UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa de Pós Graduação em Genética

Maycon Douglas de Oliveira

FAMÍLIAS E DOMÍNIOS PROTEICOS ASSOCIADOS COM A EVOLUÇÃO DA EUSSOCIALIDADE EM HYMENOPTERA

Belo Horizonte
2022

Maycon Douglas de Oliveira

FAMÍLIAS E DOMÍNIOS PROTEICOS ASSOCIADOS COM A EVOLUÇÃO DA EUSSOCIALIDADE EM HYMENOPTERA

Dissertação apresentada ao Programa de Pós-Graduação em Genética da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Genética.

Orientador: Prof. Dr. Francisco Pereira Lobo

Coorientador: Dr. José Eustáquio dos Santos Junior

Belo Horizonte
2022

043 Oliveira, Maycon Douglas de.
Famílias e domínios proteicos associados com a evolução da eussocialidade em hymenoptera [manuscrito] / Maycon Douglas de Oliveira. – 2022.
123 f. : il. ; 29,5 cm.

Orientador: Francisco Pereira Lobo. Coorientador: José Eustáquio dos Santos Junior.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Pós-Graduação em Genética.

1. Genética. 2. Himenópteros. 3. Reprodução. 4. Domínios Proteicos. 5. Análise de Variância. I. Lobo, Francisco Pereira. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 575.1

Ficha Catalográfica elaborada pelo bibliotecário Rafael Gonçalves Dias (CRB: 6 - 3277)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Genética

FOLHA DE APROVAÇÃO

"Famílias e Domínios Proteicos Associados com a Evolução da Eussocialidade em Hymenoptera"

Maycon Douglas de Oliveira

Dissertação aprovada pela banca examinadora constituída pelos Professores:

Francisco Pereira Lobo
UFMG

José Eustáquio dos Santos Júnior
UFMG

Lucas Neves Perillo
UFMG

Fernada Antunes Carvalho

UFMG

Belo Horizonte, 20 de dezembro de 2022.

Documento assinado eletronicamente por **Francisco Pereira Lobo**, **Professor do Magistério Superior**, em 20/12/2022, às 12:38, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Lucas Neves Perillo**, **Usuário Externo**, em 20/12/2022, às 12:47, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **José Eustáquio dos Santos Júnior**, **Usuário Externo**, em 20/12/2022, às 12:48, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Fernanda Antunes Carvalho**, **Professora do Magistério Superior**, em 20/12/2022, às 15:13, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1972002** e o código CRC **776EF53B**.

Referência: Processo nº 23072.273751/2022-19 SEI nº 1972002

AGRADECIMENTOS

Nenhum trabalho pode ser feito sozinho. Pelo menos, não com saúde, qualidade e não sem sacrificar uma oportunidade única de aprender e se conectar. Às vezes eu mesmo preciso me lembrar desse fato já que é tão fácil se desviar e acabar por se isolar “sem querer”. Ainda assim, os melhores frutos são aqueles que colhemos de mãos dadas com quem amamos, respeitamos e admiramos.

Este trabalho nunca seria realizado sem a compreensão e o apoio incondicional de algumas pessoas lúcidas – ou loucas – o bastante para compreender, após muito falatório, o árduo caminho de uma formação acadêmica.

Primeiramente, Rafaela, o amor da minha vida, e meus pais, Claudia e Roney. Apesar de evidências do contrário, alunos de mestrado precisam se alimentar, precisam de cuidado, carinho, compreensão e às vezes atenção. Precisam de orelhas saudáveis e dispostas a ouvir suas loucuras e excitações que muitas vezes são complicadas de compartilhar. Essas três pessoas foram essenciais em todos os aspectos para a manutenção da vida e da saúde física e mental daquele que vos fala. Sem seu apoio, nada disso seria possível: e não falo apenas de minha jornada no Mestrado. Eu poderia escrever uma dissertação inteira apenas com agradecimentos a vocês, mas seria muito difícil garantir minha integridade no fim, mesmo com suas tentativas incessantes de me manter de pé. Portanto, espero que compreendam com estas poucas palavras o quanto vocês significam para mim.

Em segundo lugar, o restante da minha família: minha irmã, Duda; Rafael, meu irmão de outra mãe, meus sogros, Célio e Jaci, que são para mim como segundos pais; e a todos as avós, avôs, tios, tias, primos e primas, que são muitos. Deste grupo, gostaria de dar espaço especial ao meu avô, Eduardo, que por pouco não pôde prestigiar este momento importante da minha vida. Tenho certeza que ficaria orgulhoso.

Terceiramente, gostaria de agradecer aos meus outros “irmãos de outras mães”, amigos dos quais nunca quero me separar e que por minha vida inteira estiveram do meu lado: Lucas Erickson, Lucas Gustavo e Vinícius. Espero que possa levar a amizade de vocês comigo para o resto da vida e mais um tanto. Agradeço também aos amigos mais recentes, também tão importantes e queridos, e que fizeram da minha jornada acadêmica uma caminhada de alegria e muito prazer:

Pedro, Flávia, Winy e Lucas. Espero que possam perdoar minha ausência e que, um dia, possamos reparar o tempo perdido.

Sem falta, gostaria de agradecer a meu orientador, Chico, primeiro por me aturar e segundo por acreditar em mim e, em 2018, apesar de toda resistência e bom senso, ter me dado esta oportunidade única. Chico, em momento algum duvidei da minha escolha em ter você como orientador, e é graças a você que hoje eu posso dizer com muita alegria que eu trabalho *exatamente* com aquilo que amo. Muito obrigado por isso.

Gostaria também de agradecer ao meu coorientador, Zé, por ter acreditado em mim e ter convencido o Chico a me aceitar, além de ter me ensinado quase tudo o que sei sobre Hymenoptera e filogenia. Conhecer você no meu curto período no LBEM foi uma das “coincidências” mais felizes que poderiam ter acontecido na minha vida acadêmica. Nesta nota, agradeço também ao Prof. Fabrício, por ter um dia me aceitado como orientando e por me iniciar na vida acadêmica e acreditar em mim em cada passo seguinte.

Também dedico um agradecimento especial aos amigos e companheiros antigos e atuais de laboratório, tanto os amigos do LAB e egressos quanto aos amigos do LBEM, que me iniciaram na vida científica. Em especial, gostaria de citar, sem ordem particular: Agnello, Igor, Thieres, Camilla, Davidson, Cayo, Rahyssa, Iza, Jean, Zandora, Thais, Alison, Aline e Giovanni. Àqueles que não foram citados nominalmente, não pensem que vocês são menos importantes pra mim, pois não são.

Foi uma jornada cansativa e desafiadora, mas, graças àqueles que sempre estiveram do meu lado, acredito que consegui percorrê-la com tranquilidade. Vocês terão para sempre minha gratidão e meu carinho.

“Esperar e ter esperança” (DUMAS, 2021).

RESUMO

A ordem Hymenoptera consiste nos insetos comumente conhecidos como abelhas, formigas, vespas e moscas-serra, importantes para a manutenção da diversidade e homeostase dos ecossistemas (*e. g.* polinização em ambientes rurais e silvestres e dispersão de sementes nativas). Os Hymenoptera possuem o maior número de ganhos e perdas independentes da eussocialidade entre os metazoários. A busca por regiões homólogas significativamente associadas com o fenótipo da eussocialidade, através da análise de famílias gênicas e domínios proteicos, pode fornecer informações importantes para a compreensão molecular desse fenômeno. Entretanto, dados oriundos de espécies não são independentes entre si já que estas compartilham ancestrais comuns. Diversos métodos comparativos foram desenvolvidos nas últimas décadas para permitir a busca por associações entre dados oriundos de organismos filogeneticamente relacionados, incorporando a filogenia como um parâmetro adicional dos modelos. Neste trabalho, integramos dados filogenéticos, fenotípicos e genômicos para construir modelos estatísticos *phylogeny-aware* e buscar grupos de homólogos cuja abundância esteja significativamente associada à eussocialidade em 62 genomas anotados de alta qualidade de Hymenoptera representativos de espécies classificadas como eussociais ou solitárias. De um total de 2.045.867 regiões homólogas pertencentes a 9.662 famílias e domínios proteicos únicos obtidos dos genomas, seis estão significativamente associados à presença/ausência da eussocialidade, representando uma grande diversidade funcional. Destacamos *THAP4*, um gene que parece estar associado com adaptações à vida em ambientes pouco oxigenados e que está expandido em formigas, que fazem ninhos subterrâneos, e o gene *Snx14*, envolvido com o desenvolvimento embrionário e armazenamento de energia, e que está presente apenas em abelhas eussociais. Observamos também genes com pouca caracterização funcional, e que compreendem alvos interessantes para avaliação funcional. Importantemente, todos estes genes estão ausentes em diversas espécies de Hymenoptera, o que sugere que os mesmos não são essenciais. Em conjunto, esses resultados providenciam um importante ponto de partida para a realização de análises funcionais através do *knockout* ou *knockdown* desses genes em Hymenoptera, visando maior compreensão de seu papel na regulação e evolução da eussocialidade.

Palavras-chave: Hymenoptera; eussocialidade; domínios proteicos; métodos comparativos; ANOVA; *phylogeny-aware*.

ABSTRACT

The order Hymenoptera consists of the insects commonly known as bees, ants, wasps and sawflies, critical for the maintenance of the diversity and homeostasis of ecosystems (*e.g.*, pollination in rural and wild environments and dispersion of native seeds). The Hymenoptera possess the largest number of independent gains and losses of eusociality among metazoans. The search for homologous regions significantly associated with the phenotype of eusociality through the analysis of gene families and protein domains may offer important information for the molecular understanding of this phenomenon. However, species-related data are not independent since they share common ancestors. A plethora of comparative methods were developed in the last decades in order to enable the search for associations between data derived from phylogenetically-related organisms, incorporating their phylogeny as an additional parameter for the models. In the present work, we integrated phylogenetic, phenotypic and genomic data in order to build phylogeny-aware statistical models and search for homologous groups whose count is significantly associated with eusociality in 62 high-quality annotated Hymenoptera genomes representative of species classified as eusocial or solitary. From a total of 2,045,867 homologous regions belonging to 9,662 unique protein domains and families gathered from those genomes, six are significantly associated with the presence/absence of eusociality and showcase a rich functional diversity. We highlight *THAP4*, a gene that seems to be associated with adaptations to life in low oxygen environments and is expanded in ants, who build underground nests, and the gene *Snx14*, involved in embryonic development and energy storage and that is only present in eusocial bees. We also observed genes with scarce functional characterization that represent interesting targets for functional evaluation. Importantly, all of those genes are absent in a number of Hymenoptera species, which suggests that they are not essential. In conjunction, these results provide an important starting point for *knockout* and *knockdown*-based functional analyses of those genes in Hymenoptera, devising a larger comprehension of their role in the control and evolution of eusociality.

Keywords: Hymenoptera; eusociality; protein domains; comparative methods; ANOVA; phylogeny-aware.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

A – Nucleotídeo adenina.

ANCOVA – Do inglês “*analysis of covariance*” (análise de covariância).

ANOVA – Do inglês “*analysis of variance*” (análise de variância).

BUSCO – Do inglês “*Benchmarking Universal Single-Copy Orthologs*” (“Comparação de Ortólogos Universais de Cópia Única”), programa de computador utilizado na avaliação da qualidade de montagem e anotação de genomas.

C – Nucleotídeo citosina.

e.g. – Do latim “e*xempli gratia*” (“por exemplo”).

G – Nucleotídeo guanina.

GB – Gigabytes, 1024 megabytes em contagem binária. Medida de capacidade de armazenamento e memória computacional.

ICB – Instituto de Ciências Biológicas da UFMG.

ID – Identificador.

IPR – Identificador único utilizado pelo banco de dados *InterPro* para a sumarização de conjuntos de termos anotadores de proteínas redundantes advindos de bancos de dados distintos.

LAB – Laboratório de Algoritmos em Biologia.

NCBI – Do inglês “*National Center for Biotechnology Information*” (Centro Nacional de Informação Biotecnológica).

PGLS – Do inglês “*phylogenetic generalized least squares*” (“mínimos quadrados generalizados filogenéticos”), modificação da técnica de quadrados ordinários.

PIC – Do inglês “*phylogenetically independent contrasts*” (“contrastes filogeneticamente independentes”).

R – Linguagem de programação e ambiente de análises estatísticas computacionais.

RAM – Do inglês “*random access memory*” (“memória de acesso aleatório”), responsável pelo armazenamento de curto prazo de dados pequenos em computadores.

RNAi – RNA de interferência.

T – Nucleotídeo timina.

UFMG – Universidade Federal de Minas Gerais.

SUMÁRIO

1 INTRODUÇÃO

Os insetos comumente conhecidos como abelhas, formigas, vespas e moscas-serra fazem parte da ordem Hymenoptera, um dos mais diversos grupos de metazoários em termos de número descrito de espécies e adaptações comportamentais (FORBES et al., 2018; GRIMALDI; ENGEL, 2005). Entre as espécies que compõem a ordem, existem grupos conhecidos por sua importância ecológica e econômica: polinizadores, aeradores do solo, dispersores de sementes, fontes de alimentação e produtos para o comércio humano (*e.g.* mel, própolis, cera e geleia real), responsáveis por causar danos a agricultura, agentes de controle biológico, entre outros. Esses insetos apresentam uma gama de comportamentos, indo do comportamento solitário à mais complexa organização social, a eussocialidade (GRIMALDI; ENGEL, 2005; HÖLLDOBLER; WILSON, 2009; HUNT, 2007).

1.1 A ordem Hymenoptera

Com aproximadamente 118.300 espécies formalmente descritas (BÁNKI et al., 2022; SHARKEY, 2007) e com uma estimativa de um número três a vinte vezes maior de espécies (GASTON; GAULD; HANSON, 1996), Hymenoptera (Insecta, Endopterygota) compreende os insetos popularmente conhecidos como abelhas (Apocrita, Anthophila); formigas (Formicoidea); vespas, clado parafilético que inclui todos os Parasitoida (Apocrita sem ferrão), Vespoidea (vespas eussociais) e demais Apocrita não classificados como formigas ou abelhas (HUNT, 2007; PETERS et al., 2017) e moscas-serra, grupo parafilético que inclui Eusymphyta e o restante dos Hymenoptera não pertencentes à subordem Apocrita (BRANSTETTER et al., 2017; PETERS et al., 2017).

Além do grande número de espécies, os Hymenoptera são conhecidos por prestarem uma considerável diversidade de serviços ecossistêmicos importantes, como a polinização, dispersão de sementes e aeração do solo; interações sinantrópicas como a produção de mel, geleia real, cera e própolis; e participação no ciclo agrícola como pragas de interesse e como agentes em controles biológicos de pragas (BRADBEAR, 2009; LESIEUR et al., 2016; MICHENER, 2007; RUNYON et al., 2002; WILSON, 1971).

Os Hymenoptera ainda possuem uma plenitude de estilos de vida altamente especializados e contrastantes no que diz respeito à sua alimentação e organização. Abelhas são comumente polinívoras (MICHENER, 2007), formigas podem ser insetívoras, cultivar fungos ou até mesmo pastorear afídeos secretores de glicose (HÖLLDOBLER; WILSON, 1990), vespas podem ser fitófagas, predadoras, parasitar plantas e outros insetos ou se alimentar de pólen (HERZNER et al., 2013; HUNT, 2007; PÉREZ-LACHAUD; BATCHELOR; HARDY, 2004; STAMP; BOWERS, 1988; WEIBLEN; BUSH, 2002; WIEMER et al., 2012), e moscas-serra normalmente são fitófagas, predadoras ou parasitas (OEYEN et al., 2020; OISHI et al., 1993; PETERSON et al., 2011; SCHMIDT et al., 2010). Além disso, diversas espécies de Hymenoptera são eussociais, fenótipo foco do presente trabalho (GRIMALDI; ENGEL, 2005; HÖLLDOBLER; WILSON, 2009; WILSON, 1971).

1.2 Eussocialidade

A eussocialidade é um traço definido pela presença de cooperação mútua de indivíduos de uma mesma espécie, com divisão comportamental, diferenças fisiológicas e morfológicas entre os indivíduos da espécie, presença de castas e divisão de trabalho, sobreposição de gerações e cuidado parental cooperativo. O traço surgiu independentemente em diversos clados de metazoários, como os Hymenoptera, cupins e ratos-toupeira-pelados (BATRA, 1966; CRESPI; YANEGA, 1995; MICHENER, 1969; WILSON, 1971). A ordem Hymenoptera possui o maior número de origens independentes conhecidas da eussocialidade, o que torna esses insetos um clado de interesse para estudos desse fenótipo (BRANSTETTER et al., 2017; CARDINAL; DANFORTH, 2011; GRIMALDI; ENGEL, 2005; LINKSVAYER; WADE, 2005). Hymenoptera eussociais formam colônias com estrutura interna complexa, podendo conter uma ou mais rainhas, e possuem intrincada hierarquia social e, em algumas espécies, a capacidade reprodutiva é um diferencial na promoção da divisão de tarefas entre as diferentes castas e subcastas de operárias (HÖLLDOBLER; WILSON, 2009). Em Hymenoptera, a eussocialidade ocorre apenas em formigas, algumas espécies de abelhas e no clado das vespas eussociais (HUNT, 2007).

Todas as formigas viventes são eussociais, possuindo colônias que podem chegar a milhões de indivíduos, complexa divisão de trabalho e castas, com a

existência de subcastas, e diversas estratégias de nutrição e forrageamento (HÖLLDOBLER; WILSON, 1990). Algumas espécies possuem comportamentos migratórios peculiares e traços indicativos de uma eussocialidade considerada mais “simples” e similar à de algumas vespas eussociais, como a ausência de rainhas e a existência de operárias com capacidade reprodutiva (BAUDIER, 2019; RAVARY; JAISSON, 2002; TSUJI; YAMAUCHI, 1995).

As espécies de abelhas eussociais são pertencentes às famílias Apidae e Halictidae (GRIMALDI; ENGEL, 2005; MICHENER, 2007). As abelhas eussociais possuem muitos traços em comum com as formigas, como a divisão complexa de castas e trabalho, além de geralmente apresentarem colônias com centenas de milhares de indivíduos (MICHENER, 2007). Em contraste com algumas formigas e vespas eussociais, as abelhas eussociais apresentam um controle da diferenciação de castas baseado na nutrição diferencial de larvas, com a presença do uso da geleia real em *A. mellifera* (MICHENER, 2007; SNODGRASS, 1925). Algumas abelhas, como as da tribo Bombini, também apresentam fenótipos que compõem o que alguns consideram uma eussocialidade “simples”, como uma divisão fluida de castas não baseada em diferenças morfológicas notáveis. Elas geralmente também possuem colônias menores, iniciadas por apenas um indivíduo: a rainha (FOSTER et al., 2004; JANDT; DORNHAUS, 2009). Algumas espécies da família Halictidae e da subfamília Xylocopinae apresentam níveis intermediários de socialidade, com ninhos iniciados por um indivíduo e o controle das fêmeas por uma mãe dominante. O ninho pode ser reutilizado por futuras gerações e a primeira fêmea a nascer controla as ações das irmãs. Em espécies com socialidade intermediária, o nível de cooperação é baixo, com apenas uma divisão rudimentar de trabalho (GRIMALDI; ENGEL, 2005; MICHENER, 2007). Algumas espécies da tribo Euglossini também apresentam um comportamento social intermediário; no entanto, nesse clado há espécies que nidificam em agregações comunais, bem como espécies solitárias de vida livre ou parasitas (GRIMALDI; ENGEL, 2005; MICHENER, 2007; REHAN; RICHARDS, 2010). Por fim, existem ainda algumas espécies de abelhas, particularmente da família Halictidae (como por exemplo a abelha *Megalopta genalis*), que apresentam eussocialidade facultativa, alternando entre um estilo de vida solitário ou eussocial de acordo com pressões ambientais como a disponibilidade de alimento (SHELL; REHAN, 2018; SMITH et al., 2010).

