros a analisar expressão com esses dados. Dentre os resultados do critério descrito, refinamos a seleção aplicando o critério de citações por ano de publicação, conforme descrito abaixo.

- Ano de publicação entre 2006 e 2017 e mais que 200 citações no *Google Scholar*;
- Ano de publicação 2018 e mais que 100 citações no *Google Scholar*;
- Ano de publicação 2019 e mais que then 50 citações no *Google Scholar*;
- Ano de publicação 2020 e mais que then 10 citações no *Google Scholar*;
- Ano de publicação 2021 e mais que 5 citações no *Google Scholar*;

As metodologias selecionadas, com base nos critérios apresentados, são listadas no Material suplementar A.1. Com base nesses dados, são desenvolvidos os itens 4, 5 e 6 da revisão, apresentados na Figura 3.1. É importante observar que os métodos desenvolvidos especialmente para a análise de *DEGs* com dados de *RNA-Seq* célula única (*scRNA-Seq: Single Cell RNA-Seq*) foram identificados com o símbolo "*" ao final de seu nome.

Para as metodologias que atendem aos critérios dos itens 1 a 3 da Figura 3.1, foi avaliado o relacionamento de dependência entre elas. Para essa avaliação, foram revisados os textos originais de cada metodologia e suas declarações de dependência ou importação em seus respectivos repositórios oficiais. Utilizando a dependência e/ou importação, foi definida uma matriz de adjacência, onde cada dependência gera uma conexão entre os nós da rede. Essas conexões são direcionais, de modo que o sentido da seta indica que o nó de origem depende do nó de destino. A rede resultante dessa análise é detalhada no Capítulo 4 e apresentada na Figura 4.2.

Após o levantamento de estudos identificados, foi realizada uma curadoria manual para a classificação de cada método, análise da relação de dependência/uso entre eles. Cada método teve sua publicação avaliada, onde foi identificada sua metodologia de identificação de genes diferencialmente expressos e os métodos em que se apoia ou faz uso para gerar tais resultados.

Observado o grande volume de dependências entre métodos selecionados, também contabilizamos a frequência dessas dependências. Foram contabilizadas as arestas direcionadas a cada vértice da rede de ferramentas, gerando assim o histograma de frequência apresentado na Figura 4.3. Para esta contagem, a rede de relações entre os estudos foi avaliada e todos os nós que recebiam arestas (eram utilizados por outra metodologia) foram contabilizados.

Ainda em busca de compreender a evolução temporal, os métodos foram organizados em uma linha do tempo, com base em mês e ano de publicação. Os resultados das análises descritas nesta seção são apresentados no Capítulo 4 e publicados no artigo (Costa-Silva et al., 2023).

3.2 IMPLEMENTAÇÃO DO PACOTE R: *CONSEXPRESSIONR*

Dentre as metodologias identificadas na "revisão de métodos para análise de expressão (Costa-Silva et al., 2023)", destaca-se o "*consexpression* (Costa-Silva et al., 2017a)". Trata-se de

uma metodologia híbrida que utiliza métodos paramétricos e não paramétricos associados para a definição de *DEGs*, mostrando-se precursora. Portanto, esta tese implementou a metodologia em formato de pacote R denominada *consexpressionR* (Costa-Silva et al., 2024), que ainda está em fase de publicação em repositório oficial.

Para a implementação do pacote R *consexpressionR*, levou-se em consideração as características do uso de consenso na definição de *DEGs*, principal contribuição da versão inicial (implementada na linguagem de programação Python (Foundation, 2024)). No entanto, algumas atualizações e necessidades apresentadas pelos usuários no repositório da versão inicial (Costa-Silva et al., 2017b) foram consideradas nesta nova versão. São elas:

1. Atualizações: novas versões de pacotes R e outras extensões utilizadas geram problemas na execução, ao alterar a forma de comandos e parâmetros, por exemplo;

2. Necessidade de análise parcial: o *consexpression* executa todos os passos da análise de expressão (mapeamento, contagem e identificação de *DEGs*), mas alguns usuários desejam apenas a identificação de *DEGs*;

3. Análises para organismos sem genoma de referência: o *consexpression* precisa do arquivo .FASTA com o genoma de referência e do arquivo .GFF de anotação do genoma para a sua execução, o que impede usuários com dados *de novo* de realizarem suas análises;

4. Usabilidade: o *consexpression* não possui interface gráfica do usuário (GUI), o que pode ser um fator limitante para o uso da ferramenta.

Nesse contexto, a implementação do *consexpressionR* segue o padrão de pacote R, conforme apresentado no livro Wickham, 2015. Para tanto, foi utilizado o pacote de apoio ao desenvolvimento de pacotes R, o *devtools* (Wickham et al., 2022), que auxilia na criação de estrutura. A documentação foi desenvolvida utilizando o pacote *roxygen2* (Wickham et al., 2024). A interface gráfica foi desenvolvida utilizando o pacote R *Shiny* (Chang et al., 2024).

Para solucionar as lacunas identificadas na primeira versão do consexpression a análise do *consexpressionR* parte dos valores inteiros da tabela de contagem, conforme apresentado na Figura 3.2. Na fase de configuração (em amarelo), espera-se que o usuário forneça os parâmetros do experimento, tais como o nome dos grupos de amostras (condições) e o número de amostras por grupo (replicatas). Na fase "Métodos para identificação de *DEGs*" (em roxo), o usuário tem a opção de parametrizar os métodos de identificação de *DEGs*. No entanto, os valores são preenchidos automaticamente com o padrão indicado no manual de cada método. Na fase de "Consenso" (em cinza), o usuário pode visualizar um gráfico que mostra a quantidade de genes indicados como diferencialmente expressos e a convergência desses resultados por meio de um gráfico para visualização de interseções de conjuntos (*Upset plot*). Nesta mesma fase da análise, o usuário também pode definir quais genes considerar, com base no número de métodos que indicaram este gene como DE. O valor definido por padrão é 5. Na fase de "Visualização" (em verde), é gerado um mapa de calor (*HeatMap*), apenas com os genes DE do consenso. Para este gráfico, utilizam-se os valores de expressão do arquivo de contagem fornecido inicialmente pelo usuário. Finalmente, na fase de "Relatórios" (em azul), fornecemos uma lista de *DEGs*, encontrados pelo consenso, contendo as métricas de cada metodologia. Na visualização, assim como na exportação, as colunas são nomeadas no seguinte formato: [nome_do_método].[métrica].

A análise de *DEGs* é executada com sete métodos implementados em formato de pacote R, são apresentado na tabela 3.1. Para o método *SAMSeq* apenas a execução com dados de

contagem é permitida, para dados de quantificação este método não é executado. O método *KnowSeq* é executado apenas para dados com genoma de referência e nomes de genes válidos em dados de anotação públicos, como *ENSEMBL* (Mudunuri et al., 2009).

A metodologia de consenso implementada no *consexpressionR* aplica as apresentados na Tabela 3.1. Algumas melhorias na escolha das metodologias também foram consideradas: o método *DESeq* (Anders e Huber, 2010) foi descontinuado, sendo substituído pela versão dois da mesma metodologia, o *DESeq2* (Love et al., 2014). O pacote *baySeq* (Hardcastle e Kelly, 2010), também utilizado na versão inicial do *consexpression*, foi descontinuado, sem uma nova versão. Para substituí-lo, selecionamos a metodologia mais recente da revisão apresentada na Seção 3.1 e implementada em R, o pacote KnowSeq (Castillo-Secilla et al., 2021a).

Tabela 3.1: Métodos de análise de *DEGs* utilizados no pacote *consexpressionR*, ordenados por ano de publicação. A classificação utiliza a metodologia da revisão apresentada na Seção 3.1.
Fonte: De autoria própria.

Método	Referência	Classificação
edgeR	(Robinson et al., 2010)	Paramétrico
NOISeq	(Tarazona et al., 2011)	Não paramétrico
EBSeq	(Leng et al., 2013)	Paramétrico
SAMSeq	(Li e Tibshirani, 2013)	Não Paramétrico
DESeq2	(Love et al., 2014)	Paramétrico
limma	(Ritchie et al., 2015)	Paramétrico
KnowSeq	(Castillo-Secilla et al., 2021a)	Paramétrico

O *consexpressionR* gera um relatório com os *DEGs* indicados por, no mínimo, cinco ferramentas. Este valor pode ser parametrizado pelo usuário entre um e sete. Além de informar o usuário sobre a concordância dos métodos na indicação de *DEGs*, apresentando um gráfico para visualização de interseções de conjuntos (*Upset plot*). O *consexpressionR* traz uma maior consistência na indicação de *DEGs*, conforme indicado pelo estudo anteriormente desenvolvido (Costa-Silva et al., 2017a). Os genes indicados como diferencialmente expressos também podem ser visualizados por meio de um mapa de calor (também denominado *Heatmap*) (Warnes et al., 2009).

Para essas indicações, os resultados de cada método são filtrados com base nos parâmetros padrão (indicados no manual de cada pacote R) ou nos valores selecionados pelo usuário. Os grupos de genes considerados como diferencialmente expressos (DE) por cada método são comparados, e cada indicação como DE é contabilizada em uma matriz binária, onde as colunas representam os métodos e as linhas são os genes analisados. Essa matriz é utilizada para gerar os gráficos e relatórios. Para o limiar de consenso, qualquer combinação de métodos é considerada. Ou seja, independentemente de quais métodos apontaram o gene como DE, ele será considerado DE, desde que seja apontado por um valor igual ou maior que o definido como limiar de consenso.

O conjunto de dados A, utilizado nos testes do pacote *consexpressionR* está inicialmente disponível no arquivo de sequências curtas SRA (do inglês, *Short-Read Archive*) do Centro Nacional de Informação Biotecnológica (NCBI, do inglês *National Center for Biotechnology Information*), acessível através do código SRA010153. Foram utilizadas duas condições: amostras de tecido cerebral (*Brain*) e um mix de células humanas (*UHR*), cada uma com sete amostras. Os *reads* foram mapeados contra a versão 19 do genoma humano (GRCh37.p13) utilizando a ferramenta *Tophat* (Trapnell et al., 2009). O genoma e o arquivo de anotação estão disponíveis na página do projeto *GENCODE* (Harrow et al., 2012). Para as métricas de desempenho, foi utilizado como *gold standard* o experimento de qPCR, desenvolvido com as mesmas amostras

biológicas de tecido cerebral (*Brain*) e mix de células humanas (*UHR*). Este conjunto de dados também foi utilizado nos testes da versão inicial do *consexpression* (Costa-Silva et al., 2017a).

O conjunto de dados B está disponível no *Gene Expression Omnibus - GEO*, repositório do Banco Mundial de Sequências *NCBI* (Barrett et al., 2011) através do código de acesso GSE95077. Este conjunto de dados fornece a tabela de contagem de um experimento realizado com o sequenciador *Illumina HiSeq 2500*. O experimento envolveu duas condições: células de mieloma múltiplo, com dois tratamentos (BM e JJ), que representam a aplicação de amilorida em diferentes dosagens, e um controle (DMSO). Cada condição contém seis amostras, resultando em uma tabela de contagem com 18 colunas e 107 linha. Para as métricas de desempenho foram utilizados os dados de *qPCR* do mesmo estudo (Corchete et al., 2020b).

3.3 AVALIAÇÃO DE METODOLOGIAS PARA INFERÊNCIA DE REDES DE REGULAÇÃO GÊNICA

Nesta seção, apresentamos os métodos e conjuntos de dados utilizados nas análises realizadas para a avaliação de métodos de inferência de redes de regulação gênica (GRNs). Para a seleção dos métodos, consideramos aqueles recentemente implementados, além dos métodos avaliados pelo estudo (Marbach et al., 2012), implementados em R e/ou Python.

Os dados utilizados para a análise dos métodos de inferência de GRNs provêm do desafio DREAM 5 (Marbach et al., 2012). Esses dados contêm análises de expressão de *microarray* de *E.coli*, oriundos do banco de dados de expressão Omnibus (GEO) (Barrett et al., 2011), contendo 4511 genes em 806 condições, presentes no conjunto de dados.

A rede de regulação utilizada como referência representa interações transcricionais conhecidas para *E. coli*, obtidas na base de dados *RegulonDB* (Gama-Castro et al., 2011). Foram considerados apenas as interações estabilizadas e anotadas com "fortes evidências", de acordo com o *RegulonDB*. O total de interações presentes no conjunto de dados é de 2066. As características das ferramentas selecionadas são descritas na Tabela 3.2, ordenadas por ano de publicação:

Tabela 3.2: Métodos de inferência de redes gênicas avaliados: Métodos avaliados ordenador por ano de publicação. A coluna Metodologia de inferência apresenta a principal análise utilizada para definir as arestas de uma rede. **Fonte:** De autoria própria.

Método	Referência	Inferência
ARACNE	(Margolin et al., 2006a)	Informação mutua
MRNET	(Meyer et al., 2007)	Informação mutua
CLR	(Faith et al., 2007)	Informação mutua
GENIE3	(Huynh-Thu et al., 2010)	Árvores
C3NET	(Altay e Emmert-Streib, 2010b)	Informação mutua
MRNETB	(Meyer et al., 2010b)	Seleção *Forward*
BC3NET	(de Matos Simoes e Emmert-Streib, 2012)	*Ensemble (bagging)*
ENNET	(Sławek e Arodź, 2013)	Árvores de Regressão Gradiente
Corto	(Mercatelli et al., 2020)	Correlação em pares otimizada
GeCoNet-Tool	(Kuang et al., 2023)	Correlação de *Pearson*

Os resultados dos métodos de inferência foram comparados a fim de definir o limiar que otimiza os resultados para as arestas a serem consideradas. Utilizando a rede inferida por cada método, consideramos as arestas com "pontuação" dentro de uma faixa de valores. Dessa forma,

é possível verificar a influência do limiar utilizado no desempenho da ferramenta. A Figura 3.3 apresenta os passos adotados para a análise aplicada às metodologias de inferência de *GRNs*.

Inicialmente, na etapa "1 - Preparação do experimento", o conjunto de dados do desafio *DREAM5* foi transformado em um objeto R. A discretização foi realizada utilizando o pacote *infotheo* R (Meyer et al., 2008), por meio da função *discretize*.

Na etapa "2 - Execução dos métodos de inferência de *GRNs*", cada método selecionado foi executado em sua forma padrão (conforme o manual do usuário). Todos os resultados de cada método incluem arestas regulatórias inferidas e seus pesos correspondentes. A *GRN* pode então ser considerada de acordo com um limite específico destes pesos.

Na etapa, "3 - Definição de limiar", a rede inferida por cada método foi filtrada com dez limiares diferentes. O limiar mais baixo foi considerado como o valor mínimo encontrado na saída do método e o limiar mais alto como o valor mais alto encontrado na saída de cada método. Os intervalos entre cada um dos dez limites aplicados foram definidos usando a Equação 3.1.

$$intervalo = \frac{(max - min)}{10} \tag{3.1}$$

Em que o $intervalo$ é a faixa, o max é o valor mais alto encontrado na saída de cada método e o min é o valor mínimo encontrado na saída do método. A partir desse ponto, consideramos apenas as arestas que têm fatores de transcrição (*TF*) em pelo menos um dos nós.

Para avaliar qual dos limiares aplicados produz os melhores resultados para o conjunto de dados adotado, as redes inferidas por cada método foram avaliadas. Utilizou-se a rede *gold standard* (padrão-ouro) para calcular as seguintes métricas:

- **Verdadeiro Positivo (TP)**: são as arestas que existem tanto na rede padrão-ouro quanto na Rede de Regulação Gênica (GRN) inferida;
- **Falso Positivo (FP)**: são as arestas que não existem na rede padrão-ouro, mas estão presentes na *GRN* inferida;
- **Verdadeiro Negativo (TN)**: são as arestas que não existem nem na rede padrão-ouro nem na *GRN* inferida;
- **Falso Negativo (FN)**: são as arestas que existem na rede padrão-ouro, mas não estão presentes na *GRN* inferida.

Neste contexto, foram consideradas apenas as arestas em que, no mínimo um dos dois genes conectados é um fator de transcrição (*TF*).

Usando as métricas mencionadas acima, foram calculados *Accuracy (ACC), Recall, Precision, F-Score e False Discovery Rate (FDR)* para cada método individualmente em todos os limiares de corte para cada método. Todos os valores de limiar e as métricas de desempenho para cada análise podem ser encontrados no Apêndice C desta tese. Os valores de limiar que otimizam o resultado da área sobre a curva de sensibilidade e precisão, foram considerados o melhor limiar, e a "melhor rede" por cada método foi usada nas próximas etapas.

Na etapa "4 - Arestas", avaliamos as especificidades de cada método e sobrepusemos os resultados, utilizando apenas a "melhor rede" de cada um. Nesta etapa, desenvolvemos *scripts* em R e Python para calcular quantos métodos identificaram quais arestas. Avaliamos também quais métodos encontraram arestas que nenhum outro método identificou, denominadas neste contexto de arestas exclusivas. Adicionalmente, analisamos a intersecção entre os melhores resultados de cada método, verificando se a associação de vários métodos melhorou o desempenho da inferência da rede. Para cada iteração dessa análise, consideramos apenas as arestas indicadas

por x ou mais métodos, independentemente da combinação. O valor definido de x variou de 1 até o número total de metodologias testadas (n).

Na etapa "5 - Caracterização de arestas", buscamos identificar o que caracteriza as arestas exclusivas (indicadas por apenas um método), comuns (indicadas por todos os métodos) e não encontradas (as que não foram indicadas por nenhum método). Para essa análise, utilizamos os dados de expressão e calculamos a entropia de *Shannon* (Shannon, 1948) para cada gene, com base nos dados de expressão de *microarray* que geraram a rede. Os valores de entropia de cada nó das arestas foram analisados por meio de gráficos de dispersão.

Conforme descrito anteriormente, as três fases desta tese compõem um *pipeline* de análise comumente utilizado em estudos de expressão. É importante ressaltar que as análises de rede foram executadas com dados de *microarray*, devido à existência de uma rede de interações padrão-ouro para esses dados.

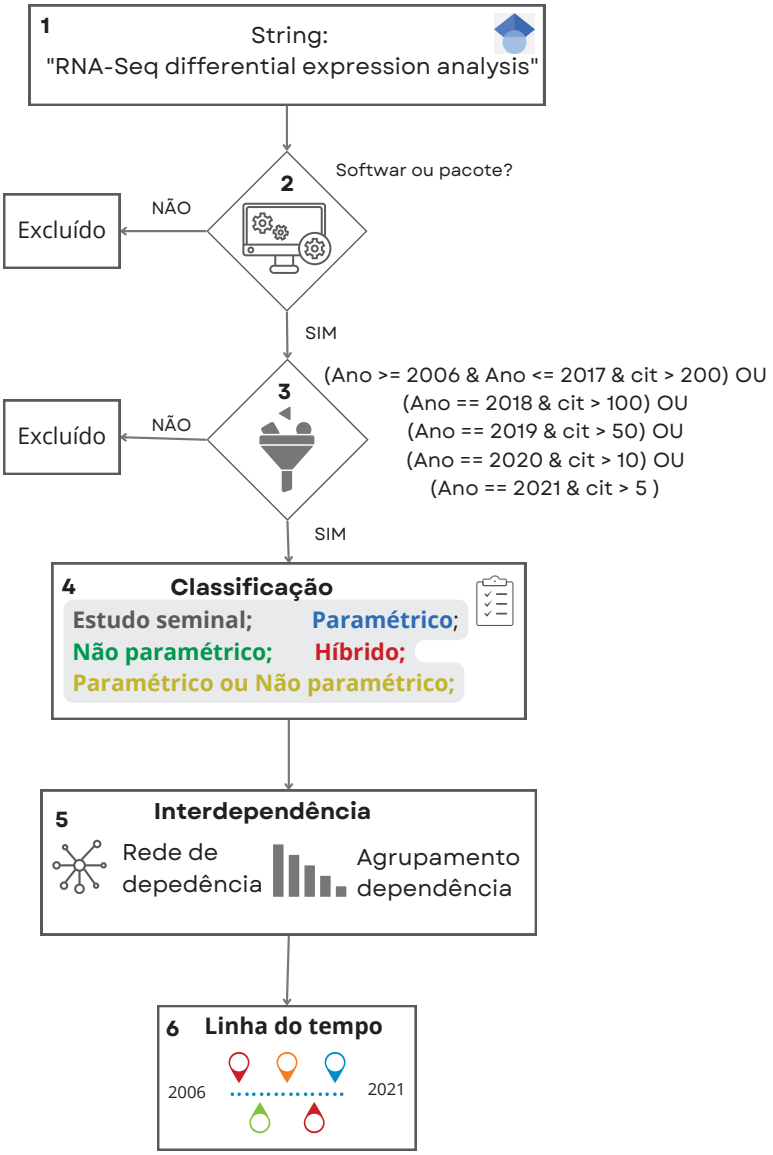


Figura 3.1: Fluxo de trabalho utilizado na revisão de métodos computacionais para análise de *DEGs*.

