UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS BIOLÓGICAS E DA SAÚDE
BACHARELADO EM BIOTECNOLOGIA

SARAH CALADO GALVÃO DE MELO

IDENTIFICAÇÃO DE ORTÓLOGOS E FAMÍLIAS GÊNICAS EM DEZ ESPÉCIES DE ABELHAS COM DIFERENTES COMPORTAMENTOS SOCIAIS

SÃO CARLOS - SP

2025

Sarah Calado Galvão de Melo

IDENTIFICAÇÃO DE ORTÓLOGOS E FAMÍLIAS GÊNICAS EM DEZ ESPÉCIES DE ABELHAS COM DIFERENTES COMPORTAMENTOS SOCIAIS

Trabalho de conclusão de curso apresentado ao curso de Bacharelado em Biotecnologia da Universidade Federal de São Carlos como requisito parcial para a obtenção do título de Bacharel em Biotecnologia.

Orientadora: Profª. Drª. Flávia Cristina de Paula Freitas

São Carlos - SP

2025

Dedico esse trabalho aos meus avós e aos meus pais, que sempre encorajaram meus estudos.

AGRADECIMENTOS

Agradeço aos meus avós que sempre acreditaram em mim e investiram na minha educação, incentivando-me a seguir em frente, obrigada por me ensinaram o valor da perseverança e da educação. Aos meus pais, que também nunca deixaram de acreditar em mim e me ensinaram a acreditar em mim mesma, não medindo esforços para abrir os caminhos e as portas do mundo que me levaram mais longe do que eu poderia imaginar, o meu mais sincero agradecimento por cada palavra de incentivo e cada sacrifício silencioso.

Aos amigos que a UFSCar me trouxe, que fizeram de São Carlos um lar longe de casa, obrigada por todas as risadas e pela companhia em tantos momentos especiais. Agradeço também a Guilherme, que compartilhou comigo cada desafio e vitória, por estar ao meu lado em todos os momentos dessa jornada, sua presença tornou tudo mais leve e mais feliz.

Gratidão aos professores John Herbert Vicente Ferreira, Dr. Eduardo José Alécio de Oliveira e Dr. Márcio Vilar França Lima, pelas oportunidades que me concederam e que tornaram possível meu ingresso na graduação e na ciência. À Profª. Drª. Flávia Cristina de Paula Freitas, minha orientadora, agradeço por depositar sua confiança em mim, por me manter motivada durante todo o processo e por me inspirar a seguir na carreira acadêmica.

Agradeço à Profª Drª Zilá Luz Paulino Simões por ter disponibilizado o servidor que utilizamos para a realização de todas as nossas análises e ao Dr. Pedro Roberto Prado pelo suporte no uso dos equipamentos.

Agradeço ao apoio financeiro da FAPESP na manutenção da estrutura do meu grupo de pesquisa por meio do Auxílio Jovem Pesquisador (Processo 2020/13719-7).

Por fim, agradeço a Deus, pois sem Ele nada seria possível.

“É preciso força pra sonhar e perceber
que a estrada vai além do que se vê”

Los Hermanos

RESUMO

As sequências gênicas que possuem ancestralidade comum são ditas homólogas e podem ser classificadas como ortólogas ou parálogas. Os parálogos são originados por duplicações do gene ancestral em determinada espécie, sendo responsáveis pela expansão das famílias gênicas, que são grupos de genes que apresentam funções semelhantes. Em contrapartida, os ortólogos surgem do processo de especiação em que um gene presente no ancestral comum é mantido nas novas espécies. Por compartilharem uma origem comum, a sequência dos ortólogos de espécies diferentes apresentam grande similaridade entre si. Desta forma, frequentemente, proteínas codificadas a partir dos ortólogos apresentam estrutura e função conservadas nas espécies. Assim, a identificação dos ortólogos auxilia na anotação gênica de novos genomas e na inferência da relação evolutiva entre as espécies. Além disso, a quantificação e identificação das expansões de famílias gênicas podem nos auxiliar na compreensão da inovação funcional em um genoma, como a eussocialidade das abelhas. Assim, esse trabalho teve como objetivo identificar os ortólogos e criar uma base de dados para dez espécies de abelhas com genomas disponíveis que diferem quanto ao comportamento social e identificar as famílias gênicas, avaliando se houve expansão ou retração nas espécies estudadas. Para isso, as sequências de proteínas das espécies foram alinhadas através do DIAMOND e a abordagem *Reciprocal Best Hit* (RBH) foi utilizada para identificar os melhores candidatos a ortólogos, totalizando 45 combinações par-a-par entre as dez espécies. Paralelamente, a ferramenta OrthoFinder foi utilizada para identificar os ortólogos e parálogos entre as espécies. A abordagem RBH identificou 4.979 e o OrthoFinder 5.755 ortólogos 1:1, sendo que 4.347 genes foram identificados em ambas as abordagens. A análise das famílias gênicas pela ferramenta CAFE identificou 200 expansões únicas para pelo menos uma das espécies de abelhas eussociais, 83 nas sociais simples e apenas 6 nas solitárias. Além da identificação dos ortólogos entre as dez espécies de abelhas, este trabalho contribuiu com a criação de um fluxo de análise automatizado que pode ser modificado para analisar os dados genômicos de outras espécies.

Palavras-chave: Bioinformática. Homologia de sequências. Famílias gênicas. Evolução molecular. Abelha.

ABSTRACT

Gene sequences that share a common ancestry are known as homologous and can be classified into orthologs or paralogs. Paralogs originate from duplications of the ancestral gene in a given species and are responsible for the expansion of gene families, which are groups of genes that have similar functions. In contrast, orthologs arise from speciation events in which a gene of the common ancestry is maintained in the novel species. As they share a common origin, orthologs sequences from different species display high similarity. Therefore, frequently, the ortholog encoded proteins display conserved structure and function in different species. Thus, the identification of orthologs helps in the gene annotation of novel genomes and in the inference of the evolutionary relationship between species. In addition, the quantification and identification of gene family expansions can help us understand functional innovation in a genome, such as the eusociality of bees. Thus, this study aimed to identify create a database of orthologs for ten bee species with available genomes that differ in social behavior and to identify gene families, evaluating whether there was expansion or retraction in the studied species. For this purpose, the protein sequences of the species were aligned using DIAMOND and the Reciprocal Best Hit (RBH) approach was used to identify the best ortholog candidates, totaling 45 pairwise combinations between the 10 species. In parallel, the OrthoFinder tool was used to identify orthologs and paralogs between species. The RBH approach identified 4.979 1:1 orthologs and the OrthoFinder 5.755, with 4.347 genes identified as 1:1 orthologs by both approachesThe evaluation of gene families using the CAFE tool identified 200 unique expansions for at least one of the eusocial bee species, 83 in the simple social bees and only 6 in the solitary bees. In addition to identifying orthologs among the ten bee species, this work contributed to the creation of an automated analysis flow that can be modified to analyze genomic data from other species.

Keywords: Bioinformatics. Sequence homology. Gene families. Molecular evolution. Bee.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Amel - *Apis mellifera*

Bimp - *Bombus impatiens*

BLAST - *Basic Local Alignment Search Tool*

Bter - *Bombus terrestris*

CAFE - *Computational Analysis of gene Family Evolution*

DAVID - *Database for Annotation, Visualization and Integrated Discovery*

DNA - Ácido desoxirribonucleico (do inglês, *deoxyribonucleic acid*)

Dnov - *Dufourea novaeangliae*

Emex - *Eufriesea mexicana*

Fvar - *Frieseomelitta varia*

Hlab - *Habropoda laboriosa*

MCL - *Markov Cluster Algorithm*

Mgen - *Megalopta genalis*

Mqua - *Melipona quadrifasciata*

Mrot - *Megachile rotundata*

NCBI - *National Center for Biotechnology Information*

RBH - Reciprocal Best Hits

UFSCar - *Universidade Federal de São Carlos*

SUMÁRIO

1 INTRODUÇÃO

1.1 As abelhas

As abelhas são insetos da ordem Hymenoptera que se destacam por características curiosas/interessantes da sua biologia, entre estas características estão a determinação do sexo e de castas. Nas abelhas, assim como na maioria dos himeópteros, o sexo é determinado pelo sistema haplodiploide, em que os machos se desenvolvem de ovos haploides e as fêmeas de ovos diploides. Em algumas espécies de abelhas existe a divisão dos indivíduos em castas, nestas as larvas fêmeas guardam o potencial de se desenvolverem em operárias e rainhas. Neste caso, o gatilho de uma ou outra via de desenvolvimento é a qualidade e/ou quantidade de alimento ofertado à larva em desenvolvimento. Nas abelhas melíferas, por exemplo, as larvas alimentadas com geleia real se desenvolvem em rainhas enquanto aquelas que recebem pólen e néctar se desenvolvem em operárias. Portanto, a alimentação diferencial culmina na diferenciação morfológica entre operárias e rainhas e assim define as funções que cada casta desempenha dentro da colônia.

O tipo de organização social das abelhas também é uma característica interessante. As abelhas corbiculadas, que possuem uma estrutura para carregar pólen em suas pernas, compreendem quatro tribos monofiléticas: as eussociais, Apini (abelhas melíferas) e Meliponini (abelhas sem ferrão), as sociais simples, Bombini (abelhas mamangavas) e as Euglossini (abelhas das orquídeas), que são em sua maioria solitárias (MICHENER, 1974). As abelhas solitárias levam a vida sem receber ajuda de outras abelhas, assim, cada uma constroi seu próprio ninho e coleta alimentos individualmente, além disso não há sobreposição de gerações nessas espécies (MICHENER, 2007). As abelhas de sociedade simples (facultativa ou obrigatória) formam pequenas colônias com uma rainha reprodutiva e uma ou mais operárias que sob influência de sinais sociais e nutricionais, renunciam à reprodução e cuidam cooperativamente de suas irmãs em uma eventual sobreposição de gerações (MICHENER, 1974). Já a eussocialidade é caracterizada pela divisão do trabalho reprodutivo, cuidado cooperativo da prole e sobreposição de gerações. A eussocialidade é considerada uma das principais inovações que permitiu não apenas abelhas, mas também que formigas e cupins se tornassem os organismos dominantes nos ecossistemas terrestres (WILSON; HÖLLDOBLER, 2005).

Na biologia, as inovações fenotípicas podem surgir a partir de alterações nas redes de regulação e de eventos de expansão ou retração de famílias gênicas. Identificar as alterações genômicas e vinculá-las às inovações fenotípicas tem sido um desafio.