Vespas eussociais são himenópteros membros das subfamílias Stenogastrinae, Vespinae e Polistinae. Esses insetos cosmopolitas e construtores de ninhos suspensos são muitas vezes caracterizados como eussociais “simples”, por possuírem uma divisão de castas fluida, muitas vezes com distinção apenas por características comportamentais (e não morfológicas), presença de múltiplas rainhas por ninho e retenção da capacidade reprodutiva em fêmeas de diversas castas (HUNT, 2007). Muitas vespas eussociais possuem conjuntos de comportamentos e sinais químicos e visuais que determinam uma hierarquia de dominância social (JANDT; TIBBETTS; TOTH, 2014). Além disso, vespas eussociais frequentemente se realocam ou formam novos ninhos utilizando duas estratégias diferentes: por meio de uma fêmea solitária ou de um enxame que abandona o ninho original (HUNT, 2007; LIEBERT; NONACS; WAYNE, 2005; NOLL et al., 2021).

A presença da eussocialidade em diversos clados dentro de Hymenoptera e os diferentes graus desse fenótipo existente entre as espécies destes insetos suscitam diversas questões sobre como se deu a evolução do comportamento eussocial nesse clado. Este fenótipo foi, inclusive, considerado por Charles Darwin como um importante desafio a sua, à época, recém-criada teoria da seleção natural (HERBERS, 2009).

1.2.1 Evolução da eussocialidade

Apesar de estar restrito à infra ordem Aculeata, a eussocialidade surgiu e foi perdida independentemente múltiplas vezes em Hymenoptera (ANDERSON, 1984; GRIMALDI; ENGEL, 2005). Apenas em abelhas, estima-se entre quatro a dez origens independentes desse fenótipo (BRANSTETTER et al., 2017; CARDINAL; DANFORTH, 2011; GRIMALDI; ENGEL, 2005; LINKSVAYER; WADE, 2005). Como todas as formigas são eussociais, o cenário mais parcimonioso sugere apenas um surgimento da eussocialidade no ancestral comum ao clado (GRIMALDI; ENGEL, 2005; HÖLLDOBLER; WILSON, 1990).

Novas evidências sugerem pelo menos duas origens independentes da eussocialidade nas vespas eussociais: um na subfamília Stenogastrinae e outro no clado formado pelas subfamílias Polistinae + Vespinae (HUNT, 2007; PETERS et al., 2017; PIEKARSKI et al., 2018).

As múltiplas origens independentes da eussocialidade em Hymenoptera sugerem uma possível existência de características genômicas comuns que predispõem a evolução do traço na ordem (DOGANTZIS et al., 2018; KAPHEIM et al., 2015; TOTH; ROBINSON, 2007; WARNER et al., 2019). Adicionalmente, a proximidade filogenética entre formigas e abelhas sugere a possível existência de uma predisposição à evolução do traço no ancestral comum dos dois clados, enquanto a existência de dois surgimentos independentes nas vespas eussociais corrobora a hipótese da presença de traços que predispõem a evolução da eussocialidade no ancestral comum de Aculeata ou a existência de características que favorecem fortemente a convergência do traço (BRANSTETTER et al., 2017).

Alguns fatores ecológicos são usualmente citados como características que podem ter favorecido uma evolução convergente da eussocialidade em Hymenoptera como, por exemplo, o altruísmo e mecanismos de defesa contra predadores e parasitas (BOURKE, 2011; NOWAK; TARNITA; WILSON, 2010; SHELL; REHAN, 2018). A construção de um ninho é um fator comum entre os Hymenoptera eussociais, e até mesmo em espécies solitárias. Este traço provavelmente estava presente na ordem antes do surgimento da eussocialidade, e estaria associado à sua evolução (GRIMALDI; ENGEL, 2005; HUNT, 2007).

Ainda assim, uma das características mais notáveis dos Hymenoptera eussociais é a determinação haplodiploide do sexo e o impacto desse sistema no grau de parentesco entre os indivíduos de uma colônia (BOURKE, 2011; HAMILTON, 1964). A haplodiploidia ocorre em todos os clados de Hymenoptera e compreende um sistema de determinação sexual no qual um dos sexos é haploide (usualmente os machos, em Hymenoptera) e o outro é diploide (KING; STANSFIELD; MULLIGAN, 2006). Os machos haploides nascem por partenogênese, *i.e.,* a partir de ovos não fecundados produzidos por fêmeas reprodutoras. Como efeito desse fenômeno, as fêmeas são mais geneticamente aparentadas com suas irmãs (75% de parentesco) do que com seus irmãos (25% de parentesco) ou suas mães ou filhos (50% de parentesco) (BOURKE, 2011).

Inicialmente, a ideia da existência de indivíduos estéreis especializados para cuidar da prole de outrem foi tida como uma potencial ressalva à teoria da seleção natural, um fato inclusive mencionado por Darwin em sua obra (DARWIN, 1859; HERBERS, 2009). Como possível explicação, foi desenvolvida uma tentadora e, inicialmente, satisfatória hipótese para a evolução do fenômeno em insetos sociais,

denominada “hipótese haplodiploide”, baseada na regra de Hamilton. Essa regra sugere que em função da assimetria de parentesco entre fêmeas de insetos haplodiploides, alelos que fossem responsáveis por fazer fêmeas haplodiploides cuidarem de suas irmãs seriam mais facilmente herdados por ditas irmãs do que por seus filhos ou filhas (HAMILTON, 1964; WEST-EBERHARD, 1975). Apesar de uma ampla adoção inicial, importantes críticas e ressalvas foram levantadas com relação à hipótese haplodiploide, como por exemplo a existência da haplodiploidia em praticamente todos os clados de Hymenoptera, contrastando com a ocorrência muito mais restrita da eussocialidade; a existência de espécies com múltiplas rainhas por colônia; rainhas que acasalam com múltiplos machos; e a necessidade, pouco evidenciada na natureza, de um grande desbalanço no investimento energético das fêmeas em suas irmãs em detrimento dos machos ou até mesmo da rainha (LINKSVAYER; WADE, 2005). Atualmente, a hipótese haplodiploide é considerada uma explicação pouco satisfatória para a evolução de castas estéreis, uma das características que predispõem a eussocialidade. Ainda há outros fatores presentes em Hymenoptera que podem ser compreendidos como características que predispõem a evolução da eussocialidade no grupo, como a presença do ferrão, a construção de ninhos e o cuidado parental (LINKSVAYER; WADE, 2005).

1.2.2 Componentes genéticos da eussocialidade em Hymenoptera

Embora haja evidências de que características genéticas que predisponham a evolução da eussocialidade em Hymenoptera possivelmente são ancestrais ao grupo, diversas evidências também sugerem que o fenótipo em si tem origem multifatorial e que cada origem da eussocialidade possivelmente se deu por vias distintas (WARNER et al., 2019).

Como exemplos de fatores gerais, os genes de receptores olfatórios, em especial os da família *9-exon* e *7-transmembrane*, parecem estar independentemente expandidos em diversas linhagens eussociais de Hymenoptera, o que pode estar associado ao importante papel da recepção olfatória no reconhecimento de indivíduos, na comunicação e na localização em espécies eussociais (KARPE et al., 2017; MCKENZIE; KRONAUER, 2018; THE HONEYBEE GENOME SEQUENCING CONSORTIUM, 2006). Como exemplos de fatores linhagem-específicos, os genes *major royal jelly protein*, responsáveis pela produção

de uma secreção mandibular usada na produção de geleia real em *A. mellifera*, estão expandidos no genoma da abelha, enquanto genes relacionados, como *major royal jelly protein-like* e os genes *yellow*, estavam presentes no ancestral comum de Apocrita e possivelmente facilitaram a expansão e diversificação posterior desta família proteica nessa linhagem (OEYEN et al., 2020; THE HONEYBEE GENOME SEQUENCING CONSORTIUM, 2006).

Resultados recentes demonstram que a evolução de genes e famílias gênicas em Hymenoptera, particularmente em Apocrita, ocorre de forma reducionista, com mais eventos de perdas do que ganhos de novos genes, contrabalanceadas por um expressivo ganho de novos arranjos de domínios proteicos (OEYEN et al., 2020). De fato, evidências demonstram uma considerável redução do número de cópias de genes de cutícula e genes envolvidos com a resposta imune em espécies eussociais de Hymenoptera (OEYEN et al., 2020; THE HONEYBEE GENOME SEQUENCING CONSORTIUM, 2006). O desenvolvimento de estratégias de imunidade de grupo, ou “imunidade social”, além de uma maior especialização do repertório gênico na reprodução, podem ser fatores explicadores para a redução dos genes de imunidade individual em Hymenoptera eussociais (LIU et al., 2019).

Em contrapartida à tendência de redução do repertório gênico, Shell et al. (2021) demonstraram que linhagens de abelhas com eussocialidade complexa possuem um maior número de grupos de genes ortólogos em expansão e sob efeito de seleção positiva do que linhagens com eussocialidade simples ou solitárias, sugerindo que a alta diversificação de genes é um fator importante para a evolução da eussocialidade.

Adicionalmente, Dogantzis et al. (2018) demonstraram que genes sob seleção positiva em espécies de Hymenoptera com diferentes níveis de eussocialidade pertencem a classes gênicas funcionais distintas. No trabalho, os autores demonstraram que genes sob evolução positiva nas vespas eussociais “simples” do gênero *Polistes* estão enriquecidos para funções de regulação transcricional, enquanto nas abelhas eussociais “simples” do gênero *Bombus* estão enriquecidos para funções metabólicas e, em contrapartida, na espécie com eussocialidade complexa *A. mellifera* observa-se um enriquecimento para funções associadas com comportamento e percepção sensorial.

Uma via comum de particular interesse é o aumento da capacidade da regulação da expressão gênica, seja pela evolução acelerada de genes regulatórios, a expansão do repertório de fatores de transcrição ou por um aumentado controle epigenético da expressão (KAPHEIM et al., 2015; KUCHARSKI et al., 2008; OLDROYD; YAGOUND, 2021; SIMOLA et al., 2016). Isso se deve ao fato de que uma maior capacidade de controle da expressão gênica e, portanto, do desenvolvimento, pode representar um mecanismo “interruptor” que permitiria a evolução da diferenciação de castas, particularmente o surgimento de castas estéreis, uma das características mais marcantes dos Hymenoptera eussociais (LAWSON; HELMREICH; REHAN, 2017; LI et al., 2010; NOWAK; TARNITA; WILSON, 2010; PATEL et al., 2007; RAMSAY; LASKO; ABOUHEIF, 2021).

Como forma de corroborar a validade destas evidências e encontrar potenciais associações entre a presença ou expansão de componentes genômicos e a presença de eussocialidade em Hymenoptera, se faz necessária a utilização de ferramentas adequadas de comparação entre diferentes genomas.

1.3 Genômica comparativa como ferramenta de estudo da eussocialidade em Hymenoptera

A genômica comparativa tradicionalmente compreende um conjunto de métodos que visa identificar elementos homólogos compartilhados entre genomas e extrair informações biologicamente relevantes em função dos padrões de conservação e variação entre os genomas (COUTINHO; FRANCO; LOBO, 2015). Tais análises podem utilizar dados quantitativos ou qualitativos, e permitem que, dentre outras, sejam feitas inferências acerca da evolução e da função dos componentes (*e.g.* domínios proteicos, famílias gênicas) codificados por estes genomas (HARDISON, 2003).

Kocher e Paxton (2014) discutem o uso de métodos comparativos em biologia para o estudo da evolução da eussocialidade em Hymenoptera, com foco especial em abelhas, táxon com o maior número de surgimentos independentes do traço (eussocialidade). Os autores chamam a atenção para a utilização da genômica comparativa nas pesquisas sobre a evolução da eussocialidade ao afirmarem que, nesse caso, é importante assumir que há um conjunto de traços genéticos comuns relacionados à organização social. Entretanto, tais traços podem tanto ter sido

herdados diretamente em função da ancestralidade comum, como adquiridos independentemente por convergência evolutiva, e a não correção deste viés pode introduzir erros do Tipo II (aceitar um resultado negativo como positivo) nas análises.

Assim, para a busca por regiões homólogas associadas à evolução da eussocialidade em Hymenoptera, faz-se necessário alguns conjuntos de dados: 1) um conjunto de regiões homólogas biologicamente relevantes para ser utilizada como elemento comum de comparação; 2) informação fenotípica sobre o fenômeno evolutivo de interesse (e.g. eussocialidade); 3) informações filogenéticas a serem utilizadas para a construção de modelos estatísticos *phylogeny-aware*, ou seja, que considerem a ordem e tempo evolutivos para realizar estudos de associação. Adicionalmente, é necessário também desenvolver metodologias estatísticas adequadas que integrem esses conjuntos de dados para produzir modelos estatísticos auditáveis, generalizáveis e reprodutíveis. As próximas seções aprofundam os conceitos necessários para o desenvolvimento destes modelos.

1.3.1 Domínios e famílias proteicas

Para a análise comparativa de genomas em nível funcional, ou seja, quais funções estão super representadas em um genoma em relação a outro, é necessário possuir uma unidade funcional comparável. Domínios proteicos são as menores unidades funcionais de uma proteína e são considerados unidades evolutivas independentes do restante da proteína que compõem. Os domínios proteicos são capazes de realizar uma função específica isoladamente e se enovelam independentemente (BAGOWSKI; BRUINS; TE VELTHUIS, 2010). Além disso, domínios proteicos comumente sofrem reordenamento/embaralhamento (*domain shuffling*) em eucariotos ao longo do tempo evolutivo, o que compreende um importante mecanismo evolutivo neste grupo (KOLKMAN; STEMMER, 2001; LONG et al., 2003; PATTHY, 1999).

Genes, proteínas e domínios proteicos possuem relações evolutivas que se traduzem em similaridades de sequência, estrutura e função e, por isso, podem ser classificados em grandes grupos, ou famílias, que englobam genes homólogos conservados (TATUSOV; KOONIN; LIPMAN, 1997). Tais domínios e famílias são identificados nos genomas através dos processos de anotação gênica e genômica (MISTRY et al., 2021; YANDELL; ENCE, 2012), e fornecem informações sobre parte

das funções de uma proteína e de um genoma como um todo. Assim, a anotação dos genes codificadores de eucariotos em diferentes unidades evolutivas e funcionais permite que estas sejam utilizadas como unidade de comparação comum em estudos de genômica comparativa.

Por sua natureza funcional e evolutivamente independente, os domínios proteicos são uma unidade de estudo aplicável em trabalhos de genômica comparativa como o presente, pois podem ser avaliados quanto a sua abundância absoluta e relativa entre genomas, bem como a sua associação com um ou mais traços ou fenótipos de interesse.

1.4 Métodos comparativos em biologia

As análises de genômica comparativa se prestam a compreender o método, o ritmo e as consequências da evolução de traços identificáveis codificados no genoma de diferentes linhagens de organismos (HARDISON, 2003; MILLER et al., 2004). Usualmente, essas inferências se dão a partir de análises estatísticas que envolvem teste de hipóteses ou de associação, como modelos de regressão linear e análises de variância (ANOVA) e covariância (ANCOVA). Entretanto, a vasta maioria dos testes estatísticos paramétricos e não paramétricos possuem uma premissa em comum: a independência dos dados amostrados.

No entanto, espécies e dados associados a elas (*e.g.* medições, presença ou ausência de traços) não podem ser tratados como pontos independentes de uma distribuição pois compartilham uma ancestralidade comum (CORNWELL; NAKAGAWA, 2017; FELSENSTEIN, 1985). Mais ainda, o nível de proximidade entre diferentes pares de espécies não é uniforme e, portanto, é de se esperar que, dados os princípios básicos de herança de traços evolutivos, espécies próximas entre si possuam valores similares para um mesmo traço por ancestralidade comum e não por adaptação convergente (CORNWELL; NAKAGAWA, 2017). No contexto de análises estatísticas que assumem a independência dos dados, isto se torna um grave problema pois, ao ferir esta importante premissa, reduz-se a credibilidade dos resultados encontrados (HONGO et al., 2021).

Visando solucionar este problema, diversos métodos estatísticos, coletivamente denominados métodos comparativos em biologia, foram desenvolvidos nos últimos quarenta anos, visando incorporar informação filogenética

aos testes estatísticos de hipóteses (CORNWELL; NAKAGAWA, 2017; FELSENSTEIN, 1985). Embora os métodos comparativos sejam amplamente utilizados em diversas áreas da biologia para avaliar associações entre os mais diversos fenótipos, há uma surpreendente ausência de métodos de genômica comparativa que incorporem a informação filogenética para buscar associações entre a abundância de elementos homólogos compartilhados entre genomas e algum fenótipo de interesse (NAGY et al., 2020). Assim, para a detecção de eventuais regiões homólogas associadas à evolução da eussocialidade em Hymenoptera, faz-se necessário o desenvolvimento de métodos capazes de integrar informações fenotípicas, genômicas e evolutivas.

Embora os fenótipos/genótipos de interesse observados nos nós terminais de uma árvore de espécies não compreendam valores independentes, as taxas de variação entre espécies (e quaisquer nós internos de uma filogenia) o são. Uma vez que se assume que os diferentes eventos de divergência em uma árvore completamente dicotômica são independentes, podemos afirmar que todos os pontos de ramificação (ancestral comum mais recente entre duas linhagens quaisquer) representados nas árvores filogenéticas são réplicas estatísticas independentes.

Assim, os métodos filogenéticos realizam transformações de dados qualitativos e quantitativos dos nós terminais e internos de uma árvore dicotômica e ultramétrica de espécies - onde o tamanho dos ramos é proporcional ao tempo de divergência - para computar as taxas de variação dos dados de interesse para cada um dos nós internos da árvore. Felsenstein (1985) chamou estas transformações de dados, particularmente aplicadas a análises de regressão linear, de “contrastes filogeneticamente independentes” (PIC). Cada um destes pontos é independente e pode, então, ser utilizado como parâmetro de entrada de modelos estatísticos tradicionais.