Fluxo de trabalho utilizado na revisão de métodos computacionais para análise de *DEGs* com dados de *RNA-Seq*. Inicialmente, foram selecionados todos os artigos com o termo "*RNA-Seq differential expression analysis*" (item 1). Desses resultados, foram selecionados apenas os trabalhos que implementam alguma metodologia em formato de software ou pacote (item 2). Dentre estes, selecionamos apenas os que realizam a análise de expressão diferencial de genes para *RNA-Seq* (total *bulk*) ou célula única (*Single-Cell*)), filtrando-os por ano de publicação e número de citações (item 3). As metodologias que atendiam aos critérios (itens 1 a 3) foram agrupadas por processo de identificação de *DEGs* (Paramétricos, Não paramétricos, Paramétricos e Não paramétricos, Híbridos e Estudos Seminais) (item 4). Também foi realizado o levantamento de dependência entre métodos (ou seja, métodos selecionados que utilizam algum outro(s) para executar sua análise). Observada a rede de dependências, também avaliamos a frequência de uso dos métodos (item 5).

Fonte: De autoria própria.

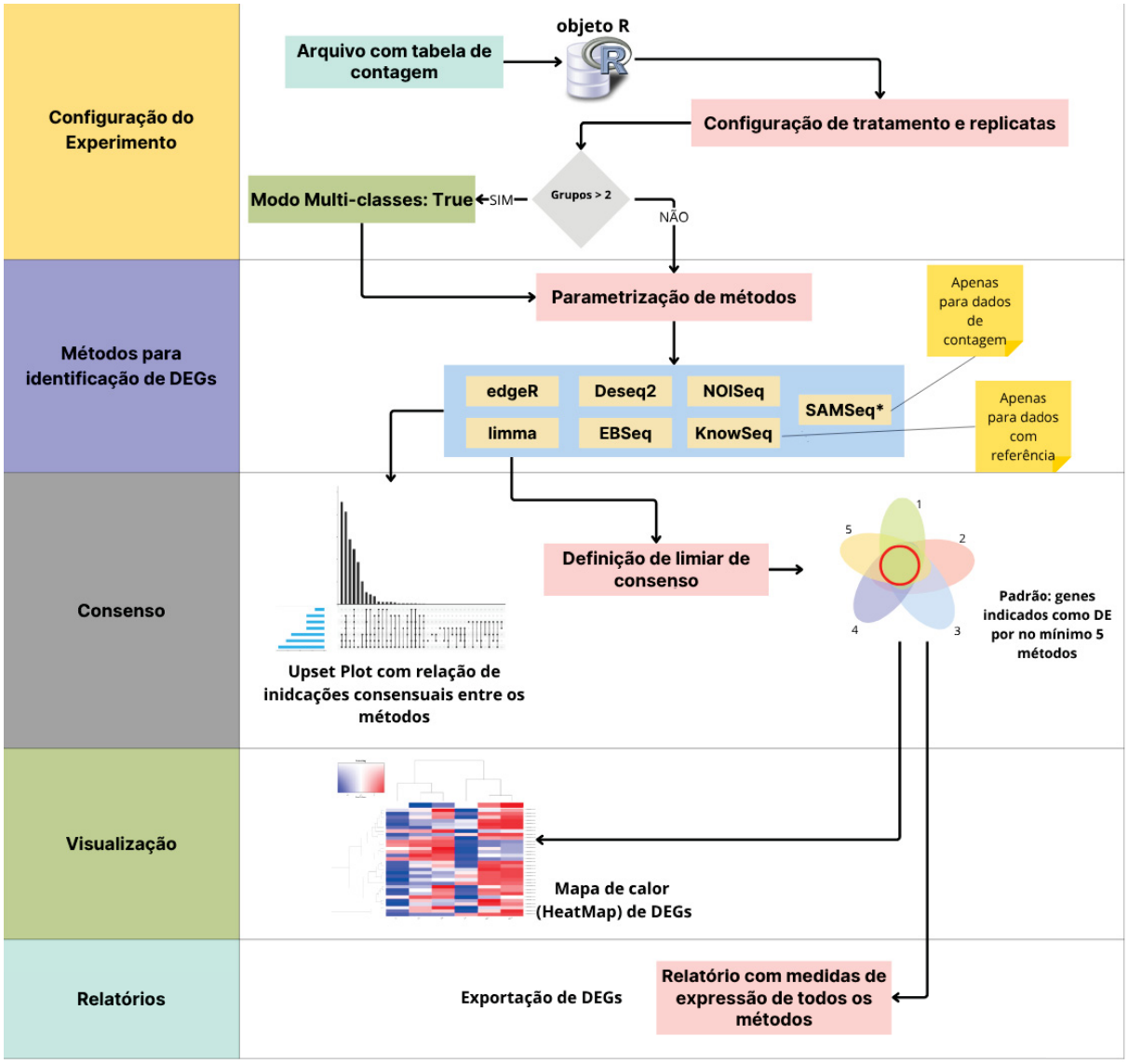


Figura 3.2: Fluxo de análise de *DEGs* do pacote *consexpressionR*.
Fluxo de análise de *DEGs* do pacote *consexpressionR*.
Fonte: De autoria própria.

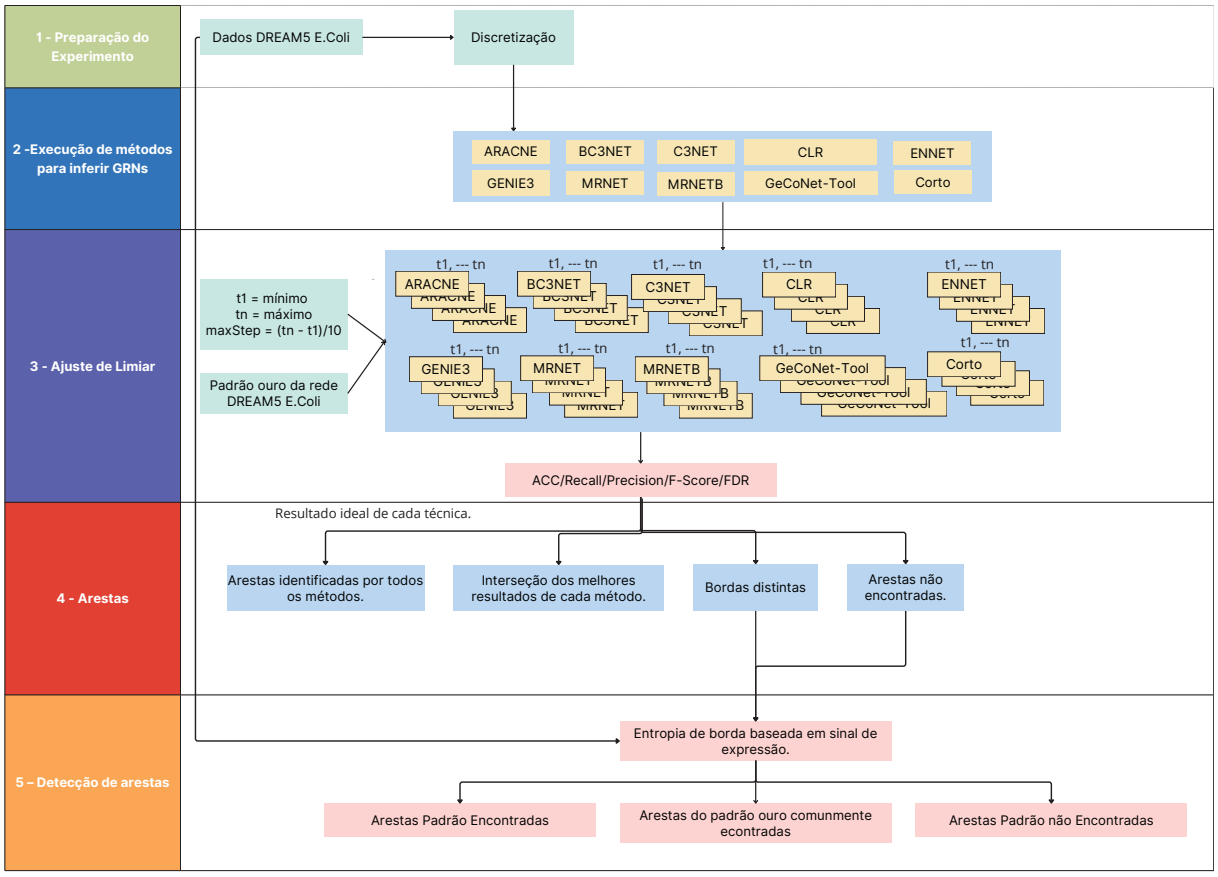


Figura 3.3: Fluxo de análise de metodologias para inferência de *GRNs*.

Fluxo de análise de metodologias para inferência de *GRNs*.

Fonte: De autoria própria.

4 RESULTADOS

Nos capítulos anteriores, apresentamos os conceitos necessários para a compreensão desta tese, bem como a metodologia desenvolvida e materiais adotados para a sua avaliação. Os resultados das análises descritas são apresentados nas seções subsequentes. Neste capítulo, seguiremos a mesma ordem de apresentação empregada nos capítulos 1, 2 e 3. Iniciaremos com os resultados da revisão de metodologias para identificação de *DEGs*, seguidos dos resultados da implementação proposta, e concluiremos com a análise e avaliação dos resultados de métodos de inferência de GRNs.

4.1 REVISÃO DE MÉTODOS PARA IDENTIFICAÇÃO DE GENES DIFERENCIALMENTE EXPRESSOS (DEGS)

Nesta seção são apresentados os resultados do estudo de revisão sobre métodos computacionais para a identificação de *DEGs*. Esses resultados também estão disponíveis no estudo publicado Costa-Silva et al., 2023.

Neste contexto, pesquisamos as principais metodologias para a identificação de *DEGs*, as quais foram implementadas em formato de software, desde os primeiros estudos até o ano de 2021. As metodologias disponíveis na literatura, publicadas em artigos científicos entre os anos de 2006 e 2021, foram selecionadas para esta tese, conforme descrito na Seção 3.1. O objetivo foi de investigar a evolução temporal dessas análises desde o início da popularização do RNA-Seq em 2008, com o trio de artigos (Bainbridge et al., 2006; Wilhelm et al., 2008; Sultan et al., 2008), até os dias atuais. As metodologias selecionadas foram organizadas de forma temporal, resultando em uma linha do tempo, que é apresentada na Figura 4.1.

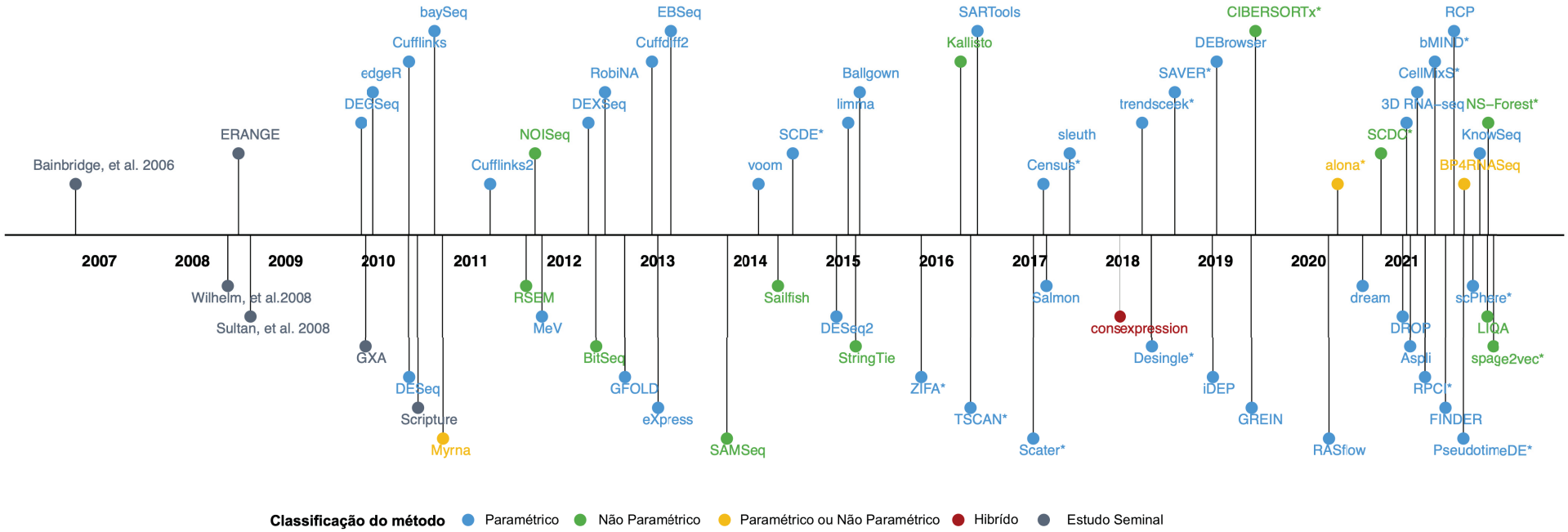


Figura 4.1: Linha do tempo, com a evolução de métodos para a identificação de *DEGs*.

Linha do tempo, com a evolução de métodos para a identificação de *DEGs*. A linha do tempo apresenta as principais metodologias e ferramentas computacionais para a análise de DEG. As ferramentas computacionais que utilizam métodos paramétricos para indicar genes diferencialmente expressos são exibidas em azul. As ferramentas não paramétricas são mostradas em verde, e as ferramentas que permitem a utilização de métodos paramétricos ou não paramétricos são apresentadas em amarelo. As ferramentas consideradas híbridas, por utilizarem métodos paramétricos e não paramétricos de forma conjunta na indicação de *DEGs*, são exibidas em vermelho. Os itens identificados em cinza representam os estudos seminais e/ou que impulsionaram as análises. A distribuição dos métodos na linha do tempo leva em consideração o mês e o ano de sua publicação. Os itens que contêm "*" indicam as ferramentas desenvolvidas no contexto de análises de sequenciamento Single-cell.

Fonte: De autoria própria.

A análise temporal das metodologias revela uma predominância de métodos paramétricos em comparação com outros tipos de métodos, de acordo com a classificação adotada nesta tese. É evidente que houve um aumento significativo no volume de métodos publicados a partir de 2020. No entanto, esses métodos mantiveram a tendência geral, sendo majoritariamente paramétricos. Outra contribuição importante fornecida pela análise temporal é a evolução das análises, no sentido de considerarem o sequenciamento *Single-cell*. A partir do ano de 2020, o número de métodos para este sequenciamento aumentou consideravelmente.

A linha do tempo permite observar algumas tendências, como a preferência por métodos paramétricos. No entanto, algumas questões ainda persistem: Existe alguma relação entre as metodologias desenvolvidas? Todas partem do princípio de que os dados seguem uma mesma distribuição?

Esta tese avaliou a relação entre os métodos, no sentido de dependência. Ou seja, verificamos se os métodos se baseiam nos algoritmos de análise uns dos outros para obter os resultados de indicações de *DEGs*. Uma maneira de visualizar relações é a aplicação de grafos. Para tanto, construímos um grafo direcionado identificando as dependências entre os métodos listados na linha do tempo da Figura 4.1. Para a construção deste grafo, foi definida como aresta a declaração de uso ou importação de biblioteca de outra metodologia, declarada na publicação de apresentação, ou na documentação da metodologia. Portanto, se a metodologia *A* declara que utiliza o conceito ou funções da metodologia *B* como parte da identificação de *DEGs*, então existe uma relação de *A* em direção a *B*.

Neste contexto, identificamos que muitas das metodologias computacionais desenvolvidas se baseiam em outras metodologias ou partes delas. Para caracterizar este grafo de relações entre metodologias, foram recuperadas da literatura atual as dependências entre essas metodologias, apresentada na Figura 4.2.

A Figura 4.2 fornece uma visão visual clara das interdependências entre as diferentes metodologias. Observa-se que apenas uma pequena parcela das metodologias pode ser considerada totalmente original (representada pelo nó em formato de losango). Isso indica que a maioria dos métodos compartilha uma metodologia comum como base. Os resultados também revelam que as metodologias edgeR (Robinson et al., 2010), DESeq2 (Love et al., 2014) e limma (Ritchie et al., 2015) são frequentemente utilizadas como componentes de outras metodologias.

Neste contexto, realizamos um mapeamento para destacar quais metodologias são mais frequentemente adotadas por outras. Os métodos edgeR (Robinson et al., 2010), DESeq2 (Love et al., 2014) e limma (Ritchie et al., 2015) são empregados em mais de 10 metodologias cada um. Considerando apenas as metodologias avaliadas nesta tese, o edgeR é adotado por 14 outras metodologias, o DESeq2 por 12 e o limma por 10 metodologias. Essas contagens foram organizadas em um histograma, apresentado na Figura 4.3.

As observações a respeito da frequência de utilização, entre as ferramentas analisadas nesta tese, são particularmente relevantes quando consideramos as metodologias paramétricas. Na Figura 4.3, fica evidente que, dentre as ferramentas mais utilizadas como base para outras, a preferência recai sobre DESeq, DESeq2, edgeR e limma. Essa preferência pode ser atribuída à manutenção, facilidade de uso e à extensa documentação disponibilizada pelos desenvolvedores e pela comunidade a respeito dessas ferramentas.

Das 56 ferramentas analisadas nesta tese, apenas 9 são utilizadas como base para outros trabalhos. Isso indica que, embora existam muitas ferramentas, a maior parte delas utiliza como base métodos previamente criados, indicando que se tratam de melhorias incrementais. Este dado também evidencia que a maior parte das soluções computacionais para análise de DEG se baseia em uma mesma metodologia de análise.

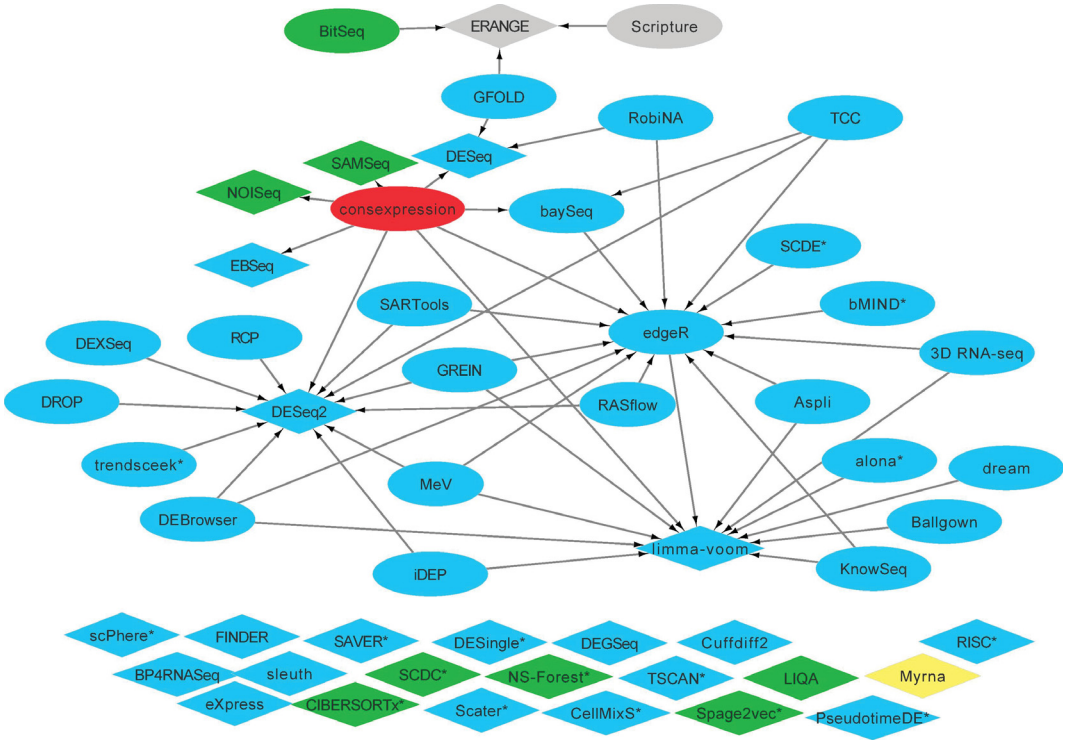


Figura 4.2: Grafo direcionado de interação entre as metodologias para a análise de expressão diferencial de genes. Grafo direcionado de interação entre as metodologias para a análise de expressão diferencial de genes. A origem das arestas na rede indicam que uma ferramenta utiliza outra, parcial ou totalmente, como base, destino das arestas. As cores dos nós representam a metodologia utilizada: azul para paramétrico; verde para não paramétrico; vermelho para híbrido e amarelo para paramétrico ou não paramétrico. Os nós em formato de losango representam ferramentas "seminais", ou seja, que não dependem de nenhuma outra para sua construção. **Fonte:** De autoria própria.

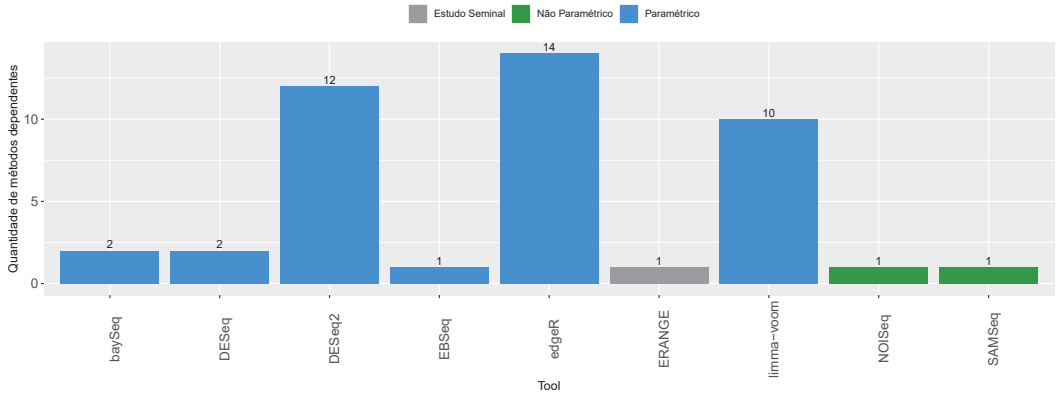


Figura 4.3: Histograma de utilização de ferramentas. Histograma de utilização de ferramentas, onde o eixo X apresenta somente as ferramentas utilizadas como base (dependência) para alguma outra. As barras são coloridas de acordo com a categoria da ferramenta, seguindo as cores de identificação da linha do tempo (Figura 4.1) e do grafo de interação (Figura 4.2), sendo que o azul indica métodos paramétricos, verde não paramétricos e cinza estudos seminais.
Fonte: De autoria própria.

Neste contexto existe um caminho claro de que as melhorias de desempenho serão obtidas apenas com inovação nas análises, entretanto temos um bom desempenho na identificação de *DEGs* com as metodologias paramétricas observadas como preferência nesta tese (Wang et al., 2009; Corchete et al., 2020b).

4.2 IMPLEMENTAÇÃO DE METODOLOGIA DE CONSENSO (CONSEXPRESSIONR)

Com base nos resultados observados na revisão de ferramentas para análise de *DEGs* apresentada na Seção 4.1, e nas solicitações de usuários da ferramenta consexpression (Costa-Silva et al., 2017a) foram identificadas algumas necessidades de melhorias no método de consenso *consexpression*. As melhorias são detalhadas nesta Seção.

O pacote *consexpressionR*, desenvolvido nesta tese, é um pacote para a linguagem R, que identifica *DEGs*, utilizando os resultados de sete análises de expressão, paramétricas e não paramétricas, e apresenta uma lista de *DEGs* que foram apontados em consenso por cinco ou mais metodologias. Atualmente encontra-se em fase de publicação no repositório Bioconductor.org (Huber et al., 2015). A ferramenta utiliza o pacote devtools (Wickham, 2015) para apoiar o desenvolvimento de pacotes R. O código fonte do pacote está disponível no repositório Github https://github.com/costasilvati/consexpressionR. O pacote desenvolvido pode ser instalado e utilizado por qualquer usuário da linguagem R, os comandos para tal uso, e manual de uso estão disponíveis na página do repositório Github https://costasilvati.github.io/consexpressionR/

Dentre as principais melhorias identificadas como necessárias, destaca-se a facilidade na instalação de dependências. Outra necessidade é a utilização do pipeline de análise para dados já mapeados, empregando uma tabela de contagem como entrada para a análise.

Foram realizadas algumas melhorias em relação aos métodos de análise de *DEGs*. A ferramenta *DESeq* (Anders e Huber, 2010) não está disponível nas novas versões do Bioconductor (repositório de pacotes R para Bioinformática) (Bioconductor org, 2022), sendo o *cDESeq2* (Anders et al., 2015) adotado como substituto natural. Outra mudança refere-se à análise de dados de expressão quantificados (*de-novo*), que não é realizada pela ferramenta *SAMSeq*; nesse caso, o *consexpressionR* executa apenas seis métodos. A lista de métodos utilizados na implementação do *consexpressionR* é apresentada na Tabela 4.1.

Tabela 4.1: Métodos de identificação de *DEGs* adotados no *consexpressionR*. Todos métodos apresentados são executados durante a identificação de *DEGs*, por padrão, apenas os genes indicados como diferencialmente expressos por cinco métodos são exibidos, porém o usuário pode parametrizar esse valor.
Fonte: De autoria própria.

Nome do pacote R	Versão	Classificação	Referência
edgeR	>= 4.1.17	Paramétrico	(Robinson et al., 2010)
NOISeq	>= 2.47.0	Não paramétrico	(Tarazona et al., 2011, 2015)
limma	>= 3.59.3	Paramétrico	(Ritchie et al., 2015)
EBSeq	>= 2.1.0	Paramétrico	(Leng et al., 2013; Ma e Leng, 2024)
DESeq2	>= 1.38.3	Paramétrico	(Love et al., 2014)
KnowSeq	>= 1.17.0	Paramétrico	(Castillo-Secilla et al., 2021b),
samr (SAMSeq)	>= 3.0	Não Paramétrico	(Li e Tibshirani, 2013)

Seguindo o fluxo descrito na Figura 3.2: A análise é iniciada na fase de "Configuração do experimento", a interface gráfica dessa fase da análise é apresentada na Figura 4.4. As configurações de tratamento devem ser preenchidas, e algumas delas definem partes importantes da análise de expressão, como nome dos grupos (nGr), em formato de lista separada por vírgula

e amostras, valor inteiro indicando quantas amostras de cada tratamento são esperadas no arquivo, onde espera-se um arquivo de texto, separado por vírgula (CSV) ou Tab (TSV), contendo dados de contagem ou quantificação organizado com linhas representado genes e colunas representando amostras, este aquivo é convertido em um objeto data.frame R. Ao executar a leitura do arquivo de contagem o número de colunas do data.frame recebido nC deve ser definido por $nC = nGr * nS$, sendo que nGr é o tamanho da lista de grupos, nS o número de amostras. Valores diferentes do esperado impedem a continuidade da execução. Para $nGr > 2$, as análises são consideradas *Multi classes*, desse modo, é necessário selecionar o grupo considerado controle, para que a análise seja de comparação dos outros grupos em relação ao grupo controle.

O experimento é configurado utilizando a interface apresentada na Figura 4.4. Na seção *Experiment Design*, o campo "*Experiment Name*" é um campo de texto, onde o usuário identifica o nome de sua análise, esse nome vai preceder os arquivos escritos pelo *consexpressionR*. O campo "*Number of replics*" é um campo inteiro onde o usuário deve colocar o número de repetições (replicatas técnicas ou biológicas) de cada tratamento/ condição avaliada pelo estudo. O campo "*Treatment Names*" espera uma lista separada por vírgula que contenha o nome dos tratamentos, na mesma ordem em que são apresentados nas colunas do arquivo de contagem. A opção "*De Novo assembly RNA-Seq data?*" é um valor booleano (Verdadeiro ou Falso), em que espera-se "*Yes*" para dados de organismos que não possuem um genoma sequenciado e anotado (principalmente devido as análises da metodologia Knowseq, que consulta dados públicos de anotação durante a análise). Na seção "*Table Count file*", a lista de seleção "*Choose a separator*", traz duas opções: i) "*Comma-separated*" para arquivo de contagem com valores separados por vírgula, geralmente utilizados com a extensão .CSV e ii) TAB, para arquivos com valores separados por TAB, geralmente utilizados com a extensão .TSV. No campo "*Select a table count file*", ao clicar no botão "*Browse...*" o usuário acessa a navegação de arquivos locais, e pode selecionar arquivos de texto, a busca é limitada a arquivos de extensão .csv, .tsv e .txt. A pós preeencher os campos anteriormente descritos e clicar no botão "*Upload Count Data*", os dados do arquivo de contagem é carregado, para tanto valida-se se o número de colunas é correspondente ao número de amostras x o número de grupos informados.

Após configurar o experimento e realizar o *upload* do arquivo de contagem, o conjunto de dados é exibido em formato de tabela, conforme apresentado na Figura 4.5, permitindo que o usuário verifique se os dados foram importados corretamente.

Com os dados de contagem transformados em objeto R, a análise prossegue para a configuração das metodologias de identificação de *DEGs*. A interface gráfica é apresentada na Figura 4.6. O *consexpressionR* gera uma GUI na qual os passos são organizados sequencialmente, de modo que, após a visualização dos dados, é apresentada a parametrização das metodologias que realizarão a identificação de *DEGs*. Na Figura 4.6, apresentamos dois grupos de ferramentas, ilustrados nas Figuras 4.6.A e 4.6.B, que estão na mesma seção de análise do *consexpressionR*, dispostas em duas "linhas" na GUI. A Figura 4.6.A corresponde à primeira linha e a Figura 4.6.B, à segunda linha, que contém o botão "*Execute Differential Expression Analysis*", utilizado para executar a análise de expressão do mesmo conjunto de dados, com as sete metodologias. Todas as metodologias possuem diversos parâmetros de execução, tais como dados de contagem, grupos de tratamento e disposição das amostras do experimento (também chamado de *design* do experimento).

A Figura 4.6 apresenta apenas os parâmetros mais comumente utilizados (*default*), baseando-se nos estudos de caso descritos no manual de cada metodologia. Na seção de parametrização das metodologias, os últimos parâmetros são destinados à filtragem de DEs após a análise, pois a maioria das ferramentas fornece valores de *P-Value* e *logFC* para todos os genes. Cabe ao usuário estabelecer os critérios para considerar um gene como diferencialmente expresso

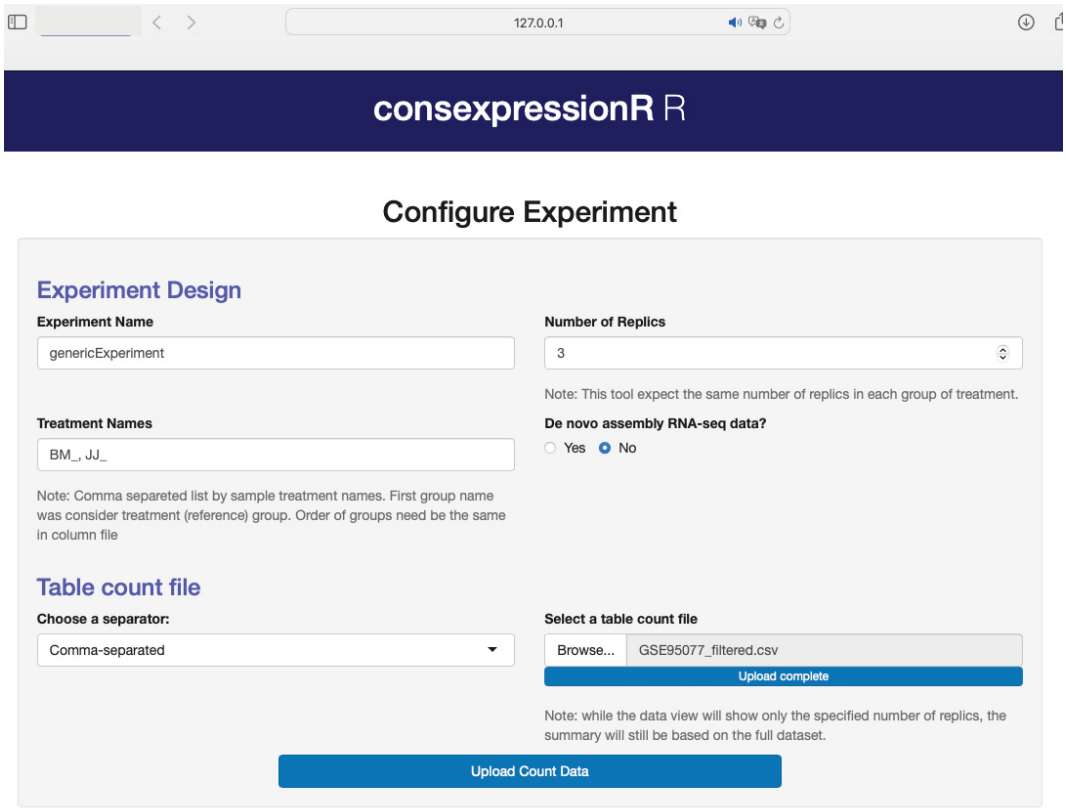


Figura 4.4: GUI *consexpressionR*: criação de experimento.
Interface gráfica do usuário *consexpressionR*, fase de configuração do experimento. Nesta seção o usuário define as especificidades de seu experimento, como número de replicatas, grupos de tratamento e origem dos dados (Humano ou não humano).
Fonte: De autoria própria.

(DE). Os parâmetros relacionados a essa filtragem são dispostos após o subtítulo "*Differential Expression Metrics*".

Durante a execução da análise de expressão, são utilizados os valores editados na interface pelo usuário. Na Figura 4.6.A, os parâmetros da ferramenta `limma` incluem a metodologia de normalização, o método de ajuste do *p-value* e o número de linhas que devem ser recuperadas da matriz de resultados. Na metodologia *SAMSeq*, o usuário deve selecionar o número de permutações utilizadas para calcular o FDR (*False Discovery Rate*), o tipo de análise, onde, para mais de dois tratamentos ou condições analisadas, deve-se utilizar o tipo "*Multiclass*". Além disso, deve-se definir o parâmetro *"Score(d)"*, disponível apenas para experimentos com a lista de grupos maior que dois. O *SAMSeq* considera que, quanto maior o *"Score(d)"*, mais significativa é a diferença na expressão do gene.

Na metodologia `DESeq2`, o usuário deve parametrizar o parâmetro *fitType*, que é utilizado para especificar o tipo de ajuste que será realizado para estimar as dispersões dos dados de contagem e a amostra de controle (que será utilizada como referência para a comparação das mudanças na expressão). Na metodologia `edgeR`, o usuário deve selecionar a metodologia de normalização dos dados de contagem. Na Figura 4.6.B, apresentamos a interface gráfica de usuário (GUI) para a parametrização das outras três metodologias. O `NOISeq` espera a seleção de um método de normalização e o tipo de replicatas utilizadas. Já o `KnowSeq` requer a definição do tipo de nomenclatura utilizada nos genes do experimento, visto que busca dados de

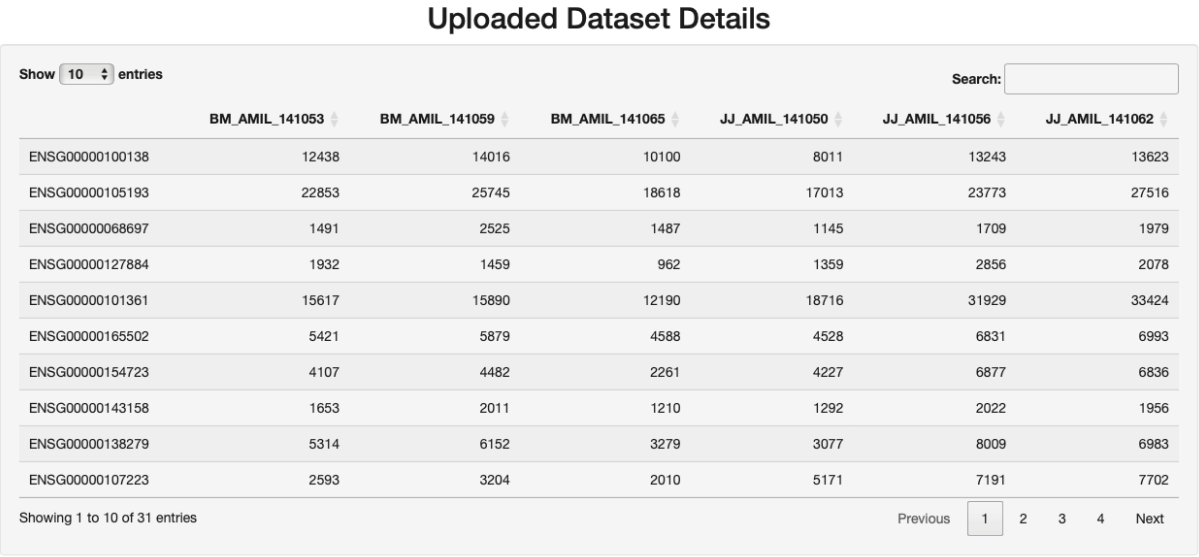


Figura 4.5: GUI *consexpressionR*: exibição de dados do experimento.
Interface gráfica do usuário *consexpressionR*, fase de exibição detalhada dos dados carregados. Esta tabela é exibida após a configuração do experimento (4.4) ser realizada com sucesso. Nesta fase, as análises são apresentadas ao usuário organizadas em duas linhas, na primeira linha são apresentados os nome das colunas (amostras de cada tratamento) identificados automaticamente (linha um do conjunto de dados), a primeira coluna apresenta o nome dos genes/ transcritos, também identificados automaticamente (coluna um do conjunto de dados). O conteúdo são os dados de contagem de cada gene
Fonte: De autoria própria.

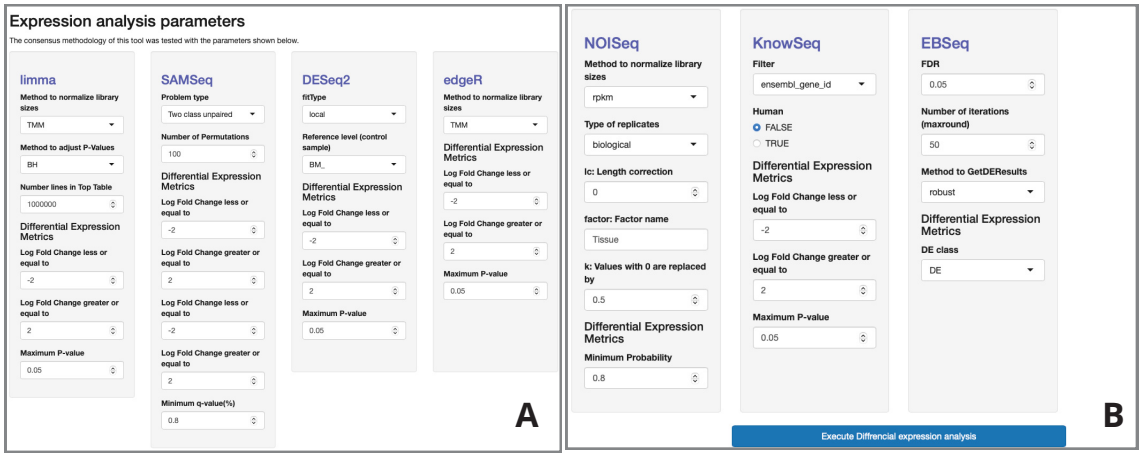


Figura 4.6: GUI *consexpressionR*: parametrização.
Interface gráfica do usuário *consexpressionR*, fase de parametrização da metodologias de identificação *DEGs*.
Fonte: De autoria própria.

anotação em bancos de dados públicos, além da especificação da origem dos dados, se são de humanos ou não.

Ao acionar o botão "*Execute Differential Expression Analysis*", todas as metodologias são executadas utilizando os dados de contagem inicialmente fornecidos. Os resultados de cada metodologia são registrados para as análises de consenso. Após a execução, o usuário visualiza

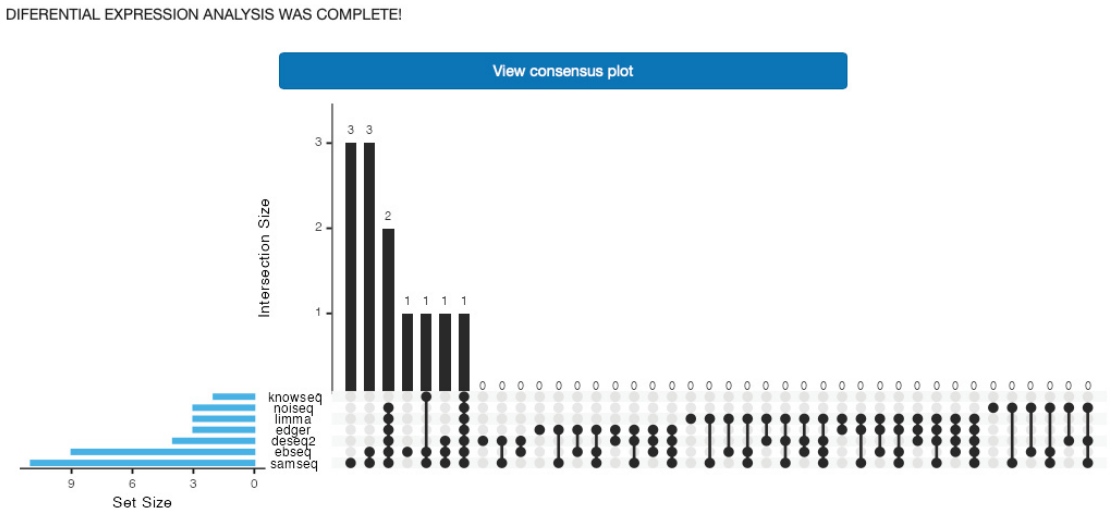


Figura 4.7: GUI *consexpressionR*: visualização de consenso entre métodos.
Interface gráfica do usuário *consexpressionR*, fase de visualização da intersecção entre as metodologias de identificação *DEGs*.
Fonte: De autoria própria.

uma mensagem abaixo do botão e pode acionar o botão "*View Consensus Plot*". Este exibe os genes indicados como diferencialmente expressos por todas as ferramentas de forma sobreposta, com o objetivo de mostrar as interseções de conjuntos (*Upset Plot*), no qual os conjuntos são as metodologias de identificação de *DEGs* e as interseções são os genes indicados como DE em consenso entre os conjuntos. A GUI desta seção, com o gráfico já disponível pode ser visualizada na Figura 4.7.

Na seção Consensus, aplicamos um filtro de seleção de modo a gerar um relatório apenas com os genes indicados por uma valor igual ou maior que o definido pelo usuário, este valor é cinco, baseando-se nos resultados de Costa-Silva et al., 2017a, mas pode ser editado pelo usuário, o resultado pode ser exportado pelo usuário e traz os valores calculados por todas as metodologias, conforme ilustrado na Figura 4.8. Na parte final da análise apresentamos um mapa de calor, que mostra através da intensidade de cor, as mudanças na expressão de um gene em relação as amostras analisadas (colunas dos dados de expressão), um exemplo desse mapa é apresentado na Figura 4.9.

Com o objetivo de avaliar o método proposto na identificação dos *DEGs*, foi realizado experimento considerando os dados apresentados na Seção 3.2. Foram consideradas diferentes métricas de assertividade conforme exibido na tabela 4.2.

A avaliação de resultados individuais de metodologias para identificação de *DEGs* deixa claro que os resultados são fortemente influenciados pelo modelo do estudo. Algumas metodologias apresentam desempenho superior com uma maior quantidade de amostras, enquanto outras têm variações nos resultados influenciadas por diferentes características do estudo (Costa-Silva et al., 2017a).

Conforme a proposta do conhecimento de multidão, muitos apontamentos indicam assertividade (Marbach et al., 2012). Com base nessa teoria, aplicamos sete técnicas de análise de expressão diferencial de genes (apresentadas na Tabela 4.1) em dados utilizados nos testes da versão inicial do consexpression (Costa-Silva et al., 2017a), aqui denominados dados A, e nos dados do estudo de Corchete et al., 2020b, aqui denominados dados B. Avaliamos o consenso

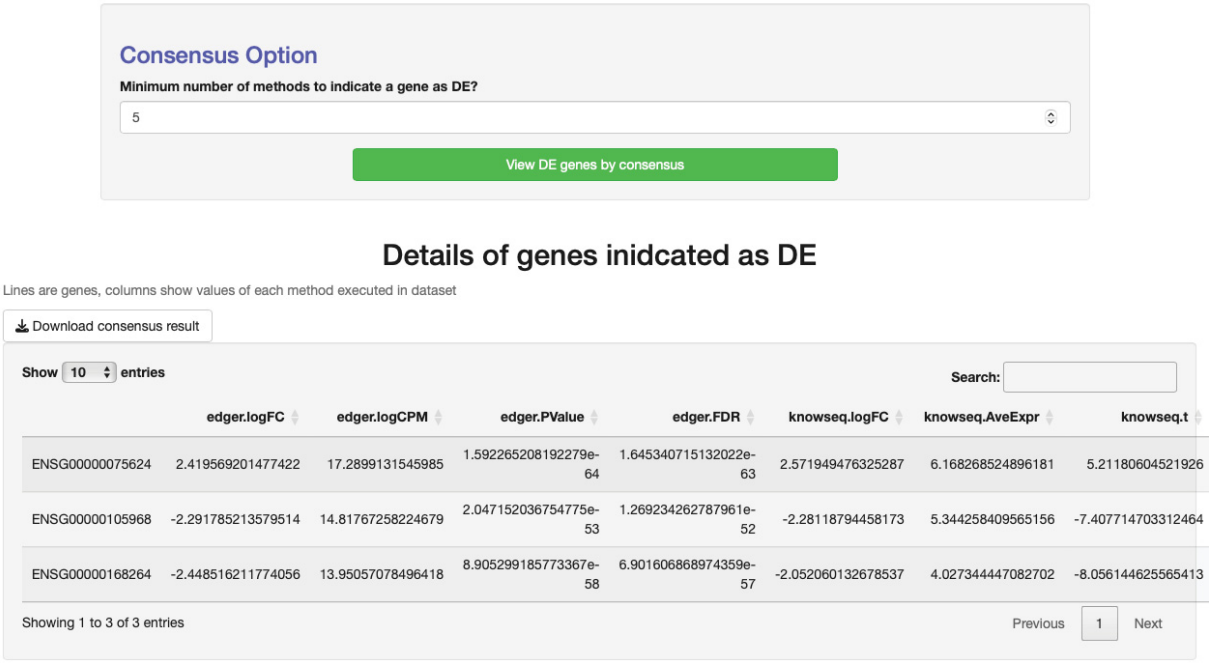


Figura 4.8: GUI *consexpressionR*: resultados da análise.
Interface gráfica do usuário *consexpressionR*, fase de resultados com base em um limiar de consenso entre as metodologias de identificação *DEGs*.
Fonte: De autoria própria.

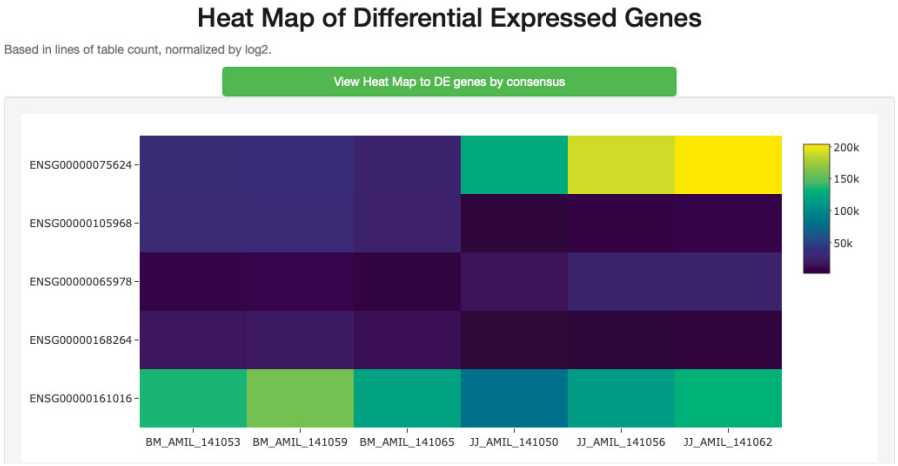


Figura 4.9: GUI *consexpressionR*: gráfico *HeatMap*.
Interface gráfica do *consexpressionR*, fase de resultados com base em um limiar de consenso entre as metodologias de identificação *DEGs*, mapa de calor dos genes identificados como DE, exibe genes e suas condições (colunas) onde a intensidade de coloração é definida pela intensidade do sinal de expressão.
Fonte: De autoria própria.

desses resultados para determinar se a teoria atualizada mantém o desempenho anteriormente registrado.

As Tabelas 4.2 e 4.3 apresentam as medidas de desempenho da análise de *DEGs* utilizando o consenso de uma a sete metodologias. O conjunto de dados A segue uma distribuição normal para experimentos de expressão com RNA-Seq, no qual a maior parte dos dados não

Tabela 4.2: Avaliação de desempenho de consenso com conjunto de dados A.

Consenso #	TP	FP	Recall	Especificidade	Precisão
1	371	354	0,93	0,36	0,51
2	359	226	0,90	0,59	0,61
3	342	80	0,86	0,85	0,81
4	333	38	0,83	0,93	0,89
5	310	20	0,78	0,96	0,94
6	263	8	0,66	0,98	0,97
7	87	6	0,22	0,99	0,93

Tabela 4.3: Avaliação de desempenho de consenso com conjunto de dados B.

Consenso #	TP	FP	Recall	Especificidade	Precisão
1	13	7	0,62	0,30	0,65
2	11	4	0,52	0,60	0,73
3	7	1	0,33	0,90	0,87
4	4	0	0,19	1,00	1,00
5	4	0	0,19	1,00	1,00
6	3	0	0,14	1,00	1,00
7	0	0	0,00	1,00	NA

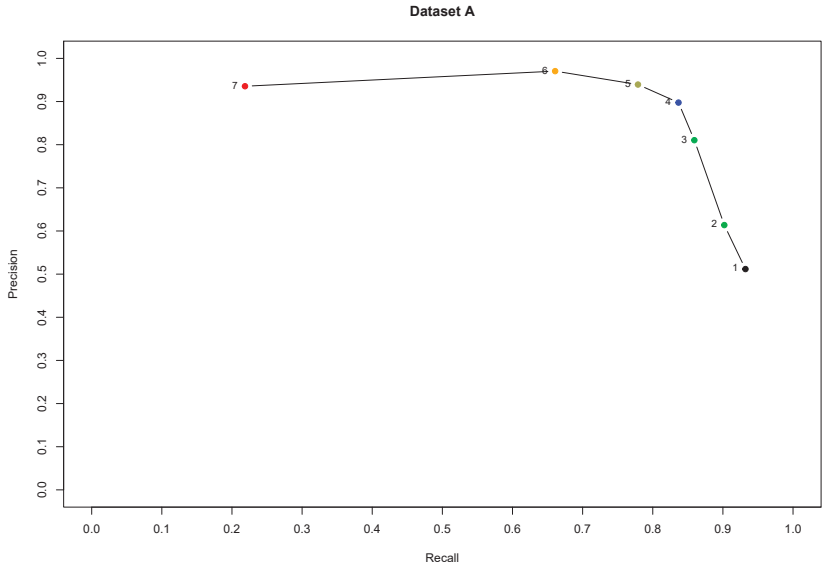
apresenta variações relevantes de expressão e um grupo menor possui uma variação evidente. Por esse motivo, mesmo com a mudança dos métodos anteriormente aplicados, os resultados seguem a mesma tendência da avaliação anterior.

O conjunto de dados B possui uma menor quantidade de dados e considera 31 genes elegíveis para análise de expressão com *qPCR*. Dentre esses genes, $\frac{1}{3}$ foi escolhido aleatoriamente e $\frac{2}{3}$ foi selecionado devido a apresentar variação considerada relevante na expressão, como superexpressos (*up regulated*) ou silenciados (*down regulated*). Para o estudo (Corchete et al., 2020b), o conjunto de dados B foi ranqueado. Entretanto, nesta tese, consideramos que os genes classificados como *up* e *down regulated* com base no valor do resultado da análise de *qPCR* são diferencialmente expressos, independentemente de sua posição no ranqueamento.

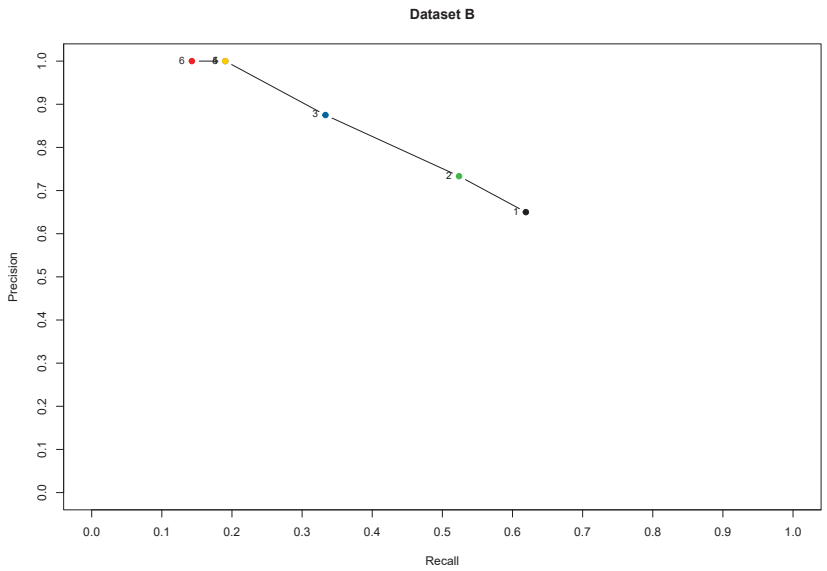
A precisão dos métodos de consenso melhora com cada metodologia adicionada, como mostrado nas Figuras 4.10(a) e 4.10(b). No entanto, quando a especificidade é aumentada, o resultado torna-se altamente restritivo, de modo que, em um conjunto de dados com poucos genes, como o conjunto de dados (B), a razão FP atinge 0 quando selecionamos apenas os genes indicados como DE por pelo menos quatro métodos.

A análise do conjunto de dados (B) mostra que, embora *EBSeq* e NOISeq atinjam alta taxa de recall, eles também apresentam as três menores taxas de precisão, indicando uma alta taxa de falsos positivos, apesar da identificação de vários genes. Esses resultados são consistentes com os observados no conjunto de dados (A). Entre os sete métodos avaliados, *edgeR, limma, DESeq2 e SAMSeq* alcançam a maior precisão, consistente com estudos anteriores (Corchete et al., 2020a; Costa-Silva et al., 2017a). Isso sugere que quase todos os genes identificados por esses métodos são verdadeiros positivos, dada a parametrização padrão usada pelo *consexpressionR*. No entanto, o número limitado de amostras leva a uma menor recuperação.

Os resultados indicam que a análise de consenso melhora a identificação de genes diferencialmente expressos, produzindo resultados mais robustos e confiáveis. Essa abordagem facilita a geração rápida de listas de genes com base na concordância de *N* métodos especificados



(a) Curva *Precision* X *Recall* conjuntos de dados B considerando o consenso dos métodos.



(b) Curva *Precision* X *Recall* conjuntos de dados B considerando o consenso dos métodos.

Figura 4.10: Relação de precisão e recuperação no consenso de metodologias.
Relação de precisão e recuperação no consenso de metodologias, aplicado ao conjunto de dados A e B.**Fonte:** De autoria própria.

pelo usuário (variando de um a sete), independentemente dos métodos específicos empregados. Notavelmente, a lista final de genes pode incluir genes identificados por *N* métodos, mesmo que nenhum desses métodos apresente desempenho individual ideal. Assim, o parâmetro *N* pode ser ajustado para produzir resultados com maior precisão (evitando falsos positivos), resultados mais permissivos com maior recuperação (recuperação) ou até mesmo um equilíbrio entre precisão e recuperação.

A análise de consenso disponibilizada através do pacote R *consexpressionR* permitirá análises de expressão com maior robustez, do mesmo modo também torna a análise mais amigável, uma vez que pode ser executada via interface gráfica com o uso da aplicação *Shiny*.

A análise de consenso disponibilizada através do pacote R *consexpressionR* permitirá análises de expressão com maior robustez, do mesmo modo também torna a análise mais amigável, uma vez que pode ser executada via interface gráfica com o uso da aplicação *Shiny*. Conforme citado anteriormente, definir os genes diferencialmente expressos é um dos passos das análises que buscam compreender a cadeia de reações gerada por estímulos do ambiente que nos cerca.

A próxima seção apresenta o passo geralmente aplicado a resultados de análise de expressão diferencial de genes: a inferência de redes de regulação genica. Essa inferência busca identificar, além da mudança na expressão, como um gene modula ou é modulado pela expressão de outros, dado um estímulo.

4.3 AVALIAÇÃO DE MÉTODOS DE INFERÊNCIA DE REDES GÊNICAS

A avaliação de desempenho de metodologias de inferência de redes de regulação genica foi aplicada a dados de *microarray*, devido à disponibilidade de uma rede de interações bem estabelecida em bancos públicos e ao conhecimento profundo dos mecanismos de regulação do organismo que originou os dados.

Para a avaliação dos métodos de inferência de Redes de Regulação Gênica (*GRNs*), foram adotadas 10 metodologias selecionadas com a parametrização padrão indicada no manual de cada método. O fluxo de trabalho estabelecido para esta avaliação é apresentado na Figura 3.3. Durante a etapa 2 da análise (aplicação de limiar), 10 limiares foram aplicados a cada saída dos métodos de inferência. Esses resultados foram comparados com a rede padrão-ouro (*gold-standard*) desenvolvida no desafio *DREAM5* (Marbach et al., 2012). Desta forma, identificamos o limiar que otimiza os resultados (para este conjunto de dados) de cada método. Os resultados dos limiares aplicados são mostrados na Tabela 4.4, e os resultados completos podem ser encontrados no Apêndice C.

Tabela 4.4: Valores de limiar associados ao o melhor *F1-Score* de cada método e outras métricas de desempenho. **Fonte:** De autoria própria.

Método	*Limiar*	*ACC*	*Recall*	*Precision*	*F-Score*	*FDR*
BC3NET	0	0.99966	0.0275	0.035121	0.030847	0.000149
C3NET	0	0.99977	0.0125	0.063939	0.020912	0.000036
ARACNE	0.1	0.999733	0.013	0.033679	0.018759	0.000073
CLR	0.1	0.999223	0.03	0.009967	0.014963	0.000586
mrnet	0.1	0.999759	0.0115	0.045908	0.018393	0.000047
mrnetB	0.2	0.999308	0.024	0.009355	0.013462	0.0005
GENIE3	0.0511	0.999701	0.07	0.106141	0.084363	0.000116
EnNET	0	0.999793	0.01	0.133333	0.018605	0.000013
GeCoNet	0	0.999438	0.048	0.024552	0.032487	0.000375
Corto	0.4546	0.999196	0.0615	0.019135	0.029188	0.00062

A aplicação de dez limiares revelou uma tendência na qual limiares mais baixos indicam o melhor *F1-Score*. O *F1-Score* foi adotado como critério decisivo nesta tese devido à sua eficácia em equilibrar precisão e sensibilidade em contextos com classes desproporcionais.

Utilizando as redes geradas pelo melhor limiar, denominado *Best Threshold Network* (*BTN*), avaliamos a interseção na indicação de arestas Verdadeiro Positivo, ou *True Positive* (*TP*), entre os métodos. Para avaliar a sobreposição de indicações, empregamos um Gráfico *Upset*, ilustrado na Figura 4.11. Este gráfico destaca a convergência entre os métodos na indicação de

certas arestas e a indicação exclusiva de outras. Nesta fase da análise, a metodologia *ENNET* foi excluída porque não indicou arestas exclusivas.

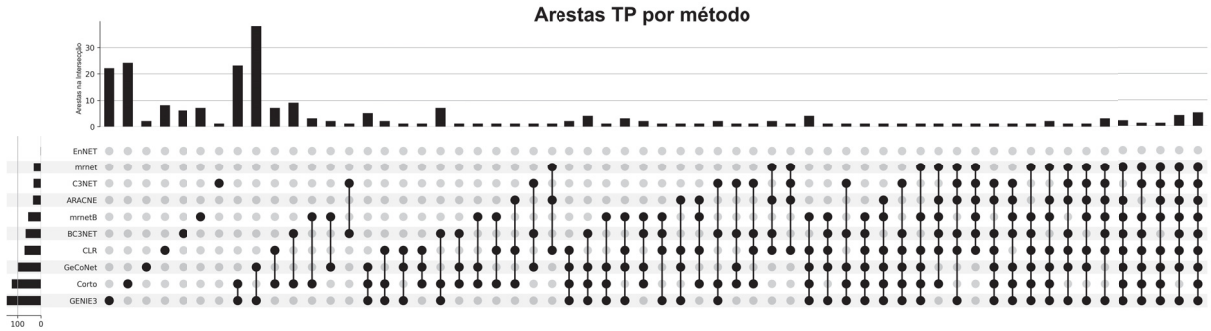


Figura 4.11: Sobreposição de indicações de arestas *TP* dos métodos.

Sobreposição de indicações de arestas *TP* dos métodos: Avaliação de concordância entre os métodos, o gráfico de sobreposição, mostra o volume de arestas encontradas pelos métodos em concordância e isoladamente **Fonte:** De autoria própria.

Para observar como o consenso entre os métodos pode otimizar os resultados, avaliamos o desempenho da inferência quando as arestas foram indicadas simultaneamente por um ou mais métodos. Os resultados foram sintetizados na Figura 4.12. A análise revelou que a indicação conjunta de arestas por pelo menos quatro ferramentas tende a melhorar a precisão, embora resulte em uma baixa taxa de recuperação. O desempenho, contudo, não superou o de estudos anteriores (Marbach et al., 2012; Zhao et al., 2021a).

Com o objetivo de avaliar as características comuns às conexões *TP* identificadas por todos os métodos (denominadas conexões comuns), foi adotada a entropia do sinal de expressão para cada fator de transcrição (*TF*) e gene (*GN*) que compõem as arestas. A análise revelou que as conexões comuns apresentam valores de entropia próximos a 6,68, como ilustrado pelas linhas de limite na Figura 4.13. Ademais, verificou-se que todas possuem um valor de entropia superior a 6 em ambos os nós. Uma representação gráfica dessa distribuição pode ser observada na Figura 4.13.

Além disso, realizamos a avaliação das arestas identificadas exclusivamente por uma única ferramenta, com o propósito de discernir as características comuns às conexões *TP* isoladas pelos métodos em questão (arestas exclusivas). Para tal, calculou-se a entropia do sinal de expressão para cada par *TF* e gene que constitui as arestas. Os resultados da análise indicaram que a maior parte das arestas exclusivas exibe uma entropia superior a 6,68 em ambos os nós, conforme delineado pelas linhas de limite na Figura 4.14. Contudo, observou-se que os métodos *MRNetB* e *GENIE3* identificaram um número significativo de conexões além dos limites estabelecidos pelas demais ferramentas, demonstrando uma maior sensibilidade na detecção de valores de entropia marginalmente inferiores aos dos outros métodos.

Em virtude da elevada proporção de conexões não detectadas, procedeu-se também à avaliação da distribuição de entropia dos sinais de expressão correspondentes. A análise indicou que a vasta maioria das conexões não identificadas por qualquer método apresenta um valor de entropia superior a 6,68, situando-se, predominantemente, próximo a 6,69, conforme ilustrado na Figura 4.15.

Adicionalmente, foi delimitada uma zona de baixa predição, evidenciada na Figura 4.16. A determinação precisa dessa zona é imperativa. Nesse sentido, a identificação de um valor de entropia que caracterize as conexões não encontradas revela-se de extrema utilidade.

Para determinar a existência de uma divisão entre as conexões identificadas e as não identificadas, que possa ser estabelecida por meio de valores de entropia, empregamos tais valores

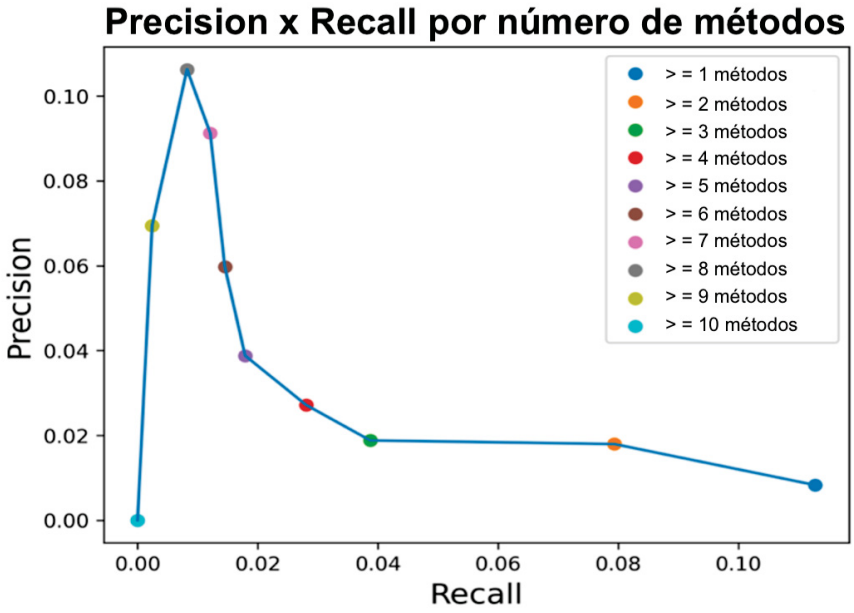


Figura 4.12: Desempenho geral de redes geradas.

Desempenho geral de redes geradas a partir de diferentes números de métodos de inferência associados. Avaliação de desempenho entre redes geradas a partir da associação de métodos, o gráfico exibe a curva de precisão e recuperação onde cada ponto indica o número de métodos utilizados para gerar a rede, e os eixos indicam o desempenho de precisão e recuperação da rede gerada.

Fonte: De autoria própria.

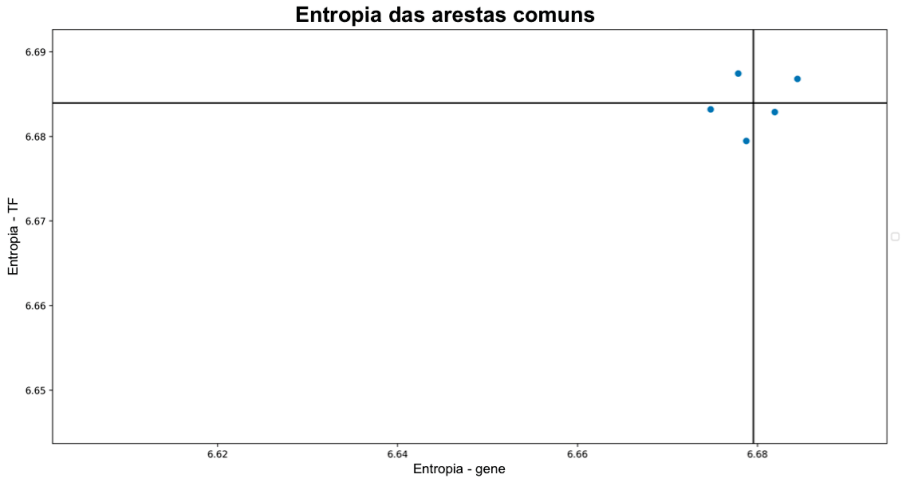


Figura 4.13: Entropia das arestas comuns.

Entropia das arestas comuns. Entropia do sinal de expressão para arestas comuns (arestas encontradas por todos os métodos).

Fonte: De autoria própria.

no contexto de uma árvore de decisão. Utilizando as conexões não detectadas e as detectadas para categorizar as conexões *TP*, aplicamos o coeficiente de correlação de *Matthews (MCC)* e o *F1-Score* para otimizar a profundidade da árvore de decisão com base nos dados.

A profundidade ótima, sugerida pelo coeficiente de correlação de *Matthews (MCC)* e pelo *F1-Score*, é de quatro níveis. A árvore de decisão gerada com essa profundidade específica

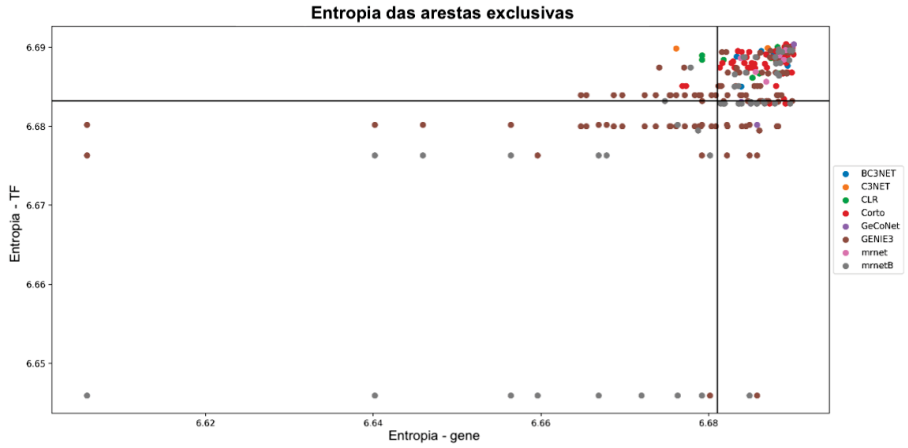


Figura 4.14: Entropia das arestas exclusivas.
Entropia das arestas exclusivas. Entropia do sinal de expressão para arestas exclusivas, agrupadas por método.
Fonte: De autoria própria.

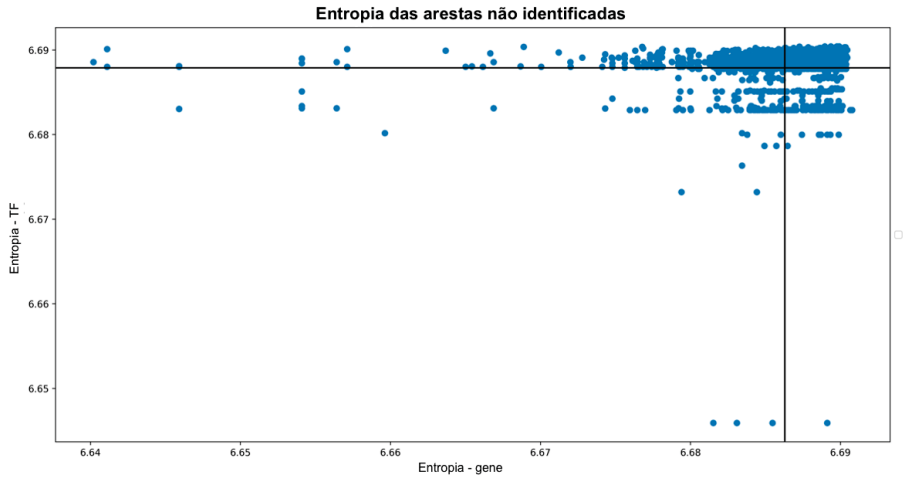


Figura 4.15: Entropia das arestas não encontradas.
Entropia das arestas não encontradas. Entropia do sinal de expressão para arestas inexistentes na inferência de todos os métodos. Estas arestas são chamadas de não identificadas nesta tese.
Fonte: De autoria própria.

é apresentada na Figura 4.17. Tal configuração evidencia a capacidade preditiva quando se consideram exclusivamente os valores de entropia.

Neste capítulo, apresentamos os resultados das três fases desta tese. Inicialmente, revisamos a literatura a respeito de métodos computacionais para identificação de *DEGs*, evidenciando que muitos métodos existentes são baseados parcialmente em outros e que a grande maioria dos métodos é paramétrica. Na segunda fase, apresentamos a metodologia de consenso "consexpressionR" em formato de pacote R e os resultados de desempenho da metodologia de consenso proposta. A aplicação do consenso ainda melhora a assertividade das análises. Na terceira fase, avaliamos métodos de inferência de redes gênicas e observamos que o consenso entre os dez métodos avaliados não teve uma contribuição significativa para a melhoria das indicações de arestas. Entretanto, ainda nessa fase do estudo, identificamos uma lacuna na identificação de arestas e caracterizamos as arestas comumente encontradas, assim como aquelas não encontradas por nenhum método avaliado.

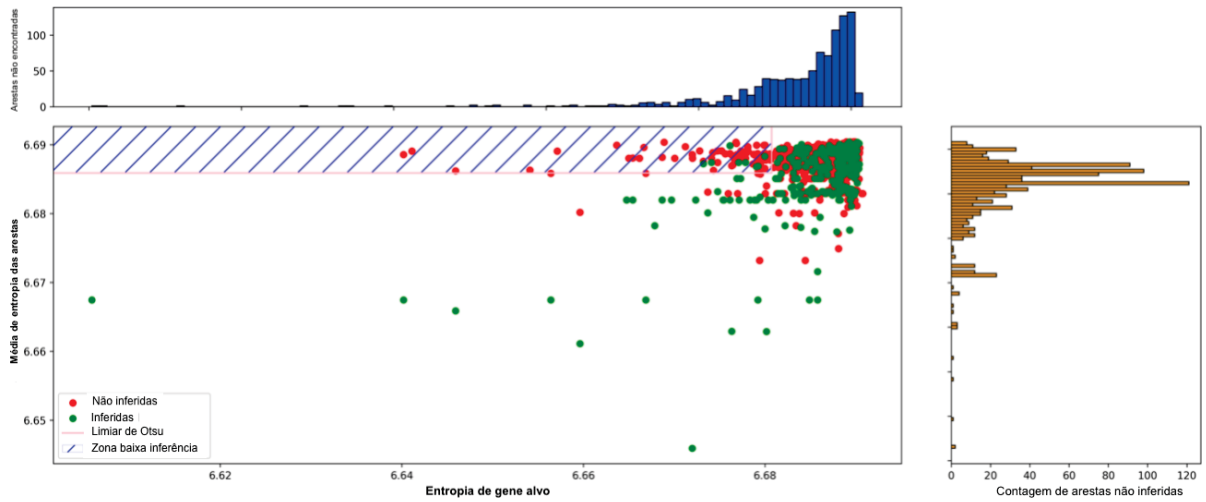


Figura 4.16: Identificação da zona de baixa predição.

Identificação da zona de baixa predição. Identificação da zona de baixa predição. Usando todas as conexões, classificadas como encontradas e não encontradas pelos métodos avaliados. Limiar de *Otsu* (região tracejada) (Otsu et al., 1975) para mapear os valores de entropia do gene-alvo em relação à entropia média da conexão.

Fonte: De autoria própria.

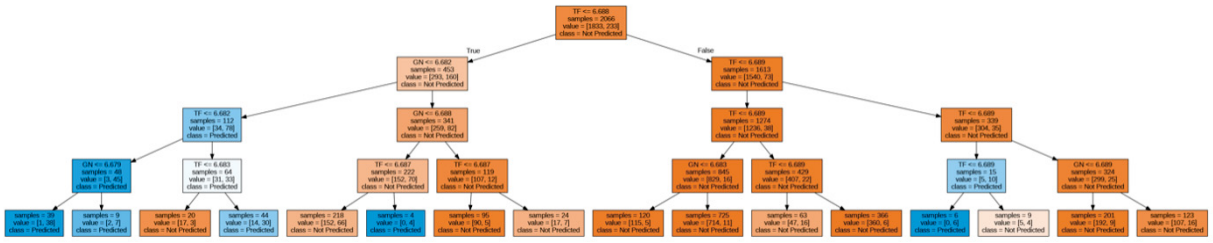


Figura 4.17: Árvore de decisão a partir da entropia das arestas.

Árvore de decisão com valores de entropia das arestas. Árvore de decisão utilizando valores de entropia da aresta. Arestas preditas (encontradas) estão em azul e não encontradas em Laranja.

Fonte: Elaboração própria.

5 DISCUSSÃO

Esta tese apresenta as principais análises computacionais em Bioinformática relacionadas à expressão gênica, bem como as análises computacionais subsequentes à expressão. O objetivo é contribuir com as análises de Bioinformática nessa área, apontando possibilidades de análises ainda não exploradas, aprimorando as análises já existentes conforme a necessidade dos usuários e propondo novas discussões sobre metodologias estabelecidas.

As interações entre organismos e suas condições de vida são, em grande parte, influenciadas pelas cadeias de reações químicas que ocorrem dentro de um organismo, de fato o estado funcional de um organismo pode ser determinado pela sua expressão gênica. A regulação dessas reações é realizada pelos mRNAs. Compreender a relação entre quais mRNAs são produzidos em situações específicas e como essa produção afeta toda a rede de mRNAs do organismo é essencial para resolver problemas como controle de patógenos em plantas, produção de vacinas e medicamentos, epidemias, infecções por vírus ou bactérias, entre outras aplicações.

Neste contexto, a presente tese desenvolveu uma revisão sobre as metodologias de identificação de Genes Diferencialmente Expressos (DEGs), utilizando dados de sequenciamento *RNA-Seq*. Adicionalmente, foi desenvolvido um pacote em R com o mesmo propósito, possibilitando análises de expressão robustas e de alta confiabilidade, com a facilidade de uso de uma interface web construída por meio do *R Shiny* (Chang et al., 2024). Por fim, investigaram-se uma análise frequentemente aplicada aos resultados de análise de expressão: inferência de redes de regulação gênica, de modo a encontrar lacunas e novos caminhos para esta análise.

Embora existam estudos que comparam as metodologias de identificação de Genes Diferencialmente Expressos (DEGs) (Rapaport et al., 2013; Trapnell et al., 2012; Bullard et al., 2010; Zhang et al., 2014; Corchete et al., 2020b), até então, não haviam levantamentos que explorassem as relações de uso entre elas.

A primeira fase desta tese publicou uma revisão sobre métodos computacionais para análise de *DEGs*, desde a popularização do sequenciamento *RNA-Seq* em 2006, relatando e discutindo as ferramentas mais relevantes desenvolvidas no período de 2006 a 2021, conforme constam na literatura. A construção do cenário atual de análises e a obtenção de uma visão clara sobre as metodologias desenvolvidas nas últimas décadas permitiram a identificação de lacunas a serem exploradas. Nesse contexto, identificamos que a associação de métodos paramétricos e não paramétricos na identificação de *DEGs* ainda é pouco explorada (Costa-Silva et al., 2023), embora os métodos exclusivamente paramétricos também apresentem resultados promissores (Corchete et al., 2020b; Costa-Silva et al., 2017a).

Identificamos também a alta frequência de uso das metodologias paramétricas, que apresentam dependência parcial ou total entre si, conforme detalhado nas Figuras 4.2 e 4.3. As indicações desta tese podem orientar o desenvolvimento de novas metodologias e contribui como referencial para pesquisas na área. Ademais, essas indicações já estão sendo utilizadas como fundamentação para novas pesquisas e revisões (Zayakin, 2024; Jiang et al., 2024; Rosati et al., 2024), contribuindo efetivamente para a compreensão das etapas envolvidas e dos métodos disponíveis, bem como de suas particularidades e aplicações.

As contribuições da revisão sobre metodologias de identificação de *DEGs* também apoiam a segunda fase desta tese, o desenvolvimento de um pacote R sobre a mesma temática. Como parte essencial da Bioinformática, o desenvolvimento de ferramentas que executem metodologias de análise para diferentes tipos de experimentos é uma necessidade. A computação

tem papel primordial nesse contexto, permitindo o acesso a análises e contribuindo com estudos cada vez mais abrangentes (Zheng et al., 2023).

O desenvolvimento de *pipeline* de análise de expressão, incluindo novas funcionalidades, tornou-se uma necessidade devido ao alto volume de dados produzidos atualmente (Zayakin, 2024; Jiang et al., 2024; Rosati et al., 2024). Evitar a necessidade de muitas replicatas, exigidas pelo sequenciamento de *RNA-Seq*, e ainda manter resultados satisfatórios, é um desafio que merece atenção no desenvolvimento de metodologias de análise de *DEGs* (Finotello e Camillo, 2015).

No contexto de revisão, apresentamos conceitos fundamentais e ferramentas computacionais para análise de expressão. É possível identificar, além da tendência de reutilização de metodologias no desenvolvimento de ferramentas computacionais (*software*), a necessidade de incorporar novas funcionalidades a metodologias existentes.

Portanto, podemos afirmar que as metodologias paramétricas apresentam um cenário mais estável, revelando convergência entre os métodos disponíveis na literatura. Em contraste, a revisão aponta um desafio no desenvolvimento de metodologias não paramétricas (orientadas a dados) e híbridas para análise de *DEG*.

Também é relevante destacar que diversos métodos abordados na revisão foram concebidos para análises específicas, tais como a identificação de *splicing alternativo* (Anders et al., 2012), e para dados de *Single-cell* (Aevermann et al., 2021; Ding e Regev, 2021; Song e Li, 2021). Dado que o sequenciamento de *Single-cell* é uma metodologia recente, a maioria das metodologias identificadas para essa finalidade são consideradas originais (sem dependência ou uso de outra metodologia), como ilustrado na Figura 4.2.

Na segunda fase desta tese, foi implementado um método computacional disponibilizado como um pacote em R para análise de expressão denominado *consexpressionR*. Este pacote permite ao usuário realizar sete diferentes análises de expressão, tanto por meio de uma interface gráfica quanto através de comandos R. A seleção dos genes por essa metodologia baseia-se no número de métodos que identificaram o gene como diferencialmente expresso (DE), sendo que, por padrão, recomenda-se os genes indicados por cinco métodos como os mais otimizados para os resultados da análise. Contudo, para conjuntos de dados menores e balanceados, genes indicados por quatro métodos também podem apresentar resultados satisfatórios, conforme demonstrado na Tabela 4.3. Embora ainda não esteja disponível em repositórios oficiais de pacotes R (CRAN ou Bioconductor), o pacote *consexpressionR* já pode ser baixado e instalado via GitHub, utilizando o pacote *devtools* (Wickham et al., 2022).

O pacote R foi escolhido devido à facilidade de uso e instalação e também pela forma padronizada de disponibilização para os usuários; no entanto, por executar sete metodologias, o *consexpressionR* possui uma lista extensa de dependências. Através da configuração do pacote, no arquivo *DESCRIPTION* (parte da estrutura de pacotes R), foram definidas as importações necessárias e suas respectivas versões. A criação de *software* e pacotes que facilitam o uso de metodologias para análise de expressão é uma atividade recorrente na literatura do tema (Sun et al., 2021; Guo et al., 2020). Entretanto, não encontramos nenhuma outra análise que associe metodologias paramétricas e não paramétricas para indicar *DEGs*. Desse modo, compreendemos a importância de disponibilizar a metodologia *consexpressionR* de forma simples e acessível.

A terceira fase desta tese aborda metodologias de inferência de redes de regulação gênica. Foram identificados métodos populares e anteriormente avaliados (Margolin et al., 2006b; Meyer et al., 2008; de Matos Simoes e Emmert-Streib, 2012; Altay e Emmert-Streib, 2010a; Huynh-Thu et al., 2010) e métodos recentes (Kuang et al., 2023; Sławek e Arodź, 2013), com um conjunto de dados muito utilizado como base para avaliação de desempenho de métodos de inferência de *GRNs* (Marbach et al., 2012).

Para iniciar a abordagem, foi utilizada uma estratégia que considera o desbalanceamento natural dos dados de inferência. Neste caso, consideramos a fração de arestas TP (verdadeiras positivas) em relação a todas as arestas positivas (sensibilidade ou *recall*) e a fração de arestas positivas verdadeiras em relação a todas as arestas positivas (precisão ou valor preditivo positivo). Naturalmente, a precisão e o *recall* dependem do limiar de corte escolhido: com um limiar muito permissivo, recuperamos muitos positivos verdadeiros (alto *recall*), ao custo de muitos falsos positivos (baixa precisão) (Huynh-Thu e Sanguinetti, 2019).

Esta tese avaliou o desempenho geral de cada método, uma análise explorada anteriormente por outros estudos (Marbach et al., 2012; Huynh-Thu e Sanguinetti, 2019; Ahmed et al., 2022). Porém, diferentemente dos estudos comparativos anteriores, também foram investigadas as arestas agrupadas por resultado de inferência (comumente encontrada, não encontrada, exclusivamente encontrada), com foco principal nas características das arestas, encontradas, não encontradas e comumente encontradas. A proposta foi identificar lacunas em métodos de inferência que podem ser exploradas e consideradas no desenvolvimento de novas metodologias para a inferência de *GRNs*. Avaliamos também o consenso (combinação de vários métodos para gerar resultados).

As arestas comumente encontradas são esperadas, uma vez que metodologias como *ARACNE, CLR* e *MRNet* utilizam métricas de associação, como informações mútuas, para quantificar a relação entre um fator de transcrição (TF) e um gene. No entanto, uma das principais limitações dessas metodologias reside no cálculo realizado para associação por pares, o qual não consegue modelar a expressão gênica como uma função de múltiplos reguladores. Métodos baseados em regressão, como o *GENIE3*, superam essa restrição ao modelar a expressão gênica como uma função de vários reguladores, permitindo uma modelagem mais precisa das relações regulatórias entre os reguladores e os genes-alvo. Contudo, uma limitação significativa desses métodos é a sua dependência exclusiva de dados transcriptômicos, desconsiderando as modificações epigenéticas que desempenham um papel crucial na regulação gênica (Kim et al., 2023).

As arestas exclusivamente encontradas foram caracterizadas com o objetivo de compreender as especificidades de cada metodologia. No entanto, ao observar a distribuição da entropia dessas arestas em um gráfico de dispersão, nota-se uma grande convergência em uma região do gráfico. Esta convergência sugere que a entropia captura as particularidades na associação por pares utilizada por cada metodologia, e que a maior parte dos métodos consegue identificar as associações em um certo intervalo do valor de entropia, entretanto também foi possível detectar que existe um intervalo de valores de entropia em que nenhum método identificou associações existentes, isso pode ser explorado em estudos futuros.

Esta tese também aponta que um subconjunto de relações (arestas) entre os genes que não foi inferido por nenhum dos métodos avaliados, sugerindo que existe um comportamento nos sinais de expressão que não são recuperados pelos métodos e apontando para oportunidade no desenvolvimento de novos métodos aumento da assertividade na inferência de *GRNs*.

Finalmente, esta tese traz alguns avanços nas análises computacionais de expressão diferencial de genes, além de indicar áreas pouco exploradas nessas análises, como métodos não paramétricos para dados de *RNA-Seq*. Do mesmo modo que apresenta uma lacuna a ser explorada nas metodologias de identificação de redes regulatórias. Nesse processo, identificamos a possibilidade de trabalhos futuros, que com certeza trarão mais contribuições a está área de estudo.

6 CONSIDERAÇÕES FINAIS E DIRECIONAMENTOS

Esta tese apresenta um conjunto articulado de contribuições no campo da Bioinformática, estruturadas em três eixos principais, com foco na análise de expressão gênica diferencial e na inferência de redes regulatórias de genes.

A primeira etapa concentrou-se na sistematização do conhecimento disponível sobre metodologias computacionais voltadas à identificação de genes diferencialmente expressos (*DEGs*). Foi realizada uma revisão abrangente da literatura, na qual os métodos foram classificados com base em suas abordagens estatísticas (paramétricas, não paramétricas e híbridas), e analisadas suas características, limitações e principais tendências de reutilização. Esta etapa permitiu além de mapear o estado da arte da área, também identificar lacunas e oportunidades de aprimoramento, especialmente no que se refere à robustez e reprodutibilidade das análises.

A segunda contribuição consistiu na implementação da metodologia *consexpressionR*. O pacote foi desenvolvido com foco em acessibilidade, transparência e reprodutibilidade, disponibilizando uma ferramenta de código aberto para a integração de múltiplos métodos de análise de expressão. A solução contempla a execução de sete algoritmos distintos, oferece uma interface gráfica interativa, e permite análises com e sem genoma de referência, promovendo maior flexibilidade para diferentes contextos experimentais, além de resultados baseados em uma análise robusta.

A terceira etapa abordou a inferência de redes regulação gênica (*GRNs*), com ênfase na avaliação do desempenho de métodos recentemente propostos e amplamente utilizados na literatura. Foram realizadas análises comparativas baseadas em métricas de precisão e cobertura, bem como caracterizações das arestas inferidas a partir dos sinais de expressão. Essa investigação permitiu identificar padrões recorrentes nas conexões mais frequentemente recuperadas, bem como nas interações não detectadas, sugerindo potenciais melhorias nos critérios de inferência adotados pelos algoritmos. A análise também incorporou medidas de entropia como recurso adicional para compreender o comportamento dos sinais de expressão nas diferentes classes de interações.

As contribuições aqui reunidas visam oferecer subsídios teóricos e práticos para o avanço da Bioinformática, especialmente nas áreas de análise de expressão gênica e inferência de *GRNs*. Entre os principais desafios enfrentados, destacam-se: a alta dimensionalidade dos dados, a escassez de conjuntos de referência (*gold-standard*), a heterogeneidade dos formatos de arquivos e a ausência de padronização nos fluxos analíticos. Ainda que tais obstáculos representem limitações importantes, o presente trabalho indica que é possível enfrentá-los por meio de estratégias integrativas e metodologias computacionais.

Como perspectivas futuras, considera-se a implementação do pacote *consexpressionR* em ambiente *web*, com o objetivo de democratizar o acesso às funcionalidades desenvolvidas e eliminar barreiras técnicas relacionadas à instalação local. Além disso, pretende-se aprofundar os achados relacionados à inferência de redes, com vistas à proposição de novos métodos baseados nas características estruturais e estatísticas das interações analisadas, bem como na adoção de estratégias guiadas pelos dados para a inferência de GRNs. Espera-se, assim, contribuir de forma contínua com o desenvolvimento de metodologias mais precisas e aplicáveis à interpretação de dados ômicos em diferentes contextos biológicos.

REFERÊNCIAS

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F. e others (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252(5013):1651–1656.

Aevermann, B., Zhang, Y., Novotny, M., Keshk, M., Bakken, T., Miller, J., Hodge, R., Lelieveldt, B., Lein, E. e Scheuermann, R. H. (2021). A machine learning method for the discovery of minimum marker gene combinations for cell type identification from single-cell RNA sequencing. *Genome Research*, 31(10):1767–1780.

Ahmed, F., Soomro, A. M., Chethikkattuveli Salih, A. R., Samantasinghar, A., Asif, A., Kang, I. S. e Choi, K. H. (2022). A comprehensive review of artificial intelligence and network based approaches to drug repurposing in Covid-19. *Biomedicine & Pharmacotherapy*, 153:113350.

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J. C., Geurts, P., Aerts, J., Van Den Oord, J., Atak, Z. K., Wouters, J. e Aerts, S. (2017). SCENIC: single-cell regulatory network inference and clustering. *Nature Methods 2017 14:11*, 14(11):1083–1086.

Alam, S., Israr, J. e Kumar, A. (2024). *Artificial Intelligence and Machine Learning in Bioinformatics*, páginas 321–345. Springer Nature Singapore.

Alawad, D. M., Katebi, A., Kabir, M. W. U. e Hoque, M. T. (2023). AGRN: accurate gene regulatory network inference using ensemble machine learning methods. *Bioinformatics Advances*, 3(1).

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. e Walter, P. (2009). *Biologia Molecular da Celula*. Artmed Editora.

Alcalá-Corona, S. A., Sandoval-Motta, S., Espinal-Enríquez, J. e Hernández-Lemus, E. (2021). Modularity in Biological Networks. *Frontiers in Genetics*, 12:1708.

Almeida, E. J. C., Vainzof, M. e Martins, P. C. (2022). O dogma central continua?

Altay, G. e Emmert-Streib, F. (2010a). Inferring the conservative causal core of gene regulatory networks. *BMC Systems Biology*, 4(1):1–13.

Altay, G. e Emmert-Streib, F. (2010b). Inferring the conservative causal core of gene regulatory networks. *BMC Systems Biology*, 4(1):1–13.

Anders, S. e Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*.

Anders, S., Pyl, P. T. e Huber, W. (2015). Htseq-a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31:166–169.

Anders, S., Reyes, A. e Huber, W. (2012). Detecting differential usage of exons from RNA-Seq data. *Nature Precedings*, páginas 1–1.

Bainbridge, M. N., Warren, R. L., Hirst, M., Romanuik, T., Zeng, T., Go, A., Delaney, A., Griffith, M., Hickenbotham, M., Magrini, V., Mardis, E. R., Sadar, M. D., Siddiqui, A. S., Marra, M. A. e Jones, S. J. (2006). Analysis of the prostate cancer cell line lncap transcriptome using a sequencing-by-synthesis approach. *BMC Genomics*, 7:246.

Banerjee, S., Bhandary, P., Woodhouse, M., Sen, T. Z., Wise, R. P. e Andorf, C. M. (2021). FINDER: an automated software package to annotate eukaryotic genes from RNA-Seq data and associated protein sequences. *BMC Bioinformatics 2021 22:1*, 22(1):1–26.

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Muertter, R. N., Holko, M., Ayanbule, O., Yefanov, A. e Soboleva, A. (2011). NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Research*, 39(suppl_1):D1005–D1010.

Bayes, T., Price, R. e Canton, J. (1763). *An essay towards solving a problem in the doctrine of chances*. C. Davis, Printer to the Royal Society of London.

Bioconductor org (2022). Bioconductor - Home.

Blake, J., Christie, K., Dolan, M., Drabkin, H., Hill, D. e L. Ni, D. S. (2015). Gene ontology consortium: going forward. *Nucleic Acids Research*, 43:D1049–D1056.

Boucher, B. e Jenna, S. (2013). Genetic interaction networks: Better understand to better predict. *Frontiers in Genetics*, 4(DEC):290.

Bullard, J. H., Purdom, E., Hansen, K. D. e Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11(1):94.

Bustin, S. (2000). Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *Journal of Molecular Endocrinology*, 25(2):169–193.

Bustin, S. A., Benes, V., Garson, J. A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M. W., Shipley, G. L., Vandesompele, J. e Wittwer, C. T. (2009). The MIQE guidelines: Minimum information for publication of quantitative real-time PCR experiments. *Clinical Chemistry*, 55(4):611–622.

Butte, A. J. e Kohane, I. S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, páginas 418–429.

Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R. e Kohane, I. S. (2000). Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences*, 97:12182–12186.

Cai, G., Bossé, Y., Xiao, F., Kheradmand, F. e Amos, C. I. (2020). Tobacco smoking increases the lung gene expression of ace2, the receptor of sars-cov-2. *American Journal of Respiratory and Critical Care Medicine*, 201:1557–1559.

Camacho, D., Licona, P. V., Mendes, P. e Laubenbacher, R. (2007). Comparison of reverse-engineering methods using an in silico network. *Annals of the New York Academy of Sciences*, 1115:73–89.

Canzar, S. e Salzberg, S. L. (2017). Short read mapping: An algorithmic tour. Em *Proceedings of the IEEE*, volume 105, páginas 436–458. Institute of Electrical and Electronics Engineers Inc.

Castillo-Secilla, D., Gálvez, J. M., Carrillo-Perez, F., Verona-Almeida, M., Redondo-Sánchez, D., Ortuno, F. M., Herrera, L. J. e Rojas, I. (2021a). KnowSeq R-Bioc package: The automatic smart gene expression tool for retrieving relevant biological knowledge. *Computers in Biology and Medicine*, 133:104387.

Castillo-Secilla, D., Gálvez, J. M., Carrillo-Perez, F., Verona-Almeida, M., Redondo-Sánchez, D., Ortuno, F. M., Herrera, L. J. e Rojas, I. (2021b). Knowseq r-bioc package: The automatic smart gene expression tool for retrieving relevant biological knowledge. *Computers in Biology and Medicine*, 133:104387.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A. e Borges, B. (2024). *shiny: Web Application Framework for R*. R package version 1.9.0.9000, https://github.com/rstudio/shiny.

Chen, S. e Mar, J. C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics*, 19(1):1–21.

Chen, Y., McCarthy, D., Baldoni, P., Ritchie, M., Robinson, M., Smyth, G. e Hall, E. (2024). edger: differential analysis of sequence read count data user's guide.

Clancy, S. e Brown, W. (2008). Translation: Dna to mrna to protein | scitable by nature education.

Cock, P. J. a., Fields, C. J., Goto, N., Heuer, M. L. e Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, 38(6):1767–71.

Corchete, L. A., Rojas, E. A., Alonso-López, D., De Las Rivas, J., Gutiérrez, N. C. e Burguillo, F. J. (2020a). Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Scientific Reports*, 10(1).

Corchete, L. A., Rojas, E. A., Alonso-López, D., Rivas, J. D. L., Gutiérrez, N. C. e Burguillo, F. J. (2020b). Systematic comparison and assessment of rna-seq procedures for gene expression quantitative analysis. *Nature Scientific Reports*, 10:19737.

Costa-Silva, J., Domingues, D. e Lopes, F. M. (2017a). RNA-Seq differential expression analysis: An extended review and a software tool. *PLOS ONE*, 12(12):e0190152.

Costa-Silva, J., Domingues, D. S., Menotti, D., Hungria, M. e Lopes, F. M. (2023). Temporal progress of gene expression analysis with rna-seq data: A review on the relationship between computational methods. *Computational and Structural Biotechnology Journal*, 21:86–98.

Costa-Silva, J., Menotti, D. G. e Lopes, F. M. (2024). costasilvati/consexpressionR: Version 2.0 of Differential Gene expression software.

Costa-Silva, J., Menotti, Domingues, D. S. e Lopes, F. M. (2017b). costasilvati/consexpression:Differential Gene expression software with consesnus.

da Rocha Vicente, F. F. e Lopes, F. M. (2014). Sffs-sw: a feature selection algorithm exploring the small-world properties of gns. Em *Pattern Recognition in Bioinformatics: 9th IAPR International Conference, PRIB 2014, Stockholm, Sweden, August 21-23, 2014. Proceedings 9*, páginas 60–71. Springer.

De Bona, F., Ossowski, S., Schneeberger, K. e Ratsch, G. (2008). Optimal spliced alignments of short sequence reads. *Bioinformatics*, 24(16):i174–i180.

de Ciências UNICAMP, B. (2020). Diagnóstico por rt-qpcr, o que é isso? - especial covid-19 | blogs unicamp.

de Matos Simoes, R. e Emmert-Streib, F. (2012). Bagging Statistical Network Inference from Large-Scale Gene Expression Data. *PLOS ONE*, 7(3):e33624.

de Souza, N. (2012). The encode project. *Nature Methods*, 9:1046–1046.

del Val, C., de la Guardia-Bolívar, E. D., Zwir, I., Mishra, P. P., Mesa, A., Salas, R., Poblete, G. F., de Erausquin, G., Raitoharju, E., Kähönen, M., Raitakari, O., Keltikangas-Järvinen, L., Lehtimäki, T. e Cloninger, C. R. (2024). Gene expression networks regulated by human personality. *Molecular Psychiatry 2024*, páginas 1–20.

Delgado, F. M. e Gómez-Vela, F. (2019). Computational methods for gene regulatory networks reconstruction and analysis: A review. *Artificial Intelligence in Medicine*, 95:133–145.

Ding, J. e Regev, A. (2021). Deep generative model embedding of single-cell RNA-Seq profiles on hyperspheres and hyperbolic spaces. *Nature Communications*, 12(1):2554.

Dong, M., Thennavan, A., Urrutia, E., Li, Y., Perou, C. M., Zou, F. e Jiang, Y. (2021). SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Briefings in Bioinformatics*, 22(1):416–427.

Edsgärd, D., Johnsson, P. e Sandberg, R. (2018). Identification of spatial expression trends in single-cell gene expression data. *Nature Methods*, 15(5):339–342.

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J. e Gardner, T. S. (2007). Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, 5(1):e8.

Feng, J., Meyer, C. A., Wang, Q., Liu, J. S., Shirley Liu, X. e Zhang, Y. (2012). GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics*, 28(21):2782–2788.

Finotello, F. e Camillo, B. D. (2015). Measuring differential gene expression with rna-seq: challenges and strategies for data analysis. *Briefings in Functional Genomics*, 14:130–142.

Fisher, R. A. (1941). The negative binomial distribution. *Annals of Human Genetics*, 11(1):182–187.

Foundation, P. S. (2024). *Python.org*.

Franzén, O. e Björkegren, J. L. M. (2020). alona: a web server for single-cell RNA-seq analysis. *Bioinformatics*, 36(12):3910–3912.

Frazee, A. C., Pertea, G., Jaffe, A. E., Langmead, B., Salzberg, S. L. e Leek, J. T. (2015). Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nature Biotechnology*, 33(3):243–246.

Gahlawat, A., R, R., Varma, T., Kamble, P., Banerjee, A., Sandhu, H. e Garg, P. (2023). Bioinformatics: Theory and applications. *The Quintessence of Basic and Clinical Research and Scientific Publishing*, páginas 539–555.

Gama-Castro, S., Salgado, H., Peralta-Gil, M., Santos-Zavaleta, A., Muniz-Rascado, L., Solano-Lira, H., Jimenez-Jacinto, V., Weiss, V., García-Sotelo, J. S., López-Fuentes, A., Porrón-Sotelo, L., Alquicira-Hernández, S., Medina-Rivera, A., Martínez-Flores, I., Alquicira-Hernández, K., Martínez-Adame, R., Bonavides-Martínez, C., Miranda-Ríos, J., Huerta, A. M., Mendoza-Vargas, A., Collado-Torres, L., Taboada, B., Vega-Alvarado, L., Olvera, M., Olvera, L., Grande, R., Morett, E. e Collado-Vides, J. (2011). RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic acids research*, 39(Database issue).

Garber, M., Grabherr, M. G., Guttman, M. e Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8(6):469–477.

Gardner, T. S., di Bernardo, D., Lorenz, D. e Collins, J. J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301:102–105.

Ge, S. X., Son, E. W. e Yao, R. (2018). iDEP: An integrated web application for differential expression and pathway analysis of RNA-Seq data. *BMC Bioinformatics*, 19(1):1–24.

Glaus, P., Honkela, A. e Rattray, M. (2012). Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, 28(13):1721–1728.

Griffith, M., Griffith, O. L., Mwenifumbo, J., Goya, R., Morrissy, A. S., Morin, R. D., Corbett, R., Tang, M. J., Hou, Y.-C., Pugh, T. J., Robertson, G., Chittaranjan, S., Ally, A., Asano, J. K., Chan, S. Y., Li, H. I., McDonald, H., Teague, K., Zhao, Y., Zeng, T., Delaney, A., Hirst, M., Morin, G. B., Jones, S. J. M., Tai, I. T. e Marra, M. A. (2010). Alternative expression analysis by RNA sequencing. *Nature Methods*, 7(10):843–847.

Guindalini, C. e Tufik, S. (2007). Uso de microarrays na busca de perfis de expressão gênica: aplicação no estudo de fenótipos complexos. *Revista Brasileira de Psiquiatria*, 29:370–374.

Guo, W., Tzioutziou, N. A., Stephen, G., Milne, I., Calixto, C. P., Waugh, R., Brown, J. W. e Zhang, R. (2020). 3D RNA-seq: a powerful and flexible tool for rapid and accurate differential expression and alternative splicing analysis of RNA-seq data for biologists. *RNA Biology*, páginas 1–14.

Hansey, C. N., Vaillancourt, B., Sekhon, R. S., de Leon, N., Kaeppler, S. M. e Buell, C. R. (2012). Maize (zea mays l.) genome diversity as revealed by rna-sequencing. *PLOS ONE*, 7(3):1–10.

Hardcastle, T. J. e Kelly, K. A. (2010). baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11(1):422.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Searle, S. e others (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research*, 22(9):1760–1774.

Hashimoto, R. F., Kim, S., Shmulevich, I., Zhang, W., Bittner, M. L. e Dougherty, E. R. (2004a). Growing genetic regulatory networks from seed genes. *Bioinformatics*, 20:1241–1247.

Hashimoto, R. F., Kim, S., Shmulevich, I., Zhang, W., Bittner, M. L. e Dougherty, E. R. (2004b). Growing genetic regulatory networks from seed genes. *Bioinformatics*, 20(8):1241–1247.

Heid, C. A., Stevens, J., Livak, K. J. e Williams, P. M. (1996). Real time quantitative PCR. *Genome research*, 6(10):986–994.

Hillmer, R. A. (2015). Systems biology for biologists. *PLOS Pathogens*, 11:e1004786.

Hoffman, G. E. e Roussos, P. (2020). Dream: powerful differential expression analysis for repeated measures designs. *Bioinformatics*.

Hong, M., Tao, S., Zhang, L., Diao, L. T., Huang, X., Huang, S., Xie, S. J., Xiao, Z. D. e Zhang, H. (2020). RNA sequencing: new technologies and applications in cancer research. *Journal of Hematology & Oncology 2020 13:1*, 13(1):1–16.

Hosseini, M., Pratas, D. e Pinho, A. (2016). A survey on data compression methods for biological sequences. *Information*, 7:56.

Howe, E. A., Sinha, R., Schlauch, D. e Quackenbush, J. (2012). RNA-Seq analysis in MeV. *Bioinformatics*, 27(22):3209–3210.

Hu, Y., Fang, L., Chen, X., Zhong, J. F., Li, M. e Wang, K. (2021). LIQA: long-read isoform quantification and analysis. *Genome Biology*, 22(1):1–21.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M. e Zhang, N. R. (2018). SAVER: gene expression recovery for single-cell RNA sequencing. *Nature Methods*, 15(7):539–542.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Ole's, A. K., Pag'es, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L. e Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):115–121.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. e Geurts, P. (2010). Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLOS ONE*, 5(9):e12776.

Huynh-Thu, V. A. e Sanguinetti, G. (2019). Gene regulatory network inference: An introductory survey. Em *Methods in Molecular Biology*, volume 1883, páginas 1–23. Humana Press Inc.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P. e von Mering, C. (2009). STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37(suppl_1):D412–D416.

Ji, Z. e Ji, H. (2016). TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Research*, 44(13):e117.

Jiang, G., Zheng, J.-Y., Ren, S.-N., Yin, W., Xia, X., Li, Y. e Wang, H.-L. (2024). A comprehensive workflow for optimizing rna-seq data analysis. *BMC Genomics*, 25:631.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. e Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462.

Kapushesky, M., Emam, I., Holloway, E., Kurnosov, P., Zorin, A., Malone, J., Rustici, G., Williams, E., Parkinson, H. e Brazma, A. (2009). Gene expression Atlas at the European Bioinformatics Institute. *Nucleic Acids Research*, 38(SUPPL.1):D690–D698.

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome research*, 12(4):656–64.

Kharchenko, P. V., Silberstein, L. e Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods 2014 11:7*, 11(7):740–742.

Kim, D., Langmead, B. e Salzberg, S. L. (2015). Hisat: a fast spliced aligner with low memory requirements. *Nature Methods*, 12:357–360.

Kim, D., Paggi, J. M., Park, C., Bennett, C. e Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. *Nature biotechnology*, 37:907–915.

Kim, D., Tran, A., Kim, H. J., Lin, Y., Yang, J. Y. H. e Yang, P. (2023). Gene regulatory network reconstruction: harnessing the power of single-cell multi-omic data. *npj Systems Biology and Applications 2023 9:1*, 9(1):1–13.

Klipp, E., Liebermeister, W., Wierling, C. e Kowald, A. (2016). *Systems Biology: A Textbook*, volume 2. Wiley.

Kuang, J., Michel, K. e Scoglio, C. (2023). GeCoNet-Tool: a software package for gene co-expression network construction and analysis. *BMC Bioinformatics*, 24(1):281.

Kucukural, A., Yukselen, O., Ozata, D. M., Moore, M. J. e Garber, M. (2019). DEBrowser: Interactive differential expression analysis and visualization tool for count data 06 Biological Sciences 0604 Genetics 08 Information and Computing Sciences 0806 Information Systems. *BMC Genomics*, 20(1):6.

Ladeira, P. R. S. d., Isaac, C. e Ferreira, M. C. (2011). Reação em cadeia da polimerase da transcrição reversa em tempo real. *Revista de Medicina*, 90(1):47.

Langmead, B., Hansen, K. D. e Leek, J. T. (2010). Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome biology*, 11(8):1–11.

Langmead, B. e Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359.

Langmead, B., Trapnell, C., Pop, M. e Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25.

Law, C. W., Chen, Y., Shi, W. e Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29.

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M. G., Haag, J. D., Gould, M. N., Stewart, R. M. e Kendziorski, C. (2013). EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, 29(8):1035–1043.

Li, B. e Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323.

Li, D. (2019). Statistical Methods for RNA Sequencing Data Analysis. Em *Computational Biology*, páginas 85–99. Codon Publications.

Li, H. e Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, J. e Tibshirani, R. (2013). Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research*, 22(5):519–536.

Li, P., Piao, Y., Shon, H. S. e Ryu, K. H. (2015a). Comparing the normalization methods for the differential analysis of illumina high-throughput rna-seq data. *BMC bioinformatics*, 16(1):347.

Li, X., Nair, A., Wang, S. e Wang, L. (2015b). Quality control of rna-seq experiments. *Methods in Molecular Biology*, 1269:137–146.

Liao, Y., Smyth, G. K. e Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930.

Liao, Y., Smyth, G. K. e Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*, 47(8):e47–e47.

Liu, F., Zhang, S. W., Guo, W. F., Wei, Z. G. e Chen, L. (2016). Inference of Gene Regulatory Network Based on Local Bayesian Networks. *PLoS Computational Biology*, 12(8):e1005024.

Liu, Y., Wang, T., Zhou, B. e Zheng, D. (2021). Robust integration of multiple single-cell RNA sequencing datasets using a single reference space. *Nature Biotechnology 2021 39:7*, 39(7):877–884.

Lohse, M., Bolger, A. M., Nagel, A., Fernie, A. R., Lunn, J. E., Stitt, M. e Usadel, B. (2012). Robina: A user-friendly, integrated software solution for rna-seq-based transcriptomics. *Nucleic Acids Research*, 40:W622–W627.

Lopes, F. M., Martins, D. C., Barrera, J. e Cesar, R. M. (2010). Sffs-mr: A floating search strategy for grns inference. Em Dijkstra, T. M. H., Tsivtsivadze, E., Marchiori, E. e Heskes, T., editores, *Pattern Recognition in Bioinformatics*, páginas 407–418, Berlin, Heidelberg. Springer Berlin Heidelberg.

Lopes, F. M., Martins, D. C. e Cesar, R. M. (2008). Feature selection environment for genomic applications. *BMC bioinformatics*, 9:1–8.

Lopes, F. M., Martins, D. C. e Cesar, R. M. (2009). Comparative study of grns inference methods based on feature selection by mutual information. Em *2009 IEEE International Workshop on Genomic Signal Processing and Statistics*, páginas 1–4. IEEE.

Lopes, F. M., Martins Jr, D. C., Barrera, J. e Cesar Jr, R. M. (2014). A feature selection technique for inference of graphs from their known topological properties: Revealing scale-free gene regulatory networks. *Information Sciences*, 272:1–15.

Lopes, M. F. (2011). *Redes complexas de expressão gênica: síntese, identificação, análise e aplicações*. Tese de doutorado, Universidade de São Paulo (USP).

Love, M. I., Huber, W. e Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550.

Lütge, A., Zyprych-Walczak, J., Kunzmann, U. B., Crowell, H. L., Calini, D., Malhotra, D., Soneson, C. e Robinson, M. D. (2021). CellMixS: quantifying and visualizing batch effects in single-cell RNA-seq data. *Life Science Alliance*, 4(6).

Ma, X. e Leng, N. (2024). *EBSeq: An R package for gene and isoform differential expression analysis of RNA-seq data*. R package version 2.2.0.

MacNeil, L. T. e Walhout, A. J. (2011). Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Research*, 21:645–657.

Madhukar, N. S., Elemento, O. e Pandey, G. (2015). Prediction of genetic interactions using machine learning and network properties.

Maetschke, S. R., Madhamshettiwar, P. B., Davis, M. J. e Ragan, M. A. (2014). Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Briefings in Bioinformatics*, 15(2):195–211.

Mahi, N. A., Najafabadi, M. F., Pilarczyk, M., Kouril, M. e Medvedovic, M. (2019). GREIN: An Interactive Web Platform for Re-analyzing GEO RNA-seq Data. *Scientific Reports*, 9(1):1–9.

Mancini, E., Rabinovich, A., Iserte, J., Yanovsky, M. e Chernomoretz, A. (2021). ASpli: integrative analysis of splicing landscapes through RNA-Seq assays. *Bioinformatics*, 37(17):2609–2616.

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., DREAM5 Consortium, Kellis, M., Collins, J. J. e Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804.

Marbach, D., Schaffter, T., Mattiussi, C. e Floreano, D. (2009). Generating Realistic In Silico Gene Networks for Performance Assessment of Reverse Engineering Methods. *Journal of Computational Biology*, 16(2):229–239.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D. e Califano, A. (2006a). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(SUPPL.1):1–15.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D. e Califano, A. (2006b). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(SUPPL.1):1–15.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. e Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. e Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517.

Marku, M. e Pancaldi, V. (2023). From time-series transcriptomics to gene regulatory networks: A review on inference methods. *PLOS Computational Biology*, 19:e1011254.

Martins Jr, D. C., Lopes, F. M. e Ray, S. S. (2016). Inference of gene regulatory networks by topological prior information and data integration. Em *Emerging Research in the Analysis and Modeling of Gene Regulatory Networks*, páginas 1–51. IGI Global.

Mason, O. e Verwoerd, M. (2007). Graph theory and networks in biology. *IET Systems Biology*, 1(2):89–119.

McCarthy, D. J., Campbell, K. R., Lun, A. T. L. e Wills, Q. F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, 33(8):1179–1186.

McGettigan, P. A. (2013). Transcriptomics in the RNA-seq era. *Current opinion in chemical biology*, 17(1):4–11.

Mendoza, M. R., Lopes, F. M. e Bazzan, A. L. C. (2012). Reverse engineering of grns: an evolutionary approach based on the tsallis entropy. Em *Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation*, GECCO '12, página 185–192, New York, NY, USA. Association for Computing Machinery.

Mercatelli, D., Lopez-Garcia, G. e Giorgi, F. M. (2020). corto: a lightweight R package for gene network inference and master regulator analysis. *Bioinformatics*, 36(12):3916–3917.

Meyer, P. (2010). Cran: Package infotheo.

Meyer, P. E., Kontos, K., Lafitte, F. e Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *EURASIP journal on bioinformatics & systems biology*, 2007(1).

Meyer, P. E., Lafitte, F. e Bontempi, G. (2008). Minet: A r/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, 9(1):1–10.

Meyer, P. E., Marbach, D., Roy, S. e Kellis, M. (2010a). Information-theoretic inference of gene networks using backward elimination. *MIT web domain*.

Meyer, P. E., Marbach, D., Roy, S. e Kellis, M. (2010b). Information-theoretic inference of gene networks using backward elimination. *MIT Open Access Articles - CSREA Press*.

Miao, Z., Deng, K., Wang, X. e Zhang, X. (2018). DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics*, 34(18):3223–3224.

Mira, N. P., Teixeira, M. C. e Sá-Correia, I. (2012). Characterization of complex regulatory networks and identification of promoter regulatory elements in yeast: "in silico" and "wet-lab" approaches. *Methods in Molecular Biology*, 809:27–48.

MORETTIN, L. G. (2009). *Estatística básica: probabilidade e inferência*. Pearson Prentice Hal, São Paulo.

Morgante, C. V., Blawid, R. e Petrolina, E. S. (2016). Análise da expressão gênica pela técnica de pcr quantitativa em tempo real: Princípios e fundamentos. Em *Documentos Online 278*.

Morin, E., Le Moigne, J.-L., Poloni, J. d. F., Feltes, B. C., Silva, F. R. e Bonatto, D. (2014). *Biologia de Sistemas*, páginas 37–67. Sociedade Brasileira de Bioquímica e Biologia Molecular - SBBq, 1 edition.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. e Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628.

Mount, D. W. (2007). Using the Basic Local Alignment Search Tool (BLAST). *Cold Spring Harbor Protocols*, 2007(7):pdb.top17.

Mudunuri, U., Che, A., Yi, M. e Stephens, R. M. (2009). bioDBnet: the biological database network. *Bioinformatics*, 25(4):555–556.

Mullis, K. (1993). Nobel Lecture: The Polymerase Chain Reaction. *Nobel Prize: Chemistry. The Royal Swedish Academy of Sciences, Sweden*, 8.

Mullis, K. B. e Faloona, F. A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods in enzymology*, 155:335.

Musilova, J., Vafek, Z., Puniya, B. L., Zimmer, R., Helikar, T. e Sedlar, K. (2024). Augusta: From RNA-Seq to gene regulatory networks and Boolean models. *Computational and Structural Biotechnology Journal*, 23:783–790.

Najjar, R. e Mustelin, T. (2023). Prediction of alternative pre-mrna splicing outcomes. *Scientific Reports 2023 13:1*, 13:1–10.

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., Khodadoust, M. S., Esfahani, M. S., Luca, B. A., Steiner, D., Diehn, M. e Alizadeh, A. A. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology*, 37:773–782. Using Seurat28, clusters were identified by (supl.1)
.

Otero, J. M. e Nielsen, J. (2010). Industrial systems biology. *Biotechnology and Bioengineering*, 105(3):439–460.

Otsu, N. et al. (1975). A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27.

Overbey, E. G., Saravia-Butler, A. M., Zhang, Z., Rathi, K. S., Fogle, H., da Silveira, W. A., Barker, R. J., Bass, J. J., Beheshti, A., Berrios, D. C., Blaber, E. A., Cekanaviciute, E., Costa, H. A., Davin, L. B., Fisch, K. M., Gebre, S. G., Geniza, M., Gilbert, R., Gilroy, S., Hardiman, G., Herranz, R., Kidane, Y. H., Kruse, C. P., Lee, M. D., Liefeld, T., Lewis, N. G., McDonald, J. T., Meller, R., Mishra, T., Perera, I. Y., Ray, S., Reinsch, S. S., Rosenthal, S. B., Strong, M., Szewczyk, N. J., Tahimic, C. G., Taylor, D. M., Vandenbrink, J. P., Villacampa, A., Weging, S., Wolverton, C., Wyatt, S. E., Zea, L., Costes, S. V. e Galazka, J. M. (2021). Nasa genelab rna-seq consensus pipeline: Standardized processing of short-read rna-seq data. *iScience*, 24:102361.

Partel, G. e Wählby, C. (2021). Spage2vec: Unsupervised representation of localized spatial gene expression signatures. *The FEBS Journal*, 288(6):1859–1870.

Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., Schneider, R. e Bagos, P. G. (2011). Using graph theory to analyze biological networks. *BioData Mining*, 4(1):1–27.

Penfold, C. A., Buchanan-Wollaston, V., Denby, K. J. e Wild, D. L. (2012). Nonparametric Bayesian inference for perturbed and orthologous gene regulatory networks. *Bioinformatics*, 28(12):i233–i241.

Pham, T. H., Ho, T. B., Nguyen, Q. D., Tran, D. H. e Nguyen, V. H. (2012). Multivariate mutual information measures for discovering biological networks. *2012 IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future, RIVF 2012*.

Pietu, G., Mariage-Samson, R., Fayein, N.-A., Matingou, C., Eveno, E., Houlgatte, R., Decraene, C., Vandenbrouck, Y., Tahi, F., Devignes, M.-D. e others (1999). The Genexpress IMAGE knowledge base of the human brain transcriptome: A prototype integrated resource for functional and computational genomics. *Genome research*, 9(2):195–209.

Pimentel, H., Bray, N. L., Puente, S., Melsted, P. e Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods*, 14(7):687–690.

Poisson, S. D. e Schnuse, C. H. (1841). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*. Meyer.

Pressman, R. e Maxim, B. (2021). *Engenharia de software - 9.ed.* McGraw Hill Brasil.

Pušnik, v., Mraz, M., Zimic, N. e Moškon, M. (2022). Review and assessment of boolean approaches for inference of gene regulatory networks. *Heliyon*, 8(8):e10222.

Quinlan, A. R. e Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D. e Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol*, 14(9):R95.

Rastogi, A. e Gupta, D. (2014). GFF-Ex: a genome feature extraction package. *BMC Research Notes*, 7(1):315.

Redhu, N. e Thakur, Z. (2022). Network biology and applications. *Bioinformatics*, páginas 381–407.

Rieu, I. e Powers, S. J. (2009). Real-time quantitative RT-PCR: design, calculations, and statistics. *The Plant Cell*, 21(4):1031–1033.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. e Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47.

Roberts, A. e Pachter, L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods*, 10(1):71–73.

Robinson, M. D., McCarthy, D. J. e Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Robinson, M. D. e Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25.

Rosati, D., Palmieri, M., Brunelli, G., Morrione, A., Iannelli, F., Frullanti, E. e Giordano, A. (2024). Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review. *Computational and Structural Biotechnology Journal*, 23:1154–1168.

Sahadeo, N. S. D., Allicock, O. M., De Salazar, P. M., Auguste, A. J., Widen, S., Olowokure, B., Gutierrez, C., Valadere, A. M., Polson-Edwards, K., Weaver, S. C. e Carrington, C. V. F. (2017). Understanding the evolution and spread of chikungunya virus in the americas using complete genome sequences. *Virus Evolution*, 3(1):vex010.

Salojärvi, J., Rambani, A., Yu, Z., Guyot, R., Strickler, S., Lepelley, M., Wang, C., Rajaraman, S., Rastas, P., Zheng, C. et al. (2024). The genome and population genomics of allopolyploid coffea arabica reveal the diversification history of modern coffee cultivars. *Nature genetics*, 56(4):721–731.

Schaarschmidt, S., Fischer, A., Zuther, E. e Hincha, D. K. (2020). Evaluation of seven different rna-seq alignment tools based on experimental data from the model plant arabidopsis thaliana. *International Journal of Molecular Sciences*, 21:1720.

Schena, M., Shalon, D., Davis, R. W. e Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470.

Schoolar, G. (2024). Página inicial do google acadêmico.

Seki, M., Narusaka, M., Ishida, J., Nanjo, T., Fujita, M., Oono, Y., Kamiya, A., Nakajima, M., Enju, A., Sakurai, T., Satou, M., Akiyama, K., Taji, T., Yamaguchi-Shinozaki, K., Carninci, P., Kawai, J., Hayashizaki, Y. e Shinozaki, K. (2002). Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. *The Plant Journal*, 31(3):279–292.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

Simoneau, J., Dumontier, S., Gosselin, R. e Scott, M. S. (2021). Current RNA-seq methodology reporting limits reproducibility. *Briefings in Bioinformatics*, 22(1):140–145.

Sławek, J. e Arodź, T. (2013). ENNET: Inferring large gene regulatory networks from expression data using gradient boosting. *BMC Systems Biology*, 7(1):1–13.

Snustad, D. P., Simmons, M. J., Jenkins, J. B. e Crow, J. F. (2000). *Principles of genetics*. John Wiley.

Soneson, C. e Delorenzi, M. (2013). A comparison of methods for differential expression analysis of rna-seq data. *BMC bioinformatics*, 14(1):91.

Song, D. e Li, J. J. (2021). Pseudotimede: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell rna sequencing data. *Genome Biology*, 22:1–25.

Steuer, R., Kurths, J., Daub, C. O., Weise, J. e Selbig, J. (2002). The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics*, 18:S231–S240.

Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H. e Yaspo, M.-L. (2008). A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science*, 321(5891):956–960.

Sun, J., Nishiyama, T., Shimizu, K. e Kadota, K. (2013). Tcc: An r package for comparing tag count data with robust normalization strategies. *BMC Bioinformatics*, 14:1–14.

Sun, S., Xu, L., Zou, Q. e Wang, G. (2021). BP4RNAseq: a babysitter package for retrospective and newly generated RNA-seq data analyses using both alignment-based and alignment-free quantification method. *Bioinformatics*, 37(9):1319–1321.

Sutter, J. M. e Kalivas, J. H. (1993). Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection. *Microchemical Journal*, 47:60–66.

Tao, Y., Zhang, Q., Wang, H., Yang, X. e Mu, H. (2024). Alternative splicing and related rna binding proteins in human health and disease. *Signal Transduction and Targeted Therapy 2024 9:1*, 9:1–33.

Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. D., Nueda, M. J., Ferrer, A. e Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research*, página gkv711.

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. e Conesa, A. (2011). Differential expression in RNA-seq: A matter of depth. *Genome Research*, 21(12):2213–2223.

ThermoFisher, S. (2024). Microarray analysis - br.

Todorov, H., Cannoodt, R., Saelens, W. e Saeys, Y. (2019). *Network Inference from Single-Cell Transcriptomic Data*, páginas 235–249. Springer New York.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L. e Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, 31(1):46–53.

Trapnell, C., Pachter, L. e Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L. e Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3):562–578.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. e Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515.

Varet, H., Brillet-Guéguen, L., Coppée, J.-Y. e Dillies, M.-A. (2016). Sartools: A deseq2- and edger-based r pipeline for comprehensive differential analysis of rna-seq data. *PLOS ONE*, 11:e0157022.

Velculescu, V. E., Zhang, L., Vogelstein, B. e Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, 270(5235):484–487.

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Hieter, P., Vogelstein, B. e Kinzler, K. W. (1997). Characterization of the yeast transcriptome. *Cell*, 88(2):243–251.

Wagner, G. P., Kin, K. e Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences*, 131(4):281–285.

Wang, J., Chen, Y. e Zou, Q. (2023). Inferring gene regulatory network from single-cell transcriptomes with graph autoencoder model. *PLOS Genetics*, 19(9):e1010942.

Wang, J., Roeder, K. e Devlin, B. (2021). Bayesian estimation of cell type–specific gene expression with prior derived from single-cell data. *Genome Research*, 31(10):1807–1818.

Wang, L., Feng, Z., Wang, X., Wang, X. e Zhang, X. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 26(1):136–138.

Wang, T. e Brown, M. J. (1999). mRNA quantification by real time TaqMan polymerase chain reaction: validation and comparison with RNase protection. *Analytical biochemistry*, 269(1):198–201.

Wang, Z., Gerstein, M. e Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S. et al. (2009). *gplots: Various R programming tools for plotting data*.

Wickham, H. (2015). *R Packages: Organize, Test, Document, and Share Your Code*. O'Reilly Media, 1 edition.

Wickham, H., Danenberg, P., Csárdi, G. e Eugster, M. (2024). *roxygen2: In-Line Documentation for R*. R package version 7.3.2, https://github.com/r-lib/roxygen2.

Wickham, H., Hester, J., Chang, W. e Bryan, J. (2022). *devtools: Tools to Make Developing R Packages Easier*. https://devtools.r-lib.org/, https://github.com/r-lib/devtools.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80.

Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C. J., Rogers, J. e Bähler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453(7199):1239–1243.

Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., Meng, J., Zhu, Z., Zhang, Z., Wang, J. et al. (2020). Genome composition and divergence of the novel coronavirus (2019-ncov) originating in china. *Cell host & microbe*, 27(3):325–328.

Wu, T. D. e Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881.

Wu, T.-J., Huang, Y.-H. e Li, L.-A. (2005). Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences. *Bioinformatics (Oxford, England)*, 21(22):4125–32.

Yamaguchi-Shinozaki, K. e Shinozaki, K. (1994). A novel cis-acting element in an arabidopsis gene is involved in responsiveness to drought, low-temperature, or high-salt stress. *The Plant Cell*, 6(2):251–264.

Yamaguchi-Shinozaki, K. e Shinozaki, K. (2006). Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. *Annual Review of Plant Biology*, 57:781–803.

Yépez, V. A., Mertes, C., Müller, M. F., Klaproth-Andrade, D., Wachutka, L., Frésard, L., Gusic, M., Scheller, I. F., Goldberg, P. F., Prokisch, H. e Gagneur, J. (2021). Detection of aberrant gene expression events in RNA sequencing data. *Nature Protocols*, 16(2):1276–1296.

Zaha, A., Ferreira, H. B. e Passaglia, L. M. P. (2014). *Biologia Molecular Básica-5*. Artmed Editora.

Zayakin, P. (2024). srnaflow: A tool for the analysis of small rna-seq data. *Non-Coding RNA*, 10:6.

Zhang, H. (2016). Overview of Sequence Data Formats. Em *Methods in Molecular Biology*, volume 1418, capítulo: 1, páginas 3–17. Humana Press, New York, NY.

Zhang, Y., Parmigiani, G. e Johnson, W. E. (2020). Combat-seq: batch effect adjustment for rna-seq count data. *NAR Genomics and Bioinformatics*, 2.

Zhang, Z. H., Jhaveri, D. J., Marshall, V. M., Bauer, D. C., Edson, J., Narayanan, R. K., Robinson, G. J., Lundberg, A. E., Bartlett, P. F., Wray, N. R. e others (2014). A comparative study of techniques for differential expression analysis on RNA-Seq data. *PloS one*, 9(8):e103207.

Zhao, M., He, W., Tang, J., Zou, Q. e Guo, F. (2021a). A comprehensive overview and critical evaluation of gene regulatory network inference technologies. *Briefings in Bioinformatics*, 22(5):1–15.

Zhao, R. F. (2012). Encode: Deciphering function in the human genome.

Zhao, Y., Li, M.-C., Konaté, M. M., Chen, L., Das, B., Karlovich, C., Williams, P. M., Evrard, Y. A., Doroshow, J. H. e McShane, L. M. (2021b). Tpm, fpkm, or normalized counts? a comparative study of quantification measures for the analysis of rna-seq data from the nci patient-derived models repository. *Journal of translational medicine*, 19(1):269.

Zheng, Q., Guo, L., Huang, J., Hao, X., Li, X., Li, N., Wang, Y., Zhang, K., Wang, X., Wang, L. e Zeng, J. (2023). Comparative transcriptomics provides novel insights into the mechanisms of selenium accumulation and transportation in tea cultivars (camellia sinensis (l.) o. kuntze). *Frontiers in Plant Science*, 14.

APÊNDICE A – DETALHAMENTO DE MÉTODOS PARA ANÁLISE DE EXPRESSÃO

Neste apêndice apresentamos os métodos para análise de expressão diferencial de genes analisados no artigo de revisão (Costa-Silva et al., 2023) desta tese. Também são detalhados os parâmetros utilizados na busca.

A.1 MÉTODO DE BUSCA APLICADO

Esta seção apresenta os critérios adotados para a seleção dos métodos computacionais considerados. O termo de pesquisa: "RNA-Seq differential expression analysis" foi realizado considerando a ferramenta de pesquisa acadêmica Google Scholar.

Os resultados da pesquisa foram selecionados pelos critérios: método implementado (software), realização da análise de expressão e citações por ano de publicação, conforme descrito a seguir.

1. Ano de publicação entre 2006 a 2017 e mais que 200 citações no Google Acadêmico;
2. Ano de publicação 2018 e mais que 100 citações no Google Acadêmico;
3. Ano de publicação 2019 e mais que 50 citações no Google Acadêmico;
4. Ano de publicação 2020 e mais que 10 citações no Google Acadêmico;
5. Ano de publicação 2021 ou maior e mais que 5 citações no Google Acadêmico;

A.2 SOFTWARES SELECIONADOS

Como resultado da busca na literatura até 2022, identificamos os métodos computacionais apresentados no trabalho e listados na tabela A.1.

Tabela A.1: Métodos computacionais identificados com base nas características da busca. O número de citações se refere a data em que a revisão foi conduzida, conforme descrito.

Método Computacional	Classificação	Data	Distribuição	Citações	Referência
Bainbridge, et al. 2006	Estudo seminal	2006	NA	192	(Bainbridge et al., 2006)
Wilhelm, et al.2008	Estudo seminal	2008	NA	1059	(Wilhelm et al., 2008)
ERANGE	Estudo seminal	2008	NA	11778	(Mortazavi et al., 2008)
Sultan, et al. 2008	Estudo seminal	2008	Poisson	1414	(Sultan et al., 2008)
DEGSeq	Paramétrica	2009	Poisson	2465	(Wang et al., 2010)
GXA	Estudo seminal	2009	NA	232	(Kapushesky et al., 2009)
edgeR	Paramétrica	2009	Binomial Negativa	18483	(Robinson et al., 2010)
DESeq	Paramétrica	2010	Binomial Negativa	11662	(Anders e Huber, 2010)
baySeq	Paramétrica	2010	Binomial Negativa	778	(Hardcastle e Kelly, 2010)
Myrna	Paramétrica ou Não Paramétrica	2010	Poisson	381	(Langmead et al., 2010)
NOISeq	Não Paramétrica	2011	NA	1222	(Tarazona et al., 2011)
MeV	Paramétrica	2011	Binomial Negativa	267	(Howe et al., 2012)
DEXSeq	Paramétrica	2012	Binomial Negativa	1089	(Anders et al., 2012)
BitSeq	Não Paramétrica	2012	NA	207	(Glaus et al., 2012)
RobiNA	Paramétrica	2012	Binomial Negativa	766	(Lohse et al., 2012)
GFOLD	Paramétrica	2012	Poisson	315	(Feng et al., 2012)
Cuffdiff2	Paramétrica	2012	Binomial Negativa	2920	(Trapnell et al., 2013)
eXpress	Paramétrica	2013	Beta-binomial	840	(Roberts e Pachter, 2013)
TCC	Paramétrica	2013	Binomial Negativa	401*	(Sun et al., 2013)
EBSeq	Paramétrica	2013	Binomial Negativa	931	(Leng et al., 2013)

Continuação da Tabela A.1 da página anterior

Método Computacional	Classificação	Data	Distribuição	Citado	Referência
SAMSeq	Não Paramétrica	2013	NA	419	(Li e Tibshirani, 2013)
limma-voom	Paramétrica	2014	Binomial Negativa	2777	(Law et al., 2014)
SCDE*	Paramétrica	2014	Poisson	925	(Kharchenko et al., 2014)
DESeq2	Paramétrica	2014	Binomial Negativa	23374	(Love et al., 2014)
Ballgown	Paramétrica	2015	Binomial Negativa	339	(Frazee et al., 2015)
TSCAN*	Paramétrica	2016	NA	302	(Ji e Ji, 2016)
SARTools	Paramétrica	2016	Binomial Negativa	246	(Varet et al., 2016)
Scater*	Paramétrica	2017	Binomial Negativa	645	(McCarthy et al., 2017)
sleuth	Paramétrica	2017	Normal	613	(Pimentel et al., 2017)
consexpression	Hibrída	2017	Binomial Negativa	241	(Costa-Silva et al., 2017a)
trendsceek*	Paramétrica	2018	Binomial Negativa	71	(Edsgärd et al., 2018)
DESingle*	Paramétrica	2018	Binomial Negativa	55	(Miao et al., 2018)
SAVER*	Paramétrica	2018	Poisson	269	(Huang et al., 2018)
iDEP	Paramétrica	2018	Binomial Negativa	139	(Ge et al., 2018)
DEBrowser	Paramétrica	2019	Binomial Negativa	64	(Kucukural et al., 2019)
CIBERSORTx*	Não Paramétrica	2019	NA	513	(Newman et al., 2019)
GREIN	Paramétrica	2019	Binomial Negativa	31	(Mahi et al., 2019)
RASflow	Paramétrica	2020	Binomial Negativa	4	(Zhang et al., 2020)
alona*	Paramétrica	2020	Normal	7	(Franzén e Björkegren, 2020)
dream	Paramétrica	2020	Binomial Negativa	12	(Hoffman e Roussos, 2020)
Spage2vec*	Não Paramétrica	2020	NA	7	(Partel e Wählby, 2021)
SCDC*	Não Paramétrica	2020	NA	47	(Dong et al., 2021)
3D RNA-seq	Paramétrica	2020	Binomial Negativa	14	(Guo et al., 2020)
DROP	Paramétrica	2021	Binomial Negativa	12	(Yépez et al., 2021)
Aspli	Paramétrica	2021	Binomial Negativa	5	(Mancini et al., 2021)
CellMixS*	Paramétrica	2021	NA	6	(Lütge et al., 2021)
RISC*	Paramétrica	2021	Binomial Negativa	4	(Liu et al., 2021)
bMIND*	Paramétrica	2021	Invertida de Wishart	6	(Wang et al., 2021)
FINDER	Paramétrica	2021	Exponencial	12	(Banerjee et al., 2021)
RCP	Não Paramétrica	2021	Binomial Negativa	6	(Overbey et al., 2021)
PseudotimeDE*	Paramétrica	2021	Binomial Negativa	10	(Song e Li, 2021)
BP4RNASeq	Paramétrica	2021	NA	52	(Sun et al., 2021)
scPhere*	Paramétrica	2021	Binomial Negativa	19	(Ding e Regev, 2021)
KnowSeq	Paramétrica	2021	NA	6	(Castillo-Secilla et al., 2021a)
NS-Forest*	Não Paramétrica	2021	NA	8	(Aevermann et al., 2021)
LIQA	Não Paramétrica	2021	NA	8	(Hu et al., 2021)

Tabela B.1: Avaliação de desempenho individual de metodologias com o conjunto de dados A.

Metodologia	TP	FP	TN	FN	TPR	Especificidade	PPV	ACC	F1-Score
limma	327	29	523	71	0.82	0.94	0.91	0.89	0.86
samseq	135	37	515	263	0.33	0.93	0.78	0.68	0.47
deseq2	324	33	519	74	0.81	0.94	0.90	0.88	0.85
edger	329	29	523	69	0.82	0.94	0.91	0.89	0.87
noiseq	299	223	329	99	0.75	0.59	0.57	0.66	0.65
knowseq	280	47	505	118	0.70	0.91	0.85	0.82	0.77
ebseq	371	354	198	27	0.93	0.35	0.51	0.59	0.66

Tabela B.2: Avaliação de desempenho individual de metodologias com o conjunto de dados B.

	TP	FP	TN	FN	TPR	Especificidade	PPV	ACC	F1-Score
limma	2	0	87	19	0.10	1.00	1.00	0.82	0.17
samseq	4	3	84	17	0.19	0.97	0.57	0.81	0.29
deseq2	3	0	87	18	0.14	1.00	1.00	0.83	0.25
edger	2	0	87	19	0.10	1.00	1.00	0.82	0.17
noiseq	5	7	80	16	0.24	0.92	0.42	0.79	0.30
knowseq	1	0	87	20	0.05	1.00	1.00	0.81	0.09
ebseq	10	30	57	11	0.48	0.66	0.25	0.62	0.33

APÊNDICE C – DESEMPENHO DOS MÉTODOS DE INFERÊNCIA DE REDES

Neste capítulo são apresentados os limiares aplicados na consideração de arestas em cada método analisado no trabalho. Em conjunto com esses resultados também são apresentadas as medidas de desempenho de cada limiar quando comparado com a rede de referência utilizada.

Tabela C.1: Limiares aplicados a métodos de inferência e medidas de desempenho em cada limiar.

Método	Limiar	FP	FN	TP	TN	ACC	Recall	Precision	F-Score	FDR
BC3NET	0	1511	1945	55	10168794	0,999660	0,027500	0,035121	0,030847	0,000149
	0,1	744	1960	40	10169561	0,999734	0,020000	0,051020	0,028736	0,000073
	0,2	532	1966	34	10169773	0,999754	0,017000	0,060071	0,026500	0,000052
	0,3	408	1969	31	10169897	0,999766	0,015500	0,070615	0,025420	0,000040
	0,4	341	1970	30	10169964	0,999773	0,015000	0,080863	0,025306	0,000034
	0,5	282	1975	25	10170023	0,999778	0,012500	0,081433	0,021673	0,000028
	0,6	243	1978	22	10170062	0,999782	0,011000	0,083019	0,019426	0,000024
	0,7	204	1978	22	10170101	0,999785	0,011000	0,097345	0,019766	0,000020
	0,8	161	1979	21	10170144	0,999790	0,010500	0,115385	0,019248	0,000016
	0,9	128	1981	19	10170177	0,999793	0,009500	0,129252	0,017699	0,000013
C3NET	0	366	1975	25	10169939	0,999770	0,012500	0,063939	0,020912	0,000036
	0,1781	309	1978	22	10169996	0,999775	0,011000	0,066465	0,018876	0,000030
	0,3562	115	1984	16	10170190	0,999794	0,008000	0,122137	0,015016	0,000011
	0,5343	33	1988	12	10170272	0,999801	0,006000	0,266667	0,011736	0,000003
	0,7124	16	1994	6	10170289	0,999802	0,003000	0,272727	0,005935	0,000002
	0,8904	6	1997	3	10170299	0,999803	0,001500	0,333333	0,002987	0,000001
	1,0685	2	1998	2	10170303	0,999803	0,001000	0,500000	0,001996	0,000000
	1,2466	0	2000	0	10170305	0,999803	0,000000	0,000000	0,000000	0,000000
	1,4247	0	2000	0	10170305	0,999803	0,000000	0,000000	0,000000	0,000000
	1,6028	0	2000	0	10170305	0,999803	0,000000	0,000000	0,000000	0,000000
ARACNE	0	1041	1973	27	10169264	0,999704	0,013500	0,025281	0,017601	0,000102
	0,1	746	1974	26	10169559	0,999733	0,013000	0,033679	0,018759	0,000073
	0,2	456	1981	19	10169849	0,999760	0,009500	0,040000	0,015354	0,000045
	0,3	237	1988	12	10170068	0,999781	0,006000	0,048193	0,010671	0,000023
	0,4	66	1994	6	10170239	0,999797	0,003000	0,083333	0,005792	0,000006
	0,5	15	1997	3	10170290	0,999802	0,001500	0,166667	0,002973	0,000001
	0,6	2	1999	1	10170303	0,999803	0,000500	0,333333	0,000999	0,000000
	0,7	0	1999	1	10170305	0,999803	0,000500	1,000000	0,001000	0,000000
	0,8	0	2000	0	10170305	0,999803	0,000000	0,000000	0,000000	0,000000
	0,9	0	2000	0	10170305	0,999803	0,000000	0,000000	0,000000	0,000000
CLR	0	565904	1120	880	9604401	0,944258	0,440000	0,001553	0,003094	0,055643
	0,1	5960	1940	60	10164345	0,999223	0,030000	0,009967	0,014963	0,000586
	0,2	468	1985	15	10169837	0,999759	0,007500	0,031056	0,012082	0,000046
	0,3	66	1992	8	10170239	0,999798	0,004000	0,108108	0,007715	0,000006
	0,4	14	1994	6	10170291	0,999803	0,003000	0,300000	0,005941	0,000001
	0,5	5	1994	6	10170300	0,999803	0,003000	0,545455	0,005967	0,000000
	0,6	3	1998	2	10170302	0,999803	0,001000	0,400000	0,001995	0,000000
	0,7	1	1999	1	10170304	0,999803	0,000500	0,500000	0,000999	0,000000
	0,8	1	1999	1	10170304	0,999803	0,000500	0,500000	0,000999	0,000000
	0,9	0	1999	1	10170305	0,999803	0,000500	1,000000	0,001000	0,000000
MRNET	0	521379	1210	790	9648926	0,948626	0,395000	0,001513	0,003014	0,051265
	0,1	478	1977	23	10169827	0,999759	0,011500	0,045908	0,018393	0,000047
	0,2	215	1986	14	10170090	0,999784	0,007000	0,061135	0,012562	0,000021
	0,3	120	1991	9	10170185	0,999792	0,004500	0,069767	0,008455	0,000012
	0,4	46	1994	6	10170259	0,999799	0,003000	0,115385	0,005848	0,000005
	0,5	11	1997	3	10170294	0,999803	0,001500	0,214286	0,002979	0,000001
	0,6	2	1999	1	10170303	0,999803	0,000500	0,333333	0,000999	0,000000
	0,7	0	1999	1	10170305	0,999803	0,000500	1,000000	0,001000	0,000000
	0,8	0	2000	0	10170305	0,999803	0,000000	0,000000	0,000000	0,000000
	0,9	0	2000	0	10170305	0,999803	0,000000	0,000000	0,000000	0,000000
MRNETB	0	521544	1210	790	9648761	0,948610	0,395000	0,001512	0,003013	0,051281
	0,1	33144	1860	140	10137161	0,996559	0,070000	0,004206	0,007936	0,003259
	0,2	5083	1952	48	10165222	0,999308	0,024000	0,009355	0,013462	0,000500
	0,3	765	1984	16	10169540	0,999730	0,008000	0,020487	0,011507	0,000075
	0,4	119	1991	9	10170186	0,999793	0,004500	0,070313	0,008459	0,000012
	0,5	11	1996	4	10170294	0,999803	0,002000	0,266667	0,003970	0,000001
	0,6	2	1999	1	10170303	0,999803	0,000500	0,333333	0,000999	0,000000
	0,7	0	1999	1	10170305	0,999803	0,000500	1,000000	0,001000	0,000000
	0,8	0	2000	0	10170305	0,999803	0,000000	0,000000	0,000000	0,000000
	0,9	0	2000	0	10170305	0,999803	0,000000	0,000000	0,000000	0,000000
GENIE3	0	1448729	0	2000	8721576	0,857581	1,000000	0,001379	0,002753	0,142447

Table C.1 continued from previous page

Método	Limiar	FP	FN	TP	TN	ACC	Recall	Precision	F-Score	FDR
	0,017	30862	1633	367	10139443	0,996806	0,183500	0,011752	0,022089	0,003035
	0,0341	5483	1799	201	10164822	0,999284	0,100500	0,035362	0,052317	0,000539
	0,0511	1179	1860	140	10169126	0,999701	0,070000	0,106141	0,084363	0,000116
	0,0681	382	1919	81	10169923	0,999774	0,040500	0,174946	0,065773	0,000038
	0,0851	181	1968	32	10170124	0,999789	0,016000	0,150235	0,028920	0,000018
	0,1022	108	1990	10	10170197	0,999794	0,005000	0,084746	0,009443	0,000011
	0,1192	51	1996	4	10170254	0,999799	0,002000	0,072727	0,003893	0,000005
	0,1362	16	1999	1	10170289	0,999802	0,000500	0,058824	0,000992	0,000002
	0,1533	3	2000	0	10170302	0,999803	0,000000	0,000000	0,000000	0,000000
ENNET	0	100994	1604	396	10069311	0,989914	0,198000	0,003906	0,007660	0,009930
	0	130	1980	20	10170175	0,999793	0,010000	0,133333	0,018605	0,000013
	0	46	1990	10	10170259	0,999800	0,005000	0,178571	0,009728	0,000005
	0	27	1994	6	10170278	0,999801	0,003000	0,181818	0,005903	0,000003
	0	16	1994	6	10170289	0,999802	0,003000	0,272727	0,005935	0,000002
	0	5	1996	4	10170300	0,999803	0,002000	0,444444	0,003982	0,000000
	0	3	1996	4	10170302	0,999803	0,002000	0,571429	0,003986	0,000000
	0	1	1998	2	10170304	0,999803	0,001000	0,666667	0,001997	0,000000
	0	0	2000	0	10170305	0,999803	0,000000	0,000000	0,000000	0,000000
	0	0	2000	0	10170305	0,999803	0,000000	0,000000	0,000000	0,000000
GeCoNet-Tool	0	3814	1904	96	10166491	0,999438	0,048000	0,024552	0,032487	0,000375
	0,0991	3814	1904	96	10166491	0,999438	0,048000	0,024552	0,032487	0,000375
	0,1982	3814	1904	96	10166491	0,999438	0,048000	0,024552	0,032487	0,000375
	0,2973	3814	1904	96	10166491	0,999438	0,048000	0,024552	0,032487	0,000375
	0,3964	3814	1904	96	10166491	0,999438	0,048000	0,024552	0,032487	0,000375
	0,4955	3814	1904	96	10166491	0,999438	0,048000	0,024552	0,032487	0,000375
	0,5946	3814	1904	96	10166491	0,999438	0,048000	0,024552	0,032487	0,000375
	0,6937	2557	1929	71	10167748	0,999559	0,035500	0,027017	0,030683	0,000251
	0,7928	483	1971	29	10169822	0,999759	0,014500	0,056641	0,023089	0,000047
	0,8919	68	1989	11	10170237	0,999798	0,005500	0,139241	0,010582	0,000007
Corto	-0,7806	1448728	0	2000	8721577	0,857581	1,000000	0,001379	0,002753	0,142447
	-0,6042	1448528	1	1999	8721777	0,857601	0,999500	0,001378	0,002752	0,142427
	-0,4277	1447861	4	1996	8722444	0,857666	0,998000	0,001377	0,002750	0,142362
	-0,2513	1447743	5	1995	8722562	0,857677	0,997500	0,001376	0,002748	0,142350
	-0,0748	1447743	5	1995	8722562	0,857677	0,997500	0,001376	0,002748	0,142350
	0,1017	7081	1867	133	10163224	0,999120	0,066500	0,018436	0,028869	0,000696
	0,2781	7081	1867	133	10163224	0,999120	0,066500	0,018436	0,028869	0,000696
	0,4546	6305	1877	123	10164000	0,999196	0,061500	0,019135	0,029188	0,000620
	0,631	2613	1933	67	10167692	0,999553	0,033500	0,025000	0,028632	0,000257
	0,8075	317	1982	18	10169988	0,999774	0,009000	0,053731	0,015418	0,000031