1.2 Genômica comparativa

A genômica comparativa em larga escala pode fornecer novos *insights* sobre a transição de genótipo para fenótipo e gerar hipóteses testáveis sobre a evolução da diversidade animal (THOMAS *et al.*, 2020). Os artrópodes constituem o filo mais rico em espécies e diversidade da Terra e habitam os principais ecossistemas (THOMAS *et al*., 2020). Isso graças a uma série de mudanças genômicas e inovações selecionadas ao longo de sua história evolutiva (THOMAS *et al*., 2020). Entre as mudanças genômicas estão o surgimento de novos genes por duplicação ou, menos frequentemente, a perda de genes por evolução genética de novo (SANTOS *et al.*, 2017).

Na última década, a genômica comparativa tem sido usada como uma ferramenta poderosa para elucidar a base genética das inovações fenotípicas e da adaptação em uma ampla gama de organismos (GOODMAN *et al.*, 2009; GOU *et al.*, 2014; WU *et al.*, 2014; ZHANG *et al.*, 2014; XU *et al.*, 2016; EXPOSITO-ALONSO *et al.*, 2018; GAITHER *et al.*, 2018; CHEN *et al.*, 2019; WANG *et al.*, 2019). Esses estudos identificaram frequentemente mudanças no repertório genético (por exemplo, expansão de certas famílias de genes) ou assinaturas genômicas da evolução molecular em ortólogos, comparando os genomas de espécies com e sem um fenótipo de interesse (GOODMAN *et al.*, 2009; GOU *et al.*, 2014; WU *et al.*, 2014; XU *et al.*, 2016; GAITHER *et al.*, 2018; WANG *et al.*, 2019).

Como é o caso de uma série de estudos genômicos e transcriptômicos comparativos em insetos eussociais que começaram a identificar estas supostas assinaturas genômicas da evolução de sociedades complexas (SIMOLA *et al.*, 2013, KAPHEIM *et al.*, 2015), que incluem a expansão de certas famílias de genes associadas a funções como comunicação química (MCKENZIE; OXLEY; KRONAUER, 2014; MCKENZIE *et al.*, 2016; HARRISON *et al.*, 2018). O que faz sentido, já que a evolução da eussocialidade exige que os insetos sejam capazes de reconhecer outros indivíduos de sua colônia como companheiros de ninho, da mesma ou de casta diferente, ou indivíduos invasores (predadores, escravistas e hospedeiros) para uma coordenação efetiva (THOMAS *et al.*, 2020).

1.3 Famílias gênicas

As famílias gênicas são conjuntos de genes que, frequentemente, apresentam funções genéticas semelhantes (RANSON *et al.*, 2002), esta que pode se expandir por meio de mais duplicações genéticas ou se contrair por meio de deleções, dessa forma, o tamanho da família de genes pode ser dinâmico ao longo do tempo (DEMUTH *et al.*, 2006) para uma mesma espécie e pode apresentar variações entre espécies. Assim, é possível que famílias gênicas que compartilham sequência e homologia funcional, apresentem diferenças no número de genes (BAILEY; EICHLER, 2006; GRUS *et al.*, 2005). Essas variações podem estar por trás de muitas diferenças morfológicas, fisiológicas e comportamentais importantes entre espécies (DEMUTH e HAHN, 2009), fornecendo assim *insights* sobre mudança funcional relativa às bases moleculares de fenótipos complexos (CASTILLO-MORALES, 2014, 2016; NIIMURA; NEI, 2005).

Em sua pesquisa, Thomas e colaboradores (2020) encontraram 41 termos funcionais enriquecidos para mudanças na família gênicas com múltiplos ganhos relacionados à percepção olfativa e ligação odorante de acordo com estudos de quimiorreceptores anteriores de espécies de abelhas, formigas e cupins. Além disso, Tong e colaboradores (2020), em seu estudo com insetos e outros organismos sociais também identificaram expansões de famílias de genes associadas ao metabolismo e às funções de comunicação química, enfatizando a importância dos genes associados ao metabolismo e à reprodução como importantes na resposta aos desafios sociais, sugerindo que mudanças nos genes com funções metabólicas podem frequentemente estar envolvidas em alterações sociais dos animais.

Uma grande preocupação ao estudar mudanças no tamanho da família de genes é a qualidade da montagem do genoma e da anotação do mesmo, pois uma baixa cobertura de sequenciamento na montagem pode levar tanto à adição quanto à subtração errôneas de genes. Genes podem estar ausentes porque há cobertura incompleta de todo o genoma (HUBISZ *et al.*, 2011), mas também cópias extras de genes podem ser inseridas na montagem se a diversidade alélica for montada incorretamente como loci duplicados (HOLT *et al.*, 2002; COLBOURNE *et al.*, 2011) ou se um único gene multi exon for dividido entre múltiplos contigs, ocasionando em múltiplos modelos de genes que podem ser previstos a partir de um único gene (COLBOURNE *et al.*, 2011).

1.4 Genes homólogos, parálogos e ortólogos

Os genes homólogos são aqueles que compartilham a evolução com um ancestral em comum, revelado por similaridade das sequências entre os genes. Estes podem ser separados em duas categorias (Figura 1): ortólogos e parálogos (BROWN, 2017). Os ortólogos são aqueles presentes em diferentes organismos, cujo ancestral comum antecede a divisão entre as espécies, esses genes apresentam funções iguais ou muito semelhantes, por exemplo os genes da mioglobina de chimpanzés e humanos (BROWN, 2017). Os parálogos por sua vez estão presentes no mesmo organismo comumente como membro da mesma família gênica em decorrência de duplicações, sendo seu ancestral comum anterior ou não à espécie em que os genes são agora encontrados, um exemplo seriam os genes da mioglobina e da β-globina dos humanos, os quais originaram-se da duplicação de um gene ancestral há cerca de 550 milhões de anos (BROWN, 2017).

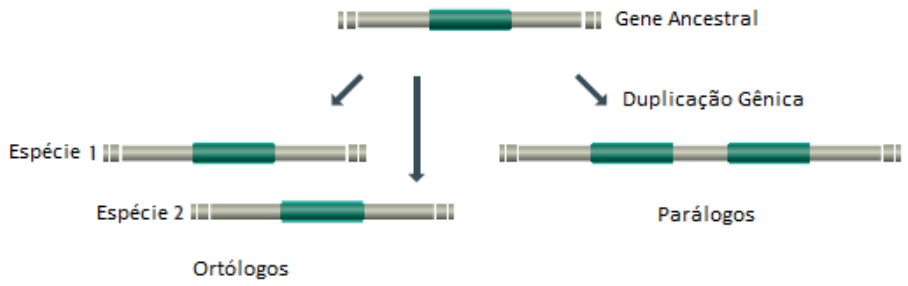

Figura 1: Origem dos ortólogos e parálogos.

Ilustração mostrando a diferença entre os ortólogos e parálogos, sendo o primeiro a presença do gene em diferentes espécies, podendo haver suas referidas alterações, mas mantendo a similaridade entre si e com o ancestral, já o segundo é dado pela duplicação dos genes ancestrais em uma mesma espécie. Fonte: BROWN, 2017 (adaptado)

Há uma hipótese de que a duplicação de genes, os genes parálogos, impulsiona a evolução de novas funções (OHNO, 1970). Isso porque, após a duplicação uma cópia do gene pode se tornar não funcional, pseudogeneizado, ou pode adquirir uma nova função, neofuncionalização, mas também é possível que as duas cópias dividam a função original, subfuncionalização (OHNO, 1970). Tickle e Urrutia (2017) reconheceram a duplicação gênica como uma fonte significativa de inovação funcional em genoma, eles consideram que

a duplicação de genes relacionados ao desenvolvimento desempenhou um papel significativo na evolução de várias características dos vertebrados.

Um par de genes homólogos geralmente não possui sequências de nucleotídeos idênticas, já que os dois genes sofrem alterações aleatórias diferentes por mutação, mas possuem sequências semelhantes porque essas alterações aleatórias operaram na mesma sequência inicial, o gene ancestral comum (BROWN, 2017). Dessa forma se um gene recentemente sequenciado for semelhante a um gene previamente sequenciado, então uma relação evolutiva pode ser inferida (BROWN, 2017), permitindo a inferência da possível função do novo gene com base na função previamente relatada de seus ortólogos (YATES, 2021).

Uma análise de homologia pode ser conduzida utilizando a sequência de DNA (quatro nucleotídeos) ou, se for uma região codificadora, a sequência em aminoácidos (20 aminoácidos). São duas as vantagens de se utilizar a sequência de aminoácidos. A primeira, é decorrente do código genético ser degenerado. A partir dos quatro nucleotídeos, A, T, C e G, são formadas 64 trincas (códons) que codificam apenas 20 aminoácidos. Com esta redundância do código genético, algumas mutações nas trincas não implicam em troca de aminoácido na sequência final, mantendo a semelhança entre as sequências apesar da mutação. A segunda vantagem do uso das sequências de aminoácidos está no fato do conjunto de aminoácidos (20) ser maior que o conjunto de nucleotídeos (quatro). O maior número de aminoácidos possíveis diminui as chances de falsos positivos na identificação de sequências homólogas, conforme exemplificado na Figura 2.

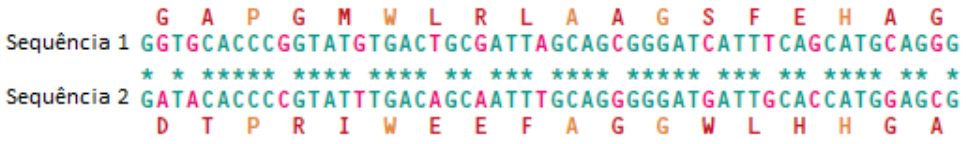

Figura 2: Diferenças nas análises das sequências de DNA e aminoácidos

Nesta figura vemos a comparação entre duas sequências. Quando a comparação é feita em nível de nucleotídeos, as sequências se mostram 76% idênticas, as similaridades estão representadas pelos asteriscos. Contudo, quando a comparação é feita em nível de aminoácidos, a identidade reduz para 28%, os aminoácidos idênticos estão em amarelo, enquanto os diferentes estão em marrom. A comparação de sequências através dos aminoácidos evita equívocos e modifica assim a conclusão quanto à homologia das sequências. Fonte: BROWN, 2017 (adaptado)

A identificação de homólogos é um processo importante na análise de padrões genéticos subjacentes às características e às relações evolutivas entre as espécies. A análise de

famílias de genes é frequentemente usada para formar e apoiar hipóteses sobre padrões genéticos, como presença, ausência ou divergência funcional de genes, que fundamentam características examinadas em estudos funcionais. Estas análises muitas vezes requerem a identificação precisa de todos os membros de uma família genética alvo sem a inclusão de membros em outras famílias de genes (FAVRE *et al.*, 2014; STEINEGGER *et al.*, 2019).

1.5 Alinhamento de sequências

Uma pesquisa em homologia se inicia com o alinhamento entre a sequência de entrada (*query*) e a sequência de *database* (*subject*), cada uma possuindo um identificador (ID), de forma que para cada alinhamento, uma pontuação (*score*) é calculada, sendo a partir dessa pontuação que se pode avaliar a possibilidade ou não de homologia (BROWN, 2017). Os programas contam o número de posições nas quais o mesmo aminoácido está presente em ambas as sequências, este número, quando convertido em percentagem, dá o grau de identidade entre duas sequências (BROWN, 2017).

Para atingir a pontuação mais alta possível, o algoritmo introduz lacunas (*gaps*) em várias posições em uma ou ambas as sequências, algo que ocorre durante a evolução dos genes, quando blocos de nucleotídeos que codificam aminoácidos individuais ou adjacentes podem ser inseridos ou deletados de um gene (BROWN, 2017). Além desses parâmetros apresentados, também temos o chamado valor esperado (*e-value*), o qual é definido como a probabilidade de um alinhamento com o *score* tão bom quanto o encontrado seja observado entre duas sequências aleatórias ou não relacionadas em uma busca numa *database* de mesmo tamanho, vale ressaltar que quanto menor o valor do *e-value* mais significativo é o alinhamento entre as sequências (BAXEVANIS; OUELLETTE, 2004).

Existem vários *softwares* para este tipo de análise, sendo o mais popular o *Basic Local Alignment Search Tool*, BLAST (ALTSCHUL *et al.*, 1990), a análise pode ser realizada simplesmente acessando o site de um dos bancos de dados de DNA e inserindo a sequência na ferramenta de busca *online* (BROWN, 2017).

Os ortólogos são genes altamente conservados nas sequências dos genomas, de forma que é esperado que as proteínas codificadas tenham a mesma estrutura e função, além de derivarem do mesmo ancestral (BAXEVANIS; OUELLETTE, 2004). Para identificar os ortólogos, cada proteína do proteoma de um organismo é usada como *query* na busca de um

similar comparando com uma *database* do proteoma de outro organismo, o melhor resultado é o mais propenso a ser ortólogo da *query* (BAXEVANIS; OUELLETTE, 2004).

Para isso, exige-se que, para cada par de ortólogos, o primeiro também seja o melhor resultado quando o segundo for usado como *query* no proteoma do primeiro (BAXEVANIS; OUELLETTE, 2004). Juntamente com isso, baixos valores de *e-value* ($<10^{-20}$) e um alinhamento que inclui 60–80% da sequência *query*, são requisitos para evitar erros de pareamento com parálogos, assim caso os parâmetros sejam mais relaxados pode haver resultados falso-positivos (BAXEVANIS; OUELLETTE, 2004).

1.6 Ferramentas da bioinformática

As ferramentas da bioinformática podem ser softwares, algoritmos ou plataformas desenvolvidas para analisar e interpretar diferentes tipos de dados biológicos. Algumas delas são DIAMOND, OrthoFinder, DAVID e CAFE.

O DIAMOND (BUCHFINK; REUTER; DROST, 2021) é um alinhador de proteínas rápido e sensível, desenvolvido para realizar alinhamentos ultra rápidos, sendo utilizado na metagenômica, a versão v2.0.7 tem uma sensibilidade de alinhamento que corresponde ao BLAST.

Já o OrthoFinder (EMMS; KELLY, 2019) infere ortogrupos, ortólogos, conjuntos completos de árvores genéticas para todos os ortogrupos, árvores de espécies enraizadas e todos os eventos de duplicação genética tendo como entrada as sequências genéticas.

O DAVID, *Database for Annotation, Visualization and Integrated Discovery* (https://davidbioinformatics.nih.gov/, HUANG; SHERMAN; LEMPICKI, 2018), consiste em ferramentas analíticas voltadas para extrair o significado biológico de grandes listas de genes/proteínas.

O CAFE, *Computational Analysis of gene Family Evolution* (DE BIE *et al*., 2006), pode ser utilizado para fazer inferências sobre a direção e a grandeza das mudanças no tamanho da família de genes, bem como se grandes mudanças no tamanho são realmente significativas evolutivamente, ou seja, é uma ferramenta para analisar mudanças no tamanho da família de genes em um contexto filogenético (DE BIE *et al*., 2006). As principais entradas do CAFE (HAN, M. V. *et al*., 2013) são uma descrição de Newick de uma árvore filogenética enraizada e bifurcada (incluindo comprimentos de ramificação em unidades de tempo) e um

arquivo de dados contendo os tamanhos de famílias de genes para os táxons existentes (DE BIE *et al.*, 2006).

2 OBJETIVOS

2.1 Objetivo geral

Identificar ortólogos para dez espécies de abelhas que diferem quanto ao comportamento social e criar uma base de dados.

2.2 Objetivos específicos

- Identificar ortólogos entre 10 espécies de abelhas com genomas disponíveis, sendo elas: *Apis mellifera, Bombus impatiens, Bombus terrestris, Dufourea novaeangliae, Eufriesea mexicana, Frieseomelitta varia, Habropoda laboriosa, Megalopta genalis, Melipona quadrifasciata, Megachile rotundata*;
- Quantificar os genes conservados e exclusivos de cada espécie com base na análise de ortólogos;
- Identificar famílias gênicas e avaliar se houve expansão ou contração nas espécies estudadas.

3 MATERIAL E MÉTODOS

3.1 Recuperação dos dados

Os arquivos no formato FASTA contendo a sequência de aminoácidos das proteínas das espécies de interesse foram recuperados da base de dados genômicos do *National Center for Biotechnology Information*, NCBI (SAYERS *et al*., 2022), baixados e salvos no servidor para posterior utilização nas análises. Nesta mesma base de dados, obtivemos informações sobre os genomas de cada espécie (Tabela 1).

Tabela 1: Informações gerais das espécies

Espécie	Abreviação	Comportamento	Banco de dados	Identificador	nº genes totais	nº genes codificadores	nº proteínas
Apis mellifera	Amel	eussocial	NCBI RefSeq	GCF_003254395.2	12.398	9.935	23.471
Bombus impatiens	Bimp	social simples	NCBI RefSeq	GCF_000188095.3	13.161	10.632	24.471
Bombus terrestris	Bter	social simples	NCBI RefSeq	GCF_910591885.1	13.398	10.310	25.755
Dufourea novaeangliae	Dnov	solitário	NCBI RefSeq	GCF_001272555.1	10.043	9.844	12.157
Eufriesea mexicana	Emex	social simples	NCBI RefSeq	GCF_001483705.2	11.090	10.278	15.659
Frieseomelitta varia	Fvar	eussocial	NCBI RefSeq	GCF_011392965.1	12.429	10.618	23.628
Habropoda laboriosa	Hlab	solitário	NCBI RefSeq	GCF_001263275.1	10.352	10.069	12.256
Megalopta genalis	Mgen	social simples	NCBI RefSeq	GCF_011865705.1	12.793	10.425	22.381
Melipona quadrifasciata	Mqua	eussocial	NCBI GenBank	GCA_001276565.1	14.711	14.257	14.257
Megachile rotundata	Mrot	solitário	NCBI RefSeq	GCF_000220905.1	11.705	10.788	26.024

Tabela contendo as informações para cada espécie como abreviação utilizada neste trabalho, comportamento social e as descrições dos genomas, o banco de dados no qual o genoma foi armazenado, o identificador, o número total de genes naquele genoma, o número total de genes codificadores, bem como o número total de proteínas provenientes dele.

A partir do NCBI (SAYERS *et al.*, 2022) também conseguimos obter arquivos chamados “gene info”, que constitui uma tabela com várias informações incluindo o

identificador do gene, uma abreviação para identificá-lo (*gene symbol*) e sua descrição. Outro arquivo semelhante obtido foi o "gene2refseq", também muito informativo, mas que adicionalmente continha informação sobre o identificador da proteína, contudo sem a descrição dos genes.

3.2 Identificação de ortólogos

Foram utilizadas duas abordagens para identificação dos ortólogos, uma delas baseada no uso do BLAST (ALTSCHUL *et al.*, 1990), do DIAMOND (BUCHFINK; REUTER; DROST, 2021) e do *Reciprocal Best Hits*, RBH (HERNÁNDEZ-SALMERÓN *et al.*, 2020), enquanto a segunda abordagem utilizou a ferramenta OrthoFinder (EMMS; KELLY, 2019).

3.2.1 Melhor alinhamento recíproco

3.2.1.1 Das espécies eussociais *A. mellifera*, *F. varia* e de *M. quadrifasciata*

Os *scripts* para executar o BLAST entre as proteínas foram desenvolvidos na linguagem Python utilizando o editor de código *Sublime Text* e salvos no formato de arquivo .py, esses arquivos foram rodados dentro de um servidor na plataforma *MobaXterm*.

O primeiro *script* realizava uma busca no multifasta de uma das espécies, salvando todas as proteínas em arquivos FASTA individuais, os quais foram utilizados como *query* para o alinhamento contra o multifasta de outra espécie, o resultado para cada proteína foi salvo em um .txt no formato tabular, resgatando informações como identificador (ID) da sequência *query*, ID da sequência *subject* (proteína de saída), bem como comprimento da sequência *query*, comprimento da sequência *subject*, comprimento da região de alinhamento entre as sequências, percentual de identidade, número de lacunas (*gaps*) e o valor esperado (*e-value*).

Ao todo, esse *script* foi utilizado seis vezes, de forma que todas as proteínas de *A. mellifera* não fossem apenas alinhadas com o multifasta de *F. varia*, mas também com o de *Melipona quadrifasciata*. Além disso, todas as proteínas de *M. quadrifasciata* foram alinhadas com o multifasta de *F. varia* e de *A. mellifera*, seguindo a mesma lógica, todas as proteínas de *F. varia* foram alinhadas com os multifastas de *A. mellifera* e *M. quadrifasciata*.

Uma vez que todas as proteínas de todas as espécies foram alinhadas com os multifastas das outras espécies, seguiu-se para a próxima etapa. Nesse novo *script* avaliamos a reciprocidade dos alinhamentos, isto é, recuperamos os pares de proteína X-Y, em que o melhor resultado da proteína X de uma espécie, foi a proteína Y da outra, e o melhor resultado para a proteína Y teve como melhor resultado a proteína X. Assim, como arquivo de saída o .txt possuía os códigos das proteínas X e Y, mas também o *status* do alinhamento, se o alinhamento foi recíproco ou não entre as espécies.

Uma vez com a tabela dos *status* montada, resgatamos os dados dos arquivos iniciais para os alinhamentos recíprocos, montando um novo arquivo .txt para cada dupla de espécie contendo os ID das proteínas, os tamanhos das sequências, o tamanho do alinhamento, percentual de identidade, número de lacunas e o valor esperado.

Com isso, foi possível seguir para a etapa de filtragem dos resultados. Para isso, utilizamos como parâmetro um percentual de identidade maior que 80% (SANTOS *et al.*, 2024), as proteínas que atenderam a esta condição foram submetidas a outro filtro, que seria o percentual de cobertura do alinhamento, ou seja, o quanto o comprimento do alinhamento corresponde ao comprimento da proteína. Para isso, esse percentual teria que ser pelo menos maior que 60% para uma das proteínas, ocorrendo isso era necessário também que fosse maior que 90% para a outra proteína, seguindo os critérios estabelecidos por SANTOS *et al.* (2024). Por fim, as proteínas restantes foram filtradas mais uma vez, agora com relação ao número de *gaps* do alinhamento, sendo necessário que fosse menor que 1% do tamanho das proteínas tanto *query* quanto *subject*, também baseado nos critérios propostos por SANTOS *et al.* (2024).

Com as proteínas filtradas, foi possível criar duas tabelas para expor os resultados finais, informado através do ID das proteínas. Em uma primeira tabela, foi exposto os resultados de reciprocidade do trio, assim a proteína de *A. mellifera* alinhou reciprocamente com a de *F. varia* e de *M. quadrifasciata*, que também alinharam reciprocamente entre si. Já na segunda tabela, quando a reciprocidade do trio não foi encontrada, mas havia reciprocidade entre duas das três espécies, esse resultado foi registrado.

Essa era uma abordagem piloto e como esse resultado foi muito mais restrito do que o esperado, excluindo inclusive proteínas sabidamente conhecidas como conservadas, outras abordagens para as análises de reciprocidade foram feitas utilizando o DIAMOND (BUCHFINK; REUTER; DROST, 2021).

3.2.1.2 Das dez espécies de abelhas

Para encontrar os ortólogos entre as dez espécies de abelhas, utilizamos a abordagem RBH proposta por HERNÁNDEZ-SALMERÓN e colaboradores (2020) em que os alinhamentos foram realizados pela ferramenta DIAMOND (BUCHFINK; REUTER; DROST, 2021) e a identificação dos melhores candidatos a ortólogos foi feita com o *script getRBH.pl* disponibilizado pelos autores. Essa abordagem foi aplicada para cada combinação de duas espécies das dez espécies analisadas, totalizando 45 combinações. Como dado de entrada foram utilizados arquivos FASTA contendo a proteína mais longa para cada gene, então mesmo as análises sendo realizadas em nível de proteínas, nos referimos muitas vezes ao gene, já que um está relacionado com o outro.

Para cada par de espécies, as sequências de proteínas de uma espécie foram alinhadas contra todas as proteínas da segunda espécie e o resultado das relações entre os ortólogos foi organizado em uma tabela. Foram considerados os alinhamentos com cobertura de pelo menos 40% das sequências *query* e *subject*. Para reunir os resultados de todas as combinações par-a-par de espécies, desenvolvemos um *script* em Python que recuperou os identificadores dos genes dos arquivos de anotação do NCBI (SAYERS *et al.*, 2022) e gerou duas tabelas. Ambas as tabelas foram organizadas com os genes nas linhas e as espécies nas colunas. A primeira tabela reuniu apenas os genes presentes em todas as dez espécies, ou seja, uma tabela com genes 1:1. A segunda tabela reuniu todos os genes de todas as espécies e nas colunas trazia a informação se o gene foi encontrado (informação do identificador do gene) ou se está ausente em determinada espécie (a depender da coluna). A fim de facilitar a leitura, o termo "RBH" foi utilizado nas próximas seções do trabalho para se referir aos resultados obtidos nesta abordagem.

3.2.2 OrthoFinder

A ferramenta OrthoFinder (EMMS; KELLY, 2019) foi utilizada nesta etapa através do servidor com a linha de comando padrão, sendo que esta análise também foi feita para as dez espécies com o arquivo FASTA contendo apenas uma proteína por gene. Essa ferramenta se difere da abordagem de RBH (HERNÁNDEZ-SALMERÓN; MORENO-HAGELSIEB, 2020) porque se baseia na pontuação do *score* do alinhamento par-a-par realizado com o

DIAMOND e aplica uma correção que elimina o viés provocado pelo comprimento das sequências analisadas. Além de identificar os ortólogos, o OrthoFinder faz uma divisão destes em grupos, os chamados ortogrupos. Os arquivos de resultados gerados foram salvos nos seus devidos formatos e pastas para utilização em outras etapas.

3.2.3 Combinação dos resultados RBH e OrthoFinder

A partir do arquivo "Orthogroups_SingleCopyOrthologues.txt", gerado pelo OrthoFinder (EMMS; KELLY, 2019), foi possível resgatar quais ortogrupos apresentam genes presentes em todas as espécies, sendo que cada espécie contribui para o ortogrupo com apenas um gene (ortólogos 1:1). Uma vez selecionando esses ortogrupos, buscou-se em "Orthogroups.tsv" quais eram tais genes. Utilizando um *script* em Python, comparamos os resultados dos genes presentes em todas as espécies gerados pela abordagem RBH (HERNÁNDEZ-SALMERÓN; MORENO-HAGELSIEB, 2020) e pelo OrthoFinder (EMMS; KELLY, 2019). As relações identificadas por ambas as abordagens foram organizadas em uma tabela final e, posteriormente, analisadas na ferramenta *online* DAVID (https://davidbioinformatics.nih.gov/, HUANG; SHERMAN; LEMPICKI, 2018), que teve como resultado anotação funcional dos genes, análise da ontologia, das vias biológicas e dos domínios proteicos.

3.3 Análise das famílias gênicas

O CAFE (DE BIE *et al.*, 2006) foi utilizado como ferramenta para analisar as mudanças no tamanho das famílias gênicas, para isso foi necessário preparar os arquivos de entrada, sendo estes uma árvore ultramétrica das espécies e um arquivo de contagem dos genes de cada espécies para as famílias gênicas. A preparação destes dados seguiu um tutorial disponibilizado pelos desenvolvedores da ferramenta CAFE (HAHNLAB, 2018).

Para a construção da árvore filogenética, foram utilizados os arquivos contendo as sequências dos ortólogos 1:1 das dez espécies determinados pelo OrthoFinder (EMMS; KELLY, 2019), pois são genes conservados e presentes em todas as espécies. Cada um dos arquivos foram alinhados (KATOH; STANDLEY, 2013) e também foi realizado um *trimming (*CAPELLA-GUTIERREZ; SILLA-MARTINEZ; GABALDON, 2009) em seguida para

remover regiões mal alinhadas ou indesejadas que podem afetar a qualidade das análises subsequentes. Para construir a árvore pelo método de supermatriz foi necessário concatenar todas as sequências para cada espécie e juntar todas as espécies em um único arquivo.

Esse arquivo foi então utilizado para construção da árvore através do IQ-TREE (NGUYEN *et al*., 2014) dentro do servidor por linha de comando padrão. Mas essa árvore não estava no formato ultramétrico necessário para o CAFE (DE BIE *et al.*, 2006), então algumas alterações foram feitas utilizando o *script* em linguagem Python fornecido pela equipe pesquisadora no tutorial disponibilizado, bem como a ferramenta r8s (SANDERSON, 2003).

Para a construção do arquivo de contagem das famílias gênicas, foram utilizados inicialmente os multifastas das dez espécies analisadas, as quais foram submetidas a outro *script* em linguagem Python para a determinação da isoforma mais longa de cada proteína, sendo criado um novo arquivo FASTA para cada espécie. Um arquivo único com todas as proteínas dos novos arquivos FASTA foi criado e utilizado para a criação de um banco de dados. O qual foi utilizado para a realização de um BLASTp "tudo contra todos", ou seja, todas as proteínas de todas as espécies.

O resultado desse BLASTp foi convertido para o formato .abc e tratado pela ferramenta *Markov Cluster Algorithm*, MCL (VAN DONGEN, 2008), para o agrupamento e rearranjo tabular dos dados, obtendo assim uma tabela que contém em cada linha uma família gênica e as proteínas pertencentes a ela separadas por tabulação. Esse arquivo foi tratado por mais um *script* Python fornecido, de forma que o arquivo de saída foi uma tabela que continha em cada linha a identificação da família gênica e nas colunas a contagem de quantos genes cada espécie tem dentro daquela família. Essa tabela foi filtrada pelo uso de *script* Python, excluindo-se as famílias que uma ou mais espécie apresentavam mais de 100 genes, isso porque famílias gênicas que apresentam uma grande variação no número de cópias podem fazer com que as estimativas dos parâmetros sejam não informativas, acrescentando ruídos que interferem na estimativa dos parâmetros de ganho e perda no modelo probabilístico do CAFE (DE BIE *et al.*, 2006), assim essa filtragem é um modo de resolver esta questão.

Com os dois arquivos de entrada obtidos, o CAFE (DE BIE *et al.*, 2006) foi utilizado através da linha de comando padrão. Gerando ao todo nove arquivos como resultado, entre eles um que continha as contagens de expansão e retração das famílias para cada clado da

árvore e outro em formato tabular, no qual para cada família havia a contagem dos genes em cada clado da árvore, possibilitando a comparação e identificação das famílias que expandiram ou contraíram para cada espécie.

3.4 Fluxograma das análises

As etapas das análises realizadas foram organizadas em um fluxograma (Figura 3) para proporcionar uma melhor visualização do processo, facilitando o entendimento das diferentes fases.

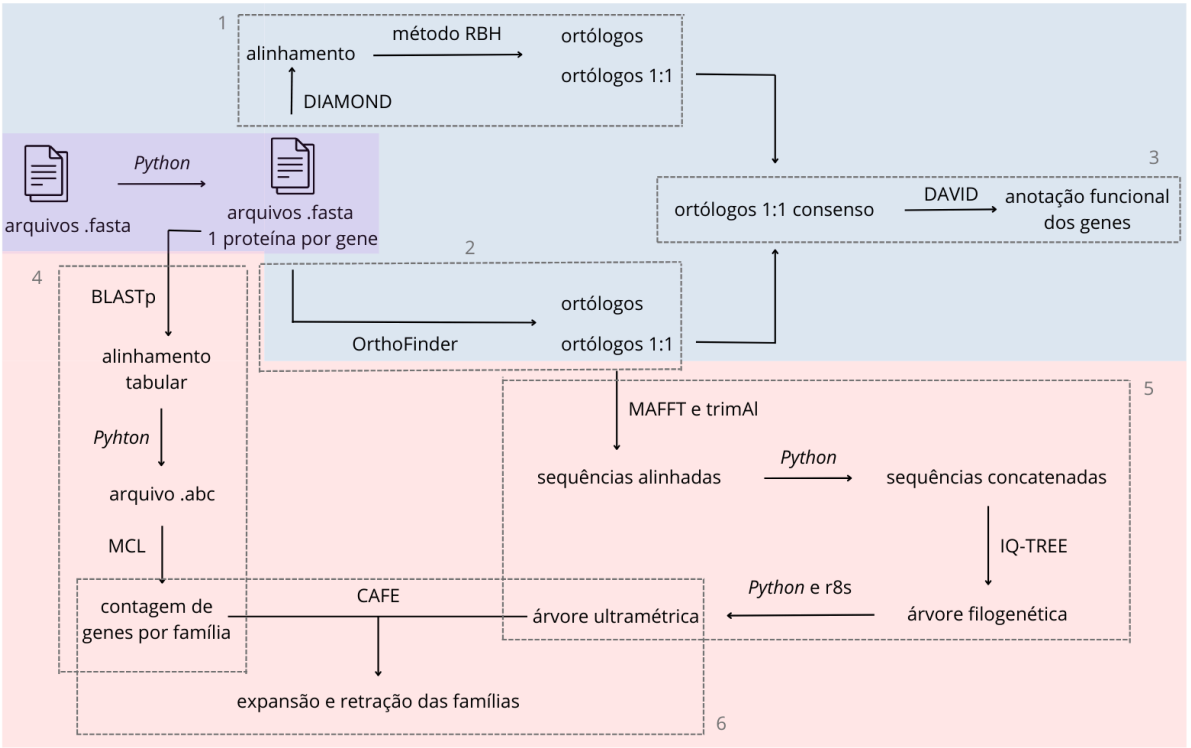

Figura 3: Fluxograma das análises realizadas
Imagem ilustrando as etapas das análises realizadas. Em azul, estão representadas as etapas de identificação dos ortólogos, enquanto em laranja encontram-se as etapas relacionadas à análise das famílias gênicas. Na identificação dos ortólogos, o item 1 descreve os passos do alinhamento recíproco realizado para as dez espécies analisadas, o item 2 refere-se à etapa conduzida com o OrthoFinder, esses resultados foram combinados e utilizados na anotação funcional dos genes, conforme descrito no item 3. A análise das famílias gênicas também pode ser dividida em três partes: no item 4, têm-se a obtenção do arquivo com a contagem de genes por família; no item 5, a árvore ultramétrica das espécies é construída; sendo ambos utilizado como entrada do CAFE para realização da etapa 6, sendo esta a análise e geração os resultados de expansão e retração das famílias gênicas.

4 RESULTADOS E DISCUSSÃO

4.1 Identificação de ortólogos

4.1.1 Das espécies eussociais

Com os *scripts* desenvolvidos e a utilização destes, foi possível comparar as sequências de proteínas entre as espécies *A. mellifera* x *F. varia*, *A. mellifera* x *M. quadrifasciata*, *F. varia* x *M. quadrifasciata*. Identificamos 392 ortólogos entre as três espécies e 2.101 reciprocidades entre duas espécies, sendo 1.776 entre *F. varia* e *M. quadrifasciata*, 309 entre *F. varia* e *A. mellifera*, e por fim 15 entre as espécies *M. quadrifasciata* e *A. mellifera*. A diferença de proteínas recíprocas encontrada entre as comparações dos dois pares com *A. mellifera* e o par de *F. varia* e *M. quadrifasciata* se deve ao grau de parentesco destas duas últimas espécies, que apresentam um ancestral comum mais próximo entre si do que entre *A. mellifera*. Como mencionado, esse resultado foi muito mais restrito do que o esperado, e por isso testamos o DIAMOND e utilizamos o *script* em PERL de HERNÁNDEZ-SALMERÓN e colaboradores (2020) para analisar as sequências das dez espécies de abelhas.

4.1.2 Das dez espécies

Para determinar os melhores alinhamentos recíprocos, utilizamos a ferramenta DIAMOND e recuperamos os pares de proteínas recíprocas entre as 45 combinações de pares de espécies. Com esta abordagem, encontramos 4.979 ortólogos 1:1 nas 10 espécies de abelhas. As quantidades de genes recíprocos por pares de espécies estão representadas na Tabela 2, enquanto a Tabela 3 traz a informação sobre a quantificação de genes exclusivos de cada espécie.

Tabela 2: Relação de reciprocidade entre espécies

Espécie 1	Espécie 2	Número de genes recíprocos
Amel	Bimp	12.163
Amel	Bter	12.397
Amel	Dnov	9.126

Amel	Emex	10.291
Amel	Fvar	8.799
Amel	Hlab	9.294
Amel	Mgen	10.864
Amel	Mqua	7.319
Amel	Mrot	10.643
Bimp	Bter	15.938
Bimp	Dnov	9.407
Bimp	Emex	10.424
Bimp	Fvar	11.946
Bimp	Hlab	9.451
Bimp	Mgen	10.867
Bimp	Mqua	7.563
Bimp	Mrot	10.903
Bter	Dnov	9.320
Bter	Emex	10.374
Bter	Fvar	12.017
Bter	Hlab	9.348
Bter	Mgen	11.042
Bter	Mqua	7.440
Bter	Mrot	10.925
Dnov	Emex	9.104
Dnov	Fvar	9.067
Dnov	Hlab	8.989
Dnov	Mgen	8.868
Dnov	Mqua	7.241
Dnov	Mrot	9.273
Emex	Fvar	10.001
Emex	Hlab	9.404
Emex	Mgen	9.579
Emex	Mqua	7.387
Emex	Mrot	9.735
Fvar	Hlab	9.159
Fvar	Mgen	10.341
Fvar	Mqua	7.560
Fvar	Mrot	10.509
Hlab	Mgen	8.941
Hlab	Mqua	7.315
Hlab	Mrot	9.182

Mgen	Mqua	7.038
Mgen	Mrot	9.979
Mqua	Mrot	7.301

Tabela contendo as quantidades de genes recíprocos para cada par de espécie de acordo com a abordagem RBH.

Tabela 3: Dados provenientes da análise RBH

Espécie	nº genes totais	nº genes exclusivos	nº ortólogos 1:1	% genes exclusivos	% ortólogos 1:1
Amel	9915	19	4.979	0,19	50,22
Bimp	10632	8	4.979	0,08	46,83
Bter	10310	14	4.979	0,14	48,29
Dnov	9844	0	4.979	0	50,58
Emex	10297	4	4.979	0,04	48,35
Fvar	10618	5	4.979	0,05	46,89
Hlab	10069	7	4.979	0,07	49,45
Mgen	10425	19	4.979	0,18	47,76
Mqua	14257	14	4.979	0,10	34,92
Mrot	10788	17	4.979	0,16	46,15

Tabela com os resultados do RBH contendo para cada espécie, o número de genes analisados, quantos genes são exclusivos, ou seja, espécie-específico, além do número de ortólogos 1:1. A tabela descreve também a porcentagem dos genes exclusivos e dos ortólogos 1:1 em relação ao número total de genes da espécie.

4.1.3 OrthoFinder

O OrthoFinder analisou a proteína mais longa codificada por todos os genes de todas as dez espécies, totalizando 107.155 genes analisados. Destes, 101.789 foram agrupados em 10.433 ortogrupos (conjunto de genes considerados ortólogos), enquanto 5.366 genes não foram agrupados. Além disso, 6.956 ortogrupos são compostos por genes de todas as dez espécies, dos quais 5.755 ortogrupos são formados por ortólogos 1:1, tivemos também 214 ortogrupos espécie-específico com 1.196 genes. Essas informações e outras complementares se encontram detalhadas por espécies na Tabela 4.

Tabela 4: Dados provenientes da análise do OrthoFinder

Espécie	nº genes totais	nº genes em ortogrupos	nº genes não agrupados	nº ortogrupos espécie específico	nº genes em ortogrupos espécie específico	% ortólogos 1:1	% genes não agrupados
Amel	9.915	9.745	170	5	34	58	1,7
Bimp	10.632	10.440	192	15	49	54	1,8
Bter	10.310	10.253	57	10	65	56	0,6
Dnov	9.844	9.760	84	7	34	58	0,9
Emex	10.297	10.143	154	20	61	56	1,5
Fvar	10.618	10.277	341	6	60	54	3,2
Hlab	10.069	9.971	98	7	30	57	1
Mgen	10.425	10.248	177	34	119	55	1,7
Mqua	14.257	10.504	3.753	74	554	40	26,3
Mrot	10.788	10.448	340	36	190	53	3,2

Tabela com os resultados do OrthoFinder contendo para cada espécie, o número de genes analisados, quantos destes genes foram agrupados e quantos não foram, além disso também relativo a cada espécie quantos ortogrupos são espécie-específico e quantos genes estão nestes grupos, bem como a porcentagem de genes caracterizados como ortólogos 1:1 e a porcentagem de genes não agrupados.

4.1.4 Combinação dos resultados RBH e Orthofinder

Foram encontrados 4.979 ortólogos 1:1 com a abordagem RBH e 5.755 com o OrthoFinder. A intersecção dos ortólogos identificados por cada abordagem revelou 4.347 ortólogos 1:1 em concordância, que correspondem a 87,3% e 75,5% dos ortólogos identificados pelo RBH e OrthoFinder, respectivamente. Os resultados individuais das duas abordagens, bem como a combinação entre elas estão disponíveis no link https://shorturl.at/lYzvS.

Para explorar a função putativa destes genes, realizamos uma análise de enriquecimento funcional na plataforma DAVID. Desta análise podemos destacar alguns pontos como a ontologia gênica dos termos classificando-os em processo biológico, componente celular e função molecular, os resultados significativos ($p<0,05$) podem ser observados nas Tabelas 5, 6 e 7, assim como a relação dos termos com a contagem dos genes ligados a ele e o p-valor referente. Um resultado inesperado foi a presença de um gene relacionado com o veneno da *Apis mellifera*, o veneno serina carboxipeptidase (LOC410451) na lista dos genes compartilhados por todas as espécies. Esse veneno é expresso no ducto, sendo secretado e apresentando capacidade de causar reação alérgica em humanos, de acordo com a descrição do DAVID.

Tabela 5: Processos biológicos

Termo	Contagem	Porcentagem (%)	p-valor
Transporte de proteínas	75	1,7	7,20E-06
Transporte	325	7,5	2,50E-03
Processamento de mRNA	45	1	4,60E-03
Via de conjugação de Ubl	54	1,2	6,10E-03
Dano ao DNA	49	1,1	1,30E-02
Reparo do DNA	44	1	1,90E-02
Ciclo celular	51	1,2	2,50E-02

Tabela contendo os resultados do DAVID para o processo biológico, relacionando o termo enriquecido com a contagem dos genes e o percentual destes sobre o total dos 4347 genes analisados, bem como o p-valor da análise.

Tabela 6: Componente celular

Termo	Contagem	Porcentagem (%)	p-valor
Núcleo	519	11,9	1,10E-10
Vesícula citoplasmática	21	0,5	9,30E-04
Proteassomo	24	0,6	1,90E-03
Retículo endoplasmático	75	1,7	2,90E-03
Complexo de Golgi	48	1,1	3,20E-03
Citoplasma	271	6,2	4,10E-03
Mitocôndria	109	2,5	7,00E-03
Membrana interna da mitocôndria	40	0,9	2,00E-02

Tabela contendo os resultados do DAVID para o componente celular, relacionando o termo enriquecido com a contagem dos genes e o percentual destes sobre o total dos 4347 genes analisados, bem como o p-valor da análise.

Tabela 7: Função molecular

Termo	Contagem	Porcentagem (%)	p-valor
Transferase	401	9,2	5,60E-05
Proteína quinase serina/treonina	46	1,1	6,80E-03
Aminoacil-tRNA sintetase	15	0,3	1,20E-02
Ativador	31	0,7	1,70E-02
Helicase	48	1,1	1,80E-02
Ligação ao RNA	66	1,5	2,40E-02
Aciltransferase	28	0,6	3,30E-02
Protease tiol	19	0,4	3,30E-02
Glicosiltransferase	29	0,7	4,70E-02

Tabela contendo os resultados do DAVID para a função molecular, relacionando o termo enriquecido com a contagem dos genes e o percentual destes sobre o total dos 4347 genes analisados, bem como o p-valor da análise.

Também tivemos como resultado o enriquecimento de vias metabólicas registrado na Tabela 8 abaixo:

Tabela 8: Vias biológicas

Termo	Contagem	Porcentagem (%)	p-valor
Transporte nucleocitoplasmático	57	1,3	4,40E-03
Endocitose	72	1,7	9,90E-03
Exportação de proteínas	21	0,5	1,10E-02
Via de vigilância do mRNA	43	1	1,60E-02
Apoptose - mosca	32	0,7	2,40E-02
Reparo por excisão de nucleotídeos	34	0,8	3,00E-02
Proteassomo	28	0,6	3,50E-02

Tabela contendo os resultados do DAVID para as vias biológicas, relacionando o termo enriquecido com a contagem dos genes e o percentual destes sobre o total dos 4.347 genes analisados, bem como o p-valor da análise.

A identificação de ortólogos contribui para a pesquisa na área em expansão que é a Genômica da Biodiversidade (LANGSCHIED *et al*., 2024). As aplicações do conhecimento sobre genes espécie-específicos, também chamados de genes taxonomicamente restritos, são diversas, sendo uma delas o desenvolvimento de marcadores para a identificação de espécies. Trine e colaboradores (2023), utilizaram genes espécie-específicos em um estudo de metagenômica, cujo objetivo era identificar e quantificar a abundância relativa das espécies (ZACHARIASEN *et al.*, 2023).

A identificação de genes espécie-específicos também contribuem com o entendimento das novidades fenotípicas. Por exemplo, na anêmona *Nematostella vectensis*, genes espécie-específicos são responsáveis pela diferenciação dos poucos tipos celulares que formam este organismo (BABONIS; MARTINDALE; RYAN, 2016). Em abelhas da espécie *A. mellifera*, é sugerido que os genes espécie-específicos tenham desempenhado um importante papel no surgimento da eussocialidade. Isto porque, foi encontrado que os transcritos desses genes estão duas vezes mais concentrados em operárias do que em rainhas, lembrando que o comportamento das operárias nas espécies eussociais representa uma novidade (JOHNSON; TSUTSUI, 2011).

Como os genes conservados frequentemente mantêm sua função em diferentes espécies, conhecer tais genes em um grupo de espécies pode contribuir com a identificação de

processos biológicos essenciais para a sobrevivência do grupo. Oz e colaboradores (2022) testaram a hipótese de que genes essenciais conservados representam uma fonte de genes pró-longevidade. Dos genes essenciais conservados entre *Saccharomyces cerevisiae*, *Homo sapiens* e *Caenorhabditis elegans*, 21% aumentaram a longevidade replicativa de *S. cerevisiae* quando superexpressos (OZ *et al.*, 2022).

4.2 Expansão e retração das famílias gênicas

Os resultados estatísticos do CAFE mostraram que o valor da máxima verossimilhança (-lnL) foi 81203,2. O valor de Lambda, que representa a probabilidade de ganho e perda de genes por gene por unidade de tempo na filogenia, foi de 0,0012981321882984, sendo o valor máximo possível de Lambda para esta topologia 0,0121951. Foram testados 37 valores de lambda, dos quais nenhum foi rejeitado (0% de rejeição).

Além disso, também como resultado tivemos o agrupamento dos genes em 9.071 famílias, que foram filtradas em 8.178, excluindo aquelas que possuíam pelo menos uma espécie com 100 genes ou mais presentes naquela família. Destas famílias filtradas, 7.564 foram analisadas pelo CAFE quanto às expansões (Figura 4) e retrações (Figura 5) nas espécies analisadas.

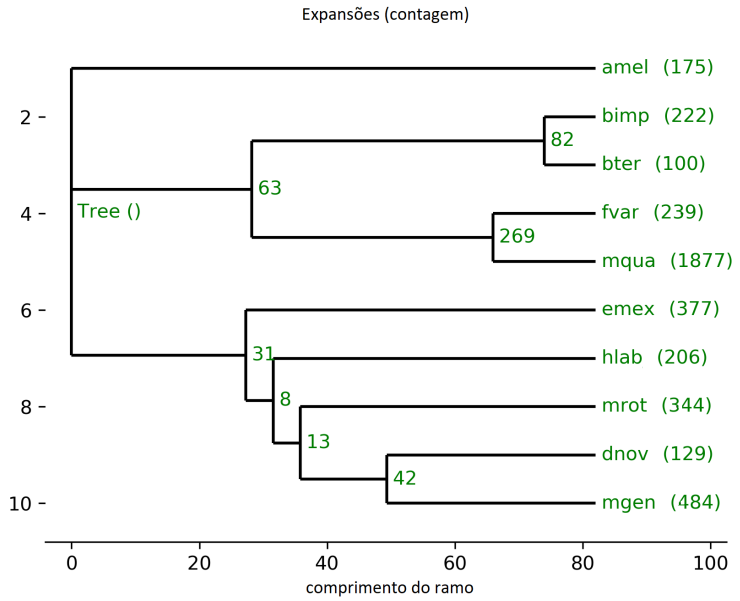

Figura 4. Árvore filogenética das espécies com as expansões das famílias

Imagem contendo a evolução das famílias gênicas ao longo da árvore filogenética, com o comprimento do ramo refletindo o tempo evolutivo entre os nós e os números em verde representando as expansões das famílias gênicas de acordo com a análise realizada pelo CAFE.

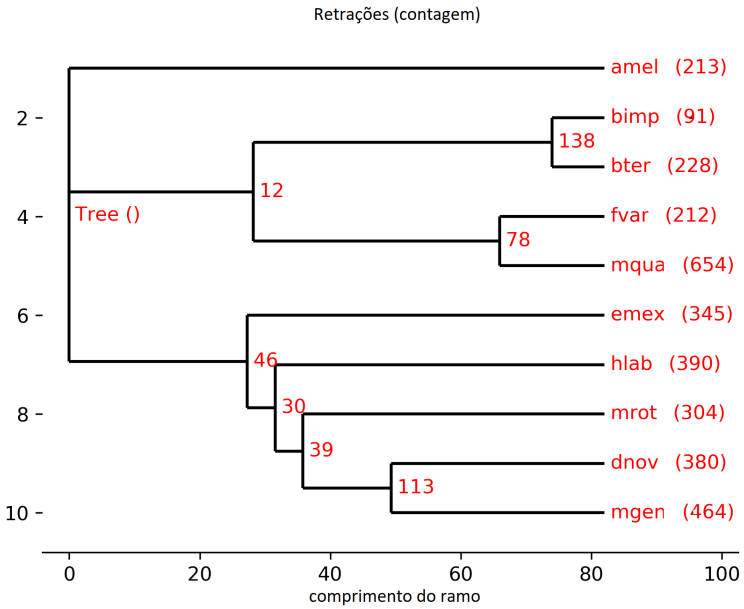

Figura 5. Árvore filogenética das espécies com as retrações das famílias

Imagem contendo a evolução das famílias gênicas ao longo da árvore filogenética, com o comprimento do ramo refletindo o tempo evolutivo entre os nós e os números em vermelho representando as retrações das famílias gênicas de acordo com a análise realizada pelo CAFE.

Das famílias gênicas analisadas pelo CAFE, 821 obtiveram resultados significativos (p-valor < 0,05) e com o uso de *scripts* em Python foi identificado que destas famílias 200 se expandiram unicamente no grupo das espécies eussociais, 83 sofreram expansão exclusiva nas sociais simples e apenas 6 se encontram expandidas apenas nas espécies solitárias. Para esta comparação, consideramos a expansão em pelo menos uma das espécies que compõem cada grupo.

Dentre as 200 famílias que tiveram expansão unicamente nas espécies eussociais, destacamos a família que compreende os genes *farnesyl pyrophosphate synthase* e *farnesyl pyrophosphate synthase-like,* que possuem sete componentes em *Apis mellifera* (genes 102654683, 551189, 551910, 726969, 727618 e 107965814, 726859) e apenas um gene nas outras espécies de abelhas.

O aumento da expansão das famílias gênicas de espécies eussociais em relação às espécies com sociedade simples e solitárias, indica uma maior complexidade do genoma destas espécies. Esse resultado está em concordância com o trabalho de SHELL e colaboradores (2021), que também encontraram expansões em espécies eussociais de abelhas, semelhantemente à TONG e colaboradores (2020), que identificaram expansões nas famílias gênicas de espécies sociais de aranha. Uma das causas de expansão das famílias é a duplicação dos genes, sendo a duplicação gênica uma fonte significativa de inovação funcional em genoma (TICKLE; URRUTIA, 2017), as 200 famílias expandidas nas espécies eussociais poderiam explicar a eussocialidade dessas espécies e sua capacidade de viver em sociedades com divisão de tarefas reprodutivas.

É importante mencionar que a identificação de eventos de expansão e retração das famílias gênicas pode ser influenciada pela qualidade da montagem dos genomas e da predição e anotação dos genes codificadores de proteína. Para avaliar o impacto destas variáveis, calculamos o valor de Epsilon (e) que estima a porcentagem de famílias gênicas que podem ter sofrido erros na identificação de seus genes membros. Considerando todas as espécies de abelhas, a estimativa do erro de distribuição das famílias gênicas (e) foi de 0,0256, o que significa que pouco mais de 2,5% das famílias gênicas podem apresentar algum erro na quantificação dos genes membros. Um estudo sobre evolução das famílias gênicas em insetos hematófagos e não-hematófagos encontrou uma taxa de 7,5% (FREITAS; NERY, 2020).

6 CONCLUSÃO E PERSPECTIVAS

Ao analisar sequências genômicas de dez espécies de abelhas que diferem quanto à organização social, este trabalho contribuiu com a identificação de 4.347 ortólogos 1:1 e da presença ou ausência de cada gene em cada uma das dez espécies (https://shorturl.at/lYzvS). O compartilhamento destes resultados beneficiará grupos de pesquisa que se dedicam aos estudos das abelhas, uma vez que estudos comparativos requerem constante avaliação do grau de conservação dos genes.

Outra contribuição deste trabalho foi o estabelecimento de um fluxo de análise automatizado. Aqui, apresentamos os resultados da análise do genoma de dez espécies de abelhas, no entanto, o fluxo de análise permite a inclusão de dados genômicos de outras espécies com facilidade. A expansão do banco de dados com sequências de proteínas de outras abelhas com genoma disponível mais recentemente, como *Tetragonisca angustula* (FERRARI *et al.*, 2024), *Melipona bicolor* (ARAUJO *et al.*, 2024) e *Trigonisca nataliae* (FRANÇOSO *et al.*, 2023) está em andamento e em breve nossos resultados serão atualizados. Além disso, a automatização desse fluxo de análise foi testada com um banco de dados formado pelas sequências genômicas das dez espécies de abelhas, mosca-das-frutas, humano e camundongo e teve um bom desempenho.

Os próximos passos incluem: (1) adequação dos parâmetros das análises de expansão e retração das famílias gênicas considerando a taxa de erro proveniente de montagens de genoma e predição de genes imprecisos; (2) integração da base de dados de ortólogos, com a classificação dos genes em famílias gênicas e a anotação funcional (em andamento) dos domínios proteicos conservados; (3) organização dos resultados e escrita do manuscrito para publicação em revista científica com revisão pelos pares.

7 REFERÊNCIAS BIBLIOGRÁFICAS

ALTSCHUL, S. F. *et al.* Basic local alignment search tool. **Journal of Molecular Biology**, out. 1990. v. 215, n. 3, p. 403–410. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/2231712/>. Acesso em: 12 dez. 2024.

ARAUJO, N. De S. *et al.* Insights from Melipona bicolor hybrid genome assembly: a stingless bee genome with chromosome-level scaffold. England: BMC genomics, Spring. 2024. v. 25, n. 1, p. 171. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/38350872/>. Acesso em: 7 jan. 2025.

BABONIS, L. S.; MARTINDALE, M. Q.; RYAN, J. F. Do novel genes drive morphological novelty? An investigation of the nematosomes in the sea anemone Nematostella vectensis. **BMC Evolutionary Biology**, 23 maio. 2016. v. 16, n. 1. Disponível em: <https://bmcecolevol.biomedcentral.com/articles/10.1186/s12862-016-0683-3>. Acesso em: 7 jan. 2025.

BAILEY, J. A.; EICHLER, E. E. Primate segmental duplications: crucibles of evolution, diversity and disease. **Nature Reviews Genetics**, 13 jun. 2006. v. 7, n. 7, p. 552–564. Disponível em: <https://www.nature.com/articles/nrg1895>. Acesso em: 31 jul. 2024.

BAXEVANIS, A. D.; OUELLETTE, F. **Bioinformatics**. 2. ed. [S.l.]: John Wiley & Sons, 2004. V. 43.

BROWN, T. A. **Genomes 3**. New York: Garland Science Pub, 2007.

BUCHFINK, B.; REUTER, K.; DROST, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. **Nature Methods**, abr. 2021. v. 18, n. 4, p. 366–368. Disponível em: <https://www.nature.com/articles/s41592-021-01101-x>. Acesso em: 5 dez. 2024.

BUCHFINK, B.; XIE, C.; HUSON, D. H. Fast and sensitive protein alignment using DIAMOND. **Nature Methods**, 17 nov. 2014. v. 12, n. 1, p. 59–60. Disponível em: <https://www.nature.com/articles/nmeth.3176>. Acesso em: 12 dez. 2024.

CAPELLA-GUTIERREZ, S.; SILLA-MARTINEZ, J. M.; GABALDON, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. **Bioinformatics**, 8 jun. 2009. v. 25, n. 15, p. 1972–1973. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2712344/>. Acesso em: 29 dez. 2024.

CASTILLO-MORALES, A. *et al.* Increased brain size in mammals is associated with size variations in gene families with cell signalling, chemotaxis and immune-related functions. **Proceedings of the Royal Society B: Biological Sciences**, 22 jan. 2014. v. 281, n. 1775. Disponível em: <https://royalsocietypublishing.org/doi/10.1098/rspb.2013.2428>. Acesso em: 31 jul. 2024.

CASTILLO-MORALES, *et al.* Neocortex expansion is linked to size variations in gene families with chemotaxis, cell–cell signalling and immune response functions in mammals. **Open biology**, 1 out. 2016. v. 6, n. 10. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5090057/#>. Acesso em: 31 jul. 2024.

CHEN, L. *et al.* Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. **Science**, 20 jun. 2019. v. 364, n. 6446. Disponível em: <https://www.science.org/doi/10.1126/science.aav6202>. Acesso em: 31 jul. 2024.

DE BIE, T. *et al.* CAFE: a computational tool for the study of gene family evolution. **Bioinformatics**, 16 mar. 2006. v. 22, n. 10, p. 1269–1271. Disponível em: <https://doi.org/10.1093/bioinformatics/btl097>. Acesso em: 4 dez. 2024.

DEMUTH, J. P. *et al.* The Evolution of Mammalian Gene Families. **PLoS ONE**, 20 dez. 2006. v. 1, n. 1. Disponível em: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0000085>. Acesso em: 31 jul. 2024.

DEMUTH, J. P.; HAHN, M. W. The life and death of gene families. **BioEssays**, jan. 2009. v. 31, n. 1, p. 29–39. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/bies.080085>. Acesso em: 8 ago. 2024.

EMMS, D. M.; KELLY, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. **Genome Biology**, 6 ago. 2015. v. 16, n. 1. Disponível em: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0721-2>. Acesso em: 5 dez. 2024.

EMMS, D. M.; KELLY, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. **Genome Biology**, 14 nov. 2019. v. 20, n. 1. Disponível em: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1832-y>. Acesso em: 5 dez. 2024.

EXPOSITO-ALONSO, M. *et al.* Genomic basis and evolutionary potential for extreme drought adaptation in Arabidopsis thaliana. **Nature Ecology & Evolution**, 1 fev. 2018. v. 2, n. 2, p. 352–358. Disponível em: <https://www.nature.com/articles/s41559-017-0423-0>.

FAVRE, P. *et al.* A novel bioinformatics pipeline to discover genes related to arbuscular mycorrhizal symbiosis based on their evolutionary conservation pattern among higher plants. **BMC Plant Biology**, dez. 2014. v. 14, n. 1. Acesso em: 31 jul. 2024.

FERRARI, R. R. *et al.* The nuclear and mitochondrial genome assemblies of Tetragonisca angustula (Apidae: Meliponini), a tiny yet remarkable pollinator in the Neotropics. BMC Genomics, 11 jun. 2024. v. 25, n. 1. Disponível em: <https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-024-10502-z>. Acesso em: 7 jan. 2025.

FRANÇOSO, E. *et al.* The complete mitochondrial genome of Trigonisca nataliae (Hymenoptera, Apidae) assemblage reveals heteroplasmy in the control region. Netherlands: Gene, 2023. v. 881, p. 147621. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/37419430/>. Acesso em: 7 jan. 2025.

FREITAS, L.; NERY, M. F. Expansions and contractions in gene families of independently-evolved blood-feeding insects. **BMC Evolutionary Biology**, 17 jul. 2020. v. 20, n. 1. Disponível em: <https://bmcecolevol.biomedcentral.com/articles/10.1186/s12862-020-01650-3>. Acesso em: 4 jan. 2025.

GAITHER, M. R. *et al.* Genomics of habitat choice and adaptive evolution in a deep-sea fish. **Nature Ecology & Evolution**, 5 mar. 2018. v. 2, n. 4, p. 680–687. Disponível em: <https://www.nature.com/articles/s41559-018-0482-x>. Acesso em: 31 jul. 2024.

GOODMAN, M. *et al.* Phylogenomic analyses reveal convergent patterns of adaptive evolution in elephant and human ancestries. **Proceedings of the National Academy of Sciences**, 8 dez. 2009. v. 106, n. 49, p. 20824–20829. Disponível em: <https://www.pnas.org/doi/abs/10.1073/pnas.0911239106>. Acesso em: 31 jul. 2024.

GOU, X. *et al.* Whole-genome sequencing of six dog breeds from continuous altitudes reveals adaptation to high-altitude hypoxia. **Genome Research**, 10 abr. 2014. v. 24, n. 8, p. 1308–1315. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/24721644/>. Acesso em: 31 jul. 2024.

GRUS, W. E. *et al.* Dramatic variation of the vomeronasal pheromone receptor gene repertoire among five orders of placental and marsupial mammals. **Proceedings of the National Academy of Sciences**, 24 mar. 2005. v. 102, n. 16, p. 5767–5772. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC556306/>. Acesso em: 31 jul. 2024.

HAHN, M. W. Estimating the tempo and mode of gene family evolution from comparative genomic data. **Genome Research**, 1 ago. 2005. v. 15, n. 8, p. 1153–1160. Disponível em: <https://genome.cshlp.org/content/15/8/1153>. Acesso em: 31 jul. 2024.

HAHNLAB. CAFE5/docs/tutorial/tutorial.md at master · hahnlab/CAFE5. **GitHub**, 2018. Disponível em: <https://github.com/hahnlab/CAFE5/blob/master/docs/tutorial/tutorial.md>. Acesso em: 4 dez. 2024.

HAN, M. V. *et al.* Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3. **Molecular Biology and Evolution**, 1 ago. 2013. v. 30, n. 8, p. 1987–1997. Disponível em: <https://academic.oup.com/mbe/article/30/8/1987/1017616?login=true>. Acesso em: 8 ago. 2024.

HARRISON, M. C. *et al.* Hemimetabolous genomes reveal molecular basis of termite eusociality. **Nature Ecology & Evolution**, 1 mar. 2018. v. 2, n. 3, p. 557–566. Disponível em: <https://www.nature.com/articles/s41559-017-0459-1>. Acesso em: 31 jul. 2024.

HERNÁNDEZ-SALMERÓN, J. E.; MORENO-HAGELSIEB, G. Progress in quickly finding orthologs as reciprocal best hits: comparing blast, last, diamond and MMseqs2. **BMC Genomics**, 24 out. 2020. v. 21, n. 1. Disponível em: <https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-020-07132-6>. Acesso em: 17 dez. 2024.

HOLT, R. A. *et al.* The Genome Sequence of the Malaria MosquitoAnopheles gambiae. **Science**, 4 out. 2002. v. 298, n. 5591, p. 129–149. Disponível em: <https://science.sciencemag.org/content/298/5591/129>. Acesso em: 8 ago. 2024.

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**, 25 nov. 2009. v. 37, n. 1, p. 1–13. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/19033363/>. Acesso em: 5 dez. 2024.

HUANG, D. W.; SHERMAN, B. T.; LEMPICKI, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. **Nature Protocols**, 18 dez. 2018. v. 4, n. 1, p. 44–57. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/19131956/>. Acesso em: 5 dez. 2024.

HUBISZ, M. J. *et al.* Error and Error Mitigation in Low-Coverage Genome Assemblies. **PLoS ONE**, 14 fev. 2011. v. 6, n. 2. Disponível em: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0017034>. Acesso em: 8 ago. 2024.

JOHNSON, B. R.; TSUTSUI, N. D. Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. **BMC Genomics**, 29 mar. 2011. v. 12, n. 1. Disponível em: <https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-12-164>. Acesso em: 7 jan. 2025.

KAPHEIM, K. M. *et al.* Genomic signatures of evolutionary transitions from solitary to group living. **Science**, 14 maio. 2015. v. 348, n. 6239, p. 1139–1143. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5471836/>. Acesso em: 31 jul. 2024.

KATOH, K. *et al.* MAFFT: a Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. **Nucleic Acids Research**, 15 jul. 2002. v. 30, n. 14, p. 3059–3066. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/12136088/>. Acesso em: 29 dez. 2024.

KATOH, K.; STANDLEY, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. **Molecular Biology and Evolution**, 16 jan. 2013. v. 30, n. 4, p. 772–780. Disponível em: <https://academic.oup.com/mbe/article/30/4/772/1073398>. Acesso em: 29 dez. 2024.

KUMAR, S. *et al.* TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Molecular Biology and Evolution, 6 abr. 2017. v. 34, n. 7, p. 1812–1819. Disponível em: <https://academic.oup.com/mbe/article/34/7/1812/3091705>. Acesso em: 10 jan. 2025.

LANGSCHIED, F. *et al.* Quest for Orthologs in the Era of Biodiversity Genomics. Genome Biology and Evolution, 1 out. 2024. v. 16, n. 10. Disponível em: <https://academic.oup.com/gbe/article/16/10/evae224/7822254>. Acesso em: 6 jan. 2025.

MCKENZIE, S. K. *et al.* Transcriptomics and neuroanatomy of the clonal raider ant implicate an expanded clade of odorant receptors in chemical communication. **Proceedings of the National Academy of Sciences**, 22 nov. 2016. v. 113, n. 49, p. 14091–14096. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5150400/>. Acesso em: 31 jul. 2024.

MCKENZIE, S. K.; OXLEY, P. R.; KRONAUER, D. J. Comparative genomics and transcriptomics in ants provide new insights into the evolution and function of odorant binding and chemosensory proteins. **BMC Genomics**, 2014. v. 15, n. 1, p. 718–732. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4161878/>. Acesso em: 31 jul. 2024.

MENDES, F. K. *et al.* CAFE 5 models variation in evolutionary rates among gene families. **Bioinformatics**, 1 dez. 2020. v. 36, n. 22-23, p. 5516–5518. Disponível em: <https://doi.org/10.1093/bioinformatics/btaa1022>. Acesso em: 4 dez. 2024.

MICHENER, C. D. **The Social Behavior of the Bees: A Comparative Study**. [S.l.]: Harvard University Press, 1974.

MICHENER, C. D. **The bees of the world**. Baltimore: Johns Hopkins University Press, 2007.

NGUYEN, L.-T. *et al.* IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. **Molecular Biology and Evolution**, 3 nov. 2014. v. 32, n. 1, p. 268–274. Disponível em: <https://academic.oup.com/mbe/article/32/1/268/2925592?login=false>. Acesso em: 5 dez. 2024.

NIIMURA, Y.; NEI, M. Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. **Proceedings of the National Academy of Sciences**, 11 maio. 2005. v. 102, n. 17, p. 6039–6044. Disponível em: <https://www.pnas.org/content/102/17/6039>. Acesso em: 31 jul. 2024.

OHNO, S. **Evolution by gene duplication**. Berlin: Springer, 1970.

OZ, N. *et al.* Evidence that conserved essential genes are enriched for pro-longevity factors. GeroScience, 13 jun. 2022. v. 44, n. 4, p. 1995–2006. Disponível em: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9616985/>. Acesso em: 7 jan. 2025.

RANSON, H. *et al.* Evolution of Supergene Families Associated with Insecticide Resistance. **Science**, 4 out. 2002. v. 298, n. 5591, p. 179–181. Disponível em: <https://www.science.org/doi/10.1126/science.1076781>. Acesso em: 31 jul. 2024.

SANDERSON, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. **Bioinformatics**, 22 jan. 2003. v. 19, n. 2, p. 301–302. Disponível em: <https://doi.org/10.1093/bioinformatics/19.2.301>. Acesso em: 23 dez. 2024.

SANTOS *et al.* SpliceProt 2.0: A Sequence Repository of Human, Mouse, and Rat Proteoforms. **International journal of molecular sciences**, 18 jan. 2024. v. 25, n. 2, p. 1183–1183. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/38256255/>. Acesso em: 29 dez. 2024.

SANTOS, M. E. *et al.* Taxon-restricted genes at the origin of a novel trait allowing access to a new environment. **Science**, 20 out. 2017. v. 358, n. 6361, p. 386–390. Disponível em: <https://www.science.org/doi/10.1126/science.aan2748>. Acesso em: 31 jul. 2024.

SAYERS, E. W. *et al.* Database resources of the National Center for Biotechnology Information in 2023. **Nucleic Acids Research**, 12 nov. 2022. v. 51, n. D1, p. D29–D38. Disponível em: <https://academic.oup.com/nar/article/51/D1/D29/6825348#google_vignette>. Acesso em: 23 dez. 2024.

SHELL, W. A. *et al.* Sociality sculpts similar patterns of molecular evolution in two independently evolved lineages of eusocial bees. **Communications Biology**, 26 fev. 2021. v. 4, n. 1. Disponível em: <https://www.nature.com/articles/s42003-021-01770-6>. Acesso em: 9 dez. 2024.

SIMOLA, D. F. *et al.* Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. **Genome Research**, 1 maio. 2013. v. 23, n. 8, p. 1235–1247. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3730098/>. Acesso em: 31 jul. 2024.

STEINEGGER, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. **BMC Bioinformatics**, 14 set. 2019. v. 20, n. 1. Acesso em: 1º mar. 2021.

THOMAS, G. W. C. *et al.* Gene content evolution in the arthropods. **Genome Biology**, 23 jan. 2020. v. 21, n. 1. Disponível em: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1925-7>. Acesso em: 31 jul. 2024.

TICKLE, C.; URRUTIA, A. O. Perspectives on the history of evo-devo and the contemporary research landscape in the genomics era. **Philosophical Transactions of the Royal Society B: Biological Sciences**, 5 fev. 2017. v. 372, n. 1713, p. 20150473. Disponível em: <https://royalsocietypublishing.org/doi/10.1098/rstb.2015.0473>. Acesso em: 31 jul. 2024.

TONG, C. *et al.* Comparative Genomics Identifies Putative Signatures of Sociality in Spiders. **Genome Biology and Evolution**, 21 jan. 2020. v. 12, n. 3, p. 122–133. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7108510/>. Acesso em: 18 dez. 2024.

VAN DONGEN, S. Graph Clustering Via a Discrete Uncoupling Process. **SIAM Journal on Matrix Analysis and Applications**, jan. 2008. v. 30, n. 1, p. 121–141. Disponível em: <https://epubs.siam.org/doi/10.1137/040608635>. Acesso em: 5 dez. 2024.

WANG, K. *et al.* Morphology and genome of a snailfish from the Mariana Trench provide insights into deep-sea adaptation. **Nature Ecology & Evolution**, 15 abr. 2019. v. 3, n. 5, p. 823–833. Disponível em: <https://www.nature.com/articles/s41559-019-0864-8>. Acesso em: 31 jul. 2024.

WILSON, E. O.; HOLLDOBLER, B. Eusociality: Origin and consequences. **Proceedings of the National Academy of Sciences**, 12 set. 2005. v. 102, n. 38, p. 13367–13371. Disponível em: <https://ncbi.nlm.nih.gov/pmc/articles/PMC1224642/>. Acesso em: 31 jul. 2024.

WU, H. *et al.* Camelid genomes reveal evolution and adaptation to desert environments. **Nature Communications**, 21 out. 2014. v. 5, n. 1. Disponível em: <https://www.nature.com/articles/ncomms6188.pdf?origin=ppub>. Acesso em: 31 jul. 2024.

XU, J. *et al.* Genomic Basis of Adaptive Evolution: The Survival of Amur Ide (*Leuciscuswaleckii*) in an Extremely Alkaline Environment. **Molecular Biology and Evolution**, 20 out. 2016. v. 34, n. 1, p. 145–159. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5854124/>. Acesso em: 31 jul. 2024.

ZACHARIASEN, T. *et al.* Identification of representative species-specific genes for abundance measurements. **Bioinformatics Advances**, 1 jan. 2023. v. 3, n. 1. Disponível em: <https://academic.oup.com/bioinformaticsadvances/article/3/1/vbad060/7156835>. Acesso em: 7 jan. 2025.

ZHANG, G. *et al.* Comparative genomics reveals insights into avian genome evolution and adaptation. **Science**, 11 dez. 2014. v. 346, n. 6215, p. 1311–1320.