Os métodos filogenéticos comparativos permitem a aplicação de uma gama de análises e modelos estatísticos no contexto de dados que podem ser representados por relações de parentesco, como relações entre espécies, e permitem responder perguntas sobre o modo e o ritmo da evolução de traços e linhagens e até mesmo a predição *ab initio* de traços precursores hipotéticos que permitiram a evolução posterior de um fenótipo em uma ou mais linhagens relacionadas (CORNWELL; NAKAGAWA, 2017)

Embora amplamente utilizados em estudos ecológicos para avaliar diversas hipóteses evolutivas, há uma surpreendente ausência de software que integrem informações genômicas, fenotípicas e filogenéticas para buscar por associações entre genótipos e fenótipos entre espécies (revisado em NAGY et al., 2020). Adicionalmente, os métodos já desenvolvidos se limitam a avaliar possíveis associações entre a presença/ausência de genes e fenótipos categóricos, não sendo capazes de analisar cenários mais complexos de variação de número de cópias de regiões homólogas (HILLER et al., 2012; NAGY et al., 2020).

2 HIPÓTESE

A eussocialidade possui múltiplas origens independentes em Hymenoptera. Adicionalmente, há diversas evidências de que expansões e contrações de conjuntos específicos de genes aparentam estar associadas à evolução deste fenótipo em algumas linhagens deste grupo. Assim, nossa hipótese é de que há elementos genômicos homólogos cuja abundância está significativamente associada à presença da eussocialidade em Hymenoptera, e que estas associações podem ser detectadas através da utilização de métodos comparativos que façam uso da informação filogenética.

3 OBJETIVO

O objetivo geral e objetivos específicos do presente trabalho estão descritos a seguir.

3.1 Objetivo geral

Desenvolver um software que integre **1)** dados fenotípicos categóricos; **2)** abundância de regiões homólogas e **3)** informações filogenéticas para buscar por regiões homólogas cujo número de cópias seja significativamente associado à eussocialidade em Hymenoptera.

3.2 Objetivos específicos

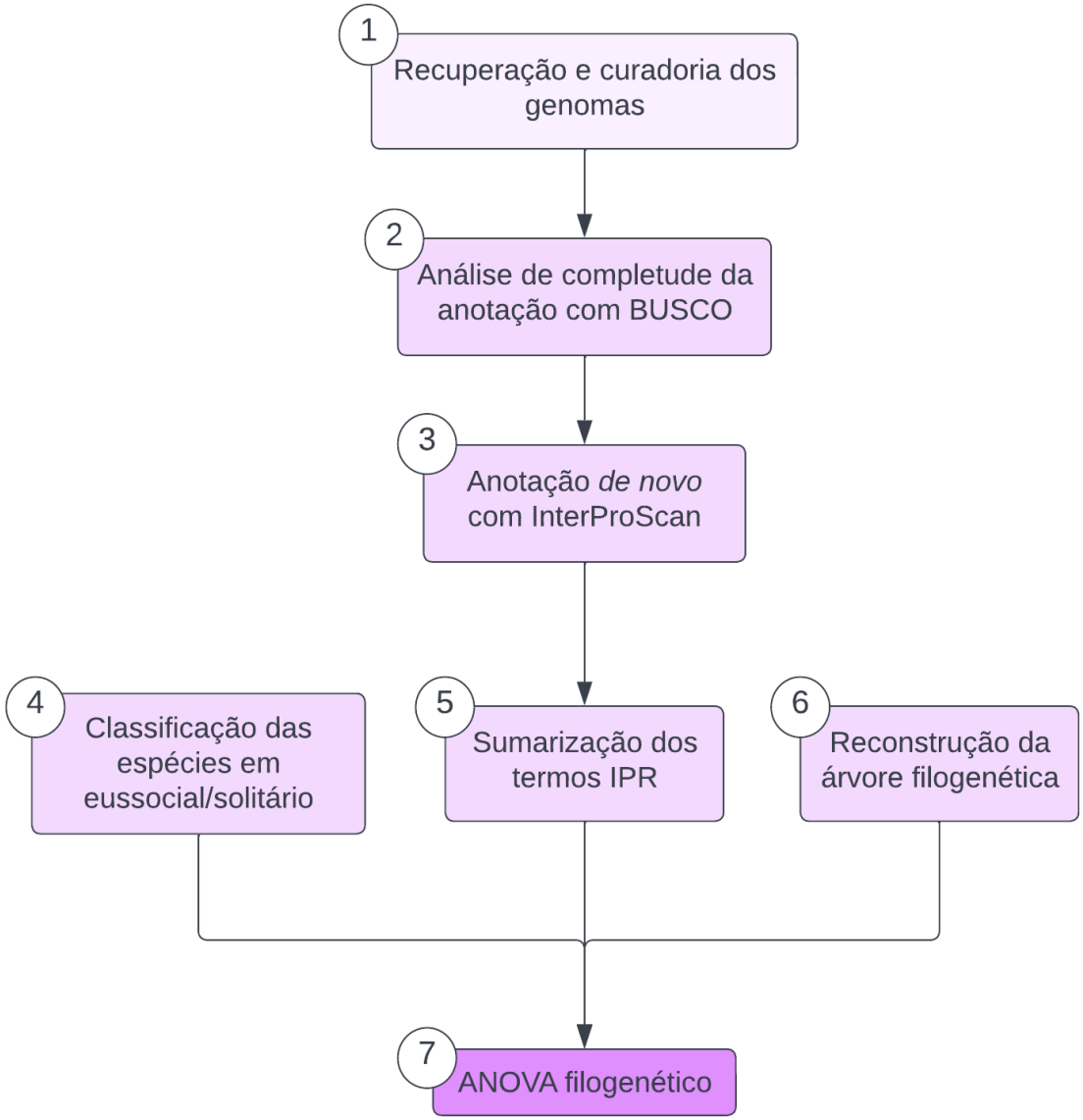
Os objetivos específicos deste trabalho são:

1. Realizar a obtenção, curadoria e tradução (convertendo em proteomas) de todos os genomas de Hymenoptera anotados disponíveis no banco de dados público *Genome* do *National Center for Biotechnology Information* (NCBI);
2. Usar os dados presentes na literatura para classificar as espécies de Hymenoptera analisadas com relação à presença ou ausência de eussocialidade;
3. Avaliar, através da análise de presença de genes ortólogos universais em Hymenoptera, a qualidade de montagem e completude de anotação dos proteomas obtidos;
4. Construção de uma árvore filogenética ultramétrica e dicotômica para as espécies onde o tamanho dos ramos seja proporcional ao tempo evolutivo;
5. Gerar contagens não redundantes da presença de domínios proteicos e famílias gênicas nos proteomas obtidos;
6. Desenvolver um programa na linguagem de programação R para integrar as informações fenotípicas, filogenéticas e genômicas e produzir modelos *phylogeny-aware* para detectar regiões homólogas cuja abundância absoluta (número de cópias) esteja significativamente associada a um fenótipo categórico com duas classes;
7. Integrar as informações obtidas para buscar por possíveis regiões homólogas associadas à eussocialidade em Hymenoptera.

4 MATERIAL E MÉTODOS

O fluxograma a seguir apresenta um resumo dos passos metodológicos seguidos para a realização do trabalho.

Figura 1 - Fluxograma da metodologia



Fluxograma resumindo os passos metodológicos seguidos para a realização do trabalho. Cada passo está referenciado no texto de acordo com a numeração na figura. Fonte: elaborada pelo autor (2022).

4.1 Infraestrutura Computacional

Todas as etapas de processamento foram realizadas em um servidor CentOS Linux com 64 núcleos de processamento e 128GB de memória RAM. O servidor contém suporte para as linguagens de programação Shell Script, R, Perl e Python e está instalado no Laboratório de Algoritmos em Biologia (LAB), bloco L3, sala 174, Departamento de Genética, Ecologia e Evolução do Instituto de Ciências Biológicas (ICB) da Universidade Federal de Minas Gerais (UFMG).

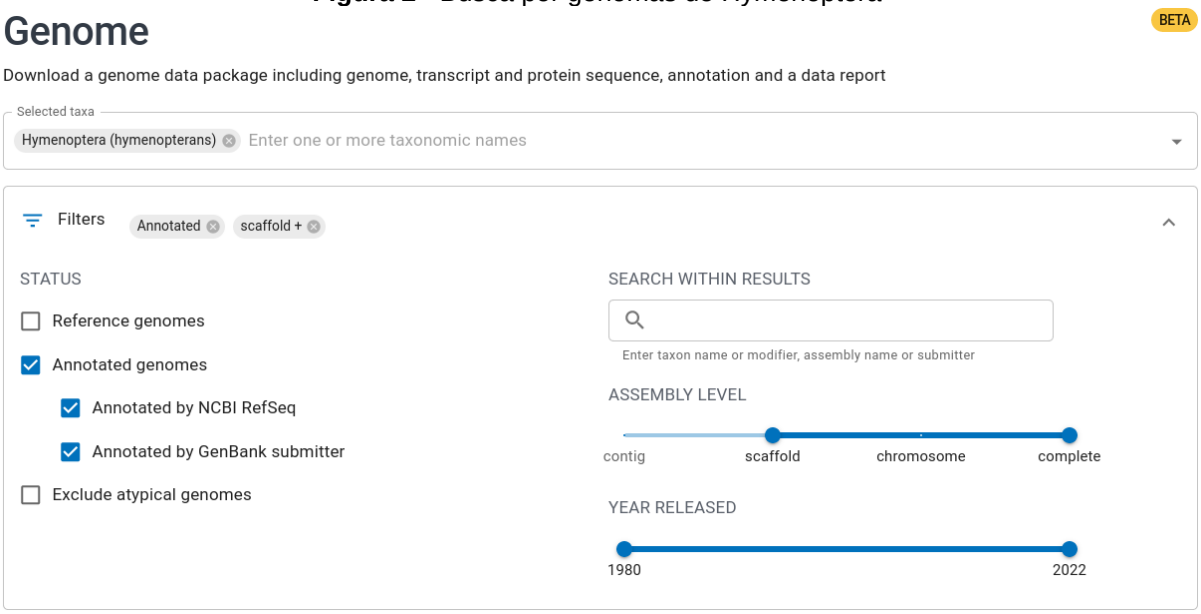
4.2 Obtenção e Controle de Qualidade dos Genomas

Os genomas de Hymenoptera foram obtidos (Figura 1, passo 1) através da plataforma *Genome*[1] pertencente ao *National Center for Biotechnology Information* (NCBI). Foram baixadas as sequências codificantes de todos os genomas encontrados disponíveis até 04 de agosto de 2022 (Figura 2). O termo "*Hymenoptera (hymenopterans)*" foi utilizado na busca, além dos filtros de anotação disponível ("*Annotated*") e montagem a nível de *scaffold* ou superior ("*scaffold*+"). Para genomas com mais de uma anotação disponível, optamos pela versão indicada como sequência de referência (*Reference Sequence*, *RefSeq*).

Em seguida, utilizando *scripts* escritos na linguagem de programação Python (Apêndice D), foram removidas de todos os genomas quaisquer sequências contendo nucleotídeos não canônicos (qualquer nucleotídeo exceto A, C, G ou T). Por se tratarem de organismos eucariotos, cada lócus gênico pode codificar um grande número de diferentes isoformas (ROMERO et al., 2006). O número de isoformas anotadas por lócus gênico pode ser influenciado pelo conhecimento atual sobre cada espécie, fazendo com que espécies mais estudadas tenham uma tendência a possuir mais isoformas anotadas (LOBB et al., 2020). Para evitar que esse possível viés influencie significativamente os resultados, todos os genomas filtrados foram sumarizados através da obtenção da maior isoforma conhecida por lócus (VOGEL; CHOTHIA, 2006) utilizando *scripts in-house* (Apêndice E) com auxílio dos pacotes da linguagem Python *pandas* (THE PANDAS DEVELOPMENT TEAM, 2022) e *Biopython* (COCK et al., 2009). As sequências restantes foram então traduzidas e filtradas através da remoção de sequências possivelmente de baixa qualidade, definidas pela: **1)** ausência de códons de início e/ou término; **2)** presença de códons de término internos. Ao fim destas etapas de controle de qualidade, os arquivos resultantes foram denominados proteomas não redundantes (Figura 1, passo 1).

[1] https://www.ncbi.nlm.nih.gov/data-hub/genome

Figura 2 - Busca por genomas de Hymenoptera



Captura de tela que demonstra os filtros utilizados para a recuperação dos genomas de Hymenoptera anotados através da ferramenta de busca da plataforma *Genome.* Fonte: elaborada pelo autor (2022).

Como última etapa de controle de qualidade, realizamos uma avaliação da completude da anotação dos proteomas não redundantes através da ferramenta BUSCO versão 5 (MANNI et al., 2021) (Figura 1, passo 2). Resumidamente, este programa produz um estimador indireto quantitativo da qualidade de um determinado conjunto de sequências (*e.g.* proteomas não redundantes) em função do número de ortólogos cópia simples esperados para o grupo taxonômico ao qual o organismo cujo proteoma está sob análise pertence. Para tal, utiliza-se um conjunto de genomas de alta qualidade para produzir listas de ortólogos quase universais (presentes em mais de 90% das espécies), denominados BUSCOs. Ao fim da análise, o programa BUSCO produz cinco métricas: a porcentagem de completude, considerando ortólogos de cópia única ou duplicados; a porcentagem de completude apenas de ortólogos de cópia única; a porcentagem de ortólogos completos duplicados; a porcentagem de ortólogos aparentemente presentes, porém fragmentados; e a porcentagem de ortólogos não encontrados entre as sequências de entrada. Para avaliar a completude da anotação dos genomas de Hymenoptera analisados neste estudo, utilizamos o banco de dados de ortólogos de Hymenoptera (*hymenoptera_odb10*). Os proteomas com completude acima de 90% de ortólogos de cópia única foram considerados de alta qualidade, e utilizados nas análises

posteriores. O tamanho total de cada proteoma foi calculado em função do número total de proteínas por proteoma não redundante (Apêndice I).

4.3 Classificação do nível de socialidade das espécies

Todas as espécies de Hymenoptera com proteomas de alta qualidade foram classificadas em relação ao comportamento social utilizando uma escala simplificada de três níveis: solitário, intermediário ou eussocial (Figura 1, passo 4). Para a classificação foram utilizadas as descrições presentes na literatura referentes ao modo de vida, tamanho das colônias, comportamento, presença de castas e divisão de trabalho, cuidado parental cooperativo, ocorrência de agregações e cooperação simples, presença de parasitismo e termos qualificadores como “eussocial” ou “solitário”.

4.4 Anotação e contagem de domínios

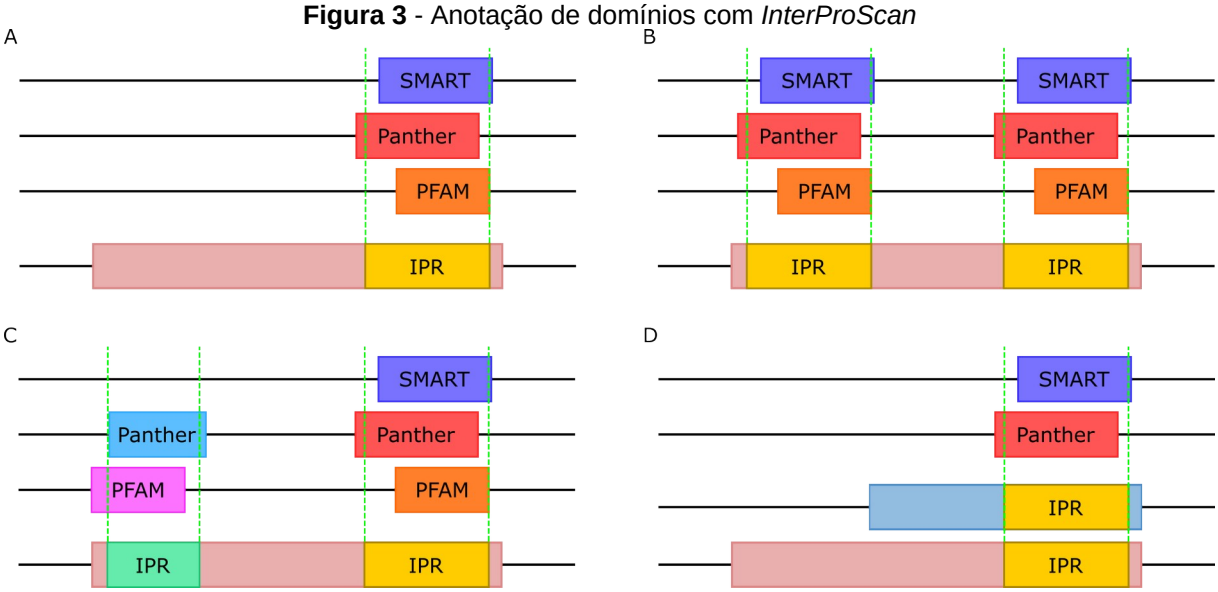
Utilizamos o software *InterProScan* (JONES et al., 2014) para gerar uma anotação *de novo* dos proteomas não redundantes de alta qualidade (Figura 1, passo 3). O *InterProScan* utiliza 13 bancos de dados construídos para diferentes finalidades e que armazenam informações de motivos, domínios, bem como arquiteturas mais complexas, predizendo assim diferentes classes de regiões homólogas. Não surpreendentemente, há um grau considerável de sobreposição entre os diferentes bancos de dados, fazendo com que uma mesma região homóloga receba identificadores distintos (THE INTERPRO CONSORTIUM et al., 2002). Para solucionar possíveis inconsistências causadas por esse fato, esse banco de dados também fornece um identificador interno (composto pelos caracteres “IPR” mais um identificador numérico único) que integra os identificadores dos diferentes bancos de dados membros do *InterProscan* quando estes representam a mesma região homóloga (Figura 3). Desta forma, o banco de dados do *InterProScan* é capaz de promover soluções práticas para cenários comuns advindos da anotação proteica através de bancos de dados e metodologias distintas: **a)** sumarização da anotação de um domínio presente em uma proteína com múltiplas anotações advindas de bancos de dados distintos que compreendem fragmentos da sequência com tamanhos diferentes (Figura 3, A); **b)**

homogeneização da anotação de um domínio presente em duas ou mais cópias em uma mesma proteína, garantindo que cópias de um mesmo domínio recebam o mesmo identificador IPR (Figura 3, B); **c)** distinção de domínios diferentes anotados em uma mesma proteína, que recebem identificadores IPR distintos (Figura 3, C); **d)** homogeneização da anotação de um mesmo domínio presente em proteínas distintas, que recebe um mesmo identificador IPR (Figura 3, D).

Adicionalmente, os identificadores IPR possuem uma hierarquia funcional, na qual uma mesma região homóloga pode ser representada por um identificador que compreende uma região homóloga mais ancestral e/ou menos especializada, denominada “pai” (*e.g.* globinas), e por regiões homólogas mais especializadas denominadas “filhas” (*e.g.* hemoglobinas e mioglobinas). Dessa maneira, é possível representar tanto a abundância de homólogos ancestrais como a abundância de homólogos recentes funcionalmente distintos (Figura 4).

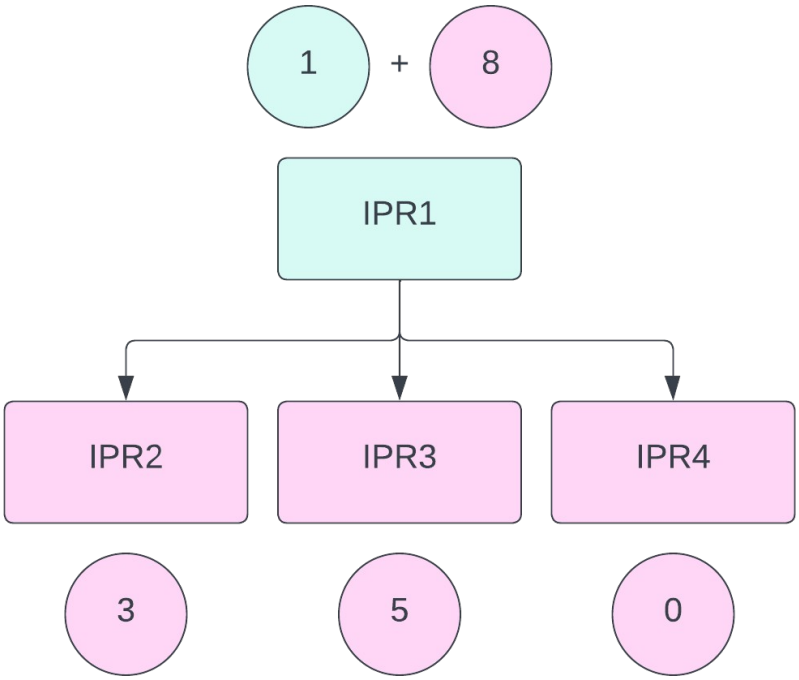
Como anotação final, consideramos cada par proteína-IPR como um elemento distinto. De posse dessa informação, computamos a abundância absoluta (número de cópias) para cada IPR em cada proteoma não-redundante. Esse valor representa o número de genes distintos que codificam a região homóloga representada pelo IPR em questão (ex.: se um proteoma possui o valor de “3” para um dado IPR, isso significa que este IPR foi observado em três regiões codificadoras distintas). A cobertura de anotação de cada genoma foi avaliada em função da razão entre o número de proteínas anotadas e o número total de proteínas por proteoma não redundante (Apêndice I).

Figura 3 - Anotação de domínios com *InterProScan*



Representação esquemática simplificada que demonstra a sumarização dos termos anotadores realizada pelo software *InterProScan*. A) Uma proteína (retângulo rosa) possui um domínio (retângulo amarelo, “IPR”) anotado por diferentes bancos de dados (retângulo laranja, “PFAM”; retângulo vermelho, “Panther”; retângulo azul, “SMART”), sumarizados em um único termo IPR. B) Uma proteína (retângulo rosa) pode possuir mais de uma cópia do mesmo domínio (retângulos amarelos, “IPR”); nesse caso, todas as cópias recebem o mesmo identificador IPR. C) Uma mesma proteína (retângulo rosa) pode possuir mais de um domínio distinto (retângulos verde e amarelo, “IPR”); nesse caso, cada um recebe um identificador IPR distinto. D) Por fim, duas ou mais proteínas distintas (retângulos rosa e azul claro) podem conter o mesmo domínio; nesse caso, os domínios iguais (retângulos amarelos, “IPR”) recebem o mesmo identificador IPR. As linhas pontilhadas verdes representam a área de sobreposição dos diferentes termos anotadores que anotam um mesmo domínio. Fonte: elaborada pelo autor (2022).

Figura 4 - Estrutura hierárquica de termos IPR



Representação esquemática da relação hierárquica entre termos IPR relacionados (retângulos coloridos) e suas respectivas contagens em um dado proteoma (círculos coloridos). Cada termo filho (rosa) tem sua contagem individualmente avaliada e adicionada à contagem do termo pai (ciano). Portanto, mesmo que o termo pai não esteja nominalmente presente, ainda assim é possível avaliar a presença do domínio/família que ele representa a partir das contagens de seus termos-filho. Dessa maneira, a contagem do termo pai “IPR1” é 9 (uma observação direta e oito indiretas a partir dos termos-filho). Fonte: elaborada pelo autor (2022).

4.5 Árvore Filogenética

Para a correção do viés filogenético dos testes estatísticos realizados, foi gerada uma árvore filogenética incluindo as espécies de Hymenoptera com genomas selecionados, e como grupo externo foram utilizadas as espécies *Drosophila melanogaster* (mosca das frutas) e *Tribolium castaneum* (besouro vermelho da farinha). Utilizamos uma metodologia adaptada da literatura (KVERKOVÁ et al., 2022) através de um script da linguagem R (Apêndice F),

assumindo como base a filogenia publicada por Peters et al. (2017). Nesta abordagem, utiliza-se uma “árvore-esqueleto” (*scaffold*) para representar relações entre grandes grupos, bem como os tempos de divergência entre estes, enquanto as espécies remanescentes são acrescentadas à árvore principal a partir de outras árvores mais específicas disponíveis na literatura que representam as relações entre grupos menores das espécies faltantes na “árvore esqueleto”. Especificamente, as espécies não representadas na árvore inicial foram incluídas utilizando a função *tree.merger* do pacote *RRphylo* (CASTIGLIONE et al., 2018, 2022), da linguagem R, a qual permite editar uma “árvore esqueleto” qualquer e incorporar espécies, juntamente com seus tempos de divergência, em pontos específicos de uma topologia (CASTIGLIONE et al., 2022). Para a inserção dessas espécies, foram consideradas as posições baseadas em filogenias publicadas que incluíssem o nível taxonômico não-ambíguo mais próximo (espécie, gênero ou níveis taxonômicos superiores). As espécies presentes na árvore original cujos genomas não se encontravam incluídos nas análises foram removidas com a função *drop.tip* do pacote *ape* (PARADIS; SCHLIEP, 2019).

O pacote *phytools* (REVELL, 2012) foi utilizado para leitura e manipulação das árvores filogenéticas. Além disso, para construção e manipulação da filogenia final, foram utilizados os pacotes *treeio* (WANG et al., 2020), *RRphylo* (CASTIGLIONE et al., 2018, 2022) e *ape* (PARADIS; SCHLIEP, 2019). As visualizações das árvores filogenéticas foram geradas com o auxílio do pacote *ggtree* (YU et al., 2017).

4.6 ANOVA filogenético

Para identificar uma relação entre a contagem de regiões homólogas e a ocorrência de eussocialidade entre as espécies de Hymenoptera analisadas, utilizamos um método comparativo *phylogeny-aware* equivalente à análise de variância (ANOVA; Figura 1, passo 7). Em sua versão tradicional, ANOVAs são conjuntos de métodos que avaliam se as diferenças entre as médias (ou medianas) de grupos de uma variável de interesse são significativas. Para tal, costuma-se estimar e representar as variâncias dos grupos através de diversos métodos, coletivamente denominados matrizes de variância/covariância.

O cálculo dos quadrados das diferenças entre os valores observados e uma estatística de tendência central (*e.g.* média de um grupo) é denominado método dos mínimos quadrados ordinários (*least squares*). Estes valores são comumente utilizados para estimar as variâncias/covariâncias utilizadas para ajustar os valores observados em diversos modelos estatísticos, como regressões lineares e ANOVAs, onde há uma variável de resposta *y* (*e.g.* contagem de um dado domínio proteico) e uma variável preditora *x* (*e.g.* status de socialidade).

Dentre diversos pressupostos, o método dos quadrados mínimos ordinários assume que os resíduos são independentes e identicamente distribuídos para diferentes observações da variável *y*. As similaridades entre espécies podem existir devido à convergência evolutiva ou à herança de caracteres por ancestralidade comum (CORNWELL; NAKAGAWA, 2017). Como os pontos do universo amostral dos nossos dados representam espécies que estão hierarquicamente relacionadas através de suas relações filogenéticas, testes paramétricos e não-paramétricos que utilizem estimadores que assumam independência dos dados não podem ser utilizados, já que os dados oriundos de organismos filogeneticamente relacionados ferem essa premissa (FELSENSTEIN, 1985; GARLAND et al., 1993). Nesse cenário, organismos filogeneticamente mais próximos tendem a apresentar resíduos mais correlacionados do que os preditos pelos modelos em função da ancestralidade comum mais recente (BININDA-EMONDS, 2014). No caso dos contrastes filogeneticamente independentes propostos por Felsenstein (1985), a independência dos pontos amostrais é obtida ao se calcular a taxa de variação de genótipos/fenótipos para todos os eventos de especiação de uma filogenia. Uma vez que estes são, teoricamente, independentes, os mesmos podem ser utilizados para produzir modelos lineares independentes.

A metodologia que utilizamos faz uso de outra estratégia para considerar a não-independência dos dados amostrais advindos de espécies filogeneticamente relacionadas. O método dos mínimos quadrados generalizados filogenéticos (*phylogenetic generalised least squares*, PGLS) é uma modificação da técnica de mínimos quadrados generalizados que utiliza informações filogenéticas para gerar uma matriz de covariância estimada através dos dados de espécies, assumindo que espécies filogeneticamente mais próximas devem possuir resíduos mais similares em função da ancestralidade comum. Ao incorporar a covariância estimada desses resíduos, os modelos produzidos a partir de PGLS efetivamente consideram a

filogenia como parâmetro e, portanto, podem ser utilizados como entrada para modelos que visam testar hipóteses entre grupos de organismos filogeneticamente relacionados (SYMONDS; BLOMBERG, 2014).

Tal correção assume que o erro amostral gerado pelo viés evolutivo, ou seja, a similaridade esperada entre dois taxa evolutivamente relacionados, evolui em movimento Browniano, i.e., os eventos de diversificação ocorreriam de maneira pseudoaleatória através do tempo (MARTINS; HANSEN, 1997). A matriz de covariância pode então ser estimada para cada par de espécies relacionadas, levando em conta sua distância filogenética, através do modelo descrito por Felsenstein (1985), e então utilizada em um modelo PGLS para estimar os parâmetros (coeficientes de regressão entre as variáveis analisadas), utilizados então como variáveis de entrada para um teste de hipóteses via ANOVA.

As análises estatísticas necessárias para realização dos testes de ANOVA filogenético foram feitas em R com o auxílio dos pacotes *geiger* (PENNELL et al., 2014)*, CALANGO* (HONGO et al., 2021) e *nlme* (PINHEIRO; BATES; R CORE TEAM, 2022), e a visualização dos resultados foi gerada com o auxílio adicional dos pacotes *cowplot* (WILKE, 2020), *dendextend* (GALILI, 2015) e *ComplexHeatmap* (GU, 2022; GU; EILS; SCHLESNER, 2016). O pacote *phytools* (REVELL, 2012) foi utilizado para leitura e manipulação das árvores nos procedimentos de correção filogenética. O código completo está disponível no Apêndice G.

Como arquivos de entrada, nosso programa utiliza três tipos de dados: **1)** os fenótipos obtidos a partir da revisão bibliográfica (solitário/eussocial); **2)** as informações de anotação dos proteomas não-redundantes de alta qualidade (pares entre proteínas e IPRs únicos); **3)** a árvore de espécies dicotômica e ultramétrica, na qual o tamanho dos ramos é proporcional ao tempo de divergência entre os nós (Figura 1, passos 4, 5 e 6).

Especificamente, utilizamos a saída do programa CALANGO (HONGO et al., 2021) para obter as sumarizações do total de contagens de cada IPR ID. Uma vez que cada par proteína-IPR ocorre uma única vez em cada proteoma, este número corresponde ao número de genes que possuem a região homóloga descrita pelo IPR.

Para cada IPR ID, nosso programa utiliza a árvore de espécies subjacente para computar as matrizes de covariância *phylogeny-aware* para os dados genômicos e fenotípicos assumindo variação temporal Browniana através da função

corBrownian do pacote *ape* (Apêndice G). De posse destes dados, utilizamos a função *gls* para computar os valores de PGLS para a tabela de covariância de fenótipos/genótipos, tendo como parâmetro adicional a matriz de covariância computada pela função *corBrownian*. O objeto *ancova* produzido representa uma matriz de covariância ajustada para o viés filogenético e, portanto, independentes do ponto de vista estatístico, podendo estas ser utilizadas como entrada para o cálculo de valores-*p* via ANOVA tradicional. Como nosso modelo avalia múltiplas hipóteses para explicar os padrões de presença e ausência de domínios encontrados, uma vez que cada IPR ID é avaliado para eventual associação, realizamos também uma correção para o teste de múltiplas hipóteses de maneira a controlar a taxa de falsos-positivos (BENJAMINI; HOCHBERG, 1995). Especificamente, os valores-*p* individuais para cada teste ANOVA são armazenados em uma lista, a qual é corrigida posteriormente pela função *p.adjust* para considerar o cenário de teste múltiplo de hipóteses (método BH).

Foram considerados como IPR significativamente associados à presença ou ausência de eussocialidade apenas aqueles IPR cujo teste estatístico compreenda um valor *p* corrigido (valor *q*) menor ou igual a 0.05 (*q* <= 0.05). O pseudocódigo representado abaixo resume e ilustra estas etapas, enquanto o código completo na linguagem R encontra-se no Apêndice G:

```
# árvore filogenética de espécies
tree

# matriz onde linhas são as espécies
# e as colunas são contagens de IPRs gene-level
df_IPR

# matriz de fenótipos
classes

# vetor para armazenar os valores-p
p.values


# itera por todos os IPR IDs para computar os valores-p
# para cada um deles, avaliando eventual associação
# com a eussocialidade


for (IPR_ID in unique(colnames(df_IPR)))  {
```

```
      # obtem a contagem para o IPR ID do laço
      table <- df_IPR[,IPR_ID]


      # cria uma coluna extra com as classes fenotípicas
      table$classes <- classes


      # lista de espécies
      spp <- rownames(table)

      # matriz de covariância considerando a árvore
      # de spp e assumindo movimento Browniano
            corBM <- corBrownian(phy=pruned.tree,form = ~spp)

      # Parâmetros estimados por mínimos quadrados ordinários
            ancova <- gls(data~classes, data = table, correlation = corBM)

      # ANOVA
            ANOVA <- anova(ancova)

      p-values[IPR_ID] <- ANOVA$p.value
}
q.values <- p.adjust(p.values, method = “BH”)
sig_IPRs <- q.values[q.values < 0.01]
```

Os procedimentos de sequenciamento, montagem e anotação genômica podem induzir erros consideráveis acerca da presença e ausência de genes, inclusive em organismos-modelo como *A. mellifera,* levando a conclusões equivocadas sobre a biologia de organismos (ELSIK et al., 2014). Assim, optamos por realizar um filtro robusto que considera tanto o número de classes dos valores de contagem dos IPR, como o número de espécies em cada uma das classes.

Para IPRs cujas contagens compreendam três ou menos classes (*e.g.* um IPR que possui uma, duas, ou nenhuma cópia), consideramos somente aqueles que possuam pelo menos três espécies em cada uma das classes com menor número de ocorrências. Assim, eventuais aumentos ou diminuições no número de cópias causadas por erros aleatórios de montagem e anotação serão efetivamente eliminados, uma vez que a chance de que estes ocorram em três ou mais espécies é certamente menor do que se considerássemos eventos em uma única espécie. Por fim, dada a natureza hierárquica da anotação por IPR, é possível que dois termos

IPR hierarquicamente relacionados representem uma informação redundante caso eles ocorram com a mesma frequência. Isso ocorre por serem termos distintos, porém anotadores dos mesmos domínios. Nesse caso, apenas o termo mais específico foi mantido em detrimento do termo menos específico, que foi removido manualmente.

4.7 Análises estatísticas e visualizações de dados adicionais

Todas as análises estatísticas e as respectivas visualizações geradas foram realizadas no ambiente estatístico e linguagem de programação R (R CORE TEAM, 2021). As análises de dados foram realizadas e as visualizações foram geradas com auxílio dos pacotes *dplyr* (WICKHAM et al., 2021)*, ggplot2* (WICKHAM, 2016), *ggpubr* (KASSAMBARA, 2020), *forcats* (WICKHAM, 2022), *RColorBrewer* (NEUWIRTH, 2022) e *Cairo* (URBANEK; HORNER, 2022).

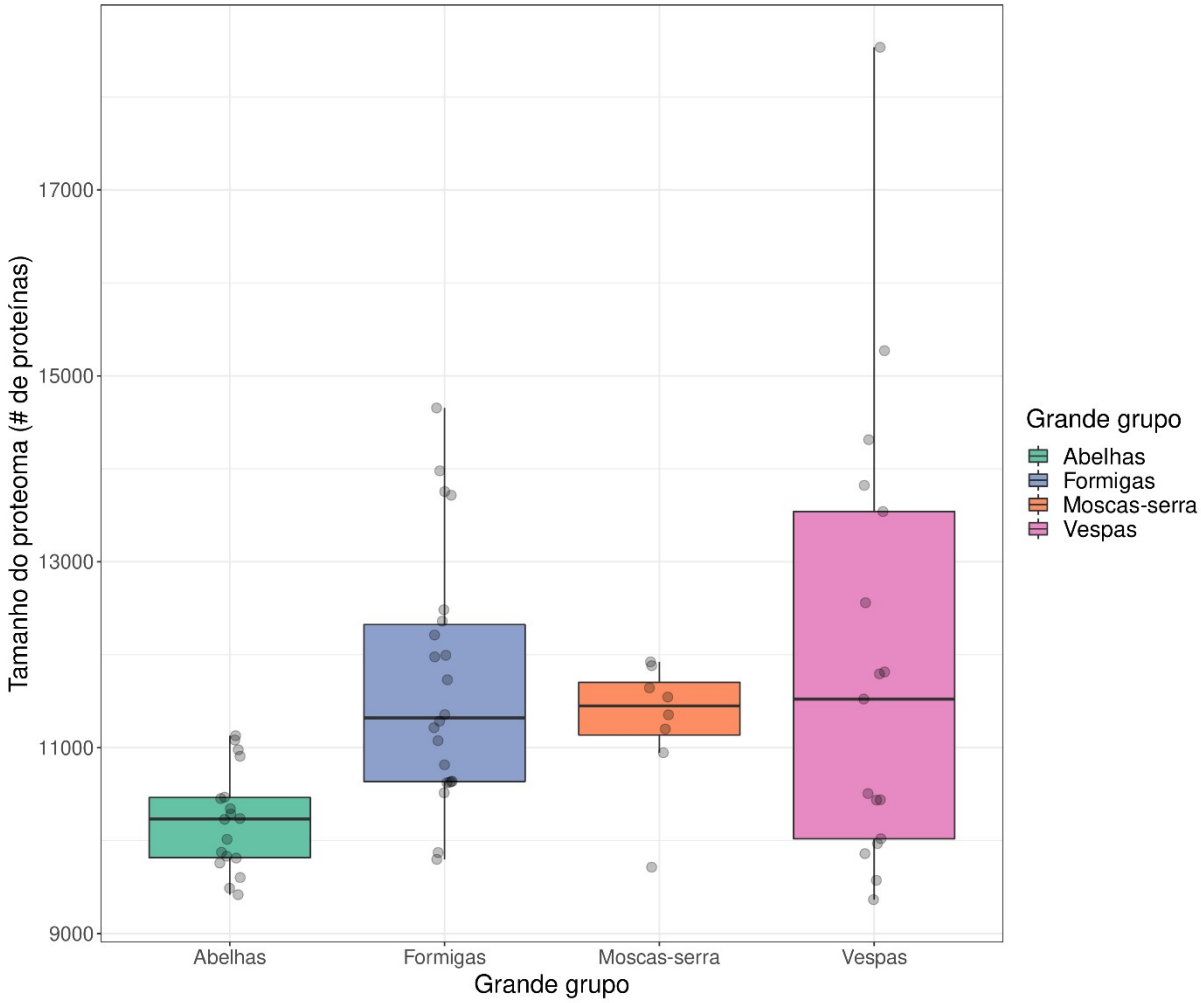
5 RESULTADOS

Inicialmente recuperamos 85 genomas de Hymenoptera, representando 29 espécies de formigas, 25 espécies de vespas, 23 espécies de abelhas e oito espécies de moscas-serra (Apêndice A).

5.1 Análises de completude e classificação de socialidade

Dentre as espécies de Hymenoptera com proteomas anotados de alta qualidade, i.e., com completude da anotação de ortólogos de cópia simples avaliada por BUSCO >= 90%, um total de 22 foram classificadas como solitárias, apenas três como intermediárias (todas abelhas) e todas as formigas e demais espécies restantes como eussociais (Apêndice B). O tamanho médio dos proteomas anotados não redundantes de alta qualidade é de 11.304,62 proteínas, sendo os proteomas das espécies *Polistes canadensis* (9.366 proteínas) e *Ampulex compressa* (18.533 proteínas) o menor e o maior proteoma, respectivamente (Figura 5).

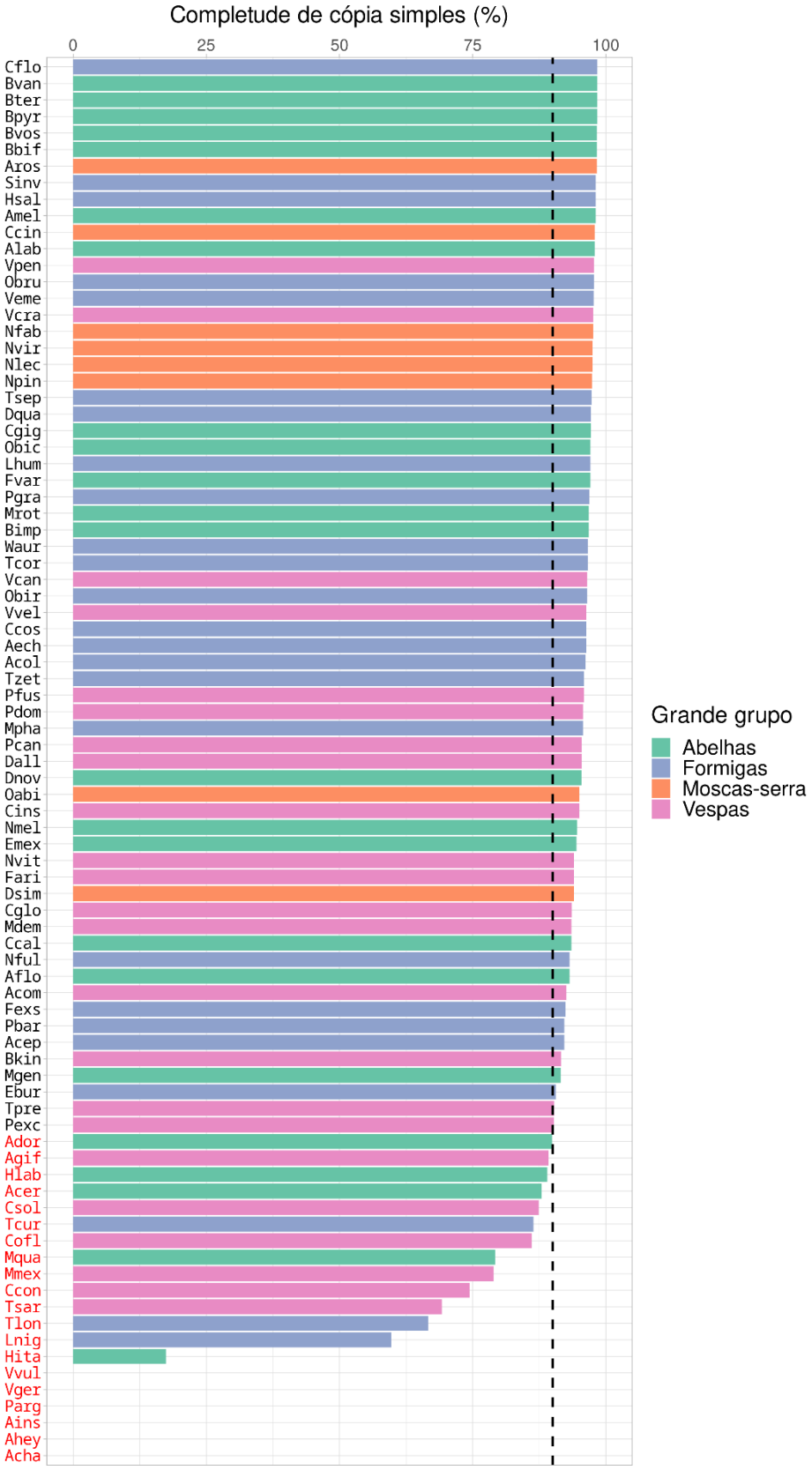
Figura 5 - Distribuição de tamanho de proteoma por grande grupo



Box plots representando a distribuição dos valores de tamanho do proteoma por grande grupo, definidos pelo número de proteínas, para cada proteoma não redundante de alta qualidade das espécies solitárias e eussociais de Hymenoptera. Cores representam o grande grupo taxonômico. O código completo está disponível no Apêndice I. Fonte: elaborada pelo autor (2022).

As análises de completude com BUSCO revelaram que 65 dos 85 (76,47% do total) proteomas de Hymenoptera obtidos possuíam valores de completude de ortólogos de cópia simples acima de 90% (Figura 6), sendo então classificados como proteomas de alta qualidade. O valor médio de completude entre todos os 85 proteomas de Hymenoptera foi de 85,7%. Entre os 65 proteomas de alta qualidade estão representadas 22 espécies de formigas, 18 espécies de abelhas, 17 espécies de vespas e todas as oito espécies recuperadas de moscas-serra, valores que proporcionam uma representação taxonômica satisfatória. Neste momento, removemos as espécies com nível de socialidade considerado intermediário (Apêndice B), de maneira a produzir dois grupos fenotipicamente contrastantes de organismos solitários ou eussociais, totalizando 22 e 40 espécies, respectivamente.

Figura 6 - Resultado da análise de completude com BUSCO



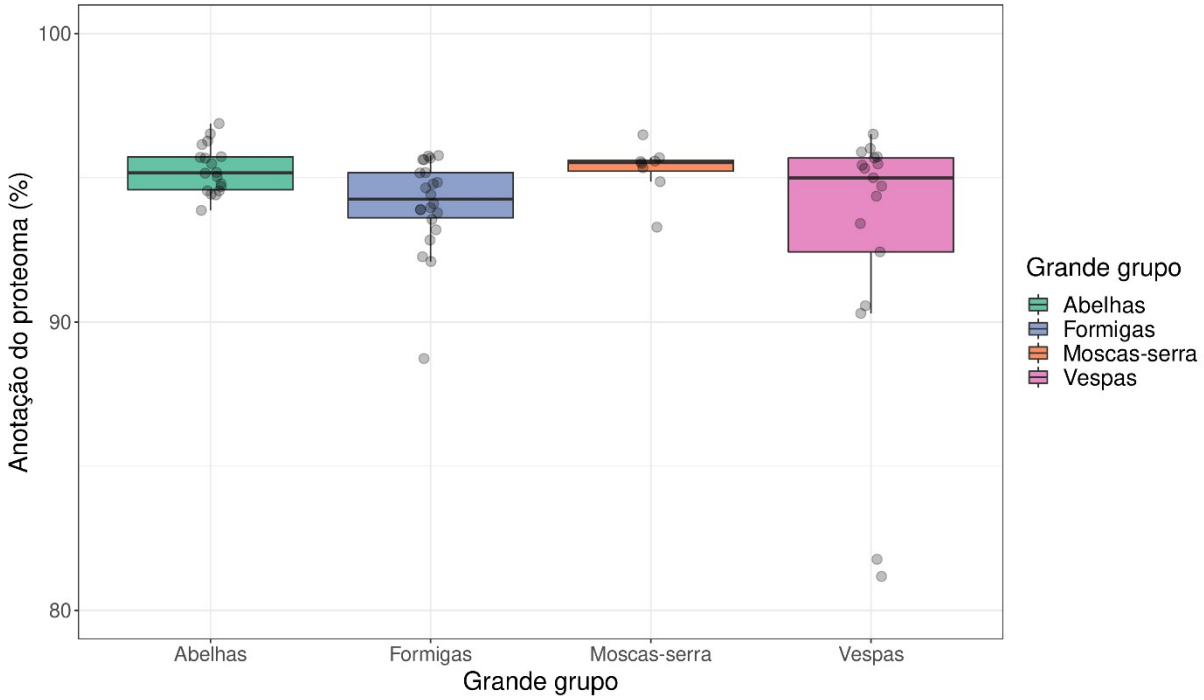
Resultado da análise de completude com BUSCO, apresentando apenas os valores de completude (em porcentagem) de ortólogos de cópia simples. As barras com os valores de completude estão coloridas de acordo com o grande grupo a que cada espécie pertence e estão ordenadas por valor de completude de cópia simples. A linha preta tracejada representa o ponto de corte de 90% de

completude de cópia simples. Espécies com código em vermelho possuem valores de completude de cópia simples < 90% e foram removidas das análises subsequentes. O significado das abreviações está disponível no Apêndice A. Figura gerada através de *script in-house* da linguagem R (Apêndice H). Fonte: elaborada pelo autor (2022).

5.2 Anotação dos genomas e contagem de IPR

Todos os 65 proteomas selecionados não redundantes de alta qualidade foram anotados *de novo*, resultando em um total de 9.662 termos IPR únicos (IPR IDs) que, em sua totalidade, produziram 2.045.867 pares proteína-IPR anotados entre todos os proteomas, com uma média de 32.997,85 pares proteína-IPR anotados por proteoma. Em média, a cobertura de anotação dos proteomas foi de 94,26%. O proteoma da vespa *Polistes exclamans* (81,18% de cobertura) representa o menor valor e o da abelha *Dufourea novaeangliae* (96,88% de cobertura) o maior valor de cobertura de anotação, respectivamente (Figura 7). As vespas *P. exclamans* e *A. compressa* são os únicos organismos entre nossos dados com uma cobertura de anotação abaixo de 85% (Figura 7). Além disso, ambas são as espécies com os maiores proteomas entre nossos dados (Figura 5). Por fim, os níveis de completude de ortólogos de cópia simples de *P. exclamans* e *A. compressa* são de 90,2% e 92,6%, respectivamente, colocando-as entre os 10 proteomas com menores valores de completude de ortólogos de cópia simples entre os proteomas considerados de alta qualidade em nossos resultados (Figura 6, Pexc e Acom, respectivamente). Em conjunto, estes resultados podem indicar potenciais falhas na predição gênica inicial dos genomas dessas espécies, o que explicaria a presença de proteínas “em excesso” que não contribuem para a completude da anotação destes proteomas. Apesar desta possibilidade, as completudes de anotação desses proteomas avaliadas por BUSCO satisfazem a qualidade mínima considerada neste trabalho e, portanto, estas espécies foram mantidas.

Figura 7 - Distribuição de cobertura de anotação por grande grupo



Box plots representando a distribuição da porcentagem do total anotado de cada proteoma por grande grupo, definida pela razão entre o número de proteínas únicas anotados pelo *InterProScan* e o tamanho do proteoma definido pelo número total de proteínas, para cada proteoma não redundante de alta qualidade das espécies solitárias e eussociais de Hymenoptera. Cores representam o grande grupo taxonômico. O código completo está disponível no Apêndice I. Fonte: elaborada pelo autor (2022).

5.3 Árvore filogenética

A árvore extraída de Peters et al. (2017), utilizada como base para a construção da filogenia deste trabalho, inclui 174 taxa que representam 173 espécies de insetos, incluindo 167 espécies de Hymenoptera (Figura 8). Aproximadamente 5% dessas espécies estão representadas em nosso conjunto de dados (Figura 8, ramos não marcados), mas diversos taxa da árvore são de grande utilidade por servirem como pontos de referência que permitem obter os padrões e os tempos de divergência para os grandes grupos basais de Hymenoptera (Figura 8, ramos em azul). Essa “árvore-esqueleto” pode, então, ser utilizada como referência para a inserção de espécies faltantes e suas respectivas topologias, obtidas a partir de árvores especializadas.

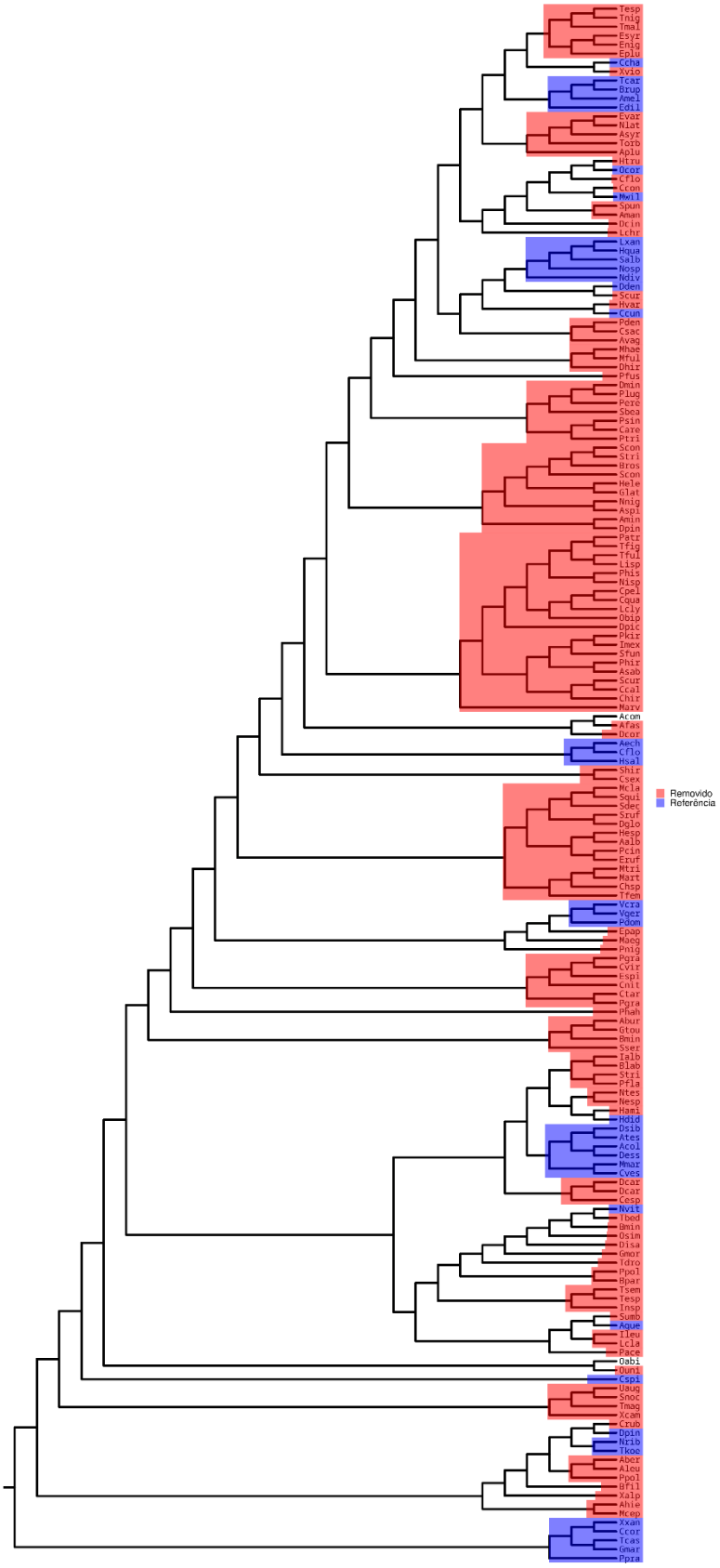
Após a remoção das espécies não informativas da “árvore-esqueleto” de Peters et al. (2017), obtivemos uma árvore intermediária consideravelmente menor (Figura 9), consistindo em 39 espécies de insetos, sendo 10 dessas espécies representativas de proteomas não redundantes de alta qualidade presentes no

nosso conjunto de dados (ramos não-marcados). As demais 29 espécies foram utilizadas como pontos de referência para a inserção dos representantes faltantes dos proteomas, os quais foram determinados a partir de uma extensa revisão literária das relações filogenéticas entre os grupos de interesse (Apêndice C).

Ao fim, 57 novas espécies representando 5 grandes grupos (abelhas, formigas, moscas, moscas-serra e vespas) foram inseridas na árvore intermediária, bem como seus respectivos tempos de divergência, seguido da remoção das espécies de referência não representadas entre os proteomas do nosso conjunto de dados (Figura 10). As espécies *Tribolium castaneum* e *Drosophila melanogaster* foram mantidas como grupo externo para o enraizamento da árvore e finalmente removidas, resultando em uma árvore final com todas as 65 espécies de Hymenoptera representadas por proteomas não redundantes de alta qualidade em nosso conjunto de dados.

A árvore ultramétrica final de espécies (Figura 11) foi construída a partir da remoção das espécies de Hymenoptera com nível intermediário de socialidade, com o intuito de modelar a eussocialidade como um fenótipo categórico binário. Compreende, portanto, um total de 62 espécies de Hymenoptera representadas por proteomas não redundantes de alta qualidade e classificáveis em “social” ou “solitária” de acordo com informações obtidas da literatura científica a partir de extensa revisão (Apêndice B). Os dados de eussocialidade ilustrados por nossa árvore estão em aparente acordo com a literatura (BRANSTETTER et al., 2017; PETERS et al., 2017), indicando: a presença de eussocialidade apenas em Aculeata; a presença de eussocialidade em todas as formigas, sugerindo um surgimento único no ancestral comum do grupo (HÖLLDOBLER; WILSON, 1990); a presença de eussocialidade em todas as espécies do clado Vespidae, corroborando um surgimento único da eussocialidade no clado formado por Vespinae + Polistinae (HUNT, 2007); a presença de eussocialidade em todas as abelhas das tribos Apinii, Bombinii e Meliponinii, de acordo com o cenário de um ou múltiplos surgimentos do traço em Apidae (CARDINAL; DANFORTH, 2011); e a presença do traço em uma abelha da família Halictidae, corroborando o cenário de um outro surgimento independente da eussocialidade em abelhas (GRIMALDI; ENGEL, 2005).

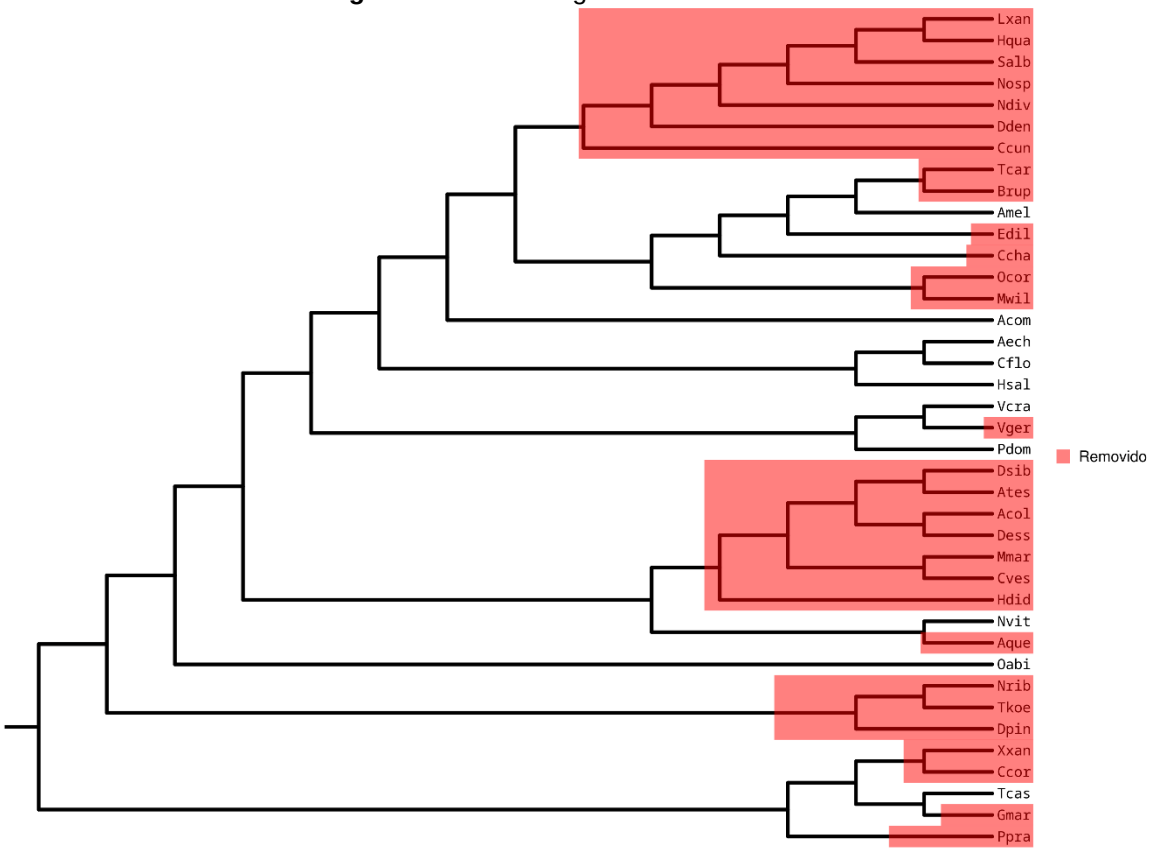
Figura 8 - Árvore filogenética “esqueleto”



Cladograma da árvore filogenética original extraída de Peters et al. (2017). Ramos destacados em vermelho são espécies não informativas para o trabalho que foram removidas da árvore. Ramos marcados em azul representam espécies utilizadas como pontos de referência para a inserção de novas espécies. Nem todas as espécies destacadas em azul foram mantidas na árvore final. Ramos

não destacados são espécies não utilizadas para a inserção de novos ramos, mas que foram mantidas na árvore final por representarem proteomas não redundantes de alta qualidade presentes em nosso conjunto de dados. Fonte: elaborada pelo autor (2022)[2].

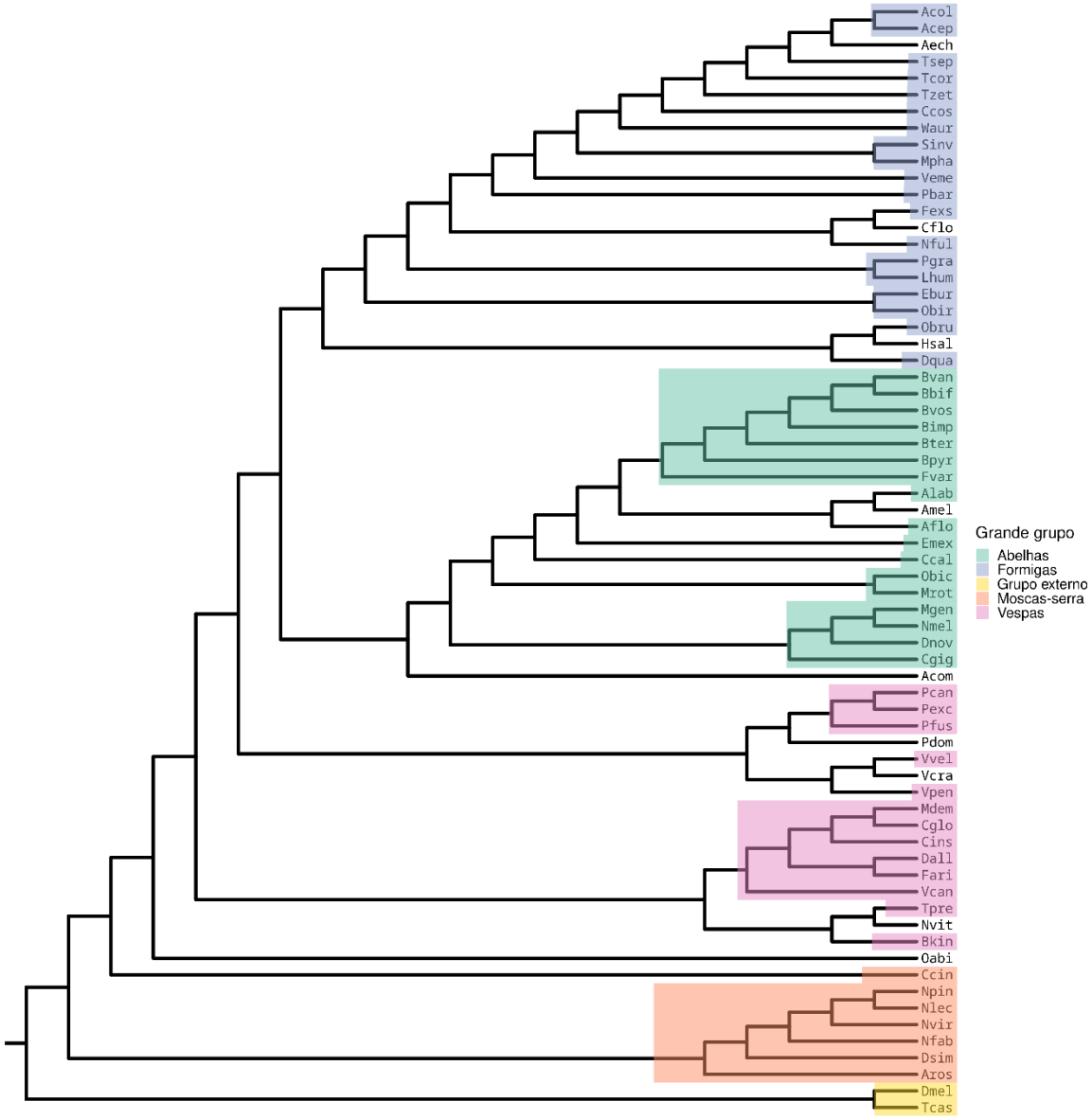
Figura 9 - Árvore filogenética intermediária



Cladograma da árvore filogenética intermediária, modificada a partir de Peters et al. (2017), antes da inserção de novas espécies. Ramos marcados em vermelho representam espécies utilizadas como pontos de referência para a inserção de novos ramos e que não foram mantidas na árvore final. Ramos não destacados representam espécies mantidas na árvore final. Fonte: elaborada pelo autor (2022).

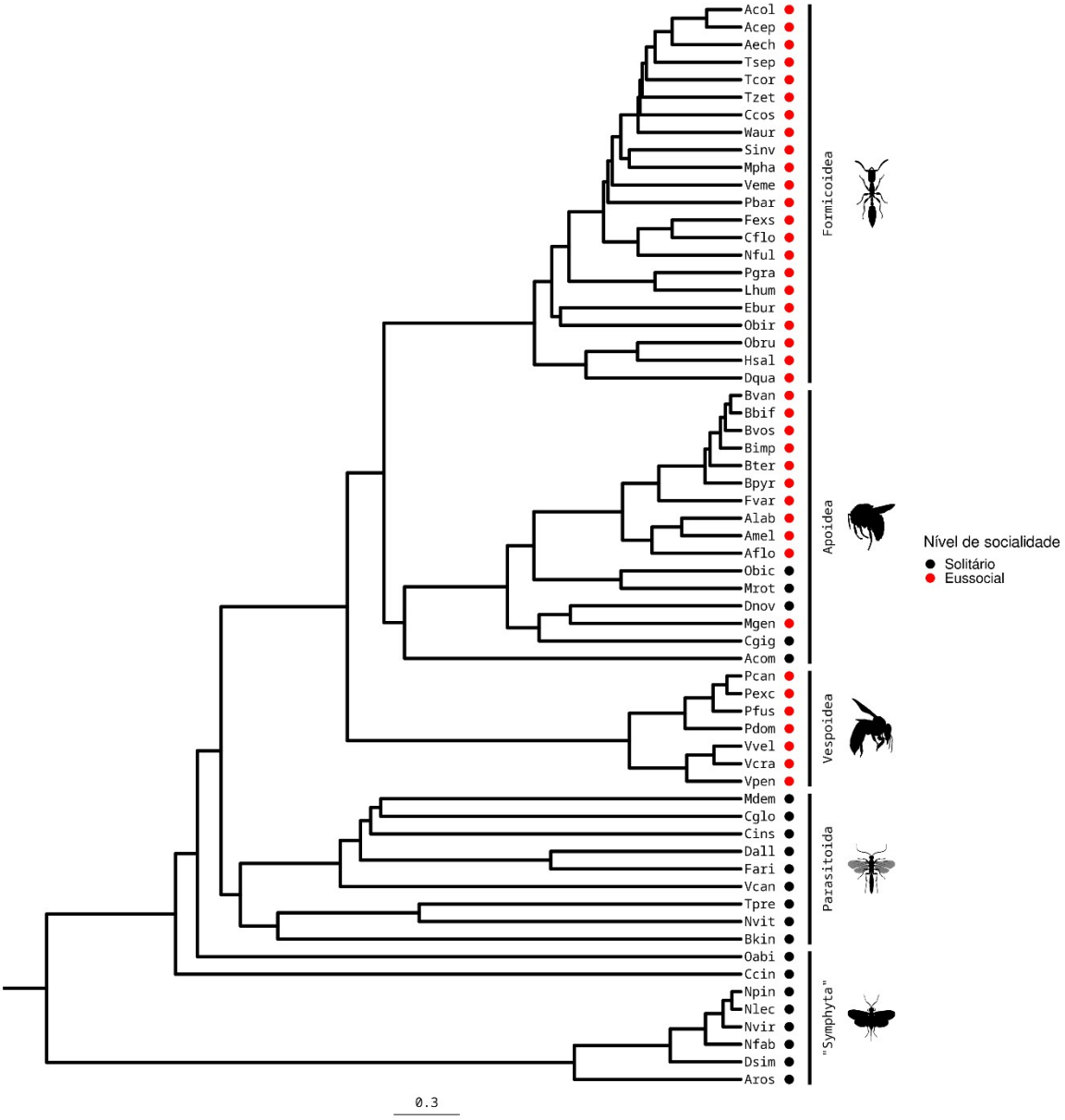
[2] Adaptada de Peters et al. (2017).

Figura 10 - Árvore filogenética intermediária após inserção de espécies



Cladograma da árvore intermediária, modificada a partir de Peters et al. (2017), após a inserção de ramos representando todas as espécies possuidoras de proteomas não redundantes de alta qualidade presentes em nosso conjunto de dados. Ramos coloridos representam o grupo externo e espécies que foram adicionadas, com posições baseadas na literatura (Apêndice C). As cores dos ramos representam o grande grupo a qual cada espécie pertence. Fonte: elaborada pelo autor (2022).

Figura 11 - Árvore filogenética final



Árvore filogenética ultramétrica contendo as 62 espécies de Hymenoptera presentes em nosso conjunto de dados que possuem proteomas não redundantes de alta qualidade e que podem ser classificadas em eussociais ou solitárias. Círculos coloridos representam o nível de socialidade de cada espécie. Barras verticais delimitam os grandes grupos referidos no trabalho. O termo “Symphyta” representa um grupo parafilético que contém as moscas-serra e não é mais utilizado na literatura. As silhuetas que ilustram cada grupo foram retiradas do site PhyloPic (http://phylopic.org/) e possuem licença de uso comum sem necessidade de atribuição. O significado das abreviações está disponível no Apêndice A. Fonte: elaborada pelo autor (2022).

5.4 ANOVA filogenético e IPR associados à eussocialidade

De posse da árvore filogenética e dos dados de contagem de IPRs para cada espécie, procedemos com a avaliação da associação entre o número de cópias dos 9.662 IPR IDs distintos e do fenótipo de eussocialidade. Ao final, encontramos 74

pares proteína-IPR associados significativamente com a presença de eussocialidade em Hymenoptera (valor-*q* < 0.05).

Após a utilização dos filtros que visam minimizar a ocorrência de eventuais associações espúrias causadas por vieses introduzidos pelo sequenciamento, montagem e anotação de genomas, sete IPR IDs possuíam as características mínimas de ocorrência para serem considerados como significativamente associados à eussocialidade. O termo IPR036658 (*CPI-17 superfamily*) foi manualmente removido por representar uma informação redundante com o termo IPR008025 (*CPI-17*). Assim, reportamos 6 IPR IDs não redundantes significativamente associados ao fenótipo de eussocialidade (Figura 12), os quais anotam um total de 176 genes nos 62 proteomas analisados.

O domínio anotado pelo termo IPR014878 encontra-se duplicado em quase todas as espécies de formigas, com exceção de *Ooceraea biroi* (Obir), e está presente em cópia simples em todas as outras espécies de Hymenoptera avaliadas neste trabalho, independentemente da presença de eussocialidade, com exceção da mosca-serra *Orussus abietinus* (Oabi), da vespa parasitoide *Trichogramma pretiosum* (Tpre) e das abelhas *Megachile rotundata* (Mrot), *Osmia bicornis bicornis* (Obic) e *Colletes gigas* (Cgig), todas solitárias (Figura 12). Assim, o padrão de ocorrência desse domínio observado em nossos resultados sugere uma associação entre a sua presença e número de cópias e a ocorrência do fenótipo de eussocialidade.

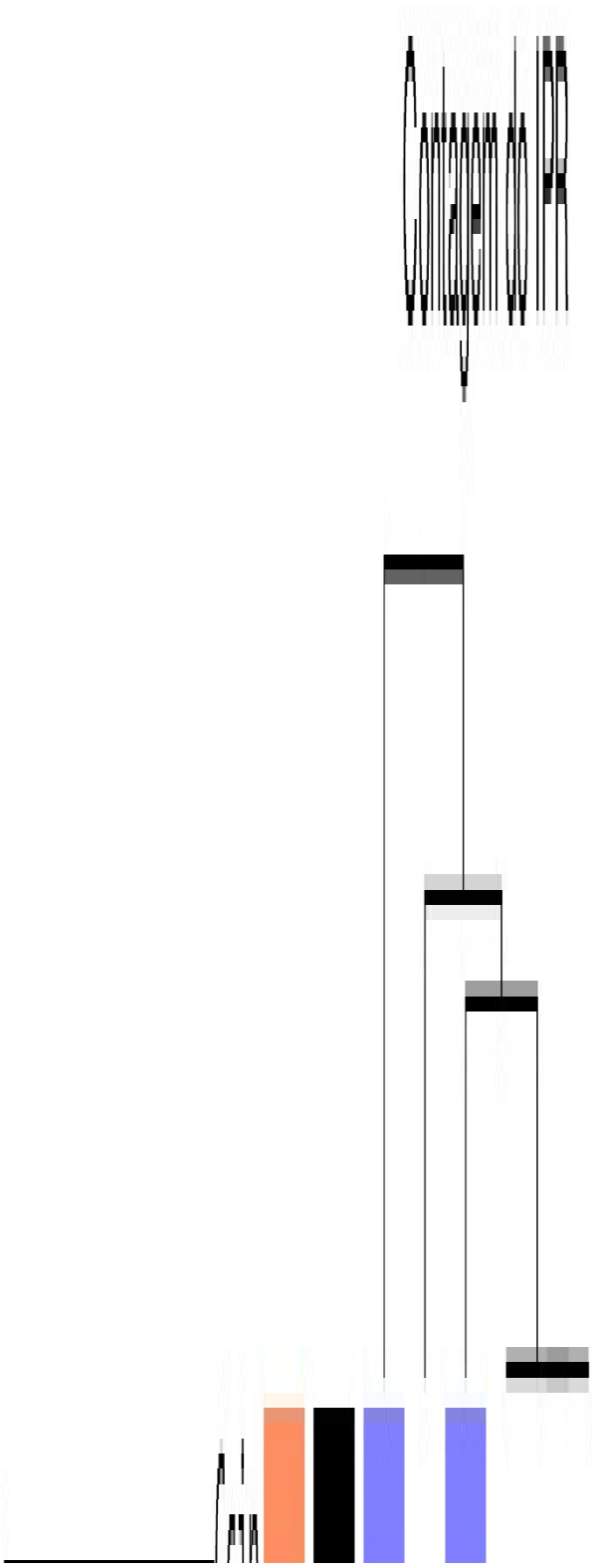
O termo IPR019023 anota um domínio presente em todas as formigas, com exceção das formigas de correição *O. biroi* e *Eciton burchellii* (Ebur), e em todas as abelhas eussociais. Também está presente em algumas espécies solitárias, nominalmente as vespas parasitoides *T. pretiosum*, *Diachasma alloeum* (Dall) e *Nasonia vitripennis* (Nvit) e na mosca-serra *Belonocnema kinseyi* (Bkin). Está ausente nas demais espécies, incluindo as vespas eussociais, único grupo eussocial com total ausência do termo (Figura 12). Concluímos que a ocorrência desse domínio também apresenta, para os dados avaliados, uma associação positiva com o fenótipo da eussocialidade.

O domínio *CPI-17*, anotado pelo termo IPR008025, compreende duas perdas independentes em Formicidae e Vespidae, ambos clados eussociais, estando presente em quase todas as abelhas, na vespa parasitoide *A. compressa* (Acom) e nas moscas-serra *Athalia rosae* (Aros) e todas as espécies de mosca-serra da

família Diprionidae (Dsim, Nfab, Nvir, Nlec e Npin) (Figura 12). Assim, concluímos que esse domínio parece estar associado negativamente com a ocorrência da eussocialidade.

Por fim, os termos IPR041989, IPR033650 e IPR037892 estão presentes apenas nas espécies de abelhas eussociais, com exceção da abelha facultativamente eussocial *Megalopta genalis* (Mgen), e estão ausentes em todas as demais espécies (Figura 12), o que sugere uma associação positiva entre sua ocorrência e o fenótipo da eussocialidade, particularmente em abelhas.

Figura 12 - Resultados das análises de ANOVA filogenético



Heatmap demonstrando termos IPR com associações significativas entre sua contagem normalizada e a presença ou ausência de eussocialidade nas 62 espécies de Hymenoptera com proteomas não redundantes de alta qualidade e classificadas em eussociais ou solitárias. Fonte: elaborada pelo autor (2022).

6 DISCUSSÃO

Conforme discutido anteriormente, diversas expansões e contrações de domínios foram previamente mencionadas como estando associadas à evolução da eussocialidade em Hymenoptera, tais como receptores olfatórios e *royal jelly protein* (KARPE et al., 2017; MCKENZIE; KRONAUER, 2018; OEYEN et al., 2020). Entretanto, nenhum destes estudos utilizou informação filogenética como um parâmetro adicional na confecção dos modelos estatísticos para buscar por eventuais associações, o que viola os pressupostos de virtualmente todos modelos estatísticos tradicionais. Após a utilização da informação filogenética, nenhuma das associações previamente mencionadas foi detectada em nossa análise.

Os IPR encontrados significativamente associados à eussocialidade em Hymenoptera representam um importante ponto de partida para uma investigação dos aspectos funcionais e evolutivos de um fenótipo multifatorial tão complexo em uma ordem de insetos tão diversa. Entretanto, apesar da aparente super-representação dos Hymenoptera entre os genomas de insetos sequenciados (HOTALING et al., 2021), a caracterização funcional da associação de genes com a eussocialidade em espécies da ordem ainda é escassa (LI et al., 2019). Uma vez que não obtivemos relatos de fenótipos mutantes dos genes contendo esses domínios em Hymenoptera, utilizamos o organismo modelo *D. melanogaster* para avaliar eventuais alterações fenotípicas desses genes, quando disponíveis e, na ausência destes, buscamos por caracterizações funcionais dos respectivos homólogos em humanos. Adicionalmente, recorremos à literatura científica para avaliar os eventuais papéis funcionais dos genes contendo estes domínios, como forma de encontrar potenciais mecanismos causais para as associações observadas.

6.1 IPR014878: Domain of unknown function DUF1794

Este IPR, cuja nomenclatura atualizada é *THAP4-like, heme-binding beta-barrel domain*, anota um domínio de ligação heme presente em proteínas *THAP domain-containing protein 4* (EMBL-EBI, 2022). A família de proteínas contendo o domínio *Thanatos-associated protein domain* (THAP) inclui genes associados com a

proliferação celular, apoptose, regulação transcricional e regulação da quiescência celular (BIANCHETTI; BINGMAN; PHILLIPS JR., 2011; GAL et al., 2021).

A proteína *THAP domain-containing protein 4* (*THAP4*) é uma das proteínas THAP com função menos conhecida. Entretanto, análises estruturais e funcionais demonstraram que há um papel crucial de *THAP4* no ciclo metabólico do óxido nítrico (NO) (BIANCHETTI; BINGMAN; PHILLIPS JR., 2011). O NO é uma molécula de sinalização celular e um importante sinalizador do sistema nervoso central em insetos (MÜLLER, 1997). A sinalização por NO é importante durante o desenvolvimento embrionário em insetos e está associada com a regulação da proliferação celular, crescimento dos axônios e maturação sináptica (BICKER, 2001). Além disso, parece existir uma associação entre a sinalização por NO e a formação de memória de longo prazo em *A. mellifera* (MÜLLER, 1996).

A memória olfativa e visual de longo prazo é um fenótipo importante para os Hymenoptera eussociais e, em formigas, é utilizada no reconhecimento e troca de informações com membros da colônia, localização e seleção de substratos adequados para o cultivo de fungos (JOSENS; ESCHBACH; GIURFA, 2009; PROVECHO; JOSENS, 2009; SAVERSCHEK et al., 2010). Formigas aprendem rápido e têm memórias de longo prazo bastante longevas (PIQUERET; SANDOZ; D'ETTORRE, 2019). Além disso, a formação rápida de memórias olfativas persistentes pode estar associada com a expansão acelerada de genes de receptores olfativos em formigas (MCKENZIE; KRONAUER, 2018).

O papel molecular mais bem caracterizado da proteína *THAP4* envolve a detoxificação do peroxinitrito, um agente oxidante gerado pela reação espontânea entre o NO e superóxidos (DE SIMONE et al., 2018). Danos comumente atribuídos ao acúmulo e a toxicidade de NO e ao estresse oxidativo são em parte causados pelo peroxinitrito, que é um potente indutor da morte celular e está associado a patologias como a neurodegeneração, alteração da sinalização celular e lesões do DNA (PACHER; BECKMAN; LIAUDET, 2007; SZABÓ; ISCHIROPOULOS; RADI, 2007; TRUJILLO; FERRER-SUETA; RADI, 2008). O estresse oxidativo está relacionado à senescência em *A. mellifera* de uma maneira casta-dependente, através da qual operárias forrageadoras envelhecem mais rapidamente, enquanto o peroxinitrito parece ser responsável pela nitração seletiva de uma proteína ainda não identificada no cérebro da abelha (SEEHUUS; KREKLING; AMDAM, 2006).

O estresse oxidativo pode ocorrer em situações de hipóxia (BURTON; JAUNIAUX, 2011). *THAP4* está potencialmente associada a adaptações à vida em ambientes pouco oxigenados (BARTH et al., 2019), o que poderia explicar sua expansão em formigas, já que esses insetos fazem ninhos subterrâneos e consequentemente sofrem com estresse oxidativo (LI-BYARLAY; CLEARE, 2020). A ausência do gene dessa proteína em *O. biroi* pode ser uma perda real presente em algumas formigas da subfamília Dorylinae conhecidas como formigas de correição, caracterizadas, entre outros fatores, por um peculiar comportamento migratório e a presença de ninhos efêmeros não subterrâneos (HÖLLDOBLER; WILSON, 1990). Entretanto, a espécie *Eciton burchellii* (Ebur), uma formiga de correição presente em nossos dados, possui o gene *THAP4* duplicado, apesar de formar ninhos superficiais. Além disso, *O. biroi* não é considerada uma formiga de correição "verdadeira" e, apesar de possuir muitos traços em comum com as mesmas (BAUDIER, 2019), *O. biroi* constrói ninhos subterrâneos (KRONAUER et al., 2013).

Portanto, nossos resultados sugerem que há uma associação significativa entre a expansão do gene *THAP4* (representado pelo termo IPR014878) em formigas e traços importantes para a evolução da eussocialidade nesses himenópteros, como a construção de ninhos subterrâneos e consequentemente a adaptação à vida em ambientes hipóxicos e a sinalização por NO na formação de memórias de longo prazo, potencialmente associadas à recepção olfatória. Além disso, apesar das aparentes contradições, a potencial perda de uma das cópias desse gene na formiga dorilínea *O. biroi* pode estar associada com o estilo de vida das formigas de correição, possibilidade corroborada por sua proximidade filogenética e similaridade comportamental.

6.2 IPR019023: Lamin-B receptor of TUDOR domain

O receptor lamina-B anotado por esse IPR é uma importante proteína transmembrana localizada no envelope nuclear capaz de se ligar à lamina e à cromatina e com um importante papel na formação do próprio envelope nuclear (MA et al., 2007). É também um receptor presente em enzimas que participam da via de biossíntese de colesterol, tendo um papel essencial na manutenção dessa via (TSAI et al., 2016). Nos proteomas pertencentes ao nosso conjunto de dados, esse domínio está presente na proteína *delta(14)-sterol reductase*, implicada diretamente

na via de biossíntese do colesterol (WATERHAM, 2006) e associada com o desenvolvimento celular em plantas (SCHRICK et al., 2000) e o desenvolvimento de células mieloides (SUBRAMANIAN et al., 2012).

Dentre muitas outras funções, o colesterol é um precursor essencial na via de biossíntese de ecdisteróides como a ecdisona (NIWA; NIWA, 2014), uma classe de hormônios fundamentais para o desenvolvimento larval e a muda em insetos e para a diapausa, determinação de castas e desenvolvimento ovariano em Hymenoptera (BLOCH; HEFETZ; HARTFELDER, 2000; HSIAO; HSIAO, 1969; SUZZONI; PASSERA; STRAMBI, 1980). Além disso, a presença de colesteróis em secreções mandibulares de formigas é um indicativo do papel dessa substância no reconhecimento e na determinação de castas (MARTINS et al., 2015).

A aparente perda dessa proteína em vespas sociais é certamente curiosa, já que ecdisteróides parecem possuir funções importantes nesse clado, como por exemplo a definição da hierarquia e dominância social (KAPHEIM, 2017; KELSTRUP et al., 2014).

Por outro lado, as formigas de correição possuem alguns traços comportamentais e fisiológicos que chamam a atenção. Operárias das espécies de formigas de correição “verdadeiras” como *E. burchellii* e espécies próximas como *O. biroi* possuem um ciclo comportamental que alterna entre duas fases: uma fase migratória, de forrageamento, e uma fase estacionária, de aninhamento (BAUDIER, 2019). Durante a fase migratória, essas formigas abandonam o ninho atual e forrageiam por alimento em regiões próximas, formando um novo ninho provisório, normalmente superficial e abrigado da luz, ao entrar na fase estacionária. Este ciclo parece ser regulado pela demanda alimentar e o estágio de desenvolvimento da prole, o que exige uma fina sincronização do ciclo de desenvolvimento larval (RAVARY; JAISSON, 2002), apesar do fato de que o tempo de desenvolvimento larval e o número de ínstares em espécies de formigas de correição não parece ser atípico quando comparado a outras formigas (HÖLLDOBLER; WILSON, 1990).

Além disso, as operárias da formiga *O. biroi* retêm a capacidade de produzir ovos por partenogênese telítoca, produzindo fêmeas diploides (TSUJI; YAMAUCHI, 1995), e colônias dessa espécie não possuem rainhas (RAVARY; JAISSON, 2004). Algumas espécies de vespas eussociais, como as da subfamília Polistinae, possuem um sistema fluido de diferenciação de castas, com a presença de castas normalmente não reprodutoras que retêm e podem reativar sua capacidade de

reprodução, havendo também pouca divisão observável entre rainhas e operárias, e possuem um número variável de rainhas por colônia (HUNT, 2007). Ainda, algumas espécies são bastante móveis, outras são parasitas sociais (se infiltram nos ninhos de outras vespas), e comumente as colônias produzem fêmeas solitárias ou enxames de vespas destinados a fundar novos ninhos (RICHARDS, 1971).

Portanto, parece haver algum grau de convergência entre características comportamentais e fisiológicas de formigas de correição, vespas sociais e até mesmo Hymenoptera solitários, como uma alta mobilidade, um caráter migratório e uma divisão de castas fluida com a existência da retenção da capacidade reprodutiva. Estas convergências podem ser um indício de uma consequência funcional da perda do gene *delta(14)-sterol reductase* e consequentemente de seu papel na síntese de hormônios ecdisteróides importantes para o desenvolvimento embrionário e maturação ovariana.

6.3 IPR008025: CPI-17

A família *CPI-17* compreende os genes conhecidos como *protein phosphatase 1 regulatory subunit 14*, responsáveis pela inibição da proteína *protein phosphatase 1* (ETO, 2009) e, nos proteomas em nosso conjunto de dados, está presente nas proteínas *protein phosphatase 1 regulatory subunit 14b* (PPP1R14B) e *protein phosphatase 1 regulatory subunit 14c* (PPP1R14C).

A inibição da proteína *protein phosphatase 1* (PP1) está relacionada com uma maior proliferação celular (CASAMAYOR; ARIÑO, 2020; ETO, 2009). O *knockdown* por RNA de interferência (RNAi) de PPP1R14B em culturas de células humanas está associado com uma redução da viabilidade celular (BURLEIGH et al., 2015; CHEUNG et al., 2011; SIMPSON et al., 2012) e com alterações na maquinaria de reparo do DNA e a presença de células cancerígenas (VIZEACOUMAR et al., 2013). Por outro lado, PPP1R14C parece estar associado com a sinalização neuronal e com a via de recompensa e vício em camundongos (GONG et al., 2005), reduzida viabilidade de células de rim humano em um mecanismo que envolve o controle da secreção da interleucina-8 (WARNER et al., 2014) e reduzida viabilidade celular em linhagens de câncer ovariano humano (CHEUNG et al., 2011). Superexpressão de PPP1R14C induz a progressão de câncer de mama triplo-negativo (JIAN et al.,

2022). Portanto, ambos os genes provavelmente possuem papéis importantes no controle da proliferação celular e no desenvolvimento.

O gene da proteína PP1 tem expressão diferencial regulada por ciclo circadiano em diferentes castas de operárias da formiga *Camponotus floridanus*, o que pode estar associado com uma maior plasticidade de castas e subcastas de operárias em formigas (DAS; DE BEKKER, 2022). Outros grupos de Hymenoptera eussociais, como por exemplo a abelha *A. mellifera*, também possuem subdivisão de castas, inclusive de modo temporal e associada à idade das operárias (ROBINSON, 2009). Entretanto, uma maior plasticidade de castas (polietismo) parece ocorrer em formigas e em espécies que apresentam eussocialidade “simples” e que possuem colônias menores, *i.e.,* com menor número de indivíduos (JEANNE; FAGEN, 1974; MERTL; TRANIELLO, 2009; TORRES; GIANOTTI; ANTONIALLI-JR, 2013). Portanto, a perda potencialmente convergente dos genes *CPI-17* em formigas e vespas eussociais pode sugerir que a ausência de inibidores da atividade de PP1 esteja associada a uma maior plasticidade de castas em espécies eussociais de Hymenoptera.

6.4 IPR041989: TOPK, catalytic domain

A anotação IPR041989 corresponde ao domínio catalítico da proteína *lymphokine-activated killer T-cell-originated protein kinase* (TOPK), também denominada *PDZ-binding kinase* (PBK) em humanos.

A proteína TOPK possui um papel essencial na mitose e na citocinese e, portanto, tem importância considerável no controle da proliferação celular (ABE et al., 2007; PARK et al., 2010). O *knockdown* de TOPK por RNAi em *Drosophila melanogaster* está associado com letalidade em larvas e pupas e com fertilidade reduzida em fêmeas (MUMMERY-WIDMER et al., 2009; SCHNORRER et al., 2010; SOPKO et al., 2014), o que corrobora o papel do gene nas vias de desenvolvimento embrionário.

Todas as abelhas que possuem o domínio pertencem à família Apidae (abelhas corbiculadas). O controle da proliferação celular é importante para o desenvolvimento embrionário e dos túbulos de Malpighi em *A. mellifera* (BARCHUK et al., 2007; CRUZ-LANDIM; PATRÍCIO; ANTONIALLI JR, 2006; GONÇALVES et al., 2018), o que pode sugerir um papel importante do gene TOPK no controle do

desenvolvimento em abelhas eussociais. Por exemplo, em vespas eussociais, diferentemente das abelhas, a definição de castas é principalmente comportamental e não reprodutiva, não sendo afetada pelo desenvolvimento ovariano diferencial, e operárias de diversas espécies de formiga não desenvolvem a espermateca, enquanto em *A. mellifera* o órgão é mantido de forma vestigial (RAMSAY; LASKO; ABOUHEIF, 2021).

Entretanto, talvez a maior diferença entre abelhas eussociais e demais Hymenoptera eussociais seja a utilização da geleia real como um sinalizador nutricional para a diferenciação de castas em *A. mellifera* (SNODGRASS, 1925). Dessa forma, o controle da proliferação celular regulado por TOPK e, portanto, da diferenciação de castas, pode estar associado a sinais nutricionais, o que explicaria sua expansão apenas em abelhas eussociais e não nos demais Hymenoptera.

6.5 IPR033650: Mitochondrial ribosomal protein L46 NUDIX

A proteína *mitochondrial ribosomal protein L46* (MRPL46) é uma hidrolase com um domínio *NUDIX motif-contaning*. Proteínas hidrolases que possuem esse domínio estão envolvidas em funções de remoção de compostos celulares tóxicos (MCLENNAN, 2006).

O gene MRPL46 parece ser essencial para o crescimento de tumores cerebrais em *D. melanogaster*, sugerindo um papel importante no desenvolvimento e na sobrevida (LOUZAO BOADO, 2019). O *knockdown* através de RNAi do gene MRPL46 em *D. melanogaster* está associado com letalidade total ou parcial e com uma resposta atípica à recepção de dor (NEELY et al., 2010; SCHNORRER et al., 2010).

Em uma análise de micro arranjo de DNA, foi encontrada uma expressão diferencial do gene MRPL46 nos cérebros de diferentes castas de *A. mellifera* (PAULA JUNIOR; BARCHUK, 2019), o que pode sugerir um papel da proteína na diferenciação de castas em abelhas eussociais. Entretanto, mais dados e análises funcionais em Hymenoptera são necessários para a identificação do papel desse gene no controle de fenótipos associados à eussocialidade.

6.6 IPR037892: SNX14, RGS domain

Presente na proteína *sorting nexin-14* (*Snx14*), o domínio *RGS* regula a atividade de sinalizadores associados à proteína G, tendo um importante papel na transdução de sinal e, na proteína *sorting nexin-13* (*SNX13*), pertencente à mesma família, possui capacidade de regular a atividade da proteína quinase A (*PKA*) e pode ter uma associação com a longevidade em *D. melanogaster* (SUH et al., 2008).

A proteína *Snx14* promove a maturação de gotículas de lipídios em contato com o retículo endoplasmático (DATTA et al., 2019). As gotículas de lipídios são a forma de armazenagem celular de ácidos graxos em excesso a partir de triacilglicerídeos. Ácidos graxos são moléculas essenciais para a biossíntese de membrana (e, portanto, possuem importância na divisão celular), sinalização celular e acúmulo de energia. Gotículas de lipídios foram encontradas em neurônios e células da glia de diversos organismos, como *D. melanogaster*, onde parecem se acumular em resposta a fatores de estresse como privação de nutrientes e hipóxia, e a desregulação do contato entre as gotículas de lipídios e o retículo endoplasmático parece ter um papel na neurodegeneração (FOWLER et al., 2019). Gotículas de lipídios também estão associadas com o desenvolvimento, estresse oxidativo, ciclo de vida de proteínas, além de serem importantes para a comunicação entre organelas e, portanto, terem um papel essencial no metabolismo celular (OLZMANN; CARVALHO, 2019; WELTE; GOULD, 2017).

As funções descritas para as gotículas de lipídios, a associação da proteína *Snx14* na maturação das mesmas e o padrão de presença do gene em nossos resultados sugere um papel no desenvolvimento e na diferenciação de castas associada à sinalização nutricional em abelhas eussociais.

7 CONCLUSÃO

Em vista da grande diversidade de modos de vida e adaptações e a extensão dos surgimentos independentes da eussocialidade em Hymenoptera, a busca por elementos genômicos comuns associados à evolução de um fenótipo multifatorial tão complexo em múltiplos clados da ordem se demonstra um grande desafio.

O recente aumento da disponibilidade de genomas anotados de alta qualidade em bancos de dados públicos providencia uma oportunidade para a busca por associações entre componentes genômicos e a evolução de fenótipos complexos como a eussocialidade. Entretanto, metodologias adequadas devem ser

aplicadas nesse tipo de análise (NAGY et al., 2020). Primeiramente, é importante levar em conta o fato de que análises comparativas filogenéticas ferem um dos princípios mais básicos de grande parte dos métodos estatísticos: a independência dos dados. Por isso, é necessária a aplicação de um método de correção do viés filogenético das amostras. Em segundo lugar, análises evolutivas comparativas que avaliam a contagem de milhares de elementos genômicos para buscar eventuais associações produz um aumento proibitivo dos erros do Tipo II (aceitar um falso-positivo como verdadeiro). Portanto, faz-se necessária também a correção para os testes de múltiplas hipóteses.

Por fim, é também importante a escolha de componentes genômicos capazes de refletir os fenômenos evolutivos esperados na evolução da eussocialidade, como o compartilhamento linhagem-específico dos elementos associados aos fenótipos da eussocialidade, a existência de componentes que permitem uma predisposição do surgimento deste fenótipo e as potenciais convergências evolutivas.

Neste trabalho fomos capazes de demonstrar, através do emprego de anotações genômicas de alta qualidade, bem como de uma metodologia estatística que emprega múltiplas correções essenciais para dados advindos de organismos filogeneticamente relacionados, que certos genes, famílias e domínios proteicos estão significativamente associados à presença da eussocialidade em Hymenoptera. Tais elementos genômicos estão associados com fenótipos importantes para insetos eussociais, como adaptações à vida nos ambientes hipóxicos dos ninhos de formigas e o armazenamento de energia em abelhas sociais, nas quais a alimentação é um importante sinal para a diferenciação de castas.

Nossos resultados também corroboram a hipótese de que muitos dos elementos genômicos responsáveis por fenótipos associados à evolução da eussocialidade parecem ser linhagem-específicos e, em alguns casos, representam convergências evolutivas, reforçando a ideia de que a eussocialidade é um fenótipo complexo e multifatorial e, apesar da possível existência de elementos que predispõem sua evolução em determinado clado, como os Hymenoptera, diferentes linhagens se tornaram eussociais por meio de adaptações evolutivas potencialmente distintas, embora compartilhadas entre algumas linhagens de maneira independente.

Finalmente, uma vez que os genes descritos como expandidos ou contraídos e associados com a evolução de Hymenoptera reportados nesse estudo apresentam ausência independente em pelo menos três espécies, os mesmos possivelmente

não caracterizam genes essenciais. Assim, acreditamos que os resultados apresentados neste trabalho proporcionam um importante ponto de partida ao sugerir alvos para o desenvolvimento de estudos funcionais por meio de *knockout* e *knockdown* que podem porventura elucidar o papel destes genes na evolução e manutenção da eussocialidade em Hymenoptera.

8 PERSPECTIVAS

Do ponto de vista bioinformático, gostaríamos ainda de avaliar associações entre as contagens de domínios nos genomas anotados de alta qualidade de Hymenoptera e diversos outros fenótipos de interesse que caracterizam a diversidade ecológica e comportamental destes insetos, como a presença de polinização, parasitismo e ferrão. Adicionalmente, pertinente aos resultados encontrados neste trabalho, gostaríamos de avaliar a expressão gênica em espécies sociais e solitárias dos genes anotados pelos domínios e famílias identificados como significativamente associados à eussocialidade, especialmente, caso dados estejam disponíveis, em diferentes castas e estágios de desenvolvimento. Temos também a pretensão de fazer estudos de sintenia desses genes para avaliar como se deram os potenciais ganhos e perdas evidenciados em nossos resultados.

No aspecto funcional, gostaríamos de, no futuro, realizar experimentos de *knockdown*, *knockout* e superexpressão destes genes com intenção de gerar um conhecimento robusto e respaldado acerca das características funcionais destes genes com respeito à eussocialidade em Hymenoptera.

REFERÊNCIAS

ABE, Y. et al. A Mitotic Kinase TOPK Enhances Cdk1/cyclin B1-dependent Phosphorylation of PRC1 and Promotes Cytokinesis. **Journal of Molecular Biology**, v. 370, n. 2, p. 231–245, 6 jul. 2007.

ANDERSON, M. The Evolution of Eusociality. **Annual Review of Ecology and Systematics**, v. 15, n. 1, p. 165–189, nov. 1984.

BAGOWSKI, C. P.; BRUINS, W.; TE VELTHUIS, A. J. W. The Nature of Protein Domain Evolution: Shaping the Interaction Network. **Current Genomics**, v. 11, n. 5, p. 368–376, 1 ago. 2010.

BÁNKI, O. et al. **Hymenoptera | COL**. Disponível em: <https://www.catalogueoflife.org/data/taxon/HYM>. Acesso em: 4 nov. 2022.

BARCHUK, A. R. et al. Molecular determinants of caste differentiation in the highly eusocial honeybee Apis mellifera. **BMC Developmental Biology**, v. 7, p. 70, 18 jun. 2007.

BARTH, J. M. I. et al. Disentangling structural genomic and behavioural barriers in a sea of connectivity. **Molecular Ecology**, v. 28, n. 6, p. 1394–1411, mar. 2019.

BATRA, S. Nests and Social Behavior of Halictine bees of India (Hymenoptera: Halictidae). **The Indian Journal of Entomology**, v. 28, p. 375–393, 1 set. 1966.

BAUDIER, K. M. Brood Stimulation Hypothesis. Em: STARR, C. (Ed.). **Encyclopedia of Social Insects**. Cham: Springer International Publishing, 2019. p. 1–4.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. **Journal of the Royal Statistical Society. Series B (Methodological)**, v. 57, n. 1, p. 289–300, 1995.

BEWICK, A. J. et al. Evolution of DNA Methylation across Insects. **Molecular Biology and Evolution**, v. 34, n. 3, p. 654–665, 1 mar. 2017.

BIANCHETTI, C. M.; BINGMAN, C. A.; PHILLIPS JR., G. N. Structure of the C-terminal heme-binding domain of THAP domain containing protein 4 from Homo sapiens. **Proteins: Structure, Function, and Bioinformatics**, v. 79, n. 4, p. 1337–

1341, 2011.

BICKER, G. Sources and targets of nitric oxide signalling in insect nervous systems. **Cell and Tissue Research**, v. 303, n. 2, p. 137–146, 1 fev. 2001.

BININDA-EMONDS, O. R. P. An Introduction to Supertree Construction (and Partitioned Phylogenetic Analyses) with a View Toward the Distinction Between Gene Trees and Species Trees. Em: GARAMSZEGI, L. Z. (Ed.). **Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology**. Berlin, Heidelberg: Springer, 2014. p. 49–76.

BLOCH, G.; HEFETZ, A.; HARTFELDER, K. Ecdysteroid titer, ovary status, and dominance in adult worker and queen bumble bees (Bombus terrestris). **Journal of Insect Physiology**, v. 46, n. 6, p. 1033–1040, 1 jun. 2000.

BOSSERT, S. et al. Combining transcriptomes and ultraconserved elements to illuminate the phylogeny of Apidae. **Molecular Phylogenetics and Evolution**, v. 130, p. 121–131, 1 jan. 2019.

BOURKE, A. F. G. **Principles of Social Evolution**. [s.l.] Oxford University Press, 2011.

BRADBEAR, N. **Bees and their role in forest livelihoods: a guide to the services provided by bees and the sustainable harvesting, processing and marketing of their products**. Rome: FAO, 2009.

BRANSTETTER, M. G. et al. Phylogenomic Insights into the Evolution of Stinging Wasps and the Origins of Ants and Bees. **Current Biology**, v. 27, n. 7, p. 1019–1025, 3 abr. 2017.

BURLEIGH, A. et al. A co-culture genome-wide RNAi screen with mammary epithelial cells reveals transmembrane signals required for growth and differentiation. **Breast cancer research: BCR**, v. 17, p. 4, 9 jan. 2015.

BURTON, G. J.; JAUNIAUX, E. Oxidative stress. **Best Practice & Research Clinical Obstetrics & Gynaecology**, Placental Bed & Maternal - Fetal Disorders. v. 25, n. 3, p. 287–299, 1 jun. 2011.

CARDINAL, S.; DANFORTH, B. N. The Antiquity and Evolutionary History of Social Behavior in Bees. **PLOS ONE**, v. 6, n. 6, p. e21086, 13 jun. 2011.

CASAMAYOR, A.; ARIÑO, J. Chapter Eight - Controlling Ser/Thr protein phosphatase PP1 activity and function through interaction with regulatory subunits. Em: KARABENCHEVA-CHRISTOVA, T.; CHRISTOV, C. (Eds.). **Advances in Protein Chemistry and Structural Biology**. Advances in Protein Chemistry and Structural Biology. [s.l.] Academic Press, 2020. v. 122p. 231–288.

CASTIGLIONE, S. et al. A new method for testing evolutionary rate variation and shifts in phenotypic evolution. **Methods in Ecology and Evolution**, v. 9, n. 4, p. 974–983, abr. 2018.

CASTIGLIONE, S. et al. Fast production of large, time-calibrated, informal supertrees with tree.merger. **Palaeontology**, v. 65, n. 1, jan. 2022.

CHEUNG, H. W. et al. Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. **Proceedings of the National Academy of Sciences of the United States of America**, v. 108, n. 30, p. 12372–12377, 26 jul. 2011.

CHHAKCHHUAK, L. et al. Complete mitochondrial genome of the Himalayan honey bee, Apis laboriosa. **Mitochondrial DNA Part A**, v. 27, n. 5, p. 3755–3756, 2 set. 2016.

COCK, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. **Bioinformatics**, v. 25, n. 11, p. 1422–1423, 1 jun. 2009.

CORNWELL, W.; NAKAGAWA, S. Phylogenetic comparative methods. **Current Biology**, v. 27, n. 9, p. R333–R336, 8 maio 2017.

COUTINHO, T. J. D.; FRANCO, G. R.; LOBO, F. P. Homology-Independent Metrics for Comparative Genomics. **Computational and Structural Biotechnology Journal**, v. 13, p. 352–357, 1 jan. 2015.

CRESPI, B. J.; YANEGA, D. The definition of eusociality. **Behavioral Ecology**, v. 6, n. 1, p. 109–115, 1995.

CRUZ-LANDIM, C. DA; PATRÍCIO, K.; ANTONIALLI JR, W. F. Cell death and ovarian development in highly eusocial bees (Hymenoptera, apidae): caste differentiation and worker egg laying. **Braz. j. morphol. sci**, p. 27–42, 2006.

DA SILVA, J. Life History and the Transitions to Eusociality in the Hymenoptera. **Frontiers in Ecology and Evolution**, v. 9, 2021.

DARWIN, C. **On the Origin of Species by Means of Natural Selection, or Preservation of Favoured Races in the Struggle for Life**. London, UK: John Murray, 1859.

DAS, B.; DE BEKKER, C. Time-course RNASeq of Camponotus floridanus forager and nurse ant brains indicate links between plasticity in the biological clock and behavioral division of labor. **BMC Genomics**, v. 23, n. 1, p. 57, 15 jan. 2022.

DATTA, S. et al. Cerebellar ataxia disease–associated Snx14 promotes lipid droplet growth at ER–droplet contacts. **Journal of Cell Biology**, v. 218, n. 4, p. 1335–1351, 14 fev. 2019.

DE PAULA FREITAS, F. C. et al. The nuclear and mitochondrial genomes of Frieseomelitta varia - a highly eusocial stingless bee (Meliponini) with a permanently sterile worker caste. **BMC genomics**, v. 21, n. 1, p. 386, 3 jun. 2020.

DE SIMONE, G. et al. Human nitrobindin: the first example of an all-β-barrel ferric heme-protein that catalyzes peroxynitrite detoxification. **FEBS Open Bio**, v. 8, n. 12, p. 2002–2010, 9 nov. 2018.

DOGANTZIS, K. A. et al. Insects with similar social complexity show convergent patterns of adaptive molecular evolution. **Scientific Reports**, v. 8, n. 1, p. 10388, 10 jul. 2018.

DUMAS, A. **O Conde de Monte Cristo**. [s.l.] Zahar, 2021.

ELSIK, C. G. et al. Finding the missing honey bee genes: lessons learned from a genome upgrade. **BMC Genomics**, v. 15, n. 1, p. 86, 30 jan. 2014.

EMBL-EBI. **THAP4-like, heme-binding beta-barrel domain (IPR014878) - InterPro entry - InterPro**. Disponível em: <https://www.ebi.ac.uk/interpro/entry/InterPro/IPR014878/>. Acesso em: 25 out. 2022.

ETO, M. Regulation of Cellular Protein Phosphatase-1 (PP1) by Phosphorylation of the CPI-17 Family, C-kinase-activated PP1 Inhibitors *. **Journal of Biological Chemistry**, v. 284, n. 51, p. 35273–35277, 18 dez. 2009.

FELSENSTEIN, J. Phylogenies and the Comparative Method. **The American Naturalist**, v. 125, n. 1, p. 1–15, jan. 1985.

FORBES, A. A. et al. Parasitoids, Hyperparasitoids, and Inquilines Associated With the Sexual and Asexual Generations of the Gall Former, Belonocnema treatae (Hymenoptera: Cynipidae). **Annals of the Entomological Society of America**, v. 109, n. 1, p. 49–63, 1 jan. 2016.

FORBES, A. A. et al. Quantifying the unquantifiable: why Hymenoptera, not Coleoptera, is the most speciose animal order. **BMC Ecology**, v. 18, p. 21, 12 jul. 2018.

FOSTER, R. L. et al. Reproductive physiology, dominance interactions, and division of labour among bumble bee workers. **Physiological Entomology**, v. 29, n. 4, p. 327–334, 2004.

FOWLER, P. C. et al. NeurodegenERation: The Central Role for ER Contacts in Neuronal Function and Axonopathy, Lessons From Hereditary Spastic Paraplegias and Related Diseases. **Frontiers in Neuroscience**, v. 13, 2019.

FOX, E. G. P.; BRESSAN-NASCIMENTO, S.; EIZEMBERG, R. Notes on the Biology and Behaviour of the Jewel Wasp, Ampulex compressa (Fabricius, 1781) (Hymenoptera; Ampulicidae), in the Laboratory, Including First Record of Gregarious Reproduction. **Entomological News**, v. 120, n. 4, p. 430–437, set. 2009.

GAL, C. et al. DREAM represses distinct targets by cooperating with different THAP domain proteins. **Cell Reports**, v. 37, n. 3, p. 109835, 19 out. 2021.

GALILI, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. **Bioinformatics**, v. 31, n. 22, p. 3718–3720, 15 nov. 2015.

GARLAND, T., Jr. et al. Phylogenetic Analysis of Covariance by Computer Simulation. **Systematic Biology**, v. 42, n. 3, p. 265–292, 1 set. 1993.

GASTON, K.; GAULD, I.; HANSON, P. The size and composition of the hymenopteran fauna of Costa Rica. **Journal of Biogeography**, v. 23, n. 1, p. 105–113, 1996.

GEERVLIET, J. B. F. et al. Learning to discriminate between infochemicals from different plant-host complexes by the parasitoids Cotesia glomerata and C. rubecula. **Entomologia Experimentalis et Applicata**, v. 86, n. 3, p. 241–252, 1998.

GHISBAIN, G. et al. Substantial genetic divergence and lack of recent gene flow support cryptic speciation in a colour polymorphic bumble bee (Bombus bifarius) species complex. **Systematic Entomology**, v. 45, n. 3, p. 635–652, 2020.

GONÇALVES, W. G. et al. Post-embryonic development of the Malpighian tubules in Apis mellifera (Hymenoptera) workers: morphology, remodeling, apoptosis, and cell proliferation. **Protoplasma**, v. 255, n. 2, p. 585–599, 1 mar. 2018.

GONG, J.-P. et al. Mouse brain localization of the protein kinase C-enhanced phosphatase 1 inhibitor KEPI (Kinase C-Enhanced PP1 Inhibitor). **Neuroscience**, v. 132, n. 3, p. 713–727, 1 jan. 2005.

GRIMALDI, D. A.; ENGEL, M. S. **Evolution of the insects**. Cambridge [U.K.] ; New York: Cambridge University Press, 2005.

GU, Z. Complex heatmap visualization. **iMeta**, v. 1, n. 3, set. 2022.

GU, Z.; EILS, R.; SCHLESNER, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. **Bioinformatics**, v. 32, n. 18, p. 2847–2849, 15 set. 2016.

HAMILTON, W. D. The genetical evolution of social behaviour. I. **Journal of Theoretical Biology**, v. 7, n. 1, p. 1–16, 1 jul. 1964.

HARDISON, R. C. Comparative Genomics. **PLOS Biology**, v. 1, n. 2, p. e58, 17 nov. 2003.

HARVEY, J. A. et al. Development of the solitary endoparasitoid Microplitis demolitor: host quality does not increase with host age and size. **Ecological Entomology**, v. 29, n. 1, p. 35–43, 2004.

HARVEY, J. A.; HARVEY, I. F.; THOMPSON, D. J. Lifetime Reproductive Success in the Solitary Endoparasitoid, Venturia canescens. **Journal of Insect Behavior**, v. 14, n. 5, p. 573–593, 1 set. 2001.

HEINRICH, B. Pheromone Induced Brooding Behavior in Bombus vosnesenskii and

B. edwardsii (Hymenoptera: Bombidae). **Journal of the Kansas Entomological Society**, v. 47, n. 3, p. 396–404, 1974.

HERBERS, J. M. Darwin's 'one special difficulty': celebrating Darwin 200. **Biology Letters**, v. 5, n. 2, p. 214–217, 23 abr. 2009.

HERZNER, G. et al. Larvae of the parasitoid wasp Ampulex compressa sanitize their host, the American cockroach, with a blend of antimicrobials. **Proceedings of the National Academy of Sciences**, v. 110, n. 4, p. 1369–1374, 22 jan. 2013.

HILLER, M. et al. A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species. **Cell reports**, v. 2, n. 4, p. 817–823, 25 out. 2012.

HÖLLDOBLER, B.; WILSON, E. O. **The ants**. Cambridge, Mass: Belknap Press of Harvard University Press, 1990.

HÖLLDOBLER, B.; WILSON, E. O. **The superorganism: the beauty, elegance, and strangeness of insect societies**. 1st ed ed. New York: W.W. Norton, 2009.

HONGO, J. A. et al. **CALANGO: an annotation-based, phylogeny-aware comparative genomics framework for exploring and interpreting complex genotypes and phenotypes**. [s.l: s.n.]. Disponível em: <https://www.biorxiv.org/content/10.1101/2021.08.25.457574v1>. Acesso em: 27 set. 2021.

HOTALING, S. et al. Long Reads Are Revolutionizing 20 Years of Insect Genome Sequencing. **Genome Biology and Evolution**, v. 13, n. 8, 1 ago. 2021.

HSIAO, C.; HSIAO, T. H. Insect hormones: Their effects on diