UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS BIOLÓGICAS (GENÉTICA)

André Luiz Molan

Construção de uma ferramenta para análise de enriquecimento funcional gênico multiespécie entre amostras comparativas

Botucatu, Junho de 2018

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS BIOLÓGICAS (GENÉTICA)

André Luiz Molan

Construção de uma ferramenta para análise de enriquecimento funcional gênico multiespécie entre amostras comparativas

Dissertação apresentada ao Instituto de Biociências, Campus de Botucatu, UNESP, em preenchimento dos requisitos para a obtenção do título de Mestre no Programa de Pós-Graduação em Ciências Biológicas (Genética).

Área de Concentração: Genética
Orientador: Prof. Dr. José Luiz Rybarczyk Filho

Botucatu, Junho de 2018.

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSANGELA APARECIDA LOBO-CRB 8/7500

Molan, André Luiz.
Construção de uma ferramenta para análise de enriquecimento funcional gênico multiespécie entre amostras comparativas / André Luiz Molan. - Botucatu, 2018

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: José Luiz Rybarczyk Filho
Capes: 20204000

1. Bioinformática. 2. Processamento eletrônico de dados. 3. Expressão gênica. 4. Transcrição genética.

Palavras-chave: atividade genica; diversidade genica; enriquecimento funcional gênico; ferramenta computacional; grupos de genes funcionalmente associados (GFAGs).

Agradecimentos

- Agradeço aos processos CNPq 458810/2013-4, 473789/2013-2 e 134469/2016-0 pelo suporte financeiro, fundamental para o desenvolvimento do trabalho;
- Agradeço ao meu orientador, Prof. Dr. José Luiz Rybarczyk Filho, pela orientação de elevada qualidade, apoio, paciência e incentivo;
- Agradeço à Prof[a]. Dr[a]. Agnes Alessandra Sekijima Takeda, fundamental para melhoria da qualidade do trabalho e apresentações orais;
- Agradeço aos amigos de laboratório, Giordano Bruno Sanches Seco, José Rafael Pilan e Carlos Alberto Oliveira de Biaggi Junior, os quais contribuíram enormemente com suas amizades e ideias para eventuais melhorias no trabalho;
- Agradeço aos meus pais, José Alberto Turini Molan e Maria Tereza de Oliveira Molan, pelo apoio e suporte incondicionais mesmo nos momentos mais difíceis.

Resumo

Sequenciar um organismo, atualmente, pode ser financeiramente custoso. A quantidade de dados gerada por uma única corrida é grande. Analisá-los não é trivial, exigindo, cada vez mais, técnicas computacionais e métodos estatísticos robustos, capazes de extrair o máximo possível de informações. Normalmente, alguns estudos visam à busca por genes diferencialmente expressos mediante diferentes condições experimentais. É importante conhecer perfis de expressão, porém, faz-se necessário entender como os genes relacionam-se entre si funcionalmente, o que só é possível por meio de uma análise de enriquecimento funcional.

Dessa forma, desenvolvemos um pacote em ambiente de programação R chamado *EntropyClusterGenes*. Ele é capaz de enriquecer conjuntos amostrais comparativos sob o ponto de vista da atividade e diversidade gênica utilizando a teoria da informação de *Shannon*. A ferramenta apresenta uma nova perspectiva de enriquecimento funcional, buscando por grupos de genes funcionalmente associados (GFAGs) através do conceito de entropia relacionado à ontologias e *KEGG pathways*. Para cada GFAG encontrado, através da técnica de *bootstrap*, calcula-se um p-valor, que é validado via FDR (*False Discovery Rate*) para determinar se o grupo encontrado é ou não significativo em uma dada comparação amostral (controle *versus* experimento). Através de uma nova análise de RNA-seq com o protocolo *Tuxedo*, quantificamos os transcritos de forma bruta e diferencial em 46 amostras de *Aedes aegypti* e 8 amostras de *Drosophila melanogaster*, reagrupadas, posteriormente, em 40 combinações (controle e experimento) para o enriquecimento funcional pela nova ferramenta. De acordo com cada combinação, encontramos diversos grupos significativos relacionados a processos biológico, funções moleculares, componentes celulares e *KEGG pathways*. Para validar a análise de enriquecimento, comparamos os resultados obtidos pelo *EntropyClusterGenes* a alguns dos principais resultados obtidos pelos pesquisadores nos estudos originais, além de realizarmos um *benchmarking* com outras três ferramentas similares, encontrando resultados semelhantes entre elas.

Abstract

An organism sequencing is still expensive. The amount of data generated by a single run is massive. Analyzing them is not trivial, requiring computational techniques and robust statistical methods capable of extracting as much information as possible. Usually, some studies aim to search for genes differentially expressed by different experimental conditions. It is important to know the expression profiles, however, it is necessary to understand how the genes are functionally related to each other, which is only possible through a functional enrichment analysis.

In this context, we have developed a package in the R programming environment called *EntropyClusterGenes*. It is able to enrich comparative sample sets from the perspective of gene activity and gene diversity using *Shannon's information theory*. The tool presents a new approach of functional enrichment, searching for groups of functionally associated genes (GFAGs) related to ontologies and KEGG pathways classifying them according to their entropy. For each GFAG found, by means of the bootstrap technique, a p-value is calculated, which is validated by FDR (False Discovery Rate) to determine if the group found is significant or not in a given sample comparison (control vs. experiment). Using a new analysis of RNA-seq with the Tuxedo protocol, we quantified the raw and differential transcripts in 46 samples of *Aedes aegypti* and 8 samples of *Drosophila melanogaster*, later regrouped in 40 combinations (control and experiment) for the enrichment with the new tool. According to each combination, we found several significant groups related to biological processes, molecular functions, cellular components and KEGG pathways. To validate the enrichment analysis, we compared the results obtained by *EntropyClusterGenes* to some of the main results obtained by the researchers in the original studies. In addition, we have run a benchmarking with three other similar tools, finding similar results between them.

Lista de Figuras

Lista de Tabelas

Sumário

1 *Introdução*

1.1 *Enriquecimento Funcional*

O avanço da biologia celular nas últimas décadas se deve, particularmente, às tecnologias de sequenciamento de alto rendimento. A quantidade de dados gerada a partir de experimentos de microarranjo e *Next Generation Sequencing* (NGS) tem aumentando substancialmente nos últimos anos. A análise e interpretação desses dados, porém, não é trivial, exigindo uma combinação de conhecimentos biológicos, modelagem estatística e técnicas computacionais. Os primeiros conjuntos de expressão gênica eram, em sua maioria, analisados considerando os genes de maneira individual. No entanto, constatou-se que estes não agem sozinhos. Processos celulares são o resultado de complexas interações entre diferentes genes e moléculas. Atualmente, grupos de genes ligados a diversas funções estão disponíveis em bases de dados públicas, podendo ser utilizados para interpretar resultados de novos experimentos (HUANG; SHERMAN; LEMPICKI, 2008).

Tais bases possibilitaram a implementação de uma série de métodos para a análise de enriquecimento gênico, com o propósito de comparar níveis de expressão sob duas condições distintas (experimento vs controle) e identificar grupos diferencialmente expressos (enriquecidos) na condição experimental (SIGNORELLI; VINCIOTTI; WIT, 2016). Hoje, o *Gene Ontology* (ASHBURNER et al., 2000) e o KEGG (OGATA et al., 1999) são duas das bases mais usadas em estudos funcionais. Contudo, estratégias baseadas em vocabulários múltiplos vêm sendo criadas, como o *Human Disease Ontology* e a *Pharmacogenomics Knowlege Base* (HOEHNDORF; DUMONTIER; GKOUTOS, 2012).

Uma série de ferramentas aplicadas à analise de enriquecimento tem sido desenvolvidas. Elas apresentam grande similaridade, uma vez que todas calculam p-valores das vias enriquecidas e utilizam diferentes métodos estatísticos, como o teste exato de *Fisher*, teste do χ^2, testes de distribuição binomial e hipergeométrica, entre outros. *Fisher* (FISHER, 1935) é mais apropriado para análise de vias que contêm um pequeno número de genes, ao passo que χ^2 é adequado quando a quantidade de genes é superior a cinco (ROSCOE; BYARS, 1971). Seme-

lhante ao teste de *Fisher*, a distribuição hipergeométrica (KEMP; KEMP, 1956) é utilizada para uma amostragem com número de genes reduzido, porém, se aproxima da distribuição binomial (mais adequada quando se considera um número elevado de genes) à medida que a quantidade de genes aumenta (BRUNK; HOLSTEIN; WILLIAMS, 1968). Além disso, é preciso ressaltar que as ferramentas de enriquecimento funcional são classificadas em 3 categorias, podendo, algumas, serem inclusas em mais de uma delas: *Singular Enrichment Analysis* (SEA), *Gene Set Enrichment Analysis* (GSEA) e *Modular Enrichment Analysis* (MEA) (MACHADO; FREITAS; COUTO, 2013). Embora estejam categorizadas, todas possuem uma estrutura em comum, conforme Figura 1.1.

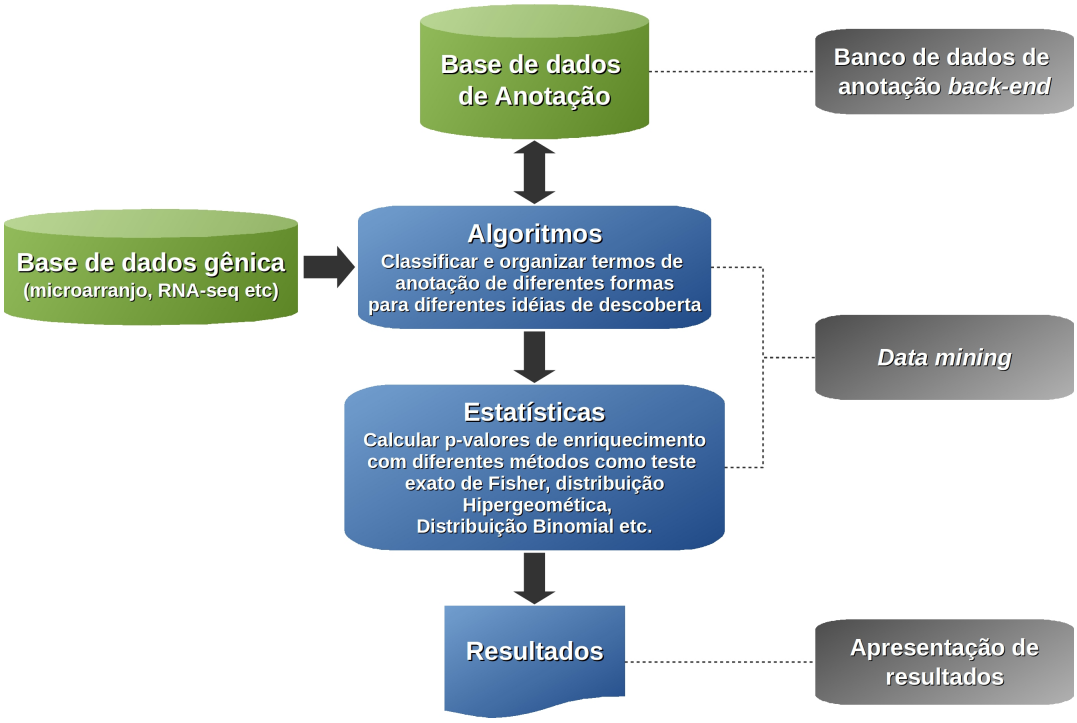

Figura 1.1: Estrutura de uma típica ferramenta de enriquecimento. Embora as ferramentas de análise de enriquecimento possuam características distintas, elas podem ser resumidas em três blocos principais: Banco de dados de anotação *back-end*, *data mining* e apresentação de resultados. [Figura adaptada de (HUANG; SHERMAN; LEMPICKI, 2008)]

1.1.1 *Singular Enrichment Analysis* (SEA)

Uma das estratégias de análise de enriquecimento mais tradicionais, SEA avalia uma relação de genes pré-selecionados, como, por exemplo, genes diferencialmente expressos, testando iterativamente cada termo de anotação de forma linear. Em seguida, os termos enriquecidos dentro do limite de p-valor estabelecido são transferidos para uma tabela, ordenados de acordo com seus respectivos p-valores de enriquecimento, calculados através de métodos estatísticos bem conhecidos (Chi-quadrado, *Fisher*, distribuição hipergeométrica, entre outros). Ferramentas

dessa classe são eficientes quanto à extração dos principais significados biológicos por trás de grandes listas de genes. Contudo, como ponto fraco desta classe, temos uma saída (*output*) de termos muito grande - de centenas a milhares - que podem dificultar a análise dos resultados gerados. Isso torna a análise via SEA instável, uma vez que diferentes *cutoffs* e métodos estatísticos são utilizados (HUANG; SHERMAN; LEMPICKI, 2008). *GOStat* (BEISSBARTH; SPEED, 2004) e *GoMiner* (ZEEBERG et al., 2003) são exemplos de ferramentas SEA.

1.1.2 *Gene Set Enrichment Analysis* (GSEA)

GSEA se assemelha à SEA, contudo, todos os genes do experimento são utilizados e não apenas um subconjunto - como genes diferencialmente expressos, por exemplo. Esta estratégia torna o enriquecimento enriquecimento funcional mais completo, uma vez que há a redução de fatores arbitrários envolvidos na seleção de grupos gênicos e toda a informação disponível é utilizada. Para obtenção dos p-valores, algumas ferramentas dessa classe calculam um *score* de enriquecimento máximo (MES) a partir do ranqueamento de todos os genes presentes em uma dada categoria, através do método estatístico *Kolmogorov-Smirnov*, (HOLLANDER; WOLFE, 1999). Outras se baseiam em métodos estatísticos paramétricos, como *z-score*, *teste t*, análise de permutação etc, levando em consideração valores experimentais - *fold change*, por exemplo - de todos os genes em cada termo de anotação encontrado (HUANG; SHERMAN; LEMPICKI, 2008).

As ferramentas dessa classe, entretanto, apresentam algumas limitações. O fato de não utilizar um valor de corte (*cutoff*) para seleção de genes, embora seja a maior vantagem de GSEA, tornou-se um grande problema em determinados estudos. O método requer como *input* um valor biológico - por exemplo, *fold change* - para cada um dos genes do experimento. Porém, de acordo com o estudo e a plataforma utilizadas, a obtenção de tais valores não é trivial. Além disso, genes com *ranking* elevado, geralmente com as maiores variações de expressão, são a força motora para a definição de p-valores, levando a assumir que estes genes são os maiores responsáveis pelos achados biológicos. Mas isso não é verdade, uma vez que pequenas alterações de sinais podem resultar em consequências mais significativas. (HUANG; SHERMAN; LEMPICKI, 2008). *GO-Mapper* (SMID; DORSSERS, 2004) e *PAGE* (KIM; VOLSKY, 2005) são exemplos de ferramentas computacionais que se encaixam no perfil definido pela classe GSEA.

1.1.3 *Modular Enrichment Analysis* (MEA)

MEA possui o mesmo cálculo de enriquecimento encontrado em SEA, acrescido, porém, de algoritmos de descoberta de redes que levam em conta as relações termo a termo. Ao considerar o relacionamento entre termos GO (ontologias referentes ao *Gene Ontology*), há um aumento na sensibilidade e especificidade do enriquecimento, sendo esta uma das maiores vantagens em se utilizar tal abordagem. Isso possibilita ao pesquisador, eventualmente, encontrar significados biológicos na ligação entre dois ou mais termos. Além disso, quando se utiliza anotações heterogêneas, os termos são altamente redundantes, apresentando fortes correlações referentes à diferentes aspectos para a mesma função biológica (HUANG; SHERMAN; LEMPICKI, 2008).

Um exemplo é a ferramenta DAVID (HUANG et al., 2007), capaz de organizar um grande conteúdo de anotações heterogêneas, como termos GO, domínios proteicos e vias metabólicas em classes gênicas. Tal organização faz uso de estatísticas tipo *Kappa*, aliada ao cálculo de p-valores, permitindo que a análise de enriquecimento varie de uma análise termo-cêntrica para uma análise biológica módulo-cêntrica. Como exemplos adicionais de ferramentas dessa classe, podemos citar *topGO* (ALEXA; RAHNENFUHRER, 2010) e *ClusterProfiler* (YU et al., 2012).

Vale ressaltar, entretanto, uma das principais limitações dessa classe. Termos "órfãos" (elementos sem uma forte relação com seus vizinhos) podem, eventualmente, ser desconsiderados pela análise. Portanto, é preciso ficar atento e examinar de forma cautelosa todo e qualquer elemento deixado de fora, pois é possível que exista significado biológico de valor elevado mesmo que o método tenha indicado o contrário (HUANG; SHERMAN; LEMPICKI, 2008).

1.2 *Big Data*

A quantidade de dados gerado em todo o mundo ao longo dos últimos anos é massiva. Crescendo de maneira exponencial, saber administrar e analisar tamanho volume armazenado digitalmente é essencial. É preciso transforma-los em informação, gerando, assim, conhecimento (MURDOCH; DETSKY, 2013). O termo *big data* se refere a conjuntos de dados cujo tamanho e complexidade demandam técnicas de análise que vão além dos métodos tradicionais existentes. Seu tamanho pode chegar à casa de petabytes (10^{15} bytes). Sua classificação é de acordo com o volume, velocidade, variedade, veracidade e valor. Tais aspectos representam a variedade dos tipos de dados, que vão de textos não estruturados a dados fisiológicos, de imagens e sequenciamento genômico. Analisar dados tão complexos requer ferramentas que se assemelham mais ao campo da informática do que ao de pesquisas clínicas, como, por exemplo, aprendizado de máquina (*machine learning*). Por meio de diferentes técnicas computacionais

é possível refinar determinas questões, gerar hipóteses e identificar grupos experimentais com maior potencial de resultados promissores (DOCHERTY; LONE, 2015).

1.3 *Meta Análise*

Dentre os significados para o prefixo *meta* destacam-se "mudança" e "reflexão crítica sobre". Dessa forma, ao tratarmos de *meta análise*, nos referimos a uma análise que transcende os resultados anteriores. É fundamental uma nova análise estatística dos dados ou resultados preexistentes. Não basta simplesmente uma análise qualitativa. Dependendo da natureza dos dados e dos objetivos do estudo, qualquer método de análise estatística, praticamente, poderá ser aplicado. Qualquer área pode fazer uso da meta análise, permitindo, eventualmente, a elucidação de problemas que, anteriormente, não tenham sido possíveis dada limitações práticas ou custos elevados. A Biologia é uma das ciências que mais se beneficiou com as técnicas de meta análise, muito em função de uma série de atividades práticas, custos e ainda implicações éticas que cercam a realização de experimentos com seres vivos (LUIZ, 2002).

Com as novas tecnologias de sequenciamento, a quantidade de dados genômicos gerados tem aumentado consideravelmente. Os diferentes designs experimentais utilizados, entretanto, geram um conteúdo que, na maioria das vezes, não é utilizado em sua totalidade. São estudos focados em determinados cenários que visam observar o comportamento de um grupo de funções biológicas de um dado organismo ou, simplesmente, a montagem do genoma deste. Dessa forma, é possível aplicar o conceito de *big data*, com o intuito de realizar novas análises e obter resultados diferentes daqueles encontrados na pesquisa inicial (MCAFEE et al., 2012).

Um exemplo de meta análise pode ser observado em um trabalho de DerSimoniane e Laird (DERSIMONIAN; LAIRD, 1986). Ao examinarem oito artigos referentes a diferentes ensaios clínicos específicos para uma condição médica específica, foram capazes de incorporar heterogeneidade dos efeitos na eficácia do tratamento. Isso permitiu que tratamentos mais específicos pudessem ser recomendados, confirmando que novas conclusões podem ser obtidas a partir de dados previamente analisados.

1.4 *Sequenciamento Next Generation Sequencing (NGS)*

Em 1944, Oswald Theodore Avery demostrou o DNA como sendo material genético. Sua estrutura em dupla hélice, composta por quatro bases, entretanto, só foi demostrada em 1953 por James D. Watson e Francis Crick, fato que levou ao dogma central da biologia molecular.

O DNA genômico, na maioria dos casos, define as espécies e indivíduos, tornando a sequência desta molécula fundamental para pesquisas sobre estruturas e funções celulares. As tecnologias de sequenciamento de DNA auxiliam em uma série de aplicações, como montagem de genomas, clonagem molecular, busca por genes patogênicos, estudos comparativos e evolucionários, entre outros. Nos últimos 30 anos, as tecnologias de sequenciamento passaram por inúmeros avanços e se tornaram a principal força na era dos genomas. Em 1977, Frederick Sanger desenvolveu uma tecnologia de sequenciamento de DNA baseada no método de *terminação de cadeias*, também conhecido como *sequenciamento de Sanger*. Devido a sua alta eficiência e baixa radioatividade, o sequenciamento de Sanger foi adotado como tecnologia primária para as aplicações de sequenciamento. Porém, utilizar este método era trabalhoso e requeria materiais radioativos. Foi então que, em 1987, uma empresa chamada *Applied Biosystems* introduziu o primeiro equipamento de sequenciamento automático, chamado de AB370. Ele adotava a eletroforese capilar, aspecto que tornava o sequenciamento mais rápido e acurado. Além disso, o AB370 era capaz de detectar 96 bases por vez - 500K diários - e *reads* com 600 bases de comprimento. Em 1998, instrumentos de sequenciamento automático com *software* associados que utilizavam máquinas de sequenciamento capilar e a tecnologia de sequenciamento Sanger se tornaram as principais ferramentas para a conclusão do projeto genoma humano em 2001. Posteriormente, surgiram as tecnlogias NGS (*Next Generation Sequencing*), diferenciando-se do método de Sanger pela análise paralela massiva, alto rendimento e custo reduzido. Embora NGS tenha tornado o sequenciamento de genomas mais acessível, as subsequentes análises dos dados e explicações biológicas ainda são o gargalo para a compreensão destes (LIU et al., 2012).

Seguindo a evolução tecnológica, em 2005 a empresa *454* lança seu sistema de sequenciamento - de mesmo nome. No ano seguinte a *Solexa* lança o *Genome Analyzer*, seguido pelo SOLiD (*Sequencing by Oligo Ligation Detection*), da *Agencourt*. Todos os três, comparados ao sequenciamento Sanger, compartilham bom desempenho quanto a rendimento, precisão e custos. Ainda em 2006 a *Agencourt* foi adquirida pela *Applied Biosystems* e, em 2007, a *Roche* comprou a *454*, enquanto a Solexa foi vendida para a *Illumina* (LIU et al., 2012).

1.4.1 *Sanger*

Em 1977, o bioquímico inglês Frederick Sanger foi responsável pelo maior avanço no que se refere ao sequenciamento de DNA, desenvolvendo a técnica de terminação de cadeias (*chain-termination*). Tal técnica faz uso de substâncias análogas aos desoxirribonucleotídeos (dNTPs) que são monômeros de fitas de DNA. Os dideoxirribonucleotídeos (ddNTPs) não possuem o grupo 3' hidroxil, necessário para a extensão das cadeias de DNA e, portanto, não são capa-

zes de formar uma ligação com o 5’ fosfato do próximo dNTP. Ao adicionar radiomarcadores ddNTPs na reação de extensão do DNA em uma fração da concentração de dNTPs padrão tem-se como resultado a produção de fitas de DNA de cada comprimento possível, com a incorporação randômica dos ddNTPs como extensão da fita e consequente interrupção do processo. Através da execução de quatro reações paralelas contendo cada uma das bases ddNTP e processando os resultados em quatro divisões de um gel de poliacrilamida, é possível utilizar autoradiografia para inferir a sequência de nucleotídeos presente no fragmento original, uma vez que existirá uma banda radioativa na correspondente divisão naquela posição do gel (HEATHER; CHAIN, 2016). Um modelo ilustrativo deste tipo de sequenciamento pode ser visto na Figura 1.2.

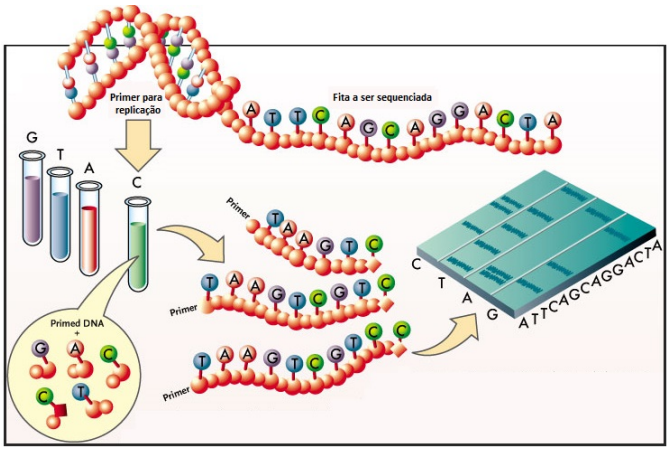

Figura 1.2: Sequenciamento Sanger. A região do DNA a ser sequenciada é amplificada e, então, desnaturada para produzir uma fita de simples, conectando-se um primer a ela. O sequenciamento se aproveita do fato de que o crescimento de uma cadeia de nucleotídeos na direção 3’ irá terminar se, ao invés de um desoxinucleotídeo convencional, um 2’3’ dideoxynucleotídeo é incorporado. Através de quatro reações separadas, cada uma contendo uma DNA polimerase e uma pequena quantidade de um dos quatro dideoxinucleotídeos além dos deoxinucleotídeos, quatro conjuntos separados de fragmentos de terminação de cadeia serão produzidos. Seguindo o passo de replicação/terminação, estes fragmentos permanecerão ligados à fita simples de DNA, a qual se comporta como um *template*. Ao aquecer essas moléculas fita dupla parcial e adicionar um agente desnaturador, como formamida, as moléculas fita simples de terminação em cadeia podem ser liberadas de seus respectivos templates e separadas por meio da utilização de eletroforese em gel desnaturador de alta resolução. A sequência da região original do DNA é, então, deduzida através da análise das posições relativas dos produtos das reações em quatro faixas do gel desnaturador. [Figura adaptada de (TECHCOUNCIL, 2013)]

1.4.2 *Illumina*

É a mais popular entre as plataformas de alto rendimento que utilizam o sequenciamento por síntese química. Após o preparo das bibliotecas, o cDNA (DNA complementar) de fita dupla é colocado sobre uma *flow cell* onde ocorre a hibridização baseada na complementariedade de bases com as sequências dos adaptadores. Em seguida, sequências em ambas as extremida-

des do adaptador são amplificadas como uma ponte. As sequências recém geradas, então, são hibridizadas próximas umas das outras e, depois de vários ciclos, uma região da *flow cell* irá conter várias cópias do cDNA original. Este processo é conhecido como geração de *clusters* e, com o surgimento destes, além da remoção de uma das fitas do cDNA, reagentes são introduzidos na *flow cell* dando início ao sequenciamento por síntese. Este tipo de sequenciamento ocorre através de uma reação na qual cada rodada de síntese, a adição de um nucleotídeo - A, C, G ou T, conforme determinado pelo sinal de fluorescência - é imageada de forma que a localização e o nucleotídeo adicionado possam ser determinados, armazenados e analisados. Vale ainda ressaltar que existem dois modos de sequenciamento. Caso o sequenciamento seja executado apenas em uma das extremidades da fita dupla de cDNA, o modo é chamado de *single-end*. Entretanto, se o sequenciamento for executado em ambas as extremidades da fita, o modo recebe o nome de *paired-end* (KORPELAINEN et al., 2014). Para melhor compreender o sequenciamento Illumina, observe a Figura 1.3.

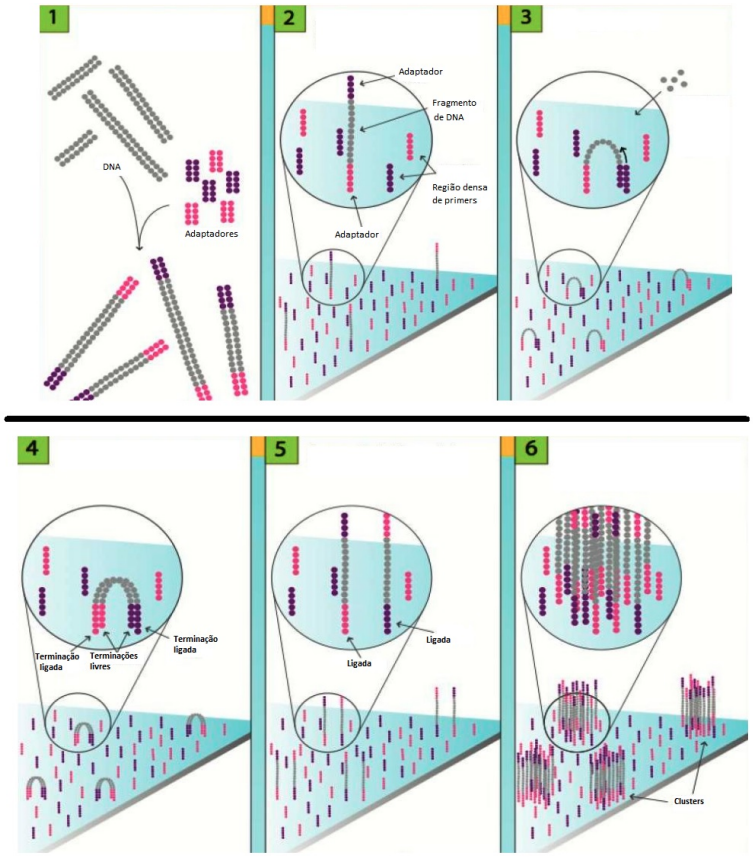

Figura 1.3: Sequenciamento Illumina. (1) Preparo da amostra de DNA genômico, com fragmentação randômica do DNA e adaptadores ligados às extremidades dos fragmentos. (2) Ligação do DNA à superfície da *flow cell*. Fragmentos de fita simples são aleatoriamente ligados à superfície de canais da *flow cell*. (3) Amplificação de ponte. Adição nucleotídeos e enzimas para iniciar a fase sólida de amplificação em ponte. (4) Fragmentos se tornam fitas duplas. (5) Desnaturação das fitas duplas. (6) Conclusão da amplificação, com a formação de densos clusters de fita dupla de DNA gerados em cada canal da *flow cell*. [Figura adaptada de (BIOLOGY NOTES HELP, 2017)]

1.4.3 *SOLiD*

SOLiD significa sequenciamento por detecção de ligações oligonucleotídeas e é uma plataforma comercializada pela *Applied Biosystems*. A química do sequenciamento, como o próprio nome diz, se dá via ligação ao invés de síntese. A biblioteca de fragmentos de DNA (derivada, originalmente, de moléculas de RNA) é associada à esferas magnéticas, sendo uma molécula por esfera. O DNA em cada esfera é,então, amplificado em uma emulsão de forma que o conteúdo amplificado permaneça na esfera. Os produtos resultantes da amplificação ligam-se de forma covalente a uma lâmina de vidro. Um exemplo gráfico deste tipo de sequenciamento pode ser observado através da Figura 1.4.

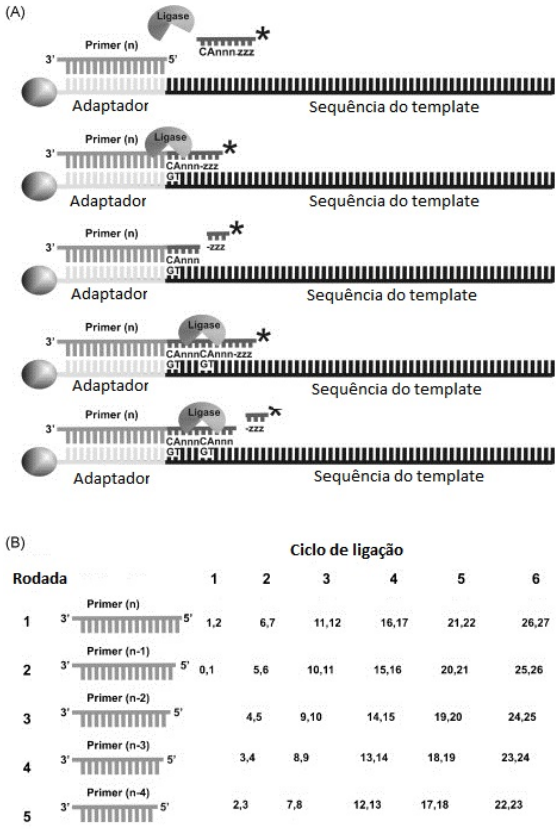

Figura 1.4: Sequenciamento SOLiD. (A) Primers hibridizam com o adaptador P1 dentro da biblioteca de template. Um conjunto de quatro sondas di-base marcadas com um agente fluorescente competem para ligaram-se à sequência do primer. Essas sondas possuem uma sequência de DNA parcialmente degenerada (indicada por *n* e *z*) e por simplicidade apenas uma sonda é mostrada (a marcação é denotada por asterisco). A especificidade de sonda di-base é obtida através da checagem da primeira e segunda base em cada reação de ligação. Seguindo a ligação, a marcação fluorescente é enzimaticamente removida juntamente com as três últimas bases do octâmero. (B) A determinação da sequência pelo sequenciamento SOLiD é feita em múltiplos ciclos de ligação, utilizando diferentes primers, cada um menor do que o anterior por uma única base. O número de ciclos de ligação (neste exemplo, 6) determina o eventual tamanho do *read*, embora para cada sequência de marcação, ocorrem seis rodadas de reinicialização do primer (do primer *n* ao primer *n-4*). [Figura adaptada de (ANSORGE, 2009)]

Através de diversos *primers* que se hibridizam com um *primer* universal, sondas di-base fluorescentes são competitivamente ligadas ao *primer*. Caso as bases na primeira e segunda posição da sonda di-base sejam complementares à sequência, a reação de ligação ocorrerá e irá gerar um sinal. *Primers* são reiniciados por um nucleotídeo simples cinco vezes, de forma que, ao final do ciclo, pelo menos quatro nucleotídeos tenham sido checados duas vezes, devido às sondas dinucleotídicas, e um nucleotídeo pelo menos uma vez. A ligação das sondas diclueotídicas subsequentes possibilitam uma segunda checagem do nucleotídeo checado anteriormente apenas uma vez e, após mais cinco ciclos, outros cinco nucleotídeos serão checados pelo menos duas vezes. Os passos de ligação continuam até que a sequência completa tenha sido lida (KORPELAINEN et al., 2014).

1.4.4 *Roche 454*

Roche 454 é uma plataforma baseada no sequenciamento de bibliotecas de DNA dupla fita com a ligação de adaptadores através de síntese química. O DNA é fixado em esferas e amplificado em uma emulsão de água e óleo. As esferas são colocadas em placas de picotituladores onde ocorrem as reações de sequenciamento. O elevado número de poços nas placas de picotituladores proporcionam o *layout* paralelo massivo utilizado, característico das tecnologias NGS. Comparado a outras plataformas, o método de detecção difere quanto à síntese química, a qual apresenta a detecção de um nucleotídeo adicionado por meio de uma reação em duas etapas. A primeira realiza a clivagem do nucleotídeo trifosfato após uma adição, liberando pirofosfato. A segunda converte o pirofosfato em adenosina trifosfato (ATP) através da enzima ATP sulfurilase. Em uma terceira etapa, o recém sitetizado ATP é utilizado para catalizar a conversão de luciferin em oxiluciferin via luciferase, gerando uma quantidade de luminescência que é capturada por uma câmera acoplada à placa picotituladora. Os nucleotídeos livres e ATPs que não reagiram são degrados por uma pirase após cada adição. Todas as etapas são repetidas até que um número predeterminado de reações, de acordo com o fabricante, tenham sido alcançadas. A gravação da luminescência gerada, bem como a localização do poço após a cada adição de nucleotídeo torna possível a reconstrução da identidade do nucleotídeo e da sequência em cada poço. Este método, visto na Figura 1.5, é conhecido como **pirosequenciamento**, apresentando, como principal vantagem, a possibilidade de *reads* mais longos quando comparado às outras plataformas - *reads* de até 1000 bases podem ser alcançados (KORPELAINEN et al., 2014)

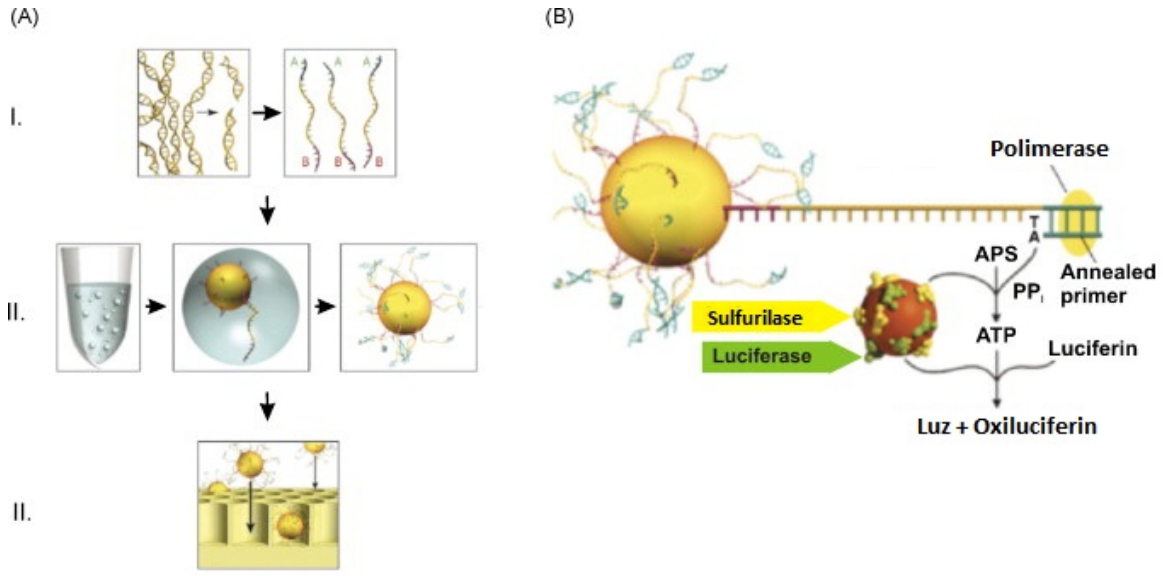

Figura 1.5: Sequenciamento Roche 454. (A) Esboço do *workflow* do sequenciador de DNA GS 454. (I) Construção da biblioteca com a ligação de adaptadores 454 específicos aos fragmentos de DNA, indicados por *A* e *B* e (II) acoplamento de DNA e esferas amplificadas em uma emulsão PCR para amplificar fragmentos antes do sequenciamento. (III) As esferas são carregados em uma placa picotituladora. (B) Ilustração esquemática da reação de pirosequenciamento na qual ocorre a incorporação para reportar o sequenciamento por síntese. [Figura adaptada de (ANSORGE, 2009)]

1.4.5 *Ion Torrent*

Esta plataforma utiliza uma biblioteca de DNA dupla fita ligada a adaptadores seguida pelo sequenciamento de síntese química utilizado por outras plataformas, mas possui uma característica única. Ao invés de detectar sinais de fluorescência ou fótons, são detectadas alterações no pH da solução em um poço quando o nucleotídeo é adicionado e prótons são produzidos. Tais mudanças são extemamente pequenas, no entanto, a plataforma faz uso de tecnologias baseadas em semicondutores a fim de alcançar alta sensibilidade. O *Ion Torrent* produz em média uma quantidade menor de reads em uma única corrida quando comparada a outras plataformas. O tempo de corrida, porém, é muito baixo, levando de 2 a 4 horas. Uma vez que não é preciso instrumentação de medidas ópticas ou nucleotídeos modificados, sua grande vantagem é a acessibilidade, tanto instrumental quanto de reagentes. O equipamento é pequeno, pode ser desligado quando não estiver em uso (sendo fácil liga-lo novamente) e demanda pouca manutenção. Diante dessas características, foram encontradas consideráveis utilizações para a plataforma, como sequenciamento de microrganismos e genômica do meio ambiente, além de aplicações na área clínica, em que o tempo é um fator crítico (KORPELAINEN et al., 2014). Um esquema desta plataforma pode ser observado na Figura 1.6.

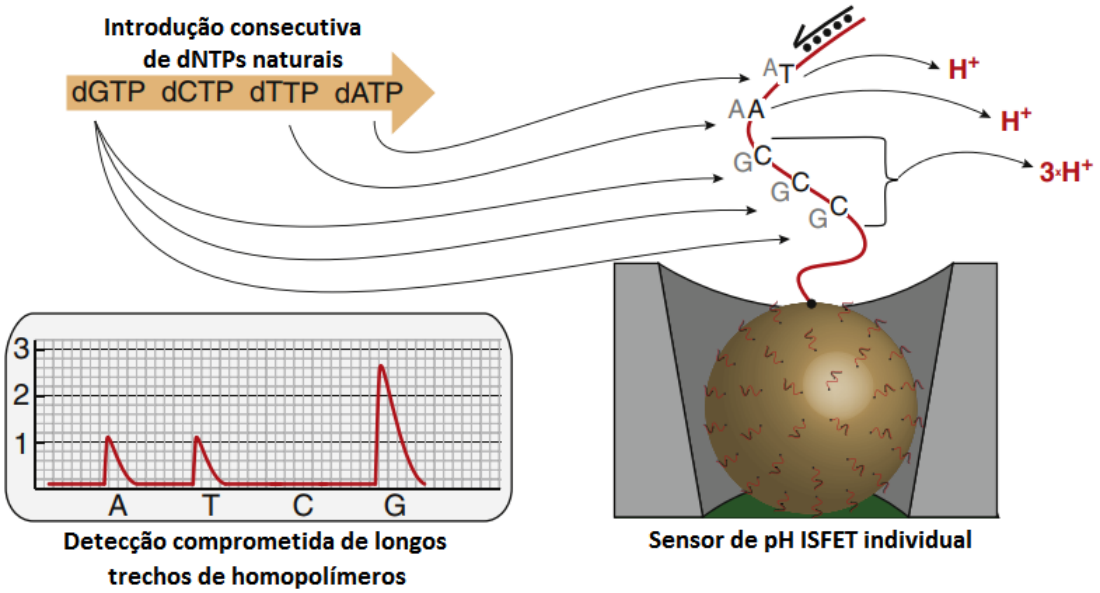

Figura 1.6: Sequenciamento Ion Torrent. Utilização de pH para a detecção da incorporação do nucleotídeo. Durante a extensão da polimerase e incorporação do nucleotídeo, um próton é liberado, alterando o pH de um micropoço. Ao fluir em um tipo de dNTP não modificado por vez (por exemplo, dATP), os sensores de pH ISFET miniaturizados reportam os micropoços que incorporaram o nucleotídeo introduzido. Sequências homopoliméricas (AAAAA, por exemplo) levam à liveração de um grande número de prótons que podem dificultar uma quantificação precisa. [Figura adaptada de (KHODAKOV; WANG; ZHANG, 2016)]

1.4.6 *Pacific Biosciences*

Trata-se de uma plataforma representante da terceira geração. A química utilizada ainda é similar às tecnologias de segunda geração, utilizando o sistema de sequenciamento por síntese. Entretanto, a maior diferença é que ela requer apenas uma única molécula, com os nucleotídeos adicionados sendo lidos em tempo real. Dessa forma, sua química recebeu o nome de **SMRT** (*Single-Molecule Real Time*). O fato de ser molécula única significa que não há a necessidade de amplificação. Nota-se que essa plataforma sequencia moléculas de DNA. SMRT utiliza *zero-mode waveguides* (ZMWs), câmaras de espaço restrito que guia a energia luminosa e reagentes para dentro de volumes extremamente pequenos que estão na ordem de zeptolitros (10^{-21} L). Fazendo uso de trifosfatos nucleotídicos fluorescentes, a adição de um A, C, G ou T a uma cadeia de nucleotídica pode ser detectada a medida que está sendo sintetizada. Como um instrumento de tempo real que mede adições a medida em que estas ocorrem, o tempo de execução pode ser bastante curto - entre uma e duas horas. O tamanho médio dos *reads* podem chegar a 5000 bases. Como consequência do sequenciamento direto de DNA de moléculas únicas, notou-se que modificações de ácidos nucleicos, tais como 5-metil citosina, ocasionaram atrasos na cinética da DNA polimerase. Tal acontecimento, porém, passou a ser explorado pela plata-

forma de forma que modificações no DNA pudessem ser sequenciadas (KORPELAINEN et al., 2014). Através da Figura 1.7.

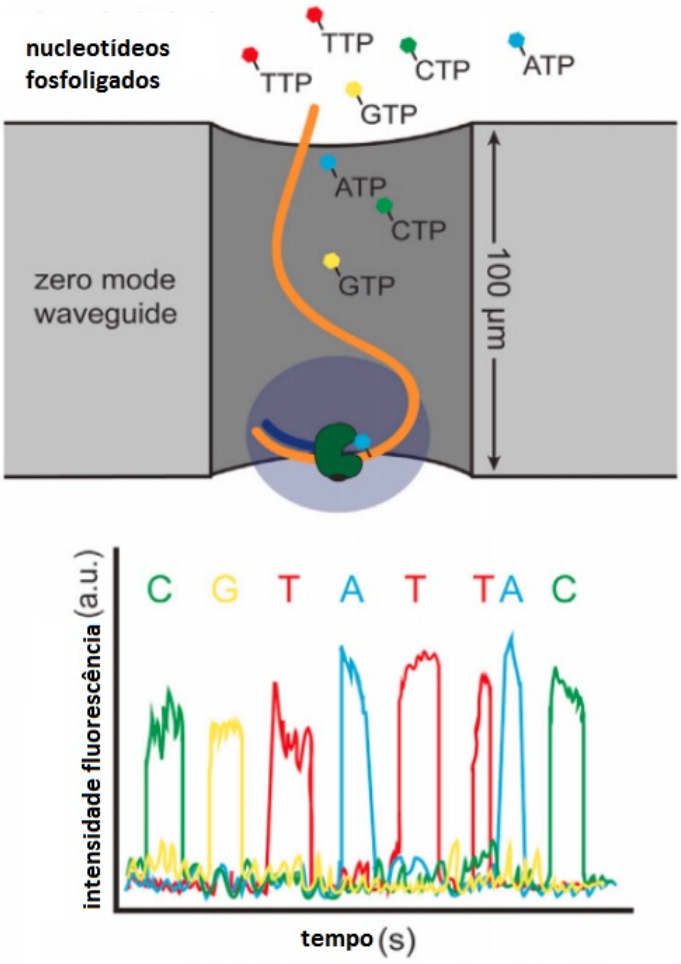

Figura 1.7: Sequenciamento *Pacific Biosciences*. Uma única polimerase é posicionada é na parte inferior de um ZMW. Versões de fosfatos marcados de todos os quatro nucleotídeos estão presentes, permitindo uma polimerização contínua de um template de DNA. A incorporação de bases aumenta o tempo de permanência de um nucleotídeo no ZMW, resultando em um sinal fluorescente detectável que é capturado em em um vídeo. [Figura adaptada de (BIOCHEMISTRIES, 2015)]

1.4.7 *Nanopore Technologies*

Apesar dos ganhos expressivos em rendimento e baixo custo por base das atuais plataformas, as tecnologias de sequenciamento continuam a evoluir. Criada pela *Oxford Nanopore Technologies*, *Nanopore*, descrita pela Figura 1.8, é uma tecnologia de sequenciamento de molécula única em que uma mesma enzima é utilizada tanto para separar a fita de DNA quanto para guia-la através de um poro proteico embutido em uma membrana. Ions atravessam o poro simultaneamente gerando uma corrente elétrica. Esta é sensível a nucleotídeos específicos que passam pelo poro, produzindo sinais distintos em função das bases A, C, G e T. A vantagem

desse sistema é a sua simplicidade, permitindo a construção de dispositivos pequenos. No entanto, ele é extremamente desafiador, uma vez que há a necessidade de medir mudanças na corrente elétrica em uma escala extremamente pequena (KORPELAINEN et al., 2014).

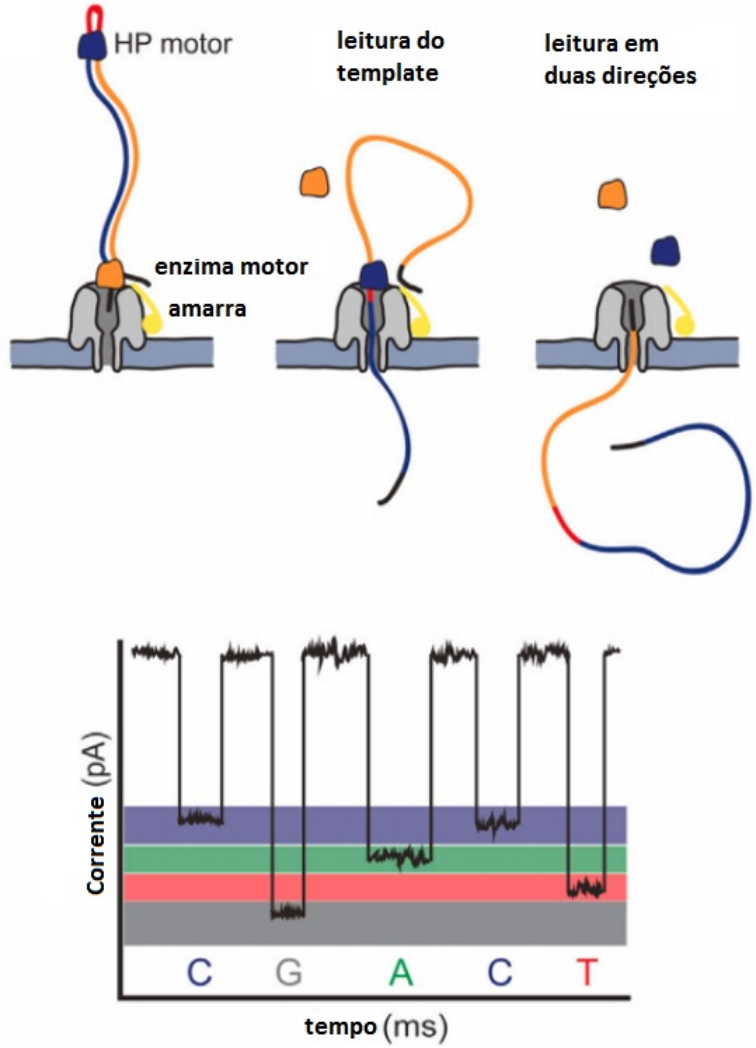

Figura 1.8: Sequenciamento Nanopore. Templates de DNA são ligados com dois adaptadores. O primeiro adaptador é ligado com uma enzima motora, assim como a amarra, enquanto o segundo é um oligo tipo grampo ligado por uma proteína motora HP. Alterações na corrente, induzidas a medida que os nucleotídeos passam pelo poro, são utilizadas para discriminar as bases. O desenho da biblioteca possibilita sequenciar ambas as fitas de DNA de uma única molécula (reads em duas direções). [Figura adaptada de (BIOCHEMISTRIES, 2015)]

1.5 *RNA-seq*

O transcriptoma é o conjunto completo de transcritos em uma célula - e a sua quantidade - em um determinado estágio de desenvolvimento. Sua compreensão é essencial para a interpretação de elementos funcionais do genoma, revelando constituintes moleculares de células e tecidos e possibilitando a compreensão do desenvolvimento de diferentes organismos

e doenças. Os principais objetivos de um transcriptoma, no entanto, consistem em catalogar todos os transcritos de uma dada espécie (mRNAs, non-coding RNAs e pequenos RNAs), determinar a estrutura transcricional de genes e quantificar alterações nos níveis de expressão durante as diversas fases de desenvolvimento do organismo e sob diferentes condições. Diversas tecnologias tem sido desenvolvidas para deduzir e quantificar o transcriptoma, incluindo abordagens baseadas em hibridização ou baseadas em sequências (WANG; GERSTEIN; SNYDER, 2009).

RNA-seq é uma coleção de métodos computacionais e experimentais que determinam a identidade e abundância de sequências de RNA em amostras biológicas. A posição de cada base nitrogenada, presente em uma fita simples de RNA, é identificada. Os métodos experimentais envolvem isolar o RNA de amostras celulares, tecidos ou animais inteiros, preparar bibliotecas que representem a espécie nas amostras utilizadas e efetuar um sequenciamento químico dessas bibliotecas, seguido de uma análise de bioinformática. Uma distinção entre RNA-seq e os métodos anteriores, está no alto rendimento das atuais plataformas, sensibilidade alcançada com as novas tecnologias e poder para descobrir novos transcritos, modelos gênicos e pequenas espécies de RNA não codificante. Os dados obtidos de um experimento de RNA-seq podem, de fato, produzir novos conhecimentos, que vão da identificação de diferentes transcritos codificadores de proteínas em linhagens de células embrionárias à caracterização de genes diferencialmente expressos em tumores de pele ou um outro órgão específico. Com base nisso, questões como diferenças nos níveis de expressão entre tecidos doentes e sadios ou após um tratamento com determinado agente mutagênico, quais genes são positivamente (*up*) ou negativamente (*down*) regulados durante o desenvolvimento do cérebro, quais transcritos estão presentes na pele e não nos músculos, entre outras, podem ser respondidas. Gerou-se grandes expectativas para a transcriptômica, quando o RNA-seq revelou que o conhecimento da estrutura dos genes e a forma como se realizava suas anotações - tanto de organismos unicelulares quanto de humanos - era muito pobre e limitada. Novos dados gerados a partir dessa técnica mostraram uma vasta diversidade para a estrutura gênica, possibilitaram a identificação de genes antes desconhecidos e jogaram luz sobre o estudo de pequenos e longos transcritos não codificantes (KORPELAINEN et al., 2014).

Fazendo uso das mais recentes tecnologias de sequenciamento, o método, em geral, utiliza uma população de RNA (total ou fracionada, como, por exemplo, *poly(A)+*) convertendo-a em uma biblioteca de fragmentos de cDNA com adaptadores conectados a uma ou ambas as extremidades da fita. Cada molécula, com ou sem amplificação, é, então, sequenciada por meio das novas tecnologias de sequenciamento de forma a obter pequenas sequências a partir de uma extremidade (*single-end sequencing*) ou ambas as extremidades (*pair-end sequencing*). O

tamanho dos *reads* obtidos dependem da tecnologia escolhida (WANG; GERSTEIN; SNYDER, 2009). Uma descrição gráfica da metodologia pode ser observada na Figura 1.9.

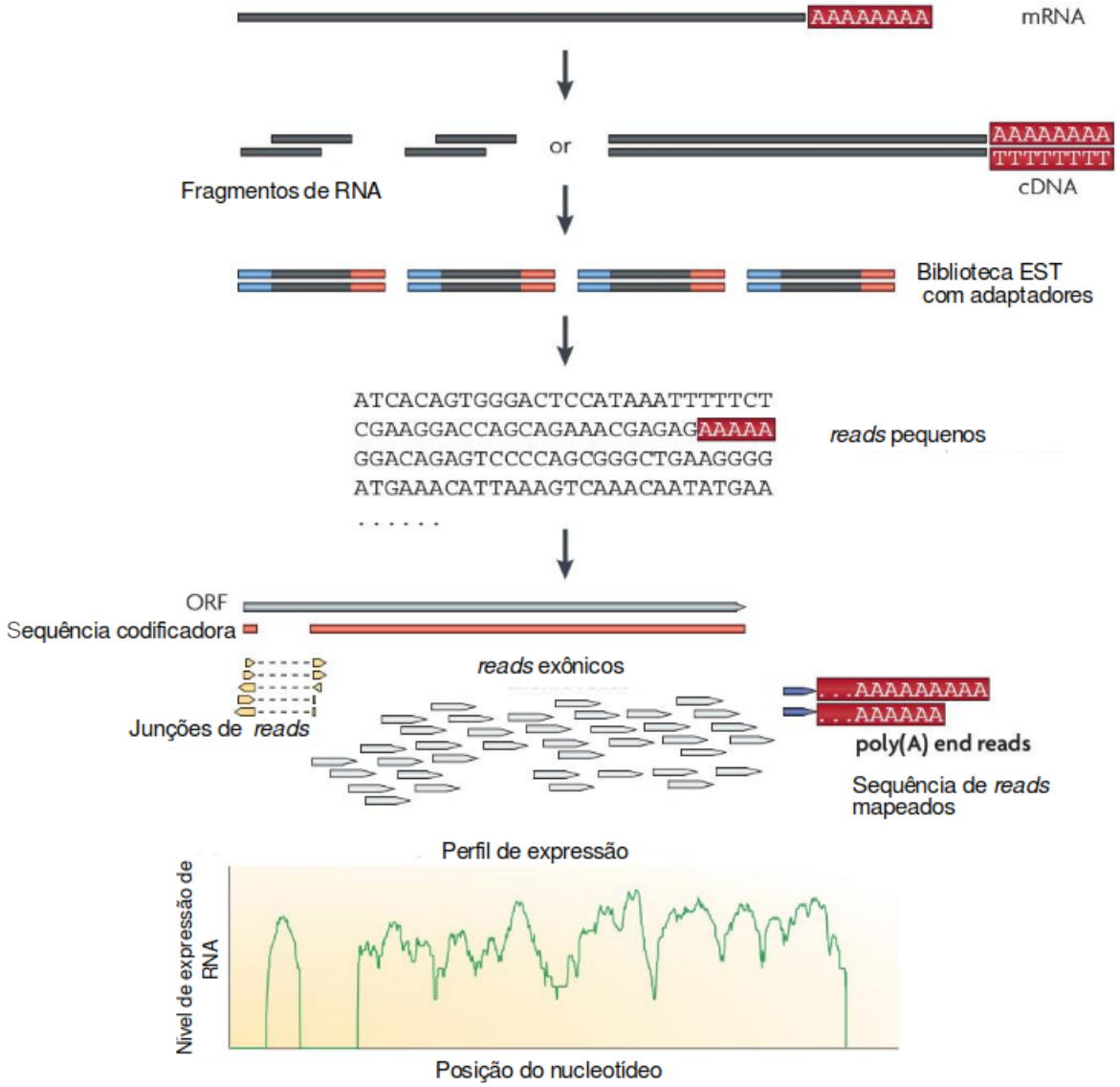

Figura 1.9: Experimento típico de RNA-seq. Inicialmente, longos RNAs são convertidos em uma biblioteca de fragmentos de cDNA através da fragmentação de RNA ou DNA. Em seguida, adaptadores de sequenciamento são adicionados a cada fragmento de cDNA e sequências pequenas são obtidas de cada cDNA utilizando uma dada tecnologia de sequenciamento de alto desempenho. Os *reads* obtidos são alinhados com um genoma ou transcriptoma de referência, e classificados em três tipos: exônicos, junção e *poly(A) end reads*. Esses três tipos são utilizados para gerar um perfil de expressão de resolução de bases para cada gene, como ilustrado no gráfico final. [Adaptada de (WANG; GERSTEIN; SNYDER, 2009)]

RNA-seq oferece uma série de vantagens em relação a outros métodos. Ao contrário das abordagens baseadas em hibridização, RNA-seq não está limitado a detectar transcritos correspondentes a uma sequência genômica existente, tornando-o extremamente atrativo para organismos não modelo, com sequências ainda a serem determinadas. Além disso, esta técnica

não possui um limite para quantificação, aspecto esse relacionado com as sequências obtidas. Consequentemente, ele possui um intervalo (*range*) de níveis de expressão através do qual os transcritos podem ser detectados. Seus resultados, ainda, mostram altos níveis de reprodutibilidade, tanto para replicatas técnicas como biológicas (WANG; GERSTEIN; SNYDER, 2009).

1.6 *Justificativa*

Um modo alternativo de análise de enriquecimento consiste na aplicação da *Teoria da Informação de Shannon* (SHANNON, 1948). A construção de comunidades celulares a partir de diferentes fenótipos com um genoma em comum é dependente de subconjuntos específicos de informações herdadas e da habilidade das células em receber e processar dados do meio que as cercam. A informação ativa em uma célula é uma soma tempo-dependente de sinais intracelulares traduzidos e/ou obtidos do meio extracelular, permitindo o controle da morfologia e demais funções das células, além de sua interação com o meio externo (GATENBY; FRIEDEN, 2002).

Em uma linguagem mais simples, informação significa conhecimento e denota uma quantidade mensurável. Tal definição, por sua vez, implica em duas propriedades adicionais, tratando-se de uma quantidade aditiva que, na ausência de ruídos, é conservada. Em genética, ruídos podem resultar de mutações aleatórias ou variações epigenéticas, provocando uma instabilidade genômica que ocasiona a redução da quantidade de informação transmitida. Vale ainda ressaltar que quanto maior a redundância da mensagem, menor é a quantidade de informação transportada. Redundância é uma consequência da estrutura sintática dentro do código, estando diretamente ligada à eficiência e precisão das mensagens transmitidas (KENDAL, 1990).

Gatenby e *Frieden* (GATENBY; FRIEDEN, 2002) aplicaram a teoria de *Shannon* para o estudo de carcinogênese, demonstrando as interações entre eventos de mutação estocásticos e pressões de seleção ambiental, responsáveis por características de fenótipos malignos. Em 2005, Castro e colaboradores (CASTRO et al., 2005) aplicaram o conceito de *Shannon* para uma análise citogenética de 14 tipos de tumores epiteliais, avaliando suas respectivas diversidades cariotípicas. Em um outro trabalho de Castro e colaboradores (CASTRO et al., 2007), a mesma teoria foi aplicada com o propósito de estudar a expressão de genes ligados à 10 vias metabólicas humanas distintas, dentre elas, o mecanismo de reparo por excisão de nucleotídeos (NER). Para tanto, foram utilizas bibliotecas SAGE (*Serial Analysis of Gene Expression*) de diferentes tecidos (sadios e com câncer) de diferentes órgãos. Através da *função de entropia normalizada de Shannon*, mediu-se a diversidade de cada uma das 10 vias presentes nas bibliotecas SAGE.

Em 2009, Castro e Rybarczyk-Filho desenvolveram um *software* chamado *ViaComplex* (CASTRO et al., 2009), implementando computacionalmente a teoria da informação de *Shannon* com o propósito de construir mapas de expressão gênica associados à redes de interação proteína-proteína. Isso permitiu, simultaneamente, verificar como os genes de um dado conjunto relacionavam-se entre si e com os seus vizinhos mais próximos e como cada gene estava expresso. Além disso, o *software* ainda disponibilizava um módulo estatístico que tornava possível a análise comparativa de dois conjuntos amostrais (controle *versus* experimento) em termos de grupos de genes funcionalmente associados (GFAGs). Os genes do conjunto eram reagrupados em subconjuntos com base em suas ontologias e vias metabólicas. Por meio do cálculo de entropia e atividade gênica, avaliava-se cada subconjunto quanto à variação da expressão entre os elementos e o quanto estes estavam expressos. Aplicava-se, então, o método de *bootstrap* para a determinação de p-valores, a fim de identificar quais funções eram significativamente expressas dentro do espaço amostral em questão. Isso possibilitava traçar um panorama funcional do experimento, encontrando, eventualmente, grupos gênicos com elementos passíveis de um estudo individual mais detalhado.

O *ViaComplex*, porém, quanto ao módulo estatístico, possui algumas limitações. Escrito na linguagem de programação *FORTRAN*, embora apresente uma interface amigável para o usuário, o preparo das entradas a serem processadas requerem conhecimentos adicionais por parte de quem o utiliza. É preciso, antes, criar uma tabela que contenha a relação entre funções (ontologias e termos KEGG, por exemplo) e genes da espécie de interesse. Tal tarefa, entretanto, nem sempre é trivial, exigindo que o usuário o acesse repositórios de dados *online* e elabore eventuais *scripts* ou programas. Para cada uma das comparações controle *versus* experimento, entretanto, todas as etapas mencionadas precisam ser repetidas, tornando-se inviável para estudos de grande porte. É preciso, ainda, salientar que a sua criação teve como base a utilização de dados do SAGE e *microarray*, ficando defasado quando leva-se em conta as diferentes técnicas de sequenciamento existentes. Dessa forma, aproveitando o conceito de análise de enriquecimento presente neste *software*, podemos desenvolver uma versão mais eficiente, otimizada e atualizada em ambiente R, capaz de processar maiores volumes de dados, a partir de diferentes métodos de sequenciamento aplicados a diferentes espécies, com um grau de automação elevado.

2 *Objetivos*

O presente trabalho tem como objetivo o desenvolvimento de um novo pacote em ambiente de programação R, capaz de efetuar o enriquecimento funcional de grandes quantidades de dados obtidas através de diferentes *designs* experimentais, como microarranjo e RNA-seq, de maneira otimizada e com elevado grau de automação. Trata-se de uma aplicação multiespécie que possibilita a busca por grupos de genes funcionalmente associados (GFAGs) significativos, relacionando ontologias (*Gene Ontology*) e KEGG *pathways* sob a perspectiva de entropia (diversidade) e valor de expressão gênica absoluto (atividade), utilizando, para tanto, a teoria da informação de *Shannon*.

2.1 Objetivos Específicos

- Condução de uma nova análise de RNA-seq em dados de *Aedes aegypti* e *Drosophila melanogaster* utilizando o protocolo *Tuxedo*;
- Desenvolvimento do *pipeline* em R;
- Encontrar grupos de genes funcionalmente associados (GFAGs) significativos a partir dos resultados da análise de RNA-seq;
- Realização de um *benchmarking* comparando a nova ferramenta com aplicativos similares disponíveis.

3 *Material e Métodos*

3.1 *Sequence Read Archive* (SRA)

As plataformas de sequenciamento NGS tem revolucionado a Biologia, permitindo o estudo de diferentes espécies e produzindo uma quantidade de dados muito superior à tecnologias como o microarranjo. Em 2009, visando ao armazenamento de tais dados, criou-se o SRA (`http://www.ncbi.nlm.nih.gov/sra`) (KODAMA; SHUMWAY; LEINONEN, 2012), um repósitório online público criado sob a tutela do *International Nucleotide Sequence Database Collaboration* (INSDC) e operado pelo *National Center for Biotechnology Information* (NCBI), *European Bioinformatics Institute* (EBI) e *DNA Data Bank of Japan* (DDBJ). Em meados de setembro de 2010, o volume de dados armazenado ultrapassava 500 bilhões de *reads* e 60 trilhões de pares de base, sendo a maior parcela derivada de sequenciamentos utilizando a plataforma Illumina e abrangendo, em sua grande maioria, organismos como *Homos sapiens* e *Mus musculus* (LEINONEN; SUGAWARA; SHUMWAY, 2010). Os dados brutos de uma vasta quantidade de sequenciamentos NGS são disponibilizados para reutilização pelo SRA, juntamente com um conjunto de ferramentas que facilitam o download e manipulação de seus arquivos.

3.2 *Conjuntos de dados brutos*

Para demonstrar o *EntropyClusterGenes*, ferramenta de enriquecimento funcional que desenvolvemos e é apresentada com mais detalhes nos próximos tópicos, comparada, na forma de um *Benchmarking* à ferramentas de enriquecimento existentes, também descritas com detalhes mais adiante, optamos por conjuntos de amostras de RNA-seq de *Aedes aegypti* e *Drosophila melanogaster* obtidas a partir de dois diferentes artigos:

- *Aedes aegypti*:
  - Projeto **PRJNA209388** (AKBARI et al., 2013): Download via SRA. São 46 corridas, com códigos compreendendo os intervalos **SRR923822**-**SRR923857**, **SRR923736**,

SRR923701-**SRR923705** e **SRR924021**-**SRR924024**. As bibliotecas foram preparadas utilizando-se ovo, larva, pupa e mosquito adulto de *Aedes aegypti*, tanto machos como fêmeas, derivados de uma cepa originária do oeste da África. No caso de fêmeas adultas, foram utilizados carcaças e ovário. Quanto aos machos, apenas carcaça e pupa. Houve variação nos tempos de cada um dos estágios de desenvolvimento dos indivíduos, com um mínimo de 2 horas e um máximo de 72 horas. Utilizou-se RNA-seq, sendo o sequenciamento dividido em *paired-end* e *single-end*. A plataforma de sequenciamento foi *Illumina*.

- *Drosophila melanogaster*:
  - Projeto **PRJNA418283** (MARXREITER; THUMMEL, 2018): Download via SRA. São 8 corridas, com códigos compreendendo o intervalo **SRR6288269** - **SRR6288276**. As bibliotecas foram preparadas utilizando o corpo inteiro de fêmeas (4 normais e 4 mutantes) de uma semana de idade. Os indivíduos mutantes apresentavam mutação no receptor nuclear DHR78. Utilizou-se a técnica de RNA-seq, implementada via *Illumina HiSeq 50 Cycle Single Read Sequencing v3*.

Realizamos o download dos arquivos brutos, no formato FASTQ, disponíveis no repositório online SRA, de forma que pudéssemos conduzir uma nova análise de RNA-seq, com um protocolo padrão e diferente daqueles utilizados nos estudos originais. Isso proporcionou a obtenção de listas de expressão gênica, bem como relação de genes diferencialmente expressos. Foi só então a partir de tais listas que pudemos fazer o enriquecimento funcional. Cada uma das amostras e combinações são descritas através da Tabela A.1 e Tabela A.2, conforme Apêndice A.

3.3 *Unidades para Quantificação de Dados de Expressão*

Em diferentes áreas da Biologia, medir a abundância de RNA em determinados organismos é de extrema importância e serve como base para uma enorme quantidade de estudos. Sendo obtidas, geralmente, a partir de tecnologias de sequenciamento de alto desempenho, como Illumina, tais medidas precisam ser normalizadas. Isso permite a remoção de dados incorretos ou com algum tipo de tendência, notadamente as que se referem ao comprimento de espécies de RNA e profundidade de sequenciamento de uma amostra (WAGNER; KIN; LYNCH, 2012). Visando à correção dessas inconsistências, diferentes formas de quantificação surgiram, como RPKM (*Reads per Kilobase per Million Reads*), FPKM (*Fragments per Kilobase per Million*) e TPM (*Transcripts per Million*).

Todas as três métricas tentam normalizar pela profundidade do sequenciamento e comprimento do gene. Desenvolvido especialmente para sequenciamentos do tipo RNA-seq *single-end*, RPKM conta o total de *reads* - equivalente a cada fragmento sequenciado - em uma amostra e o divide por 1 milhão. Isso normaliza pela profundidade do sequenciamento, dando origem ao "per million" (RPM). Em seguida, divide o RPM pelo comprimento do gene (em quilobases), dando origem ao RPKM. O FPKM foi criado para sequenciamentos que utilizam RNA-seq *paired-end*, em que dois *reads* podem corresponder a um único fragmento ou, caso um *read* no par não tenha sido mapeado, 1 *read* pode corresponder a um único fragmento. A diferença entre RPKM e FPKM está simplesmente no processo de contagem. Dois *reads* mapeiam somente um fragmento e este fragmento não é contado duas vezes. TPM, por sua vez, diferencia-se dos métodos anteriores pela ordem das operações. Primeiro ocorre a divisão da contagem de *reads* pelo comprimento de cada gene em quilobases, dando origem ao RPK. Logo após, há a contagem de todos os RPKs na amostra, dividindo-o por 1 milhão e resultando em um fator "por milhão". Para concluir, cada RPK é dividido por esse fator, o que nos dá o TPM (RNA-SEQ BLOG, 2015).

Ao usar TPM, a soma de todos os TPMs em cada amostra é a mesma, tornando mais fácil a comparação de proporções de *reads* que mapeiam um gene em cada amostra. Por outro lado, com RPKM e FPKM a soma de *reads* normalizados nas amostras pode ser diferente, tornando mais complexa a comparação direta entre amostras (RNA-SEQ BLOG, 2015).

3.4 *Protocolo Tuxedo*

O sequenciamento de alto rendimento de mRNA (RNA-seq) permite, em um único experimento, descobrir novos genes e transcritos e mensurar a expressão destes. A quantidade de dados gerados a partir do sequenciamento de uma única amostra pode ser superior a 500 gigabases em uma só corrida. Dessa forma, os dados gerados a partir de experimentos que utilizam RNA-seq necessitam de algoritmos robustos e eficientes para serem analisados. As ferramentas para análise de RNA-seq, normalmente, podem ser divididas em três categorias: alinhamento, montagem e quantificação de expressão. Tais categorias também podem ser interpretadas como uma sequência padrão de análise, que recebe o nome de Protocolo *Tuxedo* (TRAPNELL et al., 2012), composto por duas ferramentas: *Tophat* (TRAPNELL; PACHTER; SALZBERG, 2009) e *Cufflinks* (TRAPNELL et al., 2010). *Tophat* é um *software* que permite o alinhamento dos dados sequenciados com um genoma de referência. Para tanto, ele precisa de um segundo *software*, chamado *Bowtie* (LANGMEAD et al., 2009). A função do *Bowtie* é indexar o genoma de referência com base na transformada de *Burrows-Wheeler* (BWT) (BURROWS; WHEELER,

1994), a qual permite que grandes quantidades de dados possam ser varridas com alta eficiência e baixo uso de memória. O *Cufflinks*, um pacote de programas, utiliza o mapa gerado pelo *Tophat* para a montagem e quantificação de transcritos. O funcionamento do protocolo *Tuxedo* pode ser observado através da Figura 3.1.

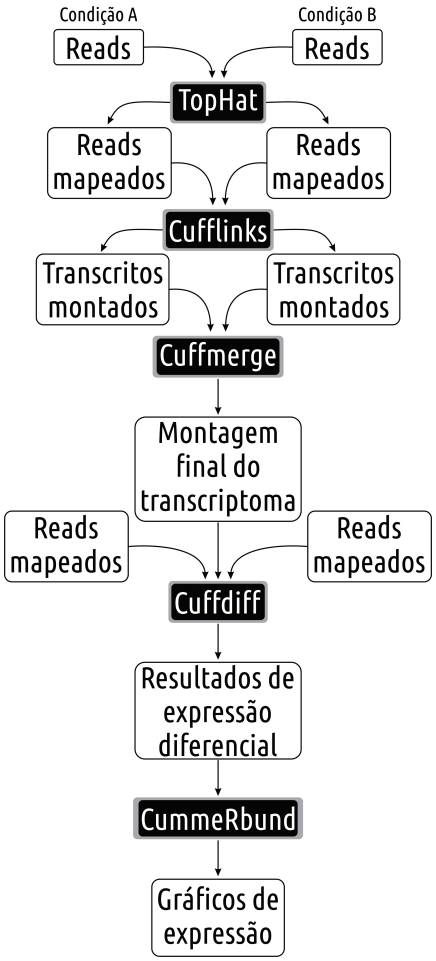

Figura 3.1: Visão global do protocolo *Tuxedo*. Alinhamento contra o genoma de referência e mapeamento usando o Tophat, montagem e quantificação dos transcritos com o Cufflinks, agrupamento de dados de diferentes condições experimentais para a montagem do genoma final com o Cuffmerge, cálculo de expressão diferencial com o Cuffdiff e extração de dados e construção de gráficos de expressão com o pacote em R cummeRbund . [Figura adaptada de (TRAPNELL et al., 2012)]

Uma análise de RNA-seq seguindo este protocolo baseia-se na comparação entre amostras submetidas a diferentes condições e tem o propósito de avaliar variações nos níveis de expressão gênica. Inicialmente, *reads* de diferentes amostras, submetidas a condições experimentais específicas e disponíveis no formato FASTQ (formato que combina sequências e parâmetros de qualidade do sequenciamento) (COCK et al., 2010) são alinhados contra um genoma de referência através do *software Tophat*. Este, por sua vez, gera um mapa com os *reads* alinhados permitindo que o *Cufflinks* realize a montagem dos transcritos, comparando o mapa com um arquivo de anotações no formato GTF (*General Transfer Format*). Os arquivos de anotações

armazenam informações sobre a estrutura do gene, incluindo *coding sequences* (CDS), exons e códons de início e parada (DORAN; CREEVEY, 2013). O *Cufflinks* engloba ainda dois outros programas: *Cuffmerge* e *Cuffdiff*. O primeiro é responsável por unir duas ou mais saídas do *Cufflinks*, ou seja, o genoma de cada condição é montado individualmente e, para compara-los, faz-se necessário agrupa-los, gerando um genoma final. O segundo utiliza a saída do *Cuffmerge* para comparar os níveis de expressão de cada transcrito entre as amostras. Isso, por exemplo, permite a verificação de quais genes são positivamente ou negativamente regulados de acordo com cada condição, além de identificar aqueles que são diferencialmente expressos. Por fim, tem-se o pacote em linguagem *R* conhecido como *cummeRbund* (GOFF; TRAPNELL; KELLEY, 2012). Trata-se de um pacote que permite navegar por todo o conteúdo gerado pelo *Cuffdiff*, possibilitando a manipulação de dados estatísticos, utilização de filtros para obtenção de relações mais apuradas de genes, construção de diferentes modelos gráficos para a comparação dos dados sequenciados, entre outras.

3.5 *Gene Ontology*

O termo função é vago quando aplicado a genes ou proteínas e é coloquialmente utilizado para descrever atividades bioquímicas, objetivos biológicos e estrutura celular. É comum se referir à função de uma proteína, por exemplo a tubulina, como “GTPase” ou “constituinte de fuso mitótico”. Por essa razão, o *Gene Ontology Consortium* criou 3 categorias independentes de ontologias: Processos Biológicos (BP), Funções Moleculares (MF) e Componentes Celulares (CC) (ASHBURNER et al., 2000).

Processos biológicos se referem a um objetivo biológico ao qual um gene ou produto gênico contribuem. Os processos normalmente estão relacionados à uma transformação física ou química, passando a idéia de que um elemento passa por um processo originando algo diferente. “Crescimento e manutenção celular” ou “metabolismo de pirimidina” são exemplos de termos BP. As funções moleculares podem ser definidas como sendo a atividade bioquímica de um produto gênico. Ela descreve apenas o que é feito, sem especificar onde ou quando o evento ocorre. Alguns exemplos são “enzima”, “ligante” e “adenilato-ciclase”. Por fim, componentes celulares indicam um local na célula onde o produto gênico está ativo e refletem nossa compreensão acerca da estrutura celular eucariótica. Alguns exemplos de tais termos são “ribossomo”, “proteossomo” e “membrana nuclear” (ASHBURNER et al., 2000).

A relação entre um produto gênico com processos biológicos, funções moleculares e componentes celulares são de um para muitos, o que reflete a realidade biológica na qual uma dada

proteína funciona em diversos processos, contém diferentes domínios e desempenha diferentes funções moleculares. Além disso, participam de múltiplas interações alternativas com outras proteínas, organelas ou localizações específicas na célula (ASHBURNER et al., 2000). Os termos, ainda, possuem uma relação pai-filho, em que o termo filho sempre é mais específico que o pai, sendo estruturados na forma de um gráfico acíclico direto (DAG) (GENTLEMAN, 2016). Através dele, um conjunto de genes anotados para um certo termo - ou um nodo - é um subconjunto daqueles anotados para os termos pais (GOEMAN; MANSMANN, 2008). Em um gráfico acíclico direto, nunca um mesmo nodo será visitado duas vezes. Ele apresenta uma topologia ordenada em que o nodo inicial possui valor inferior ao nodo subsequente - no caso de ontologias, entendemos os valores como especifidade. Um exemplo desse gráfico pode ser visto por meio da Figura 3.2.

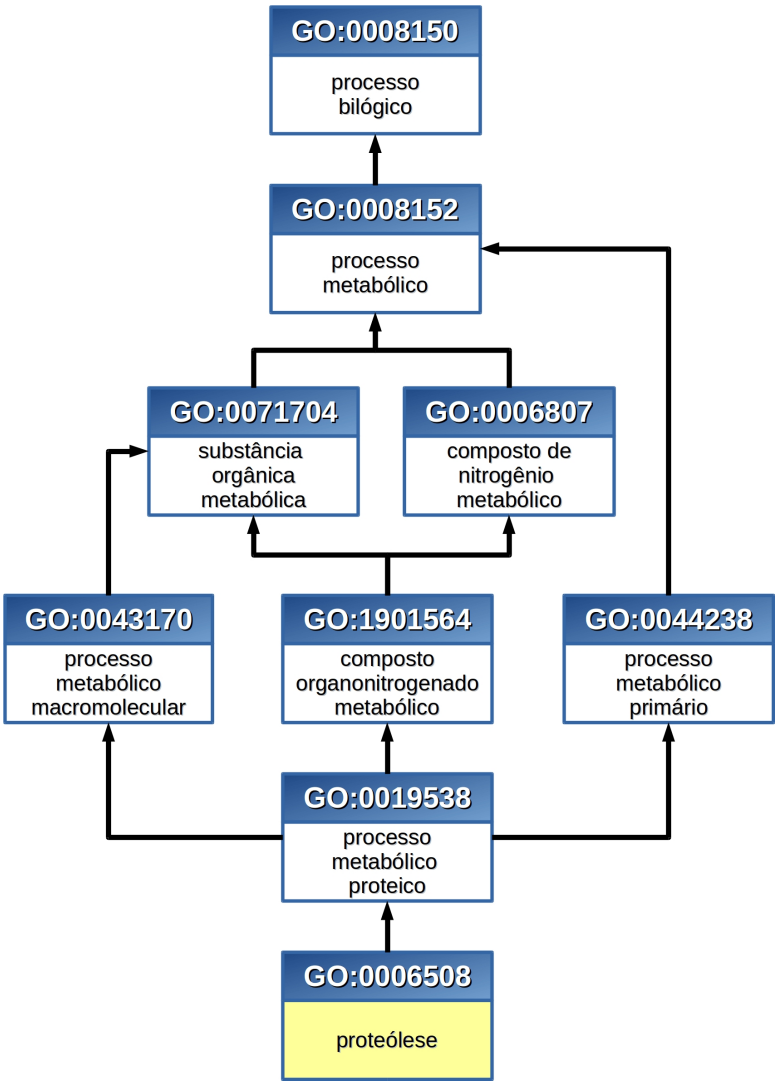

Figura 3.2: Gráfico acíclico direto (DAG), tendo como exemplo o termo GO:0006508. "Pais" se referem à nodos próximos da raiz do gráfico, ao passo que "filhos" são nodos mais próximos às extremidades. [Figura adaptada de `https://www.ebi.ac.uk/QuickGO/term/GO:0006508`. Acesso em 09/10/2017.]

As relações pai-filho podem ser "*is-a*", onde o termo filho é mais específico que o pai, "*has-a*" ou "*part-of*", em que o filho é parte do pai, como, por exemplo, o telômero sendo parte de um cromossomo. Outras relações ainda encontradas são "*regulates*", "*negatively regulates*" e "*positively regulates*" (`http://www.geneontology.org/page/ontology-relations`. Acesso em 10/10/2017). Um exemplo pode ser visto na Figura 3.3.

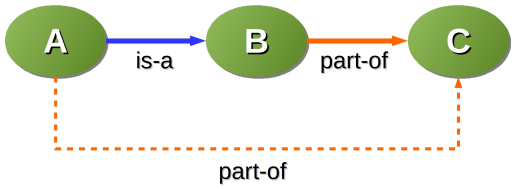

Figura 3.3: Relação entre termos GO (A, B e C). As setas indicam a direção da relação. Linhas pontilhadas representam um relacionamento de inferência. [Figura adaptada de `http://www.geneontology.org/page/ontology-relations`. Acesso em 10/10/2017.]

Mapear um gene a um termo pode ser baseado em diferentes características. Existe um conjunto de códigos de evidência, conforme Tabela 3.1. Isso permite que o pesquisador, eventualmente, deixe de utilizar genes ligados a determinadas evidências, podendo escolher as que melhor lhe atendem (CARLSON, 2017).

Tabela 3.1: Códigos de evidência de ontologias.

IMP	*inferred from mutant phenotype*	inferido de fenótipo mutante
IGI	*inferred from genetic interaction*	inferido de interação genética
IPI	*inferred from physical interaction*	inferido de interação física
ISS	*inferred from sequence similarity*	inferido de sequência similar
IDA	*infererred from direct assay*	inferido de análise direta
IEP	*inferred from expression pattern*	inferido de padrão de expressão
IEA	*inferred from electronic annotation*	inferido de anotação eletrônica
TAS	*traceable author statement*	declaração de autoria rastreável
NAS	*non-traceable author statement*	declaração de autoria não rastreável
ND	*no biological data available*	sem dados biológicos disponíveis
IC	*inferred by curator*	inferido por curador

Fonte: Tabela modificada de (CARLSON, 2017).

As ontologias, também chamadas de *GO terms*, são divididas, ainda, em quatro classes: *parents* (pais), *ancestor* (ancestrais), *children* (filhos) e *offspring* (descendentes). Para cada termo existe um nível hierárquico. É possível saber exatamente como cada ontologia está conectada às demais. Vale ainda ressaltar que o nome de cada termo é composto pelas iniciais GO seguidas de " : " e 7 dígitos como, por exemplo, GO:0000587 e GO:0089578.

3.6 *KEGG: Kyoto Encyclopedia of Genes and Genomes*

KEGG é uma base de dados online que possibilita o estudo e compreensão de funções e sistemas biológicos como células, organismos e ecossistemas a partir de dados genômicos e moleculares. Trata-se de uma representação de biologia de sistemas e tem como propósito a construção de blocos de genes, proteínas e substâncias químicas - informações genômicas e químicas, portanto - integradas a diagramas de interação escritos, reações e relacionamento com redes. Possui, ainda, informações acerca de doenças e drogas, assim como perturbações do sistema biológico como um todo (KEGG, 2017). Um esquema ilustrativo pode ser observado a partir da Figura 3.4. O banco de dados KEGG foi desenvolvido em 1995 pelo *Kanehisa Laboratories*, tornando-se referência na integração e interpretação de grandes conjuntos de dados originados do sequenciamento de genomas e outros experimentos com tecnologias de alto desempenho.

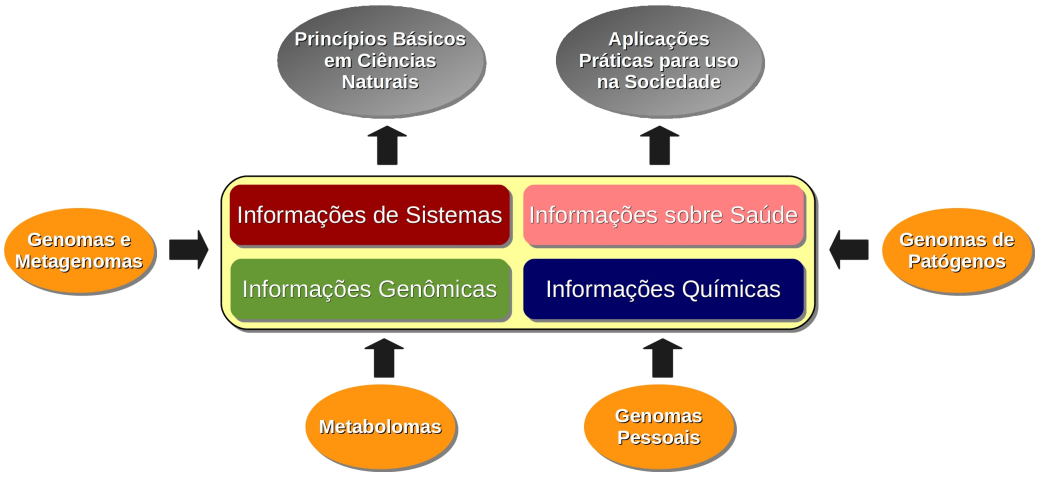

Figura 3.4: *Overview* da base de dados KEGG, mostrando o conteúdo de entrada e o que pode ser obtido através dele. [Figura adaptada de `http://www.genome.jp/kegg/kegg1a.html`. Acesso em 13/10/2017.]

Um dos módulos mais bem organizados no KEGG é a base de metabolismo, representada através de diagramas com vias metabólicas referenciadas. Cada referência pode ser visualizada como uma rede de enzimas ou *EC numbers* (*Enzyme Commission numbers*, esquema de classificação numérica para enzimas, relacionando os genes em genoma, juntamente com seus produtos gênicos, a uma via metabólica) (KANEHISA; GOTO, 2000). Um exemplo de diagrama pode ser visto na Figura 3.5.

Os genes enzimáticos são identificados com base no genoma, por meio de similaridade de sequências e posicionamento correlacional dos genes. Logo após, são definidos os *EC numbers* e, então, rotas metabólicas para organismos específicos são construídos computacionalmente. Algumas rotas metabólicas são bem conservadas entre a maioria dos organismos - de mamíferos

à bactérias. Com isso, é possível desenhar manualmente uma rota e, então, utiliza-la para diferentes espécies (KANEHISA; GOTO, 2000).

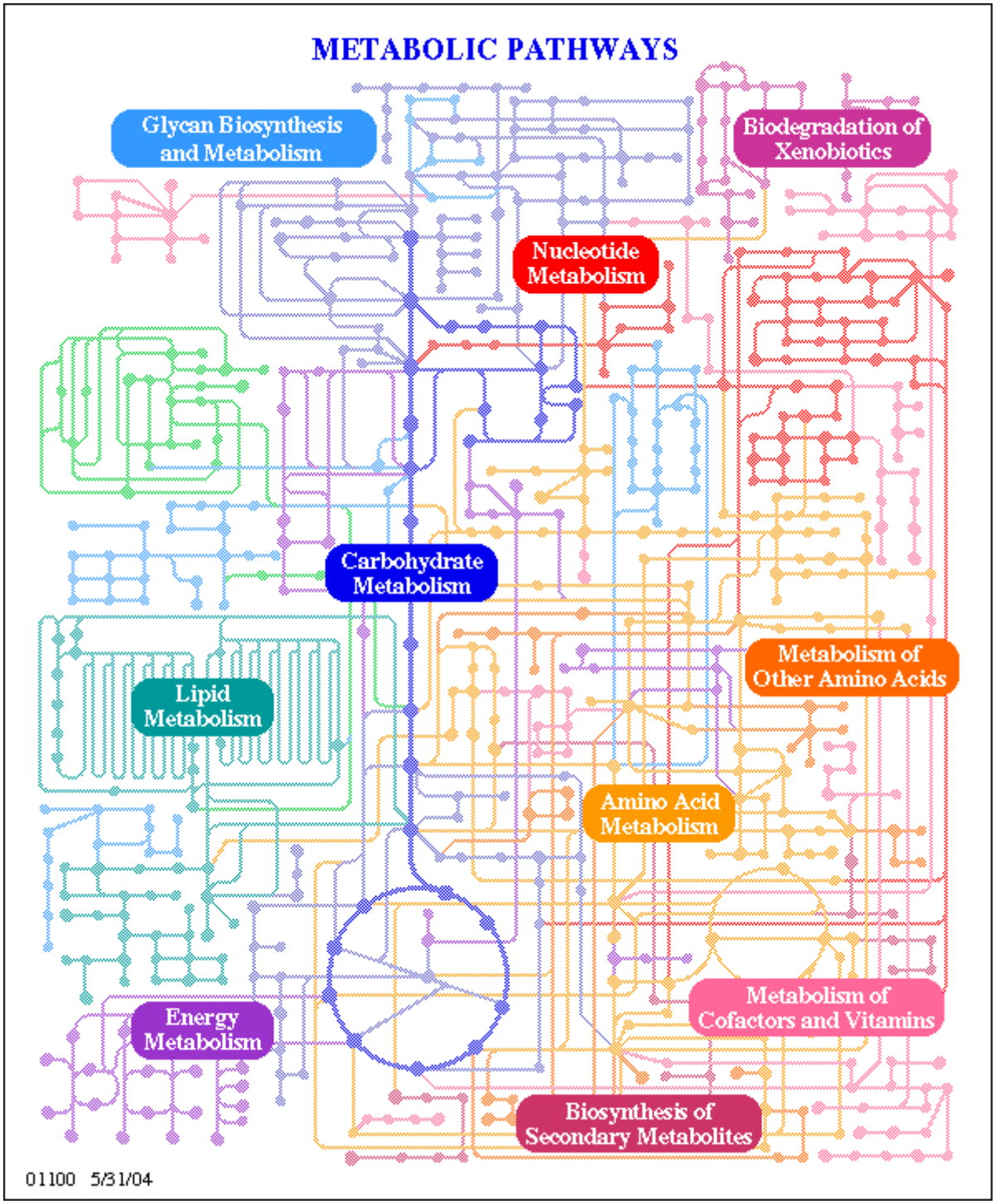

Figura 3.5: Exemplo de diagrama de vias metabólicas. Diferentes cores representam diferentes vias metabólicas com seus respectivos componentes (pontos).[Fonte: `http://www.genome.jp/kegg-bin/show_pathway?scale=1.0&query=&map=aag01100&scale=0.35&auto_image=&show_description=hide&multi_query=&show_module_list=`. Acesso em 13/10/2017.]

Dados e conhecimento quanto aos sistemas moleculares que gerenciam processos celulares e o comportamento de organismos são coletados de forma manual a partir da literatura disponível, dando origem aos mapas metabólicos. Estes representam o conhecimento relacionado a diversas redes moleculares, como redes de interação/reação para metabolismo, processamento de informações genéticas, processamento de informações ambientais e outros processos celulares, redes perturbadas de interação/reação para doenças humanas e redes de transformação de estrutura química para o desenvolvimento de drogas (KANEHISA et al., 2009).

3.7 *EntropyClusterGenes*

Com base no módulo estatístico do *software ViaComplex* (CASTRO et al., 2009), desenvolvemos um pacote em linguagem de programação R capaz de agrupar os genes de duas amostras comparativas de maneira funcional, identificando os grupos significativos com base em um valor de corte estabelecido pelo usuário. Uma visão global acerca do seu funcionamento é descrita pela Figura 3.6.

A entrada de dados consiste em um arquivo texto de, no mínimo, 3 colunas, conforme a Tabela B.1, presente no Apêndice B. A primeira refere-se aos nomes dos genes, devendo estes utilizarem a nomenclatura *Entrez ID* ou *Gene Symbol*, enquanto as demais contêm os valores de expressão de cada uma das amostras envolvidas na análise. O nome de cada uma dessas colunas funciona como identificador único amostral. Como se tratam de análises comparativas (controle vs experimento), utiliza-se, como argumento, um vetor de comparações, em que cada elemento se refere ao nome de duas colunas ligadas a duas amostras, separadas, por exemplo, por uma vírgula ou um hífen (o usuário indica o separador).

O *EntropyClusterGenes* trabalha com múltiplas espécies, permitindo a utilização de pacotes de dados de espécies, presentes no repositório *online Bioconductor* (GENTLEMAN et al., 2004), ou, caso não exista um pacote, o usuário pode montar tabelas auxiliares relacionando ontologias e vias metabólicas com os genes da espécie de interesse. O *Aedes aegypti* não possui pacote de dados associado. Dessa forma, com o auxilio dos pacotes *GO.db* (CARLSON, 2013) e *biomaRt* (DURINCK et al., 2005; DURINCK et al., 2009), além do repositório online KEGG, construímos uma base específica para essa espécie.

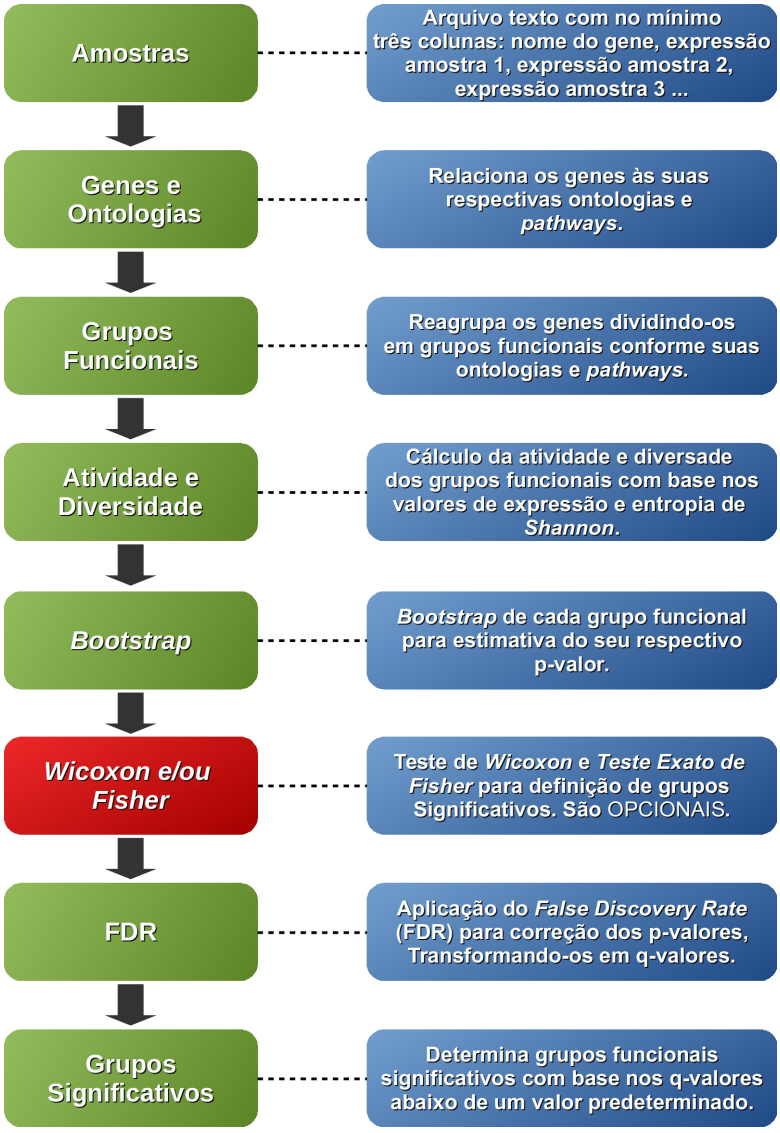

Figura 3.6: *Workflow* do pacote *EntropyclusterGenes*. A primeira coluna, à esquerda, representa os módulos principais, enquanto a segunda coluna apresenta quadros explicando brevemente cada um dos módulos.

Com a relação entre genes e seus papéis biológicos estabelecida, reagrupamos os dados iniciais em *GFAGs (Group of Functionally Associated Genes)*, estabelecendo diferentes grupos de genes associados entre si por um processo BP, função MF, componente CC ou KEGG *pathway*. Para cada grupo, afim de obter uma expressão quantitativa quanto à distribuição das amostras, conforme (CASTRO et al., 2007), medimos a informação do conjunto utilizando a Teoria da Informação de Shannon. O cálculo foi realizado da seguinte forma:

Seja M o número de genes em um dado grupo α ($\alpha = 1,...,N$). Para um dado grupo, é possível definir $S(i,\alpha)$ como sendo o sinal de um dado gene i,($i = 1,...,M_\alpha$), cuja a soma para um dado α vai até N_α . A contribuição $p(i,\alpha)$ do sinal do gene i para o sinal total do α-GFAG

é:

$$p(i,\alpha) = \frac{S(i,\alpha)}{N_\alpha} \tag{3.1}$$

de maneira que ao somarmos todos os $p(i,\alpha)$ teremos o somatório igual a 1. A função entropia normalizada H_α , aqui chamada de diversidade, é definida por:

$$H_\alpha = \frac{\sum_{i=1}^{N_\alpha} p(i,\alpha)\ln p(i,\alpha)}{\ln(M_\alpha)} \tag{3.2}$$

onde dividimos todos os termos pelo fator de normalização $\ln(M_\alpha)$ para garantir que $0 \leq H_\alpha \leq 1$. Desta maneira poderemos comparar diferentes GFAGs que podem apresentar diferentes números de genes. Finalmente, para normalizar as quantidades por grupos de genes tendo como referência o sinal de algum controle, definimos a diversidade relativa h_α para um dado GFAG como:

$$h_\alpha = \frac{H_\alpha^e}{H_\alpha^e + H_\alpha^\gamma} \tag{3.3}$$

onde H_α^e e H_α^γ são, respectivamente, a diversidade induzida pelo estímulo e pelo controle. Observe que $0 \leq h_\alpha \leq 1$, e $h_\alpha < 0.5$ implicam $H_\alpha^e < H_\alpha^\gamma$, isto é, a distribuição do sinal dos genes no α-ésimo GFAG é mais estreita para o estímulo do que para o controle, enquanto que $h_\alpha > 0.5$ representa o caso inverso. Em analogia, a atividade de expressão gênica relativa do α-GFAG é definida como

$$n_\alpha = \frac{N_\alpha^e}{N_\alpha^e + N_\alpha^\gamma} \tag{3.4}$$

onde N_α^e e N_α^γ são respectivamente, a expressão da atividade gênica induzida pelo estímulo e pelo controle. Novamente $0 \leq n_\alpha \leq 1$, e $n_\alpha < 0.5$ implica $N_\alpha^e < N_\alpha^\gamma$, isto é, neste GFAG o estímulo induz uma atividade de sinal que é menor do que a induzida pelo controle. Esta análise permite correlacionar os grupos de genes funcionalmente associados e suas funções em cada amostra.

Definido os valores de atividade e diversidade, partimos agora para uma amostragem dos dados utilizando *bootstrap*. De acordo com os grupos funcionais formados, a partir das amostras iniciais criamos conjuntos de genes aleatórios com o mesmo tamanho dos conjuntos iniciais. Para cada novo conjunto, calculamos novamente atividade e diversidade, comparando-as

com às do grupo original. Tal processo, para cada grupo, é repetido um determinado numero de vezes, número este maior que zero e definido pelo usuário. Este processo tem como objetivo estabelecer um p-valor para o conjunto gênico, o qual, através do método estatístico BH (Benjamini - Hochberg), é corrigido por um FDR (*False Discovery Rate*) padrão de 0.05 - ou um outro também estabelecido pelo usuário. Essa correção leva a um q-valor que, estando abaixo do FDR, indica que o grupo é significativo no contexto da comparação amostral. Um exemplo de arquivos de saída, apresentando os resultados, podem ser observados pela Tabela B.2 e Tabela B.3, presentes no Apêndice B.

Além da análise de diversidade e atividade, via *bootstrap*, o *EntropyClusterGenes* oferece uma análise de significância de GFAGs através de dois outros testes estatísticos: teste *Wilcoxon rank-sum* e teste exato de *Fisher*. Ambos são opcionais - podendo ou não serem realizados - e também retornam um p-valor, corrigido posteriormente por FDR, indicando se o GFAG é ou não significativo dentro de um valor de corte preestabelecido.

Por fim, o *EntropyClusterGenes* apresenta um módulo gráfico que permite uma análise individual de ontologias ou *pathways*, relacionando-as aos seus respectivos genes. A partir de um gráfico, é possível verificar quais genes pertencentes a um determinado grupo são *up* ou *down* regulados e quais deles são eventualmente significativamente expressos. Para tanto, além do resultado gerado pela nova ferramenta, o módulo gráfico necessita de um arquivo contendo os valores de expressão gênica relacionados à comparação (controle vs experimento), bem como os q-valores e logFC (*fold change*) de cada um deles.

3.8 *Ferramentas de Enriquecimento Funcional para Benchmarking*

Com o propósito de validar o *EntropyClusterGenes*, além de compararmos seus resultados com os resultados dos artigos referentes a cada uma das amostras de RNA-seq utilizadas, analisamos os mesmos conjuntos de dados utilizando outras três ferramentas de enriquecimento funcional, comparando todos os seus respectivos *outputs* com os da nova ferramenta. Para tanto, como critérios de seleção, optamos, exclusivamente, por pacotes desenvolvidos em linguagem de programação R e que utilizassem o método GSEA para análise funcional, características essas presentes, também, no *EntropyClusterGenes*.

3.8.1 *clusterProfiler*

A análise de grandes quantidades de dados requer o desenvolvimento de ferramentas de mineração de dados para a captura de informações biológicas. Uma abordagem comum consiste na clusterização gênica, agrupando diferentes genes com base em suas similaridades, como padrão de expressão e estrutura de redes proteicas. A análise de *clusters* (grupos) gênicos é utilizada para revelar padrões ocultos, como a busca por promotores ou reguladores compartilhados, classificação de processos biológicos, predição de novos genes ou genes não tão bem caracterizados e detecção de comunidades de proteínas. Uma outra forma de procurar por funções compartilhadas é a incorporação de conhecimentos biológicos através de ontologias. *Gene Ontology* (GO), capaz de associar genes a processos biológicos, funções moleculares e componentes celulares por meio de uma estrutura gráfica acíclica, *Kyoto Encyclopedia of Genes and Genomes* (KEGG), que relaciona genes à vias metabólicas, e *Disease Ontology* (DO), base de dados responsável por fazer a ligação entre genes e doenças humanas. Para tanto, surge *clusteRprofiler*, um pacote desenvolvido em linguagem R e que possibilita a análise estatística de GO e KEGG através da comparação de temas biológicos entre grupos gênicos (YU et al., 2012).

O pacote oferece um método de classificação gênica - *groupGO* - para identificar genes baseado em níveis específicos de GOs, além de funções como *enrichGO* e *enrichKEGG* para realizar o teste de enriquecimento para termos GO e vias KEGG, baseado em uma distribuição hipergeométrica. Para evitar uma elevada taxa de falsos positivos em múltiplos testes, q-valores são estimados por meio de um controle via FDR (*False Discovery Rate*). Além disso, *clusterProfiler* ainda possui uma função - *compareCluster* - que, automaticamente, calcula categorias funcionais enriquecidas referentes à cada grupo gênico, e dispõe de vários métodos para a visualização dos resultados. Para o seu funcionamento, pacotes auxiliares são necessários. **GO.db** e **KEGG.db** para o relacionamento com KEGGs e GOs, e algumas bases de dados específicas para o organismo em estudo, como **org.Hs.eg.db** (*Homo sapiens*) e **org.Mm.eg.db** (*Mus musculus*) (YU et al., 2012).

3.8.2 *Gene set variation analysis for microarray and RNA-Seq data (GSVA)*

Para facilitar a análise de enriquecimento funcional tipo GSEA, foi desenvolvido, em linguagem de programação R, o *Gene Set Variation Analysis* (GSVA), o qual permite a avaliação da variação da atividade de *pathways* subjacentes transformando uma matriz "*gene versus amostras*" em uma matriz "*conjunto de genes versus amostras*" sem um conhecimento prévio do *design* experimental. O método é não-paramétrico e sem supervisão, e ignora a abordagem

convencional de fenótipos de modelagem explícita dentro de algoritmos de enriquecimento com *score*. O foco é então substituído por um enriquecimento relativo de *pathways* através do espaço amostral ao invés de enriquecimento absoluto relacionado a um fenótipo. Um ponto positivo desta abordagem é que ela possibilita o uso de métodos analíticos tradicionais como classificação, análise de sobrevivência, clusterização e análise de correlação em uma dada via (*pathway*). Ela também facilita comparações amostrais entre vias e outros tipos de dados complexos tais como expressão de microRNA ou "*binding data*". Contudo, para experimentos caso-controle ou dados com um tamanho amostral moderado (inferior a 30), outros métodos de enriquecimento de conjunto de genes que incluem explicitamente um fenótipo em seus modelos disponibilizam maior poder estatístico para detectar enriquecimento funcional (HÄNZELMANN; CASTELO; GUINNEY, 2013).

A ferramenta parte de uma matriz de entrada em que as linhas representam os genes e as colunas representam as amostras. A saída será uma matriz na qual as linhas representam um conjunto de genes funcionalmente associados e as colunas indicam as amostras. Cada conjunto gênico dentro de uma dada amostra é ranqueado de acordo com sua significância, ou seja, a cada grupo é atribuído um peso (*score*) (HÄNZELMANN; CASTELO; GUINNEY, 2013).

3.8.3 *GAGE: Generally Applicable Gene-set Enrichment*

GAGE é um método desenvolvido em R e que pode ser aplicado a *datasets* com diferentes números de amostras, sendo baseado em um processo de randomização gênica paramétrica com a utilização de *log fold change* conforme os genes presentes. É capaz de realizar ajustes para diferentes designs experimentais e tamanhos amostrais através da decomposição de comparações *grupo a grupo* em comparações *um a um* entre amostras de diferentes grupos, funcionando tanto para experimentos de microarranjo quanto RNA-seq. Com base nos p-valores obtidos a partir de tais comparações, calcula-se um p-valor global, para cada conjunto gênico, usando-se um meta teste (LUO et al., 2009).

3.9 *Hipóteses e p-valores*

A abordagem tradicional para tornar válidas alguns tipos de inferência tem sido estabelecer uma questão a ser respondida na forma de duas hipóteses estatísticas contrastantes. A primeira, representando a inexistência de diferenças entre os parâmetros populacionais de interesse, é chamada de hipótese nula H_0, enquanto a segunda recebe o nome de hipótese alternativa, representada por H_1 . Um teste estatístico é realizado a partir de dados amostrais e comparado à

hipótese nula com o intuito de avaliar a consistência dos dados com H_0. A obtenção de valores mais extremos a partir do teste sugerem que os dados amostrais não são consistentes com H_0 (ANDERSON; BURNHAM; THOMPSON, 2000).

O p-valor é uma medida de discrepância do ajuste de um modelo ou hipótese nula para um certo conjunto de dados **X**. Trata-se, na teoria, de uma medida contínua de evidência que, na prática, resume-se à evidência forte, fraca ou sem evidência (altamente significante, marginalmente significante ou sem significância estatística), com pontos de corte de acordo (*cutoffs*) com o pesquisador (GELMAN, 2013). P-valores são normalmente usados para testar uma H_0, a qual, geralmente, afirma que não há diferença entre dois grupos ou que não existe correlação entre duas determinadas característica. Quanto menor o p-valor, menor a probabilidade de que um conjunto de dados observado tenha ocorrido ao acaso, garantindo a veracidade da hipótese nula. Um p-valor de 0.05 ou menos, normalmente, leva a acreditar na autenticidade da significância estatística obtida. No entanto, recentemente, a Associação Americana de Estatística (ASA) alertou sobre o perigo de tal prerrogativa. Um p-valor de 0.05 não significa que há uma chance de 95% de que uma dada hipótese esteja correta. Ao invés disso, ela mostra que se um hipótese nula é verdadeira, e todas as outras suposições feitas são válidas, existe 5% de chance de se obter um resultado incoerente com H_0 (BAKER et al., 2016).

3.10 *FDR: False Discovery Rate*

A quantidade de dados gerados pela comunidade científica tem aumentado consideravelmente ao longo dos últimos anos. As pesquisas por trás desse aumento, normalmente, envolvem tentativas de inferir conclusões por meio de vários testes de hipóteses. Os pesquisadores, tipicamente, são levados a realizarem ajustes de significância com o intuito de reduzirem a probabilidade de resultados falso positivos. Tais ajustes destinam-se a controlar taxas de erro e reduzir as chances de rejeitar incorretamente hipóteses nulas verdadeiras. No entanto, a desvantagem em fazer uso de tais técnicas consiste na perda de poder em detectar efeitos reais. Visando reduzir esse problema, surge, então, o método *False Discovery Rate* (FDR). O FDR é a probabilidade de que uma hipótese nula seja verdadeira dado que esta tenha sido rejeitada (GLICKMAN; RAO; SCHULTZ, 2014).

As tecnologias NGS mais recentes permitem aos pesquisadores executarem varreduras em genomas e monitorar os níveis de expressão de milhares de genes simultaneamente. O problema do teste de múltiplas hipóteses surge quando são comparados um grande número de genes entre diferentes grupos, por exemplo, pacientes com câncer de mama vs pacientes saudáveis. Nesse

contexto, o método de Bonferroni se torna conservativo e de baixo poder. Diante de tal situação, a melhor alternativa é o uso de FDR, o qual leva em conta a proporção de falsos positivos entre as hipóteses rejeitadas. Existe, atualmente, diferentes *software* e pacotes que aplicam a técnica. Basicamente é fornecido como entrada um conjunto de p-valores referentes à diferentes genes ou conjuntos gênicos, tendo, como saída, os correspondentes q-valores (p-valores corrigidos). Isso permite, então, identificar os q-valores significativos, que são aqueles menores ou iguais a um determinado valor (normalmente, 0.05). Caso exista, por exemplo, um total de **N** genes identificados como significantes pelo FDR, a maioria dos pesquisadores considerariam que o número de genes falso positivos não seria superior ao produto entre o corte estabelecido (0.05, por exemplo) e N (LIN; LEE, 2015).

3.11 *Bootstrap*

Introduzido em 1977, o termo Bootstrap (EFRON, 1979) faz referência a um método ou técnica de simulação cujo propósito consiste em obter intervalos de confiança de forma que determinados parâmetros de interesse possam ser estimados através da reamostragem do conjunto original de dados (MARTINEZ-ESPINOSA; SANDANIELO; LOUZADA-NETO, 2006). O método, ainda, é capaz de obter, com qualidade, estimativas dos erros padrão consequentes de estimativas de parâmetros ligados às iterações de reamostroagem (LEPAGE; BILLARD, 1992), além de permitir encontrar p-valores para testes estatísticos com base em uma hipótese nula (BOOS et al., 2003). A Figura 3.7 ilustra o funcionamento clássico de um bootstrap de amostra única, modelo aplicado no desenvolvimento do módulo estatístico da ferramenta *EntropyClusterGenes*, a qual ainda descreveremos neste capítulo.

Para melhor compreender o Bootstrap, é preciso lembrar que um modelo estatístico é, basicamente, um conjunto de distribuições de probabilidades com o propósito de descrever o real estado da natureza e os dados aleatórios disponíveis relacionados à compreensão desse estado. Com isso, é possível afirmar que a inferência estatística está na escolha de uma dessas distribuições, dando uma noção da incerteza sobre ela. Fazemos inferências sobre populações desconhecidas, representadas por modelos estatísticos, a partir de dados de amostragem. O design da verdadeira amostragem de dados é reproduzida o mais fiel possível, enquanto aspectos desconhecidos do modelo são substituídos por estimativas amostrais (BOOS et al., 2003).

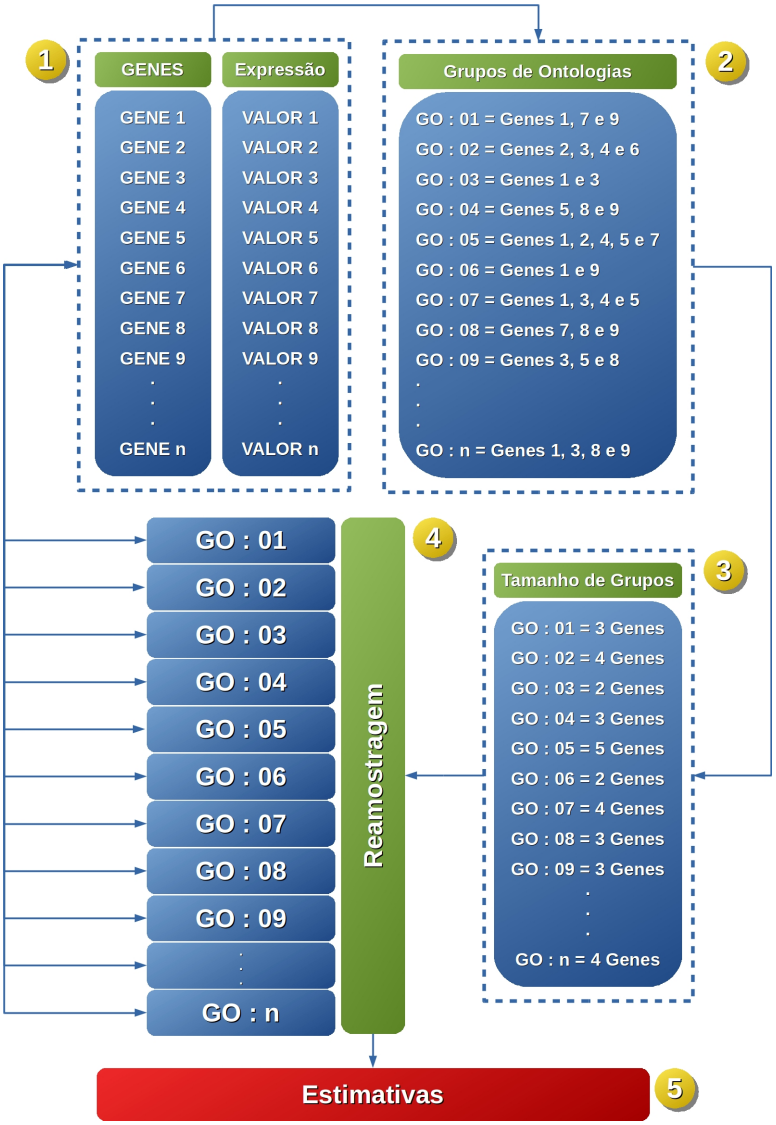

Figura 3.7: Esquema básico do funcionamento de um bootstrap em um contexto biológico. Partimos de um conjunto de genes com os seus respectivos valores de expressão (1). Em seguida, o conjunto é reagrupado em subgrupos (2) que segregam os genes de acordo com suas ontologias, em que um gene pode estar presente em mais de um subgrupo. Cada subgrupo (3) apresenta uma determinada quantidade de genes, a qual servirá como parâmetro para a reamostragem aleatória (4) com reposição a partir do conjunto inicial gênico. Após a reamostragem (5), tem-se, então as estimativas desejadas, como, por exemplo, p-valores.

3.12 *Teste Wilcoxon rank-sum*

A comparação entre duas amostras geralmente se divide em duas categorias: (i) podemos ter um determinado número de réplicas para cada uma das amostras, as quais seriam *não-pareadas* ou (ii) dispomos de uma quantidade de comparações *pareadas* que levam a diferenças cujos valores podem ser positivos ou negativos. *Frank Wilcoxon* (WILCOXON, 1945) propôs um método que permitia a utilização de *rankings*, em que *scores* (1, 2, 3, ..., n) eram substituídos

por dados numéricos reais de forma que, rapidamente, uma ideia aproximada da significância das diferenças em experimentos pudesse ser obtida.

Para duas amostras independentes, uma das variações da metodologia de Wilcoxon é o teste *rank-sum* (*Wilcoxon rank-sum test*). Neste tipo de teste, os dados de ambas as amostras são combinados e ranqueados, sendo, na sequência, separados, mas mantendo-se o *ranking* de cada uma das observações. A hipótese nula comum (H_0) é que ambas as amostras pertencem a populações idênticas, ao passo que a hipótese alternativa (H_1) afirma que as populações diferem quanto à média ou mediana. Caso as amostras tenham como origem a mesma população, espera-se uma mistura de *ranks* altos, médios e baixos em cada uma das amostras. Entretanto, considerando-se que (H_1) seja satisfeita, espera-se que *ranks* baixos dominem uma das populações enquanto valores elevados predominem na outra. Ao compararmos, por exemplo, um grupo de pacientes tratados com uma determinada dose de um certo medicamento e um grupo, tratado com o mesmo medicamento, porém com um dose diferente, a mudança de centralidade observada com a satisfação de (H_1) frequentemente se refere a um efeito aditivo do tratamento, ou seja, existe uma diferença constante entre os tratamentos (SPRENT; SMEETON, 2000).

A função *wilcox.test*, pertencente ao pacote *stats*, versão 3.6, da linguagem de programação R, permite a aplicação do teste de Wilcoxon para duas amostras pareadas ou não-pareadas. Os dados de cada amostra são armazenados em vetores numéricos finitos. Um argumento lógico (Verdadeiro ou Falso) indica se o teste será ou não pareado. Como resultado, um p-valor é retornado, indicando a significância do teste.

3.13 *Teste Exato de Fisher*

Proposto por *Ronald Aylmer Fisher*, o *teste Exato de Fisher* (FISHER, 1934) é um teste de independência em oposição a associação em tabelas de contingência 2x2, utilizado quando se dispõe de duas variáveis nominais e desejamos saber se as proporções de uma variável são diferentes entre os valores de outra. Uma situação típica para o uso dessas tabelas é quando temos contagens de indivíduos categorizada de forma dicotômica como, por exemplo, tipos de tratamento (uso da droga A ou B) e as respostas em função de cada um deles (melhora da condição clínica do paciente ou condição inalterada) (SPRENT, 2011). Um exemplo pode ser observado através de Tabela 3.2.

Tabela 3.2: Exemplo de tabela de contingência para o teste exato de *Fisher*. Sob a hipótese de independência, se considerarmos os totais marginais da tabela como fixos, a distribuição de números na primeira célula (ou outra célula qualquer) apresenta uma distribuição hipergeométrica sob independência para qualquer modelo que esteja associado à tabela. As respostas à cada droga são binomialmente distribuídas com um valor em comum para o parâmetro p (probabilidade de sucesso).

	Melhora	**Sem Melhora**	**TOTAL**
Droga A	8	1	9
Droga B	3	9	12
TOTAL	11	10	21

Fonte: Tabela modificada de (SPRENT, 2011), página 525.

Segundo Fisher, sob a hipótese de independência, se considerarmos os totais marginais da tabela como fixos, a distribuição de números na primeira célula (ou outra célula qualquer) apresenta uma distribuição hipergeométrica sob independência para qualquer modelo que esteja associado à tabela. As respostas à cada droga, como na Tabela 3.2, são binomialmente distribuídas com um valor em comum para o parâmetro p (probabilidade de sucesso) (SPRENT, 2011).

Se uma tabela de contingência apresenta entradas $n_{\mathrm{i,j}}$ (i,j = 1, 2), com o total das linhas $n_{\mathrm{i+}}$, total das colunas $n_{\mathrm{+j}}$ e o total de entrada das 4 células como n, a distribuição hipergeométrica apresenta probabilidade associada conforme Equação 3.5.

$$\frac{(n_{1+})!(n_{2+})!(n_{+1})!(n_{+2})!}{(n_{11})!(n_{12})!(n_{21})!(n_{22})!n!} \tag{3.5}$$

Para a execução do teste, é preciso calcular todas as probabilidades para todos os possíveis n_{11} consistentes com os totais marginais fixos e computar um p-valor como sendo a soma de tais probabilidades que sejam menores ou iguais à aquele associado com a configuração observada (SPRENT, 2011).

O teste exato de Fisher pode ser calculado através da função *fisher.test*, pertencente ao pacote *stats*, versão 3.6, da linguagem de programação R. Para tanto, basta fornecer como parâmetro uma matriz de contingência de de dimensões 2x2 tendo, como resultado, um p-valor mostrando a significância do teste.

4 *Resultados e Discussão*

4.1 *Desempenho*

Uma das principais características relacionadas à automação do *EntropyClusterGenes*, ferramenta que desenvolvemos especificamente para o enriquecimento funcional comparativo de grupos gênicos, é a sua função multiespécie. Através dela, é possível relacionar genes a ontologias e/ou *KEGG pathways* sem a necessidade da construção manual de tabelas relacionais auxiliares. No entanto, é obrigatória a existência de um banco de dados específico para a espécie, disponível em forma de pacote no repositório online *Bioconductor*. Embora tenhamos, até o momento, além de *Drosophila melanogaster*, demonstrado o uso da ferramenta aplicando-a à dados de RNA-seq de *Aedes aegypti*, tal espécie ainda não possui uma base de dados devidamente construída para a linguagem R. Dessa forma, através dos repositórios *Gene Ontology*, KEGG, *VectorBase* e *BioMart*, montamos nossa própria base.

Como desafio das aplicações voltadas para grandes volumes de dados, surge a questão do tempo de processamento e uso de memória computacional. No *EntropyClusterGenes*, o ponto em que nos deparamos com tal aspecto foi na determinação dos p-valores para os grupos de genes funcionalmente associados (GFAGs). O cálculo de cada p-valor é feito utilizando-se o método de *Bootstrap*. São realizados milhões de amostragens aleatórias, seguidas por diversas operações matemáticas, para cada GFAG encontrado. Quanto maior a precisão exigida de um p-valor, maior o número de amostragens e, portanto, maior o tempo de processamento necessário.

A primeira versão do pacote que desenvolvemos utilizava uma técnica de programação estruturada convencional. Rotinas executadas mais de uma vez se davam por meio de laços de repetição do tipo *FOR* concatenados. Embora o *R* disponibilize este tipo de recurso, por se tratar de uma linguagem interpretada, e não compilada (como o *C++*, por exemplo), o tempo demandado para o *Bootstrap* era extremamente elevado. A fim de aumentarmos a velocidade do programa, mudamos a estratégia de codificação. Passamos a utilizar funções da família *Apply* em detrimento dos laços *FOR*, associadas à rotinas de paralelização. Além do ganho de velocidade com a substituição dos laços, o ganho de *performance* com a utilização de mais de

um núcleo do processador para a realização de uma mesma tarefa foi elevado, como mostra Tabela 4.1. Vale ressaltar que todos os testes foram realizados sempre no mesmo equipamento, uma máquina com processador *i7* de 3.4GHz de velocidade e 8 núcleos e 16GB de Memória RAM. Quanto mais rápido e mais núcleos a máquina dispuser, melhor o desempenho, dando destaque para o processamento em *clusters*. Entretanto, o ganho com a paralelização não foi suficiente. Uma saída encontrada e que reduziu significativamente o tempo de processamento foi a criação de um código híbrido, mesclando a linguagem R com o C++ por meio da biblioteca Rcpp.

Tabela 4.1: Comparação de *performance* entre o código paralelizado, não paralelizado e híbrido. Foi sempre utilizado a mesma amostra nos três códigos, definindo como parâmetro para o *bootstrap* 10 mil passos. No teste foram utilizados 7 núcleos de um processador *i7* de 3.4GHz, entretanto, de acordo com o equipamento, mais *cores* poderiam ser usados.

Técnica	**Uso de CPU (%)**	**Uso de Memória (MB)**	**Tempo de Processamento (minutos)**
não paralelizada	12	405,9	79
paralelizada	21	391,3	37
híbrida	12	400,1	2

Fonte: O autor (2018).

4.2 Otimização de *Bootstraps*

O *EntropyClusterGenes* é um pacote para linguagem de programação R que visa ao enriquecimento funcional de diferentes conjuntos sob o ponto de vista de *diversidade* e *atividade* gênica, efetuando o cálculo de p-valores para cada grupo funcional por meio de *bootstrap*. Como mencionado no item anterior, a etapa de *bootstrap* é a que mais consome tempo de processamento. Sendo assim, além de escrever um código adequadamente otimizado, é preciso identificar parâmetros próximos de uma condição ótima para que resultados aceitáveis sejam obtidos com a menor quantidade de passos possível. De forma a encontrar este número mínimo, conforme os gráficos da Figura 4.1, aplicamos a ferramenta em quatro diferentes combinações, variando o número de passos e observando a variação de GFAGs significativos encontrados quando da utilização de cada passo.

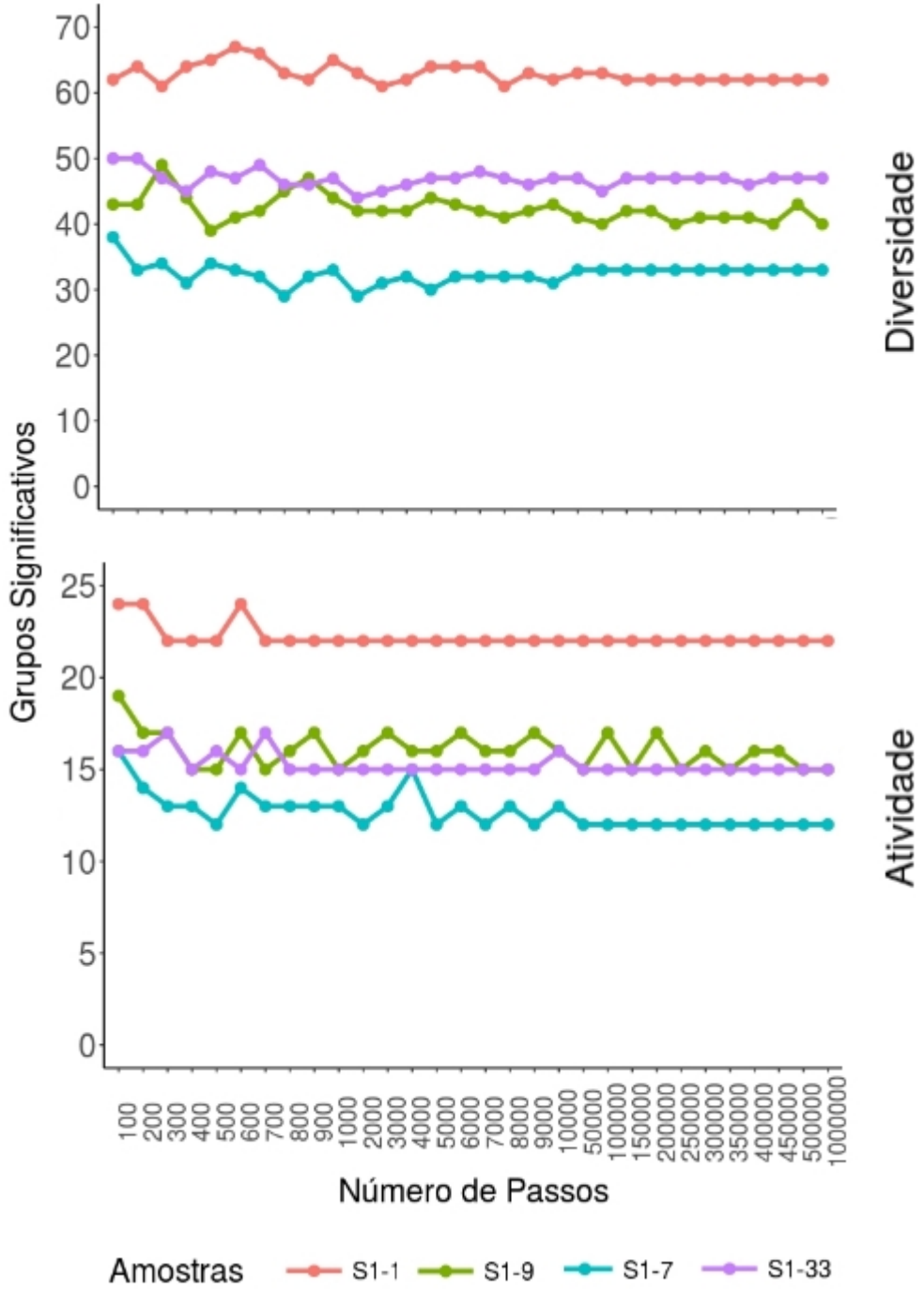

Figura 4.1: Variação do número de passos utilizados para bootstrap em função do número de GFAGs obtidos para quatro diferentes amostras de *Aedes aegypti*, referenciadas como S1-1, S1-7, S1-9 e S1-33, sob o ponto de vista de atividade e diversidade gênica. [Fonte: O autor (2018).]

Como mostra a Figura 4.1, verifica-se uma leve oscilação (para mais ou para menos) de grupos significativos encontrados à medida em que aumentamos o número de passos. Isso ainda não nos dá um valor ideal e que possa ser utilizado como padrão em futuras análises, entretanto, mostra que, eventualmente, podemos obter resultados praticamente idênticos com, por exemplo, 10 mil ou 1 milhão de passos. Ambos os gráficos nos passam uma idéia de

quantidade de significância. Porém, como saber se os grupos significantes encontrados com a utilização de um passo são os mesmos encontrados com o de outro? Para isso, observe o diagrama da Figura 4.2.

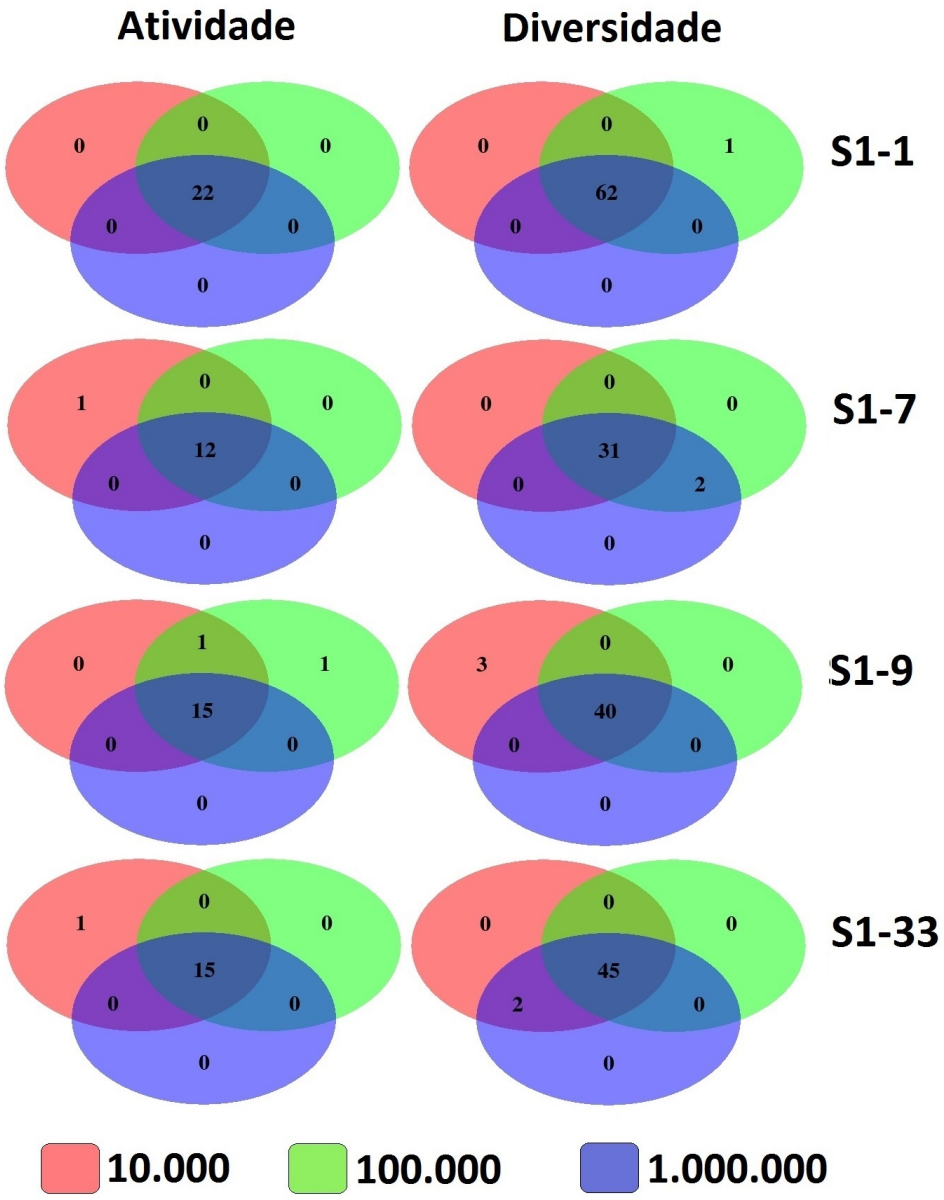

Figura 4.2: Verificação dos GFAGs encontrados por pasos de bootstrap. Com base nos resultados obtidos a partir da Figura 4.1, verificamos se os grupos encontrados com um determinado passo de bootstrap é o mesmo à medida que este número varia. [Fonte: O autor (2018).]

A partir dos diagramas de *Venn* da Figura 4.2 nota-se que para três diferentes passos (10 mil, 100 mil e 1 milhão), considerando as mesmas quatro combinações presentes na Figura 4.1, tanto para atividade quanto para diversidade gênica, uma quantidade de grupos próximo a 100% aparece em comum aos três passos em todas as combinações utilizadas. Isso mostra, portanto, que podemos considerar um número reduzido de passos (10000, por exemplo) de *bootstrap* e

obter um resultado muito próximo ao de quando usamos um valor elevado, permitindo, assim, uma demanda de tempo de processamento bem menor.

4.3 *Análise de Expressão Diferencial Gênica*

O funcionamento da nova ferramenta de enriquecimento funcional que desenvolvemos foi demonstrado por meio do enriquecimento funcional de grupos gênicos baseados em dois diferentes experimentos de RNA-seq de duas diferentes espécies: *Drosophila melanogaster* e *Aedes aegypti*. Entretanto, como mencionado anteriormente, partimos dos dados brutos oriundos dos sequenciamentos e aplicamos um protocolo de análise distinto daqueles utilizados nos estudos originais. Usamos o protocolo *Tuxedo* e reanalisamos todos os dados, do alinhamento com um genoma de referência, passando pela montagem e quantificação dos transcritos até chegarmos aos genes diferencialmente expressos, descritos pela Tabela 4.3 e Tabela 4.2. A descrição detalhada de cada uma das amostras e combinações são descritas no Apêndice A através da Tabela A.1 e Tabela A.2.

A expressão diferencial foi realizada sempre com base na comparação entre dois conjuntos amostrais - controle vs experimento. A partir das amostras disponíveis, optamos por comparações de grupos pertencentes ao mesmo experimento, buscando sempre condições contrastantes. A identificação de genes diferencialmente expressos, para mesmas combinações, pode variar de acordo com protocolo de análise utilizado. Diferentes ferramentas computacionais se baseiam em diferentes métodos estatísticos. Entretanto, a contagem de *reads* encontrada para cada transcrito deve ser constante entre os protocolos de análise, diferindo-se, apenas, pelo tipo de unidade de expressão utilizada (FPK, RPKM, TPM, entre outras). No caso deste trabalho, os arquivos de entrada para o enriquecimento funcional continham o valor de expressão gênica bruta, pré expressão diferencial, em FPKM.

Tabela 4.2: Genes diferencialmente expressos (DEG) em função das combinações do conjunto de amostras S2, conforme Tabela A.2, Apêndice A, referente à *Drosophila melanogaster*.

Comparações	**DEG**
S2-1	60
S2-2	147
S2-3	57
S2-4	70
S2-5	851

Fonte: O autor (2018).

Tabela 4.3: Genes diferencialmente expressos (DEG) em função das combinações do conjunto de amostras S1, conforme Tabela A.1, Apêndice A, referente à *Aedes aegypti*.

Comparações	DEG
S1-1	63
S1-2	87
S1-3	53
S1-4	51
S1-5	99
S1-6	41
S1-7	37
S1-8	18
S1-9	34
S1-10	54
S1-11	99
S1-12	13
S1-13	116
S1-14	160
S1-15	16
S1-16	53
S1-17	42
S1-18	12
S1-19	0
S1-20	34
S1-21	123
S1-22	55
S1-23	62
S1-24	0
S1-25	13
S1-26	22
S1-27	15
S1-28	26
S1-29	0
S1-30	0
S1-31	0
S1-32	0
S1-33	163
S1-34	0
S1-35	0

Fonte: O autor (2018).

Com base na Tabela 4.3 e Tabela 4.2, observamos uma grande oscilação no número de DEGs encontrados. Nas comparações referentes à *Aedes aegypti* verifica-se que o maior volume de expressão diferencial ocorreu na comparação S1-33, enquanto em outras, como S1-35, não foram encontrados genes diferencialmente expressos. Em relação à *Drosophila melanogaster*, todas as comparações apresentaram DEGs, com destaque para S2-5 (851 genes).

4.4 *Análise de Grupos Significativos*

O enriquecimento funcional apresentado pela nova ferramenta é realizado com base na atividade e diversidade gênica, utilizando, de forma opcional, os testes de Wilcoxon e Fisher. Cada grupo gênico funcionalmente associado possui p-valores associados que indicam sua significância no espaço amostral estudado. Além disso, cada comparação pode ser associada à *KEGG pathways*, Processos Biológicos (BP), Funções Moleculares (MF) e Componentes Celulares (CC). Portanto, uma mesma comparação pode ser analisada sob até 16 diferentes aspectos, como apresentado na Tabela 4.5 e Tabela 4.5. Nessas tabelas, além de explorarmos todas as combinações possíveis de análise, mostramos a quantidade de GFAGs significativos encontrados quanto a BP, CC, MF e KEGG.

Com o intuito de validar o *EntropyClusterGenes*, realizamos um benchmarking comparando-o a três outras ferramentas de enriquecimento similares: *GAGE*, *GSVA* e *ClusterProfiler*. A escolha de cada uma delas se deu em função de suas respectivas relevâncias na classe GSEA (*Gene Set Enrichment Analisys*) e por serem pacotes desenvolvidos para linguagem R, assim como nossa ferramenta. Todas elas foram aplicadas aos mesmos conjuntos amostrais (*Aedes* e *Drosophila*) e seus resultados quanto a grupos gênicos significativamente expressos estão descritos na Tabela 4.6, Tabela 4.7, Tabela 4.8, Tabela 4.9, Tabela 4.10 e Tabela 4.11.

Tabela 4.4: Grupos significativamente expressos encontrados via *EntropyClusterGenes* de acordo com a Diversidade (h), Atividade (n), Wilcoxon (w) e Fisher (f) para Processos Biológicos (BP), Componentes Celulares (CC), Funções Moleculares (MF) e KEGG *Pathways* referente à *Drosophila melanogaster* (amostra S2, conforme comparações descritas pela Tabela A.2, Apêndice A).

Comparações	**BP**				**CC**				**MF**				**KEGG**			
	h	**n**	**w**	**f**	**h**	**n**	**w**	**f**	**h**	**n**	**w**	**f**	**h**	**n**	**w**	**f**
S2-1	168	341	0	56	29	91	1	16	96	164	0	15	9	14	1	2
S2-2	591	426	0	148	136	91	0	62	271	194	0	45	11	9	0	12
S2-3	146	690	0	4	33	173	0	11	90	292	0	3	0	32	0	16
S2-4	202	360	0	3	46	90	0	8	91	178	0	6	3	19	0	4
S2-5	201	438	1	138	41	117	0	28	117	207	1	42	4	12	0	9

Fonte: O autor (2018).

Tabela 4.5: Grupos significativamente expressos encontrados via *EntropyClusterGenes* de acordo com a Diversidade (h), Atividade (n), Wilcoxon (w) e Fisher (f) para Processos Biológicos (BP), Componentes Celulares (CC), Funções Moleculares (MF) e KEGG *Pathways* referente à *Aedes aegypti* (amostra S1, conforme comparações descritas pela Tabela A.1, Apêndice A).

Comparações	**BP**				**CC**				**MF**				**KEGG**			
	h	**n**	**w**	**f**	**h**	**n**	**w**	**f**	**h**	**n**	**w**	**f**	**h**	**n**	**w**	**f**
S1-1	29	62	0	7	8	20	0	4	36	60	0	8	11	22	0	8
S1-2	45	30	0	9	8	14	0	8	53	38	0	14	10	6	0	11
S1-3	115	70	0	9	29	38	0	5	161	74	0	19	38	8	0	3
S1-4	121	144	1	11	46	58	1	7	134	149	2	19	22	22	0	13
S1-5	46	20	1	5	8	8	1	11	42	26	2	4	2	4	0	14
S1-6	60	23	2	10	25	5	1	4	72	36	2	10	4	2	0	9
S1-7	63	48	0	5	18	28	0	9	72	50	0	7	24	11	0	12
S1-8	71	28	0	8	34	8	1	10	68	35	0	5	7	6	1	13
S1-9	59	51	0	9	22	38	0	8	98	80	0	11	9	3	0	17
S1-10	25	83	0	15	13	28	0	8	52	113	0	20	11	28	0	21
S1-11	42	116	1	22	20	37	1	10	61	167	2	34	7	42	0	23
S1-12	48	147	0	1	14	55	0	5	58	166	0	7	6	50	0	10
S1-13	48	76	0	18	20	30	0	10	62	98	0	14	3	25	0	10
S1-14	64	169	0	13	27	55	0	7	71	195	0	21	9	37	0	19
S1-15	25	36	0	12	12	12	0	12	56	47	0	24	3	0	0	10
S1-16	36	58	0	4	17	18	0	9	36	79	0	7	4	14	0	2
S1-17	82	73	0	20	36	14	0	14	69	107	0	38	5	41	0	11
S1-18	23	21	0	11	7	13	1	12	27	15	0	21	1	1	0	7
S1-19	84	27	0	1	30	12	0	6	108	29	0	11	36	2	0	3
S1-20	20	66	0	5	14	15	0	5	27	75	0	17	3	41	0	6
S1-21	35	79	0	13	6	43	0	3	26	94	0	17	6	12	0	13
S1-22	8	56	2	20	1	13	2	9	14	85	2	33	2	20	0	41
S1-23	59	94	0	21	29	32	1	11	87	97	0	40	23	23	0	52
S1-24	49	141	0	12	19	35	0	10	40	164	0	21	7	70	0	6
S1-25	32	117	1	15	14	39	1	12	28	136	2	19	3	54	0	28
S1-26	34	130	0	32	16	37	1	11	32	148	0	35	5	70	0	25
S1-27	34	132	0	29	14	40	1	24	34	162	0	32	4	68	0	23
S1-28	34	148	0	14	17	38	0	16	38	171	0	21	3	69	0	9
S1-29	47	148	0	8	19	42	0	8	40	155	0	16	3	70	0	7
S1-30	77	56	0	26	33	16	0	6	59	70	0	42	4	5	0	53
S1-31	82	41	0	24	34	9	0	9	54	56	0	41	3	6	0	45
S1-32	82	99	0	22	33	33	0	6	53	121	0	44	3	11	0	52
S1-33	75	79	0	30	39	20	0	8	49	88	0	46	3	11	0	46
S1-34	74	66	0	25	39	15	0	8	49	75	0	42	3	10	0	49
S1-35	81	64	0	23	38	16	0	9	59	63	0	49	4	6	0	54

Fonte: O autor (2018).

Tabela 4.6: Grupos significativamente expressos encontrados via GAGE para Processos Biológicos (BP), Componentes Celulares (CC), Funções Moleculares (MF) e KEGG *Pathways* referente à *Aedes aegypti* (amostra S1, conforme comparações descritas pela Tabela A.1, Apêndice A).

Comparações	BP	CC	MF	KEGG
S1-1	1	2	1	1
S1-2	1	2	1	1
S1-3	1	2	1	1
S1-4	1	3	1	1
S1-5	1	3	1	2
S1-6	1	2	1	1
S1-7	1	3	1	2
S1-8	1	3	1	2
S1-9	1	3	1	1
S1-10	1	2	1	1
S1-11	1	2	1	1
S1-12	1	2	1	1
S1-13	1	2	1	1
S1-14	1	2	2	1
S1-15	1	3	2	2
S1-16	1	3	2	2
S1-17	1	3	2	2
S1-18	1	2	2	2
S1-19	1	4	1	2
S1-20	2	4	1	2
S1-21	2	4	4	2
S1-22	1	3	2	1
S1-23	2	3	2	2
S1-24	1	4	1	2
S1-25	1	4	2	2
S1-26	2	4	2	2
S1-27	2	3	1	2
S1-28	1	4	2	2
S1-29	1	4	1	2
S1-30	1	2	1	1
S1-31	1	2	1	1
S1-32	1	2	1	1
S1-33	1	2	1	1
S1-34	1	2	1	1
S1-35	1	2	1	1

Fonte: O autor (2018).

Tabela 4.7: Grupos significativamente expressos encontrados via GAGE para Processos Biológicos (BP), Componentes Celulares (CC), Funções Moleculares (MF) e KEGG *Pathways* referente à *Drosophila melanogaster* (amostra S2, conforme comparações descritas pela Tabela A.2, Apêndice A). Vale ressaltar que a combinação S2-2 não apresentou grupos significativos para nenhum domínio ou via metabólica.

Comparações	BP	CC	MF	KEGG
S2-1	1	2	0	1
S2-3	1	1	0	1
S2-4	0	1	0	1
S2-5	1	3	0	1

Fonte: O autor (2018).

Tabela 4.8: Grupos significativamente expressos encontrados via GSVA para Processos Biológicos (BP), Componentes Celulares (CC), Funções Moleculares (MF) e KEGG *Pathways* referente à *Aedes aegypti* (amostra S1, conforme comparações descritas pela Tabela A.1, Apêndice A). No caso desta ferramenta, a análise é realizada individualmente com cada amostra, e não como uma comparação (controle *versus* experimento).

Amostra	BP	CC	MF	KEGG
SRR923701	37	10	51	7
SRR923702	25	13	33	14
SRR923703	44	9	51	10
SRR923704	35	10	57	9
SRR923705	22	15	30	13
SRR923736	28	10	42	12
SRR923822	27	12	32	10
SRR923823	42	9	45	2
SRR923824	41	6	31	12
SRR923825	47	11	46	12
SRR923826	31	9	43	6
SRR923827	40	14	45	6
SRR923828	32	8	46	8
SRR923829	29	15	41	9
SRR923830	36	7	47	7
SRR923831	33	11	36	10
SRR923832	30	9	35	7
SRR923833	40	12	36	6
SRR923834	48	15	47	12
SRR923835	35	8	34	8
SRR923836	35	11	41	6
SRR923837	33	7	32	4
SRR923838	46	19	52	16
SRR923839	54	24	51	9
SRR923840	47	21	58	15
SRR923841	34	10	43	9
SRR923842	34	12	45	6
SRR923843	36	10	33	8
SRR923844	34	7	37	10
SRR923845	33	9	44	10
SRR923846	44	18	54	11
SRR923847	33	8	37	7
SRR923848	23	14	38	8
SRR923849	44	18	48	10
SRR923850	42	12	41	8
SRR923851	35	6	32	5
SRR923852	28	12	42	11
SRR923853	30	8	28	4
SRR923854	27	10	36	12
SRR923855	30	9	34	15
SRR923856	39	10	36	7
SRR923857	34	20	30	1
SRR924021	26	16	39	1
SRR924022	28	16	39	1
SRR924023	27	15	41	1
SRR924024	27	14	42	1

Fonte: O autor (2018).

Tabela 4.9: Grupos significativamente expressos encontrados via GSVA para Processos Biológicos (BP), Componentes Celulares (CC), Funções Moleculares (MF) e KEGG *Pathways* referente à *Drosophila melanogaster* (amostra S2, conforme comparações descritas pela Tabela A.2, Apêndice A). No caso desta ferramenta, a análise é realizada individualmente com cada amostra, e não como uma comparação (controle *versus* experimento).

Amostras	**BP**	**CC**	**MF**	**KEGG**
Controle_1	343	71	124	12
Controle_2	270	64	107	7
Controle_3	360	71	156	7
Controle_4	302	57	138	12
Controle_média	378	77	161	13
Mutante_média	304	69	136	20
Mutante_1	233	61	124	11
Mutante_2	291	61	145	14
Mutante_3	288	64	143	14
Mutante_4	286	82	126	24

Fonte: O autor (2018).

Tabela 4.10: Grupos significativamente expressos encontrados via ClusterProfiler para Processos Biológicos (BP), Componentes Celulares (CC), Funções Moleculares (MF) e KEGG *Pathways* referente à *Drosophila melanogaster* (amostra S2, conforme comparações descritas pela Tabela A.2, Apêndice A). No caso desta ferramenta, a análise é realizada individualmente com cada amostra, e não como uma comparação (controle *versus* experimento).

Amostras	**BP**	**CC**	**MF**	**KEGG**
Controle_1	3	6	3	2
Controle_2	2	7	3	2
Controle_3	2	6	3	1
Controle_4	2	7	3	2
Controle_média	2	7	3	2
Mutante_média	1	8	2	1
Mutante_1	1	8	3	1
Mutante_2	1	8	2	1
Mutante_3	1	7	2	1
Mutante_4	1	7	2	1

Fonte: O autor (2018).

Tabela 4.11: Grupos significativamente expressos encontrados via ClusterProfiler para Processos Biológicos (BP), Componentes Celulares (CC), Funções Moleculares (MF) e KEGG *Pathways* referente à *Aedes aegypti* (amostra S1, conforme comparações descritas pela Tabela A.1, Apêndice A). No caso desta ferramenta, a análise é realizada individualmente com cada amostra, e não como uma comparação (controle *versus* experimento). Vale ressaltar ainda que algumas amostras não apresentaram grupos significativos para nenhum domínio ou via metabólica.

Amostras	BP	CC	MF	KEGG
SRR923704	1	3	1	1
SRR923705	1	2	1	1
SRR923736	1	2	1	1
SRR923823	1	2	1	1
SRR923824	1	3	3	1
SRR923825	1	2	1	1
SRR923826	2	2	1	1
SRR923827	1	3	2	1
SRR923828	1	2	2	1
SRR923831	1	3	2	1
SRR923832	2	2	1	2
SRR923833	1	2	2	1
SRR923834	1	2	1	1
SRR923835	1	2	1	1
SRR923836	1	2	1	1
SRR923837	2	2	1	1
SRR923838	1	3	1	1
SRR923839	1	3	2	1
SRR923840	1	2	2	1
SRR923841	1	2	1	2
SRR923842	1	2	2	1
SRR923843	1	2	1	1
SRR923844	1	2	1	1
SRR923845	1	2	1	1
SRR923846	1	2	2	1
SRR923847	2	2	2	2
SRR923848	1	1	2	1
SRR923849	1	2	1	2
SRR923850	1	2	1	2
SRR923851	1	2	1	1
SRR923852	1	2	1	1
SRR923853	1	2	1	1
SRR923855	2	2	1	1
SRR923856	1	2	1	1

Fonte: O autor (2018).

Assim como o *EntropyClusterGenes*, o *GAGE* também permite que as amostras sejam comparadas duas a duas (controle vs experimento). Entretanto, *GSVA* e *ClusterProfiler* apresentam um resultado em que os grupos significativos são visualizados individualmente em cada uma das amostras.

A partir das tabelas, nota-se que a quantidade de grupos significativos encontrados com base na diversidade (h) e atividade gênica (n), na maior parte dos casos, é superior aos obtidos com os

testes de *Wilcoxon* e *Fisher*. A diversidade se baseia na entropia interna do GFAG, levando em conta o quão similares entre si são os valores de expressão gênica dos elementos que compõe o grupo. A atividade nada mais é do que um somatório dos valores de expressão de cada gene dentro do grupo. Uma atividade alta depende, exclusivamente, de valores de expressão elevados dos componentes do grupo, ao passo que a diversidade não depende exclusivamente da expressão em si, e sim da maneira como esta varia dentro do GFAG. Dessa forma, uma possível explicação para o padrão de variação entre atividade e diversidade encontrado pode estar relacionado ao design experimental utilizado nos estudos. As diferentes condições a que os mosquitos foram submetidos antes do sequenciamento podem ocasionar variações de expressão em grupos específicos de genes, refletindo diretamente sobre a atividade e diversidade.

4.5 *Benchmarking entre Ferramentas GSEA*

Como mencionado anteriormente, o *EntropyClusterGenes* foi comparado a outras três ferramentas GSEA: *GAGE*, *GSVA* e *ClusterProfiler*. Uma comparação entre seus resultados, para *A. aegypti* e *D. melanogaster*, pode ser observado através da Tabela 4.12 e Tabela 4.13.

Tabela 4.12: Similaridade de ferramentas com relação ao *EntropyClusterGenes* com base nos resultados da análise envolvendo *Aedes aegypti*.

Ferramentas	**BP**		**CC**		**MF**		**KEGG**	
	h	***n***	***h***	***n***	***h***	***n***	***h***	***n***
GAGE	1 (50%)	2 (100%)	3 (75%)	4 (100%)	3 (75%)	4 (100%)	2 (100%)	2 (100%)
GSVA	252 (39.68%)	285 (44.88%)	83 (38.07%)	103 (47.25%)	299 (44.63%)	336 (50.15%)	87 (79.82%)	104 (95.41%)
ClusterProfiler	4 (100%)	4 (100%)	3 (100%)	3 (100%)	6 (100%)	6 (100%)	2 (100%)	2 (100%)

Fonte: O autor (2018).

Tabela 4.13: Similaridade de ferramentas com relação ao *EntropyClusterGenes* com base nos resultados da análise envolvendo *Drosophila melanogaster*.

Ferramentas	**BP**		**CC**		**MF**		**KEGG**	
	h	***n***	***h***	***n***	***h***	***n***	***h***	***n***
GAGE	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
GSVA	394 (18.92%)	602 (28.91%)	94 (21.08%)	143 (32.06%)	195 (21.22%)	241 (26.22%)	11 (14.86%)	37 (50%)
ClusterProfiler	2 (66.67%)	1 (33.33%)	1 (12.5%)	2 (25%)	1 (25%)	1 (25%)	1 (50%)	0 (0%)

Fonte: O autor (2018).

Ao analisarmos os resultados obtidos a partir do *Aedes aegypti*, verificamos que todos os grupos significativos encontrados pelo *ClusterProfiler* também foram identificados pelo *EntropyClusterGenes*. Entretanto, assim como o *GAGE*, que também apresentou um percentual de similaridade elevado, a quantidade de GFAGs significativos é baixa. Quanto ao *GSVA*, ve-

mos uma taxa de similaridade abaixo ou próximo de 50% para BP, CC e MF. Porém, quando observamos KEGG, essa taxa sobe para mais de 79%.

Quando analisamos os resultados referentes à *Drosophila melanogaster*, o cenário para *ClusterProfiler* e *GAGE* se inverte. Enquanto a primeira apresenta taxas que variam de zero a pouco mais de 66%, a segunda, para todos os domínios e KEGG mostra similaridade zero. Os percentuais de similaridade de *GSVA* também reduzem. No entanto, a exemplo do observado em *A. aegypti*, a maior taxa ainda é encontrada na atividade de KEGG (50%).

Dessa forma, embora, na maioria dos casos, a taxa de similaridade é sempre superior a zero, notamos variações significativas de uma espécie para outra. Mesmo o protocolo de análise de RNA-seq utilizado ser o mesmo tanto para *Aedes* quanto para *Drosophila*, a forma como as amostras foram preparadas para o sequenciamento, ou mesmo o foco do estudo, impacta diretamente nos resultados de enriquecimento funcional. Além disso, mesmo pertencentes à mesma classe (GSEA), os métodos estatísticos utilizados variam, aspecto que também influencia no resultado final.

4.6 *Diversidade Gênica*

A diversidade gênica é o principal diferencial do *EntropyClusterGenes* em relação às demais ferramentas de enriquecimento funcional da classe GSEA (*Gene Set Enrichment Analysis*). Como uma medida de entropia, embasada nos conceitos da *Teoria da Informação de Shannon*, é possível classificar grupos de genes funcionalmente associados (GFAGs) com base na variação de expressão gênica de cada conjunto. A partir dos resultados gerados pela nova análise, sobretudo as que se referem a *Aedes aegypti*, notamos um padrão de diversidade ao longo de amostras ligadas às fases iniciais de desenvolvimento do mosquito, como pode ser observado pelas Figura 4.3 e Figura 4.4.

Considerando os processos biológicos referenciados pelos códigos GO:0007275 (*Multicellular organism development*) e GO:0007179 (*Transforming growth factor beta receptor signaling pathway*), associados ao desenvolvimento embrionário e aos quatro estágios de desenvolvimento larval, verificamos a existência de um aumento na diversidade com o passar das etapas, embora, em momentos específicos, tenham ocorrido algumas quedas. Nota-se, ainda, que os valores de diversidade ficaram numa faixa, entre 0.39 e 0.72. Ao considerarmos as funções moleculares representadas pelos códigos GO:0003676 (*Nucleic acid binding*) e GO:0005515 (*Protein binding*), entretanto, referentes às mesmas fases de desenvolvimento, constatamos um padrão de decréscimo na diversidade, ficando num intervalo entre 0.56 e 0.83. Vale ainda res-

saltar que essas duas funções moleculares estão relacionadas aos principais genes encontrados por Akbari e colaboradores (AKBARI et al., 2013) no mesmo conjunto amostral utilizado neste trabalho, descritos de forma mais detalhada na Seção 4.7.

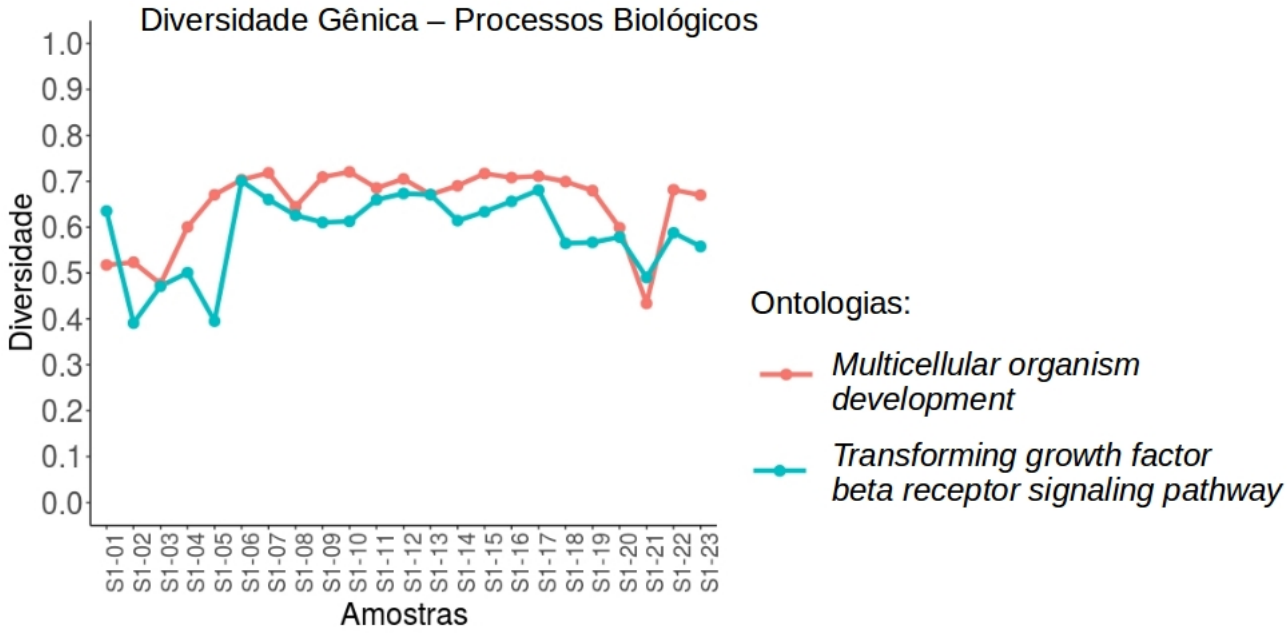

Figura 4.3: Estudo de diversidade gênica referente aos estágios iniciais de desenvolvimento do mosquito *Aedes aegypti* referente a processos biológicos (BP).

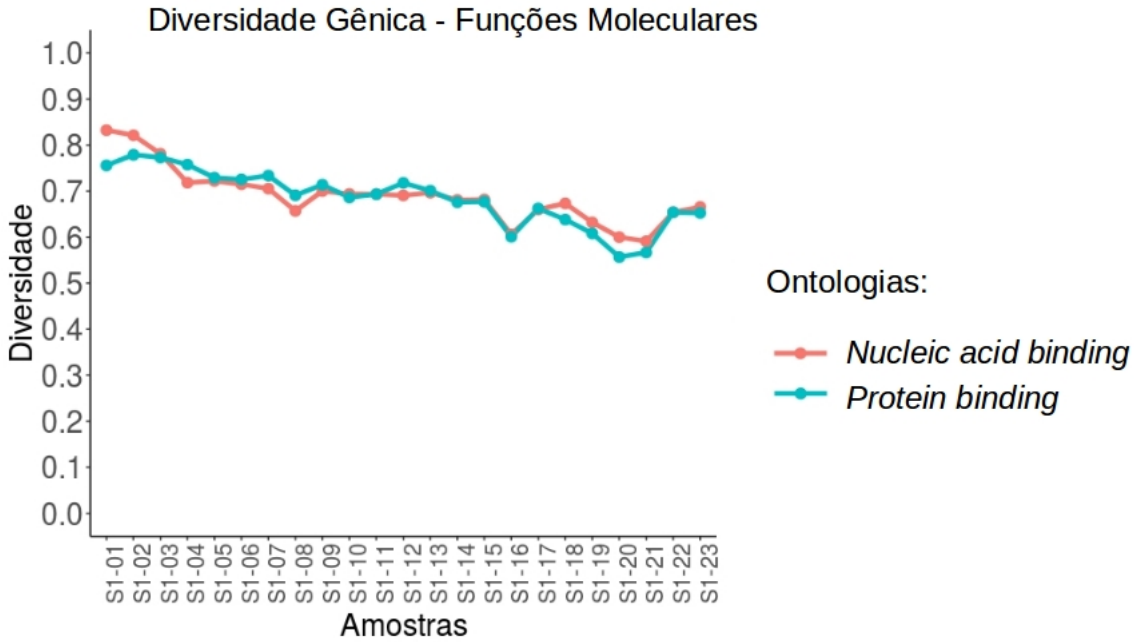

Figura 4.4: Estudo de diversidade gênica referente aos estágios iniciais de desenvolvimento do mosquito *Aedes aegypti* referente a funções moleculares (MF).

4.7 *Comparação com Estudos Originais - Aedes aegypti*

Como descrito anteriormente na Seção 3.2, utilizamos conjuntos amostrais de *A. aegypti* baseados em experimentos de RNA-seq referentes a um trabalho conduzido por Akbari e colaboradores (AKBARI et al., 2013), cujo propósito era a construção de um transcriptoma com foco no desenvolvimento do mosquito. Foram utilizadas 46 amostras de indivíduos da cepa Liverpool, considerando carcaça e ovários de fêmeas alimentadas com sangue e solução de sacarose, além de embriões, larvas e pupas. Dentre os seus principais resultados, damos destaque à aqueles relacionados a pequenos RNAs, conforme a Figura 4.5. No ovário, por exemplo, em resposta à alimentação com sangue, houve a elevação dos níveis de expressão de determinados genes.

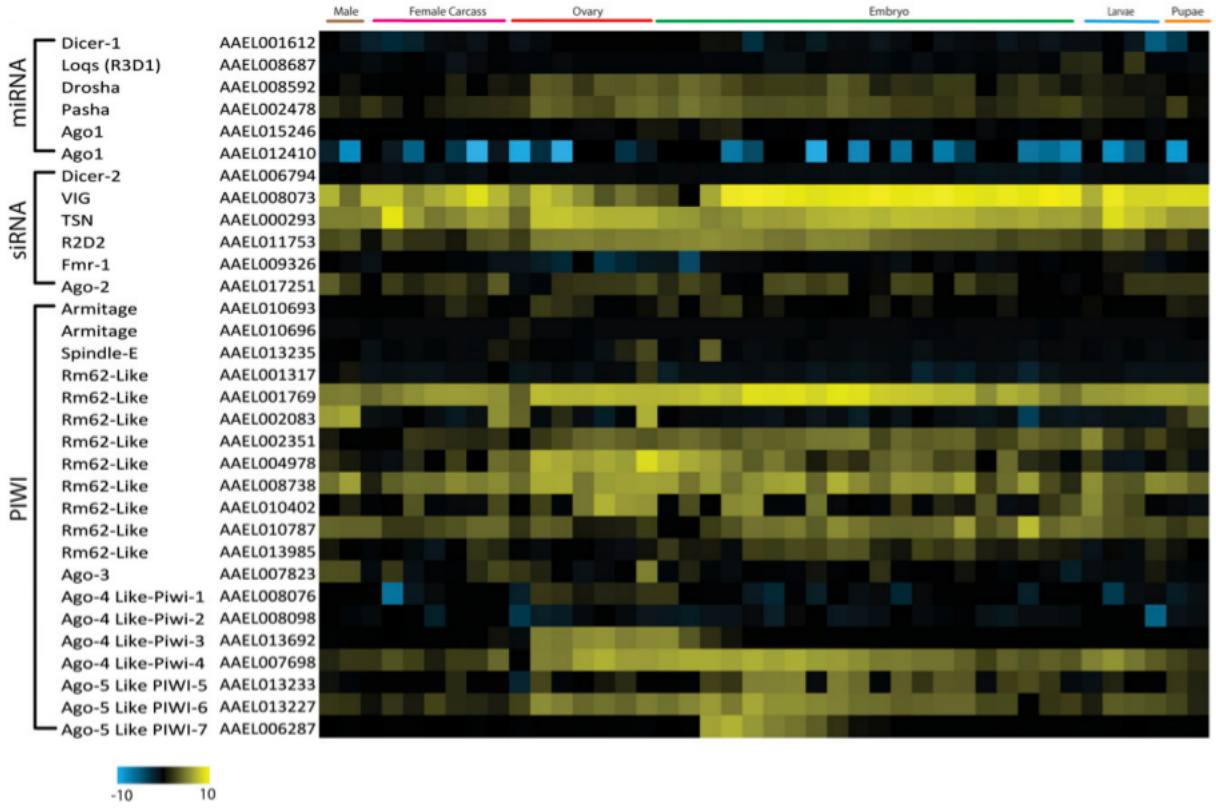

Figura 4.5: Perfis de expressão de genes envolvidos na produção de pequenos RNAs. Através de um *heatmap*, é possível observar a dinâmica de expressão, ao longo do desenvolvimento do mosquito, de genes importantes para o processamento de pequenos RNAs, incluindo miRNAs, siRNAs e piRNAs. [Figura adaptada de (AKBARI et al., 2013).]

Com base nos genes da Figura 4.5, buscamos por grupos significativos a eles relacionados a partir do enriquecimento funcional referente à atividade gênica, feita via *EntropyClusterGenes*, conforme a Tabela 4.14.

A partir da Tabela 4.14, observamos maior predominância da GO:0003676, detalhada pela

Figura 4.6, em duas diferentes comparações ligadas ao estado larval do inseto. Trata-se de uma função molecular associada à interação seletiva e não covalente com qualquer ácido nucleico, neste caso, mostrando-se mais ativa nas fases iniciais de desenvolvimento do organismo. O estudo menciona um aumento no nível de expressão de alguns genes quando analisado ovários de fêmeas alimentadas com sangue. Entretanto, a partir da análise de enriquecimento funcional, notamos, na verdade, uma redução na atividade gênica em alguns grupos significativos, como, por exemplo, a GO:0008255 (*ecdysis-triggering hormone activity*), uma função molecular com 5 genes a ela associados.

Tabela 4.14: Relação entre genes obtidos por Akbari e colaboradores (AKBARI et al., 2013) e Grupos de Genes Funcionalmente Associados e significativos (GFAGs).

Gene	**ID Grupo**	**Descrição GO**	**Comparação**	**Dominio**
AAEL008076	GO:0005515	*protein binding*	*4-8h Embryo x 8-12h Embryo*	MF
AAEL001317	GO:0003676	*nucleic acid binding*	*2nd Instar Larvae x 3rd Instar Larvae*	MF
AAEL001769	GO:0003676	*nucleic acid binding*	*2nd Instar Larvae x 3rd Instar Larvae*	MF
AAEL002083	GO:0003676	*nucleic acid binding*	*1nd Instar Larvae x 2rd Instar Larvae*	MF
AAEL002351	GO:0003676	*nucleic acid binding*	*2nd Instar Larvae x 3rd Instar Larvae*	MF
AAEL006287	GO:0003676	*nucleic acid binding*	*1nd Instar Larvae x 2rd Instar Larvae*	MF
AAEL006287	GO:0003676	*nucleic acid binding*	*2nd Instar Larvae x 3rd Instar Larvae*	MF
AAEL006794	GO:0003676	*nucleic acid binding*	*2nd Instar Larvae x 3rd Instar Larvae*	MF
AAEL007823	GO:0003676	*nucleic acid binding*	*1nd Instar Larvae x 2rd Instar Larvae*	MF
AAEL007823	GO:0003676	*nucleic acid binding*	*2nd Instar Larvae x 3rd Instar Larvae*	MF
AAEL008076	GO:0003676	*nucleic acid binding*	*1nd Instar Larvae x 2rd Instar Larvae*	MF
AAEL008076	GO:0003676	*nucleic acid binding*	*2nd Instar Larvae x 3rd Instar Larvae*	MF
AAEL008098	GO:0003676	*nucleic acid binding*	*1nd Instar Larvae x 2rd Instar Larvae*	MF
AAEL008098	GO:0003676	*nucleic acid binding*	*2nd Instar Larvae x 3rd Instar Larvae*	MF
AAEL013235	GO:0003676	*nucleic acid binding*	*1nd Instar Larvae x 2rd Instar Larvae*	MF
AAEL013235	GO:0003676	*nucleic acid binding*	*2nd Instar Larvae x 3rd Instar Larvae*	MF
AAEL013692	GO:0003676	*nucleic acid binding*	*2nd Instar Larvae x 3rd Instar Larvae*	MF
AAEL013692	GO:0003676	*nucleic acid binding*	*1nd Instar Larvae x 2rd Instar Larvae*	MF
AAEL013985	GO:0003676	*nucleic acid binding*	*2nd Instar Larvae x 3rd Instar Larvae*	MF
AAEL013985	GO:0003676	*nucleic acid binding*	*1nd Instar Larvae x 2rd Instar Larvae*	MF
AAEL015246	GO:0003676	*nucleic acid binding*	*2nd Instar Larvae x 3rd Instar Larvae*	MF
AAEL017251	GO:0003676	*nucleic acid binding*	*1nd Instar Larvae x 2rd Instar Larvae*	MF
AAEL017251	GO:0003676	*nucleic acid binding*	*2nd Instar Larvae x 3rd Instar Larvae*	MF
AAEL008592	aag03008	*Ribosome biogenesis in eukaryotes*	*2nd Instar Larvae x 3rd Instar Larvae*	KEGG

Fonte: O autor (2018).

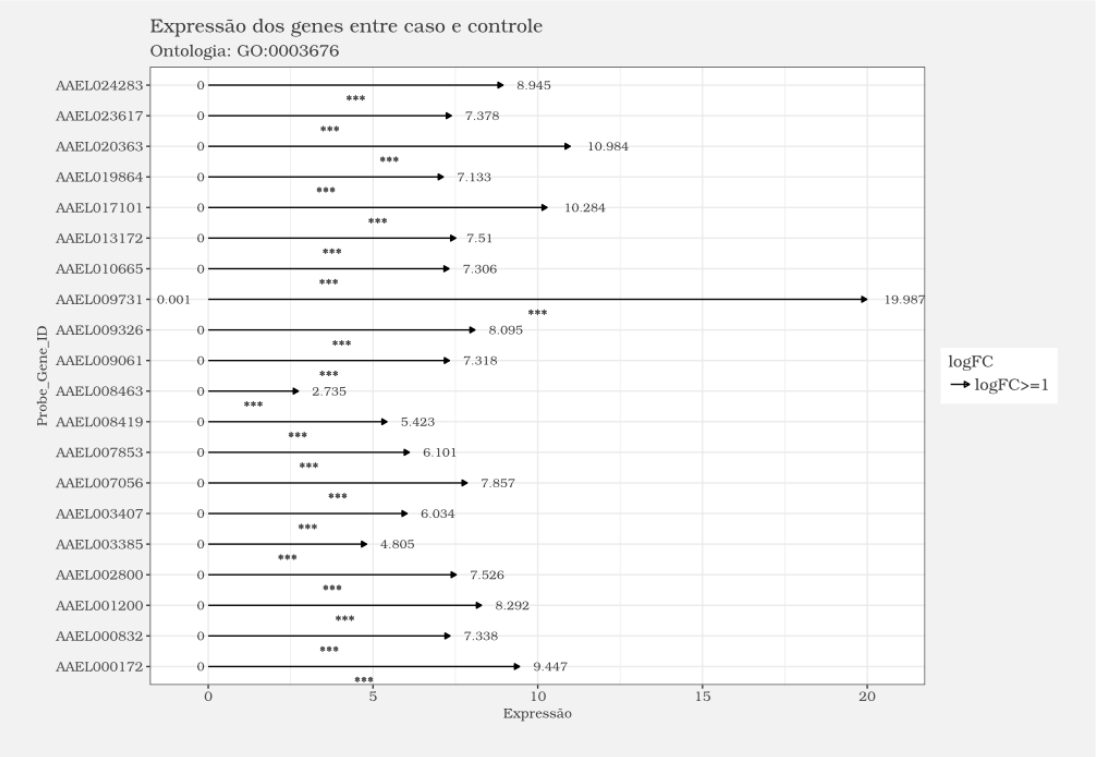

Figura 4.6: Conjunto detalhando um grupo de genes pertencentes à GO:0003676. As setas indicam a variação de expressão do gene (controle à esquerda e experimento à direita). A quantidade de asteríscos em cada uma das setas indica a significância estatística (* = q-valor menor que 0.05; ** = q-valor menor que 0.01 e *** = q-valor menor que 0.001). [Figura obtida via módulo gráfico do *EntropyClusterGenes*.]

Levando-se em consideração a diversidade gênica, damos destaque para a função molecular GO:0042302 (*structural constituent of cuticle*), presente em 7 das comparações envolvendo as fases embrionárias do mosquito, composta por 204 genes e com diversidade relativa (*h*) sempre acima de 0.5, mostrando que a diversidade do experimento é sempre maior do que no controle, como mostra a Tabela 4.15.

Tabela 4.15: Diversidade relativa fase embrionária *Aedes aegypti*, dando destaque para a GO:0042302, (*structural constituent of cuticle*).

Comparações	**Diversidade Relativa (h)**
4-4hr Embryo x 4-8hr Embryo	0.5824525
12-16hr Embryo x 16-20hr Embryo	0.5305274
20-24hr Embryo x 24-28hr Embryo	0.5191541
28-32hr Embryo x 32-36hr Embryo	0.5385458
32-36hr Embryo x 36-40hr Embryo	0.6219089
36-40hr Embryo x 40-44hr Embryo	0.5436838
48-52hr Embryo x 52-56hr Embryo	0.5228558

Fonte: O autor (2018).

Por se tratar de uma fase inicial de desenvolvimento, com a formação das estruturas necessárias para a vida do mosquito, faz sentido que este tipo de função seja manifestada de forma significativa. Dos 204 genes componentes desta ontologia, de acordo com os autores, 159 são diferencialmente expressos, apresentando *fold change* sempre superior a 1, tanto para *up-regulated* quanto para genes *down-regulated*. Alguns deles podem ser visualizados a partir da Figura 4.7. Desta forma, podemos supor que expressão diferencial é um dos fatores envolvidos na diversidade, uma vez que a mesma função, quanto à atividade, não se mostra superexpressa como na diversidade.

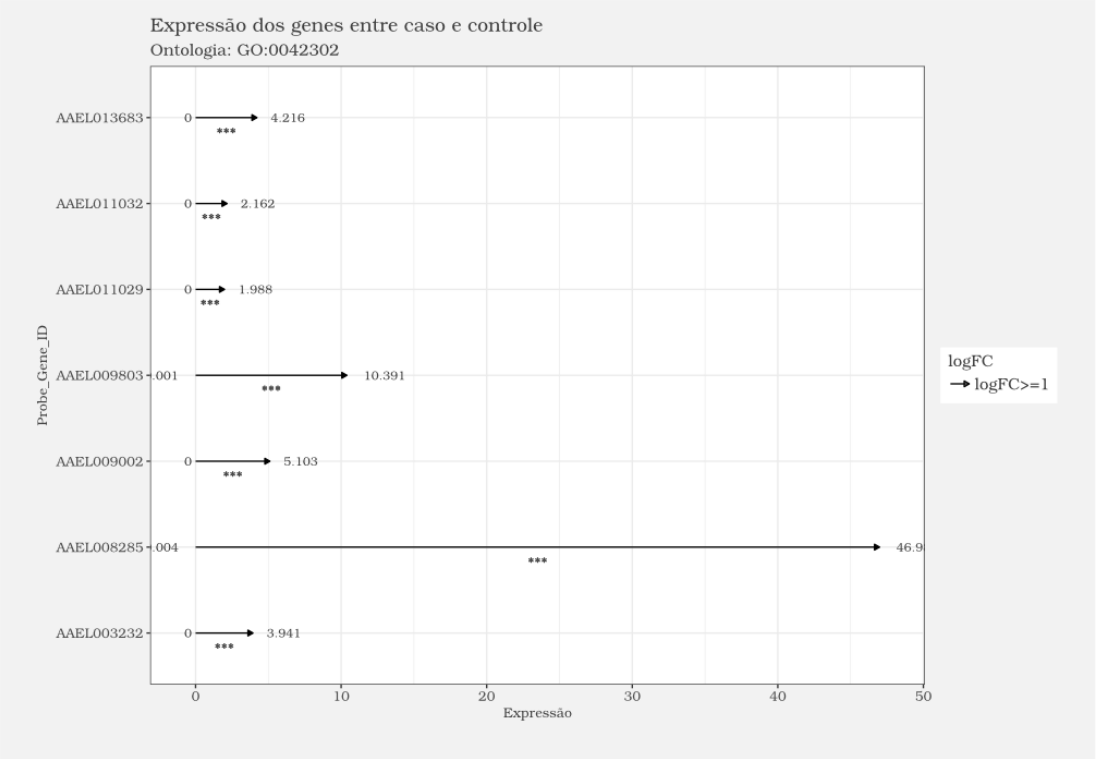

Figura 4.7: Conjunto detalhando um grupo de genes pertencentes à GO:0042302. As setas indicam a variação de expressão do gene (controle à esquerda e experimento à direita). A quantidade de asteríscos em cada uma das setas indica a significância estatística (* = q-valor menor que 0.05; ** = q-valor menor que 0.01 e *** = q-valor menor que 0.001). [Figura obtida via módulo gráfico do *EntropyClusterGenes*.]

4.8 *Comparação com Estudos Originais - Drosophila melanogaster*

Fatores de transcrição são proteínas que se ligam à sequências de DNA regulatório (*enhancers* ou *silencers*) e modulam a taxa de transcrição gênica. Eles podem ser ativados no interior do núcleo, frequentemente com o fator previamente ligado ao DNA, ou no interior do citoplasma, resultando na exposição de sinais de localização nuclear e direcionamento para o

núcleo. Existem várias famílias de fatores transcricionais, podendo, cada uma delas, compartilhar características estruturais (BARNES, 2000). Alguns exemplos são mencionadas a seguir:

- *helix-turn-helix* (Oct-1);
- *helix-loop-helix* (E2A);
- *zinc finger* (receptores glucocorticoides, proteínas GATA);
- *basic protein-leucine zipper* (CREB, AP-1);
- β*-sheet motifs* (NF-kB).

Receptores nucleares agem como fatores de transcrição ligante-dependentes que regulam diretamente a transcrição do gene alvo em resposta a um sinal hormonal (MANGELSDORF et al., 1995). O genoma da *drosophila* codifica 18 membros dessa família de receptores, sendo, em sua maioria, receptores nucleares órfãos, uma vez que não possuem ligantes hormonais conhecidos. Um exemplo é o DHR78, que participa da muda de cutícula traqueal regulando a expressão gênica durante o terceiro estágio larval (ASTLE; KOZLOVA; THUMMEL, 2003).

Em um estudo de Marxreiter e Thummel (MARXREITER; THUMMEL, 2018) sobre funções adultas do receptor nuclear DHR78 em *Drosophila*, trabalhando com fêmeas normais e mutantes para o receptor, foram encontrados 510 genes diferencialmente expressos nos mutantes DHR78, conforme os autores, (*fold change* maior que 1.5 e FDR 0.01), com 283 genes apresentando abundância reduzida e 227 aumentada. Uma análise de ontologias mostrou que muitos genes codificadores de enzimas com atividade de oxidoredutase predita são expressos em níveis mais elevados em mutantes sendo, a maior parte, referente à enzimas citocromo P450. Outras categorias de genes *up-regulated* correspondem à vias de defesa. Quanto a genes *down-regulated*, a análise de GOs mostrou uma predominância de endopeptidases ao longo dos transportadores transmembrânicos e componentes da matriz peritrófica. As ontologias de destaque podem ser observadas através da Tabela 4.16.

Dentre as ontologias descritas na Tabela 4.16, a partir dos resultados do *EntropyCluster-Genes* quanto à diversidade, encontramos a GO:0004175 (*Endopeptidase activity*), 47 genes, dos quais 19 são diferencialmente expressos (7 *down*-regulados e 12 *up*-regulados), alguns podendo ser vistos a partir da Figura 4.8. Tal expressão diferencial mostra mais uma vez, a exemplo do ocorrido em *Aedes aegypi*, a relação entre diversidade e DEGs. Vale ainda destacar a GO:0003012 (*muscle system process*), processo biológico significativamene expresso quanto à diversidade gênica, presente em 4 das 5 comparações, com 7 genes dos quais 2 são diferencialmente expressos.

Tabela 4.16: Ontologias de destaque conforme estudo sobre *Drosophila*.

Ontologias de Destaque
Endopeptidase activity
Urea cycle intermediate
Transmembrane transporter activity
Peritrophic matrix
Lysozymes
Hydrolase acting on glycosyl bonds
Arginine metabolism
Oxidoreductase activity
Defense response
Electron carrier (P450s)
Defense response to bacteria
Response to stimulus
Peptidase activity

Fonte: Adaptada de (MARXREITER; THUMMEL, 2018).

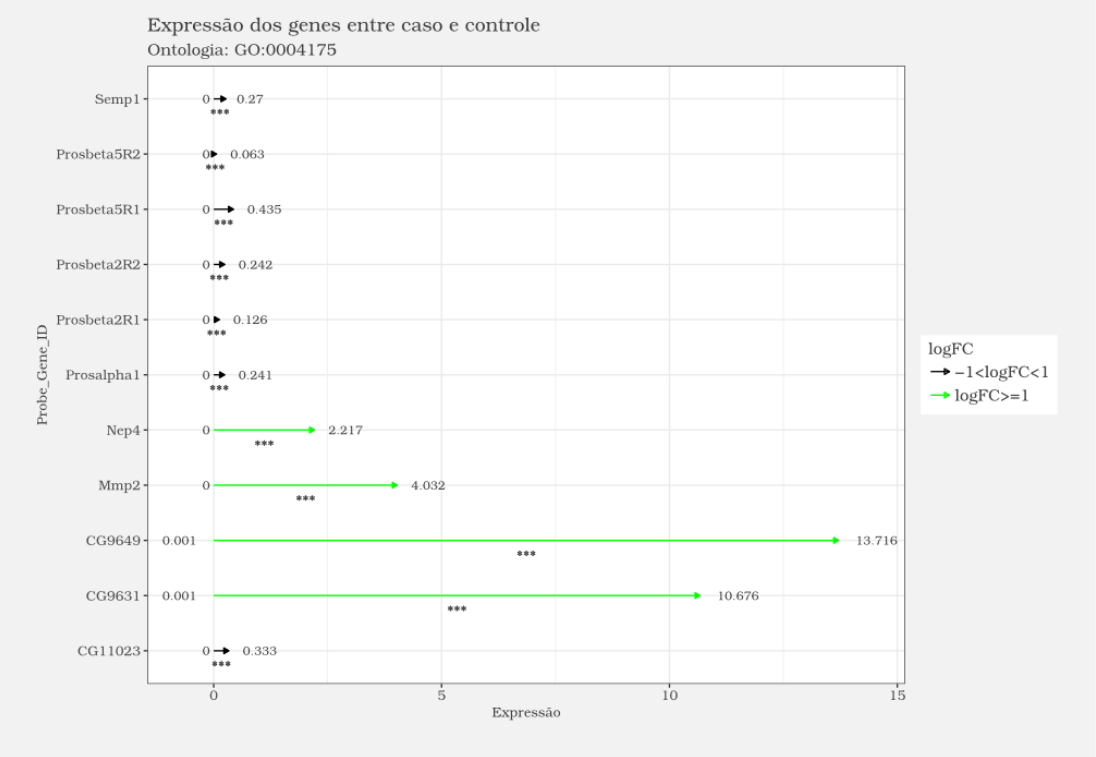

Figura 4.8: Conjunto detalhando um grupo de genes pertencentes à GO:0004175. As setas indicam a variação de expressão do gene (controle à esquerda e experimento à direita). A quantidade de asteríscos em cada uma das setas indica a significância estatística (* = q-valor menor que 0.05; ** = q-valor menor que 0.01 e *** = q-valor menor que 0.001). [Figura obtida via módulo gráfico do *EntropyClusterGenes*.]

5 *Conclusão*

Desenvolvemos um pacote em ambiente de programação R que se mostrou eficiente na detecção de GFAGs significativos em grandes volumes de dados de RNA-seq de *Aedes aegypti* e *Drosophila melanogaster*. O módulo de detecção de espécie funcionou adequadamente, relacionando de maneira correta os genes amostrais às suas respectivas ontologias (processos biológicas, funções moleculares e componentes celulares) e KEGGs *pathways*. Construímos, ainda, um conjunto de tabelas interligadas, a partir de repositórios online, para que pudéssemos demonstrar a ferramenta e, ao mesmo tempo, estudar uma espécie em evidência no cenário nacional atual por se tratar do principal vetor do vírus causador da dengue.

Uma das dificuldades encontradas quanto à implementação do pacote foi a velocidade de cálculo do *bootstrap*. Por utilizarmos, inicialmente, uma versão codificada seguindo uma linha estrutural e laços de repetição do tipo *FOR*, o tempo gasto para a determinação dos p-valores era elevado, o que deixava a aplicação lenta. Entretanto, ao alterarmos a forma de programação, passando para um código híbrido (R e C++), reduzimos substancialmente o tempo de processamento, passando de 37 minutos com uma versão paralelizada para 2 minutos com a nova forma de codificação em linguagem C++.

Com a realização do *benchmarking*, notamos que os resultados alcançados pelo *EntropyClusterGenes* se assemelham aos resultados das ferramentas GAGE, GSVA e *ClusterProfiler*. Entretanto, tal taxa de similaridade reduz quando passamos de *A. aegypti* para *D. melanogaster*. A ferramenta GAGE, por exemplo, que apresentava 100% de equidade na primeira espécie, caiu para zero na segunda. Tal ocorrência se deve, muito provavelmente, ao preparo das amostras para sequenciamento, tipo de estudo em questão e diferentes testes estatísticos utilizados pelos outros pacotes.

Em relação à comparação dos resultados do *EntropyClusterGene* com os estudos referentes às amostras utilizadas, dentre os GFAGs de maior relevância para *Aedes aegypti*, destacamos a GO:0003676 (*nucleic acid binding*), função molecular presente de maneira significativa principalmente nas fases larvais do inseto, e a GO:0042302 (*structural constituent of cuticle*),

significativa quanto à diversidade, presente na fase embrionária do inseto e com 159 genes diferencialmente expressos de um total de 204. Quanto à *Drosophila melanogaster*, nossa ferramenta também foi capaz de encontrar grupos significativos idênticos aos grupos destacados pelo estudo. A GO:0004175 (*Endopeptidase activity*), com 47 genes dos quais 19 são diferencialmente expressos, apresentou-se signicativa quanto à diversidade, corroborando que a presença de genes difencialmente expressos contribuem para grupos significativos sob o aspecto da diversidade gênica.

Como perspectivas futuras, transformaremos o *EntropyClusterGenes* em uma biblioteca (*library*) nos termos do repositório *online Bioconductor*.

Referências Bibliográficas

AKBARI, O. S. et al. The developmental transcriptome of the mosquito aedes aegypti, an invasive species and major arbovirus vector. *G3: Genes— Genomes— Genetics*, Genetics Society of America, v. 3, n. 9, p. 1493–1509, 2013.

ALEXA, A.; RAHNENFUHRER, J. *topGO: Enrichment Analysis for Gene Ontology. R Package Version 2.18. 0.; 2010*. 2010.

ANDERSON, D. R.; BURNHAM, K. P.; THOMPSON, W. L. Null hypothesis testing: problems, prevalence, and an alternative. *The journal of wildlife management*, JSTOR, p. 912–923, 2000.

ANSORGE, W. J. Next-generation dna sequencing techniques. *New biotechnology*, Elsevier, v. 25, n. 4, p. 195–203, 2009.

ASHBURNER, M. et al. Gene ontology: tool for the unification of biology. *Nature genetics*, Nature Publishing Group, v. 25, n. 1, p. 25–29, 2000.

ASTLE, J.; KOZLOVA, T.; THUMMEL, C. S. Essential roles for the dhr78 orphan nuclear receptor during molting of the drosophila tracheal system. *Insect biochemistry and molecular biology*, Elsevier, v. 33, n. 12, p. 1201–1209, 2003.

BAKER, M. et al. Statisticians issue warning on p values. *Nature*, Macmillan Publishers Ltd., London, England, v. 531, n. 7593, p. 151–151, 2016.

BARNES, P. J. Mechanisms in copd: differences from asthma. *Chest*, Elsevier, v. 117, n. 2, p. 10S–14S, 2000.

BEISSBARTH, T.; SPEED, T. P. Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, Oxford University Press, v. 20, n. 9, p. 1464–1465, 2004.

BIOCHEMISTRIES. *Of Nanopores and Isoforms*. maio 2015. DisponÃvel em: http://biochemistri.es/post/119865709426/of-nanopores-and-isoforms. Acesso em: 08-Jan-2018.

BIOLOGY NOTES HELP. *DNA Sequencing*. 2017. DisponÃvel em: http://www.biologynoteshelp.com/human-genome-projectillumina-sequence/. Acesso em: 8-Jan-2018.

BOOS, D. D. et al. Introduction to the bootstrap world. *Statistical science*, Institute of Mathematical Statistics, v. 18, n. 2, p. 168–174, 2003.

BRUNK, H.; HOLSTEIN, J. E.; WILLIAMS, F. The teacher's corner: A comparison of binomial approximations to the hypergeometric distribution. *The American Statistician*, Taylor & Francis Group, v. 22, n. 1, p. 24–26, 1968.

BURROWS, M.; WHEELER, D. J. A block-sorting lossless data compression algorithm. Citeseer, 1994.

CARLSON, M. Go. db: A set of annotation maps describing the entire gene ontology. 2015. *R package version*, v. 3, n. 0, 2013.

CARLSON, M. *org.Rn.eg.db: Genome wide annotation for Rat*. [S.l.], 2017. R package version 3.4.1.

CASTRO, M. A. et al. Viacomplex: software for landscape analysis of gene expression networks in genomic context. *Bioinformatics*, Oxford Univ Press, v. 25, n. 11, p. 1468–1469, 2009.

CASTRO, M. A. et al. Impaired expression of ner gene network in sporadic solid tumors. *Nucleic acids research*, Oxford University Press, v. 35, n. 6, p. 1859–1867, 2007.

CASTRO, M. A. et al. Profiling cytogenetic diversity with entropy-based karyotypic analysis. *Journal of theoretical biology*, Elsevier, v. 234, n. 4, p. 487–495, 2005.

COCK, P. J. et al. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic acids research*, Oxford Univ Press, v. 38, n. 6, p. 1767–1771, 2010.

DERSIMONIAN, R.; LAIRD, N. Meta-analysis in clinical trials. *Controlled clinical trials*, Elsevier, v. 7, n. 3, p. 177–188, 1986.

DOCHERTY, A. B.; LONE, N. I. Exploiting big data for critical care research. *Current opinion in critical care*, LWW, v. 21, n. 5, p. 467–472, 2015.

DORAN, A. G.; CREEVEY, C. J. Snpdat: Easy and rapid annotation of results from de novo snp discovery projects for model and non-model organisms. *BMC bioinformatics*, BioMed Central, v. 14, n. 1, p. 1, 2013.

DURINCK, S. et al. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, v. 21, p. 3439–3440, 2005.

DURINCK, S. et al. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature Protocols*, v. 4, p. 1184–1191, 2009.

EFRON, B. The 1977 rietz lecture. *The Annals of Statistics*, v. 7, n. 1, p. 1–26, 1979.

FISHER, R. A. *Statistical methods for research workers*. 5th. ed. Edinburgh: Oliver & Boyd, 1934.

FISHER, R. A. The logic of inductive inference. *Journal of the Royal Statistical Society*, JSTOR, v. 98, n. 1, p. 39–82, 1935.

GATENBY, R. A.; FRIEDEN, B. R. Application of information theory and extreme physical information to carcinogenesis. *Cancer Research*, AACR, v. 62, n. 13, p. 3675–3684, 2002.

GELMAN, A. Commentary: P values and statistical practice. *Epidemiology*, LWW, v. 24, n. 1, p. 69–72, 2013.

GENTLEMAN, R. Basic go usage. 2016.

GENTLEMAN, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, BioMed Central, v. 5, n. 10, p. R80, 2004.

GLICKMAN, M. E.; RAO, S. R.; SCHULTZ, M. R. False discovery rate control is a recommended alternative to bonferroni-type adjustments in health studies. *Journal of clinical epidemiology*, Elsevier, v. 67, n. 8, p. 850–857, 2014.

GOEMAN, J. J.; MANSMANN, U. Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics*, Oxford University Press, v. 24, n. 4, p. 537–544, 2008.

GOFF, L. A.; TRAPNELL, C.; KELLEY, D. Cummerbund: visualization and exploration of cufflinks high-throughput sequencing data. *R package version*, v. 2, n. 0, 2012.

HÄNZELMANN, S.; CASTELO, R.; GUINNEY, J. Gsva: The gene set variation analysis package for microarray and rna-seq data. 2013.

HEATHER, J. M.; CHAIN, B. The sequence of sequencers: the history of sequencing dna. *Genomics*, Elsevier, v. 107, n. 1, p. 1–8, 2016.

HOEHNDORF, R.; DUMONTIER, M.; GKOUTOS, G. V. Identifying aberrant pathways through integrated analysis of knowledge in pharmacogenomics. *Bioinformatics*, Oxford University Press, v. 28, n. 16, p. 2169–2175, 2012.

HOLLANDER, M.; WOLFE, D. A. Nonparametric statistical methods. Wiley-Interscience, 1999.

HUANG, D. W.; SHERMAN, B. T.; LEMPICKI, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, Oxford University Press, v. 37, n. 1, p. 1–13, 2008.

HUANG, D. W. et al. The david gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome biology*, BioMed Central, v. 8, n. 9, p. R183, 2007.

KANEHISA, M.; GOTO, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, Oxford University Press, v. 28, n. 1, p. 27–30, 2000.

KANEHISA, M. et al. Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, Oxford University Press, v. 38, n. suppl_1, p. D355–D360, 2009.

KEGG. *KEGG Overview*. 2017. Disponivel em: http://www.genome.jp/kegg/kegg1a.html. Acesso: 13-Out-2017.

KEMP, C. D.; KEMP, A. W. Generalized hypergeometric distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 202–211, 1956.

KENDAL, W. S. The use of information theory to analyze genomic changes in neoplasia. *Mathematical biosciences*, Elsevier, v. 100, n. 2, p. 143–159, 1990.

KHODAKOV, D.; WANG, C.; ZHANG, D. Y. Diagnostics based on nucleic acid sequence variant profiling: Pcr, hybridization, and ngs approaches. *Advanced drug delivery reviews*, Elsevier, v. 105, p. 3–19, 2016.

KIM, S.-Y.; VOLSKY, D. J. Page: parametric analysis of gene set enrichment. *BMC bioinformatics*, BioMed Central, v. 6, n. 1, p. 144, 2005.

KODAMA, Y.; SHUMWAY, M.; LEINONEN, R. The sequence read archive: explosive growth of sequencing data. *Nucleic acids research*, Oxford Univ Press, v. 40, n. D1, p. D54–D56, 2012.

KORPELAINEN, E. et al. *RNA-seq Data Analysis: A Practical Approach*. [S.l.]: CRC Press, 2014.

LANGMEAD, B. et al. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, BioMed Central, v. 10, n. 3, p. 1, 2009.

LEINONEN, R.; SUGAWARA, H.; SHUMWAY, M. The sequence read archive. *Nucleic acids research*, Oxford Univ Press, p. gkq1019, 2010.

LEPAGE, R.; BILLARD, L. *Exploring the limits of bootstrap*. [S.l.]: John Wiley & Sons, 1992.

LIN, Y.-T.; LEE, W.-C. Importance of presenting the variability of the false discovery rate control. *BMC genetics*, BioMed Central, v. 16, n. 1, p. 97, 2015.

LIU, L. et al. Comparison of next-generation sequencing systems. *BioMed Research International*, Hindawi Publishing Corporation, v. 2012, 2012.

LUIZ, A. J. B. Meta-análise: definição, aplicações e sinergia com dados espaciais. *Cadernos de Ciência & Tecnologia*, v. 19, n. 3, p. 407–428, 2002.

LUO, W. et al. Gage: generally applicable gene set enrichment for pathway analysis. *BMC bioinformatics*, BioMed Central, v. 10, n. 1, p. 161, 2009.

MACHADO, C. M.; FREITAS, A. T.; COUTO, F. M. Enrichment analysis applied to disease prognosis. *Journal of biomedical semantics*, BioMed Central, v. 4, n. 1, p. 21, 2013.

MANGELSDORF, D. J. et al. The nuclear receptor superfamily: the second decade. *Cell*, Elsevier, v. 83, n. 6, p. 835–839, 1995.

MARTINEZ-ESPINOSA, M.; SANDANIELO, V. L. M.; LOUZADA-NETO, F. O método de bootstrap para o estudo de dados de fadiga dos materiais. *Revista de Matemática e Estatística*, v. 2, p. 41–54, 2006.

MARXREITER, S.; THUMMEL, C. S. Adult functions for the drosophila dhr78 nuclear receptor. *Developmental Dynamics*, Wiley Online Library, v. 247, n. 2, p. 315–322, 2018.

MCAFEE, A. et al. Big data. *The management revolution. Harvard Bus Rev*, v. 90, n. 10, p. 61–67, 2012.

MURDOCH, T. B.; DETSKY, A. S. The inevitable application of big data to health care. *Jama*, American Medical Association, v. 309, n. 13, p. 1351–1352, 2013.

OGATA, H. et al. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, Oxford University Press, v. 27, n. 1, p. 29–34, 1999.

RNA-SEQ BLOG. *RPKM, FPKM and TPM clearly explained*. jul. 2015. Disponivel em: http://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/. Acesso: 18-Out-2017.

ROSCOE, J. T.; BYARS, J. A. An investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic. *Journal of the American Statistical Association*, Taylor & Francis, v. 66, n. 336, p. 755–759, 1971.

SHANNON, C. E. A mathematical theory of communication, part i, part ii. *Bell Syst. Tech. J.*, v. 27, p. 623–656, 1948.

SIGNORELLI, M.; VINCIOTTI, V.; WIT, E. C. Neat: an efficient network enrichment analysis test. *BMC bioinformatics*, BioMed Central, v. 17, n. 1, p. 352, 2016.

SMID, M.; DORSSERS, L. C. Go-mapper: functional analysis of gene expression data using the expression level as a score to evaluate gene ontology terms. *Bioinformatics*, Oxford University Press, v. 20, n. 16, p. 2618–2625, 2004.

SPRENT, P. Fisher exact test. In: *International Encyclopedia of Statistical Science*. [S.l.]: Springer, 2011. p. 524–525.

SPRENT, P.; SMEETON, N. C. *Applied nonparametric statistical methods*. [S.l.]: Chapman and Hall/CRC, 2000.

TECHCOUNCIL. *The Sanger DNA Sequencing Methods*. 2013. DisponÃvel em: http://www.techcouncil.org/the-sanger-dna-sequencing-methods-1. Acesso: 8-Jan-2018.

TRAPNELL, C.; PACHTER, L.; SALZBERG, S. L. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, Oxford Univ Press, v. 25, n. 9, p. 1105–1111, 2009.

TRAPNELL, C. et al. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, Nature Publishing Group, v. 7, n. 3, p. 562–578, 2012.

TRAPNELL, C. et al. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, Nature Publishing Group, v. 28, n. 5, p. 511–515, 2010.

WAGNER, G. P.; KIN, K.; LYNCH, V. J. Measurement of mrna abundance using rna-seq data: Rpkm measure is inconsistent among samples. *Theory in biosciences*, Springer, v. 131, n. 4, p. 281–285, 2012.

WANG, Z.; GERSTEIN, M.; SNYDER, M. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, Nature Publishing Group, v. 10, n. 1, p. 57–63, 2009.

WILCOXON, F. Individual comparisons by ranking methods. *Biometrics bulletin*, JSTOR, v. 1, n. 6, p. 80–83, 1945.

YU, G. et al. clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 16, n. 5, p. 284–287, 2012.

ZEEBERG, B. R. et al. Gominer: a resource for biological interpretation of genomic and proteomic data. *Genome biology*, BioMed Central, v. 4, n. 4, p. R28, 2003.

APÊNDICE A – Descrição das amostras e combinações

Tabela A.1: Descrição das comparações realizadas a partir do conjunto de amostras S1, referente à *Aedes aegypti*, conforme (AKBARI et al., 2013).

ID SRA	Novo ID	Descrição		
SRR923702_SRR923826	S1-1	0-2hr Embryo	X	2-4hr Embryo
SRR923826_SRR923837	S1-2	2-4hr Embryo	X	4-8hr Embryo
SRR923837_SRR923853	S1-3	4-8hr Embryo	X	8-12hr Embryo
SRR923853_SRR923704	S1-4	8-12hr Embryo	X	12-16hr Embryo
SRR923704_SRR923824	S1-5	12-16hr Embryo	X	16-20hr Embryo
SRR923824_SRR923827	S1-6	16-20hr Embryo	X	20-24hr Embryo
SRR923827_SRR923828	S1-7	20-24hr Embryo	X	24-28hr Embryo
SRR923828_SRR923831	S1-8	24-28hr Embryo	X	28-32hr Embryo
SRR923831_SRR923833	S1-9	28-32hr Embryo	X	32-36hr Embryo
SRR923833_SRR923834	S1-10	32-36hr Embryo	X	36-40hr Embryo
SRR923834_SRR923838	S1-11	36-40hr Embryo	X	40-44hr Embryo
SRR923838_SRR923839	S1-12	40-44hr Embryo	X	44-48hr Embryo
SRR923839_SRR923840	S1-13	44-48hr Embryo	X	48-52hr Embryo
SRR923840_SRR923844	S1-14	48-52hr Embryo	X	52-56hr Embryo
SRR923844_SRR923845	S1-15	52-56hr Embryo	X	56-60hr Embryo
SRR923845_SRR923846	S1-16	56-60hr Embryo	X	60-64hr Embryos
SRR923846_SRR923849	S1-17	60-64hr Embryos	X	64-68hr Embryos
SRR923849_SRR923850	S1-18	64-68hr Embryos	X	68-72hr Embryos
SRR923850_SRR923851	S1-19	68-72hr Embryos	X	72-76hr Embryos
SRR923851_SRR923825	S1-20	72-76hr Embryos	X	1st Instar Larvae
SRR923825_SRR923832	S1-21	1st Instar Larvae	X	2nd Instar Larvae
SRR923832_SRR923836	S1-22	2nd Instar Larvae	X	3rd Instar Larvae
SRR923836_SRR923843	S1-23	3rd Instar Larvae	X	4th Instar Larvae
SRR923856_SRR923736	S1-24	Non-BF Female Carcass	X	12hr-BF Female Carcass
SRR923856_SRR923823	S1-25	Non-BF Female Carcass	X	24hr-BF Female Carcass
SRR923856_SRR923830	S1-26	Non-BF Female Carcass	X	36hr-BF Female Carcass
SRR923856_SRR923835	S1-27	Non-BF Female Carcass	X	48hr-BF Female Carcass
SRR923856_SRR923841	S1-28	Non-BF Female Carcass	X	60hr BF-Female Carcass
SRR923856_SRR923847	S1-29	Non-BF Female Carcass	X	72hr BF female Carcass
SRR923857_SRR923705	S1-30	Non-BF Female Ovary	X	12hr-BF Female Ovary
SRR923857_SRR923822	S1-31	Non-BF Female Ovary	X	24hr-BF Female Ovary
SRR923857_SRR923829	S1-32	Non-BF Female Ovary	X	36hr-BF Female Ovary
SRR923857_SRR923842	S1-33	Non-BF Female Ovary	X	48hr-BF Female Ovary
SRR923857_SRR923848	S1-34	Non-BF Female Ovary	X	60hr-BF Female Ovary
SRR923857_SRR923852	S1-35	Non-BF Female Ovary	X	72hr BF female Ovary

Fonte: O autor (2018).

Tabela A.2: Descrição das comparações realizadas a partir do conjunto de amostras S2, referente à *Drosophila melanogaster*, conforme (MARXREITER; THUMMEL, 2018)

ID SRA	Novo ID	Descrição
SRR6288269_SRR6288273	S2-1	Controle_1 X Mutante_1
SRR6288270_SRR6288274	S2-2	Controle_2 X Mutante_2
SRR6288271_SRR6288275	S2-3	Controle_3 X Mutante_3
SRR6288272_SRR6288276	S2-4	Controle_4 X Mutante_4
*media-controles_media-mutantes	S2-5	Controle X Mutante

Fonte: O autor (2018). (*) Trata-se de uma combinação utilizando a média dos controles e média dos experimentos (mutantes), uma vez que cada um dos conjuntos são compostos por 4 replicatas.

APÊNDICE B – Formatos de arquivos de input e output do EntropyClusterGenes

Tabela B.1: Exemplo de arquivo de entrada. Trata-se de um arquivo no formato texto, em três colunas, no qual a primeira representa os genes em comum a ambas as amostras, enquanto a segunda e terceira mostram o valor de expressão gênica, respectivamente, controle e experimento.

Gene ID	**Expressão Controle**	**Expressão Experimento**
AAEL000001	5.01145e+01	6.43933e+01
AAEL000003	5.16328e+01	4.93843e+01
AAEL000004	2.72980e+01	4.98836e+01
AAEL000007	6.06254e-01	5.07011e-01
AAEL000008	2.86432e+01	2.26705e+01
AAEL000009	1.25703e+00	6.67903e-01
AAEL000010	2.53398e+03	2.27778e+03
AAEL000011	1.52473e-01	2.66412e-01
AAEL000012	1.97315e-01	0.00000e+00
AAEL000013	3.16935e+01	2.17052e+01
AAEL000014	2.73344e-01	1.81946e-01
AAEL000016	1.19602e+02	9.24366e+01
AAEL000018	0.00000e+00	1.34582e-01
AAEL000019	2.14244e+00	1.42607e+00
AAEL000020	1.17783e+01	1.91750e+01
AAEL000021	2.06526e+00	2.09729e+00
AAEL000022	3.52633e+01	3.57112e+01
AAEL000024	3.62078e-01	7.02944e-01
AAEL000026	3.71710e+01	5.76333e+01
AAEL000027	0.00000e+00	3.68193e-02
AAEL000028	3.58435e+01	6.02657e+01
AAEL000029	1.23189e-01	3.18880e-02
AAEL000031	4.47429e-02	1.04237e-01
AAEL000033	2.19490e+01	2.27757e+01
AAEL000034	2.43059e+01	1.08683e+01
AAEL000035	3.13863e-01	2.43736e-01
AAEL000036	0.00000e+00	0.00000e+00
AAEL000037	2.32613e+00	3.22570e+00
AAEL000039	0.00000e+00	7.84726e-02
AAEL000041	0.00000e+00	6.48400e-02
AAEL000042	3.11833e+01	3.91751e+01
AAEL000043	2.68099e-01	5.72540e-01
AAEL000044	5.02163e-02	0.00000e+00
AAEL000046	2.89141e+01	3.33588e+01
AAEL000047	0.00000e+00	0.00000e+00

Fonte: O autor (2018).

Tabela B.2: Primeiro arquivo de saída, relacionando os genes e suas respectivas ontologias. A coluna *Evidência* representa a forma como as ontologias foram relacionadas aos genes dentro do próprio *Gene Ontology*, conforme descrito em Tabela 3.1.

Gene ID	GO ID	GO Name	Evidência GO	Expressão Controle	Expressão Experimento
AAEL000010	GO:0006412	*translation*	IEA	2.53398e+03	2.27778e+03
AAEL000012	GO:0050912	*detection of chemical stimulus involved in sensory perception of taste*	IEA	1.97315e-01	0.00000e+00
AAEL000013	GO:0006457	*protein folding*	IEA	3.16935e+01	2.17052e+01
AAEL000013	GO:0000413	*protein peptidyl-prolyl isomerization*	IEA	3.16935e+01	2.17052e+01
AAEL000014	GO:0055085	*transmembrane transport*	IEA	2.73344e-01	1.81946e-01
AAEL000014	GO:0006811	*ion transport*	IEA	2.73344e-01	1.81946e-01
AAEL000028	GO:0006508	*proteolysis*	IEA	3.58435e+01	6.02657e+01
AAEL000033	GO:0016180	*snRNA processing*	IEA	2.19490e+01	2.27757e+01
AAEL000034	GO:0051321	*meiotic cell cycle*	IEA	2.43059e+01	1.08683e+01
AAEL000034	GO:0006302	*double-strand break repair*	IEA	2.43059e+01	1.08683e+01
AAEL000037	GO:0006508	*proteolysis*	IEA	2.32613e+00	3.22570e+00
AAEL000041	GO:0006351	*transcription, DNA-templated*	IEA	0.00000e+00	6.48400e-02
AAEL000041	GO:0006355	*regulation of transcription, DNA-templated*	IEA	0.00000e+00	6.48400e-02
AAEL000043	GO:0050912	*detection of chemical stimulus involved in sensory perception of taste*	IEA	2.68099e-01	5.72540e-01
AAEL000044	GO:0006596	*polyamine biosynthetic process*	IEA	5.02163e-02	0.00000e+00
AAEL000048	GO:0050912	*detection of chemical stimulus involved in sensory perception of taste*	IEA	0.00000e+00	5.21911e-02
AAEL000057	GO:0007165	*signal transduction*	IEA	5.59433e-01	5.79249e-01
AAEL000057	GO:0045087	*innate immune response*	IEA	5.59433e-01	5.79249e-01
AAEL000059	GO:0006508	*proteolysis*	IEA	0.00000e+00	6.49521e-02
AAEL000060	GO:0050912	*detection of chemical stimulus involved in sensory perception of taste*	IEA	0.00000e+00	5.10761e-02
AAEL000062	GO:0009190	*cyclic nucleotide biosynthetic process*	IEA	7.21974e-01	7.60897e-01
AAEL000062	GO:0006171	*cAMP biosynthetic process*	IEA	7.21974e-01	7.60897e-01
AAEL000062	GO:0035556	*intracellular signal transduction*	IEA	7.21974e-01	7.60897e-01
AAEL000069	GO:0050912	*detection of chemical stimulus involved in sensory perception of taste*	IEA	0.00000e+00	6.63030e-02
AAEL000072	GO:0006508	*proteolysis*	IEA	0.00000e+00	0.00000e+00
AAEL000074	GO:0006508	*proteolysis*	IEA	1.21749e+01	1.62168e+01
AAEL000075	GO:0050912	*detection of chemical stimulus involved in sensory perception of taste*	IEA	0.00000e+00	0.00000e+00
AAEL000076	GO:0000154	*rRNA modification*	IEA	1.93515e+01	2.28492e+01
AAEL000076	GO:0006364	*rRNA processing*	IEA	1.93515e+01	2.28492e+01
AAEL000077	GO:0055085	*transmembrane transport*	IEA	4.04625e+01	1.99700e+01
AAEL000077	GO:0006812	*cation transport*	IEA	4.04625e+01	1.99700e+01
AAEL000080	GO:0006094	*gluconeogenesis*	IEA	4.38750e+01	6.92638e+01
AAEL000082	GO:0050912	*detection of chemical stimulus involved in sensory perception of taste*	IEA	0.00000e+00	0.00000e+00
AAEL000088	GO:0007030	*Golgi organization*	IEA	1.31493e+01	1.27105e+01
AAEL000088	GO:0015031	*protein transport*	IEA	1.31493e+01	1.27105e+01

Fonte: O autor (2018).

Tabela B.3: Segundo arquivo de saída, mostrando todos os GFAGs encontrados. **N** indica atividade, **n** atividade relativa, **H** diversidade e **h** diversidade relativa. São definidos p-valores e q-valores para atividade e diversidade. Grupos com q-valor abaixo de um dado corte (FDR de 0.05, por exemplo) são considerados significativos.

GO ID	Número de Genes	GO Name	H Controle	H Experimento	N Controle	N Experimento	h	n	p-valor h	p-valor n	q-valor h	Significância h	q-valor n	Significância n
GO:0034220	2	*ion transmembrane transport*	0.003531295937326 17	0.00285810731354198	108.5047572	106.7773318	0.447319914133396	0.495987995545695	0.107436	0.354052	0.128692509202454	not significative	0.75993919706499	not significative
GO:0030154	2	*cell differentiation*	0.00916369487918723	0.00483668010861789	66.9942642	107.320509	0.345467897312095	0.6156707606008	0.042442	0.719434	0.0559918952702703	not significative	0.823867967741936	not significative
GO:0009072	4	*aromatic amino acid family metabolic process*	0.00974027068277756	NaN	76.3336349	72.964305	0	0.488716087099873	0	0.323401	0	not significative	0.75993919706499	not significative
GO:0007040	2	*lysosome organization*	0.0254804874705879	NaN	33.384801	46.6761	0	0.583007428307608	0	0.60502	0	not significative	0.775895927750411	not significative
GO:0008654	3	*phospholipid biosynthetic process*	0.0387832002450277	0.0477510495017177	38.1212498	40.1201944	0.551816762050493	0.512774205668356	0.171286	0.40587	0.190833617689016	not significative	0.75993919706499	not significative
GO:0006541	3	*glutamine metabolic process*	0.0476869377776489	0.0267064426512887	120.790626	161.655776	0.358989502793187	0.572341424267816	0.014079	0.625184	0.0195653007117438	significative	0.780135543859649	not significative
GO:0043066	2	*negative regulation of apoptotic process*	0.050277882491789	0.0733979088348233	27.535119	54.254474	0.5934702988153	0.663342022010062	0.288347	0.817096	0.29208690920882	not significative	0.871792316939891	not significative
GO:0006535	2	*cysteine biosynthetic process from serine*	0.0550603234979499	0.0295847465160689	25.0303	40.574709	0.349515293816511	0.618469681179375	0.044254	0.728704	0.0581313394957983	not significative	0.82622998838897	not significative
GO:0019343	2	*cysteine biosynthetic process via cystathionine*	0.0550603234979499	0.0295847465160689	25.0303	40.574709	0.349515293816511	0.618469681179375	0.044287	0.728902	0.0581313394957983	not significative	0.82622998838897	not significative
GO:0006165	2	*nucleoside diphosphate phosphorylation*	0.0734661115492317	0.368211053450978	2366.8447	461.5327	0.833665587965824	0.163179319704648	0.338915	0.053966	0.339785128369705	not significative	0.572589094594595	not significative
GO:0006183	2	*GTP biosynthetic process*	0.0734661115492317	0.368211053450978	2366.8447	461.5327	0.833665587965824	0.163179319704648	0.340041	0.054253	0.340123	not significative	0.572589094594595	not significative
GO:0006228	2	*UTP biosynthetic process*	0.0734661115492317	0.368211053450978	2366.8447	461.5327	0.833665587965824	0.163179319704648	0.340123	0.054229	0.340123	not significative	0.572589094594595	not significative
GO:0051013	2	*microtubule severing*	0.0805365673916346	0.176636917513999	36.321712	19.261756	0.686839537827213	0.34653749924348	0.320664	0.140368	0.32231477992278	not significative	0.709487005494505	not significative
GO:0007016	2	*cytoskeletal anchoring at plasma membrane*	0.0916956094434861	0.0746775474943246	26.751795	41.602843	0.448855746136011	0.608632335965264	0.108955	0.697751	0.130112928134557	not significative	0.816638386227545	not significative
GO:0010468	3	*regulation of gene expression*	0.0947410423792538	0.300752678736791	428.55752	77.12175	0.760448681430634	0.152511195485629	0.19972	0.032588	0.212179448369565	not significative	0.403987746031746	not significative
GO:0006417	2	*regulation of translation*	0.12150748290224	0.145830462956615	13.647135	16.85954	0.545491073061542	0.552650854280252	0.257066	0.5093	0.263475782152231	not significative	0.769244648333333	not significative
GO:0016043	3	*cellular component organization*	0.123171471670393	0.0629275539557175	9.206019	14.3544808	0.338140158144323	0.609260453804125	0.010479	0.777061	0.0147994556962025	significative	0.854187575630252	not significative
GO:0009263	3	*deoxyribonucleotide biosynthetic process*	0.134249184774501	0.135488703117143	774.94553	1363.1492	0.502297634849021	0.637553229458641	0.122765	0.868612	0.143963160660661	not significative	0.898524466225166	not significative
GO:0006013	3	*mannose metabolic process*	0.13787291547405	NaN	4.0366666	3.317499	0	0.451104745316042	0	0.257702	0	not significative	0.75993919706499	not significative
GO:0006801	6	*superoxide metabolic process*	0.140830722902623	0.187450401142507	552.511033	510.504993	0.571005724705443	0.480242047639647	0.038777	0.26963	0.051504824829932	not significative	0.75993919706499	not significative
GO:0018279	4	*protein N-linked glycosylation via asparagine*	0.157997155441473	0.102119211670229	59.324135	53.771899	0.392590488649935	0.475453445166786	0.007623	0.287773	0.0108443770491803	significative	0.75993919706499	not significative
GO:0006685	5	*sphingomyelin catabolic process*	0.165897931014114	0.0985056887430658	4.177555	4.9796662	0.372558018810523	0.543796648703867	0.001718	0.524823	0.00247556826568266	significative	0.769244648333333	not significative
GO:0019236	2	response to pheromone	0.168355761867782	0.469243004782192	10.697418	8.022807	0.735953438629839	0.428563598995204	0.328818	0.223448	0.330085935732648	not significative	0.75993919706499	not significative
GO:0042438	2	melanin biosynthetic process	0.173227551796874	0.194102682026135	518.0264	281.61213	0.52841466384621	0.352174288049877	0.240058	0.144827	0.248324898013245	not significative	0.709487005494505	not significative
GO:0060179	2	male mating behavior	0.173227551796874	0.194102682026135	518.0264	281.61213	0.52841466384621	0.352174288049877	0.240778	0.144955	0.248740235449735	not significative	0.709487005494505	not significative
GO:0000256	2	allantoin catabolic process	0.180247236371894	0.204371137040093	2.7026842	1.7847075	0.531360827167711	0.397716005045871	0.243983	0.187386	0.251386178100264	not significative	0.735191643564356	not significative
GO:0003341	2	cilium movement	0.181346143839934	NaN	1.3395878	0.571855	0	0.299174529313668	0	0.106091	0	not significative	0.709487005494505	not significative
GO:0000122	2	negative regulation of transcription from RNA polymerase II promoter	0.187604902251905	0.145932988166948	40.04471	70.72478	0.437530464630831	0.638486102987384	0.098027	0.780909	0.119065454121306	not significative	0.854187575630252	not significative
GO:0007178	2	transmembrane receptor protein serine/threonine kinase signaling pathway	0.191866948822615	0.200626920173794	32.886478	52.55076	0.511159373487256	0.615080276822619	0.216937	0.718322	0.226810973226238	not significative	0.823867967741936	not significative
GO:0006171	2	cAMP biosynthetic process	0.196474872905909	0.261171739409158	0.7446226	0.7960733	0.570684306146146	0.516697227532052	0.275957	0.405567	0.279899242857143	not significative	0.75993919706499	not significative
GO:0006241	3	CTP biosynthetic process	0.21991647766691	0.692529063423736	2486.6627	622.5257	0.758981256674186	0.20022128604365	0.199954	0.041762	0.212179448369565	not significative	0.486807791044776	not significative
GO:0007411	7	axon guidance	0.222445775258024	0.103945901841712	42.9847538	62.891429	0.318469829762078	0.594009222251636	4.8e-05	0.809789	7.1e-05	significative	0.86755172702332	not significative
GO:0030833	2	*regulation of actin filament polymerization*	0.252500521366109	0.456921492384519	57.6517	68.79441	0.644075717313635	0.544061102393739	0.308991	0.483623	0.311383188387097	not significative	0.766647342799189	not significative
GO:0006388	2	*tRNA splicing, via endonucleolytic cleavage and ligation*	0.253140444047947	0.157879471085531	25.19781	36.089361	0.384116353666853	0.588856695637004	0.061761	0.62376	0.0790743295081967	not significative	0.780135543859649	not significative
GO:0006730	5	*one-carbon metabolic process*	0.254021769552366	NaN	298.0801611	207.053621	0	0.409898581993889	0	0.141001	0	not significative	0.709487005494505	not significative
GO:0030431	2	*sleep*	0.255659688729299	NaN	1.6826191	0	0	0	0	0	0	not significative	0	significative
GO:0032222	2	*regulation of synaptic transmission, cholinergic*	0.255659688729299	NaN	1.6826191	0	0	0	0	0	0	not significative	0	significative
GO:0045187	2	*regulation of circadian sleep/wake cycle, sleep*	0.255659688729299	NaN	1.6826191	0	0	0	0	0	0	not significative	0	significative
GO:1903818	2	*positive regulation of voltage-gated potassium channel activity*	0.255659688729299	NaN	1.6826191	0	0	0	0	0	0	not significative	0	significative
GO:0019439	2	*aromatic compound catabolic process*	0.269519257201788	0.342895349810612	139.25322	113.15315	0.559907203199138	0.448297521175872	0.268433	0.267841	0.272976787760417	not significative	0.75993919706499	not significative
GO:0006071	2	*glycerol metabolic process*	0.270563454084328	0.210115206682043	71.72936	104.80275	0.437121977387234	0.59367528094464	0.098867	0.64008	0.11989926552795	not significative	0.781596377916019	not significative
GO:0051016	5	*barbed-end actin filament capping*	0.31213914519257	0.145102733481477	56.552694	85.277153	0.317343489844498	0.601263801687666	0.000518	0.809544	0.000756183177570094	significative	0.86755172702332	not significative
GO:0007205	2	*protein kinase C-activating G-protein coupled receptor signaling pathway*	0.316226242143487	0.393089188455051	15.957565	15.82995	0.554181075862614	0.497992686751387	0.26468	0.358877	0.269511186440678	not significative	0.75993919706499	not significative
GO:0006325	2	*chromatin organization*	0.329184592868324	0.372172347204327	124.57944	109.20836	0.530646131719714	0.467126000586857	0.242888	0.295952	0.250588544253633	not significative	0.75993919706499	not significative
GO:0007631	2	*feeding behavior*	0.330797924868158	0.201256590112313	2.60884	5.238143	0.378263099824836	0.667535917944515	0.057833	0.820479	0.0745339488448845	not significative	0.871854918478261	not significative
GO:0035999	2	*tetrahydrofolate interconversion*	0.356180360975869	0.313200971359	291.6542	194.6856	0.46789618447608	0.400307768354554	0.129075	0.190152	0.150011272321429	not significative	0.735191643564356	not significative
GO:0007507	5	*heart development*	0.360767116460598	0.672590843307374	2.2238538	1.535059	0.650878852724371	0.408378454536109	0.068314	0.139075	0.0860674596774194	not significative	0.709487005494505	not significative
GO:0035385	5	*Roundabout signaling pathway*	0.360767116460598	0.672590843307374	2.2238538	1.535059	0.650878852724371	0.408378454536109	0.068325	0.138756	0.0860674596774194	not significative	0.709487005494505	not significative
GO:0032502	3	*developmental process*	0.4078987859589	0.65756891865406	1.1617415	0.587884	0.61716457083317	0.336005619488285	0.188631	0.10775	0.205182188022284	not significative	0.709487005494505	not significative
GO:0042773	3	*ATP synthesis coupled electron transport*	0.411702279471234	0.229106010844591	66.90255	75.03297	0.357526602428435	0.528641244982229	0.013733	0.45899	0.0191526303571429	significative	0.75993919706499	not significative
GO:0006779	2	*porphyrin-containing compound biosynthetic process*	0.424809000181825	0.469953588799587	64.70311	70.98152	0.525227132411266	0.523136039800529	0.237008	0.423374	0.245495023872679	not significative	0.75993919706499	not significative
GO:0007169	4	*transmembrane receptor protein tyrosine kinase signaling pathway*	0.424881811930739	NaN	15.1924718	6.984253	0	0.31493618029656	0	0.070621	0	not significative	0.665806988505747	not significative
GO:0051014	5	*actin filament severing*	0.442054248752921	0.511461347107594	34.258894	16.123753	0.536395366083151	0.320025920829448	0.060931	0.057984	0.0782682746710526	not significative	0.595861894736842	not significative
GO:0006635	3	*fatty acid beta-oxidation*	0.442316982481958	0.763551604600568	107.12587	57.17119	0.633196363832565	0.347974516403398	0.191196	0.11748	0.20739455	not significative	0.709487005494505	not significative
GO:0006816	4	*calcium ion transport*	0.453281379834379	NaN	9.66758071	7.2452041	0	0.428386228607139	0	0.18991	0	not significative	0.735191643564356	not significative

Fonte: O autor (2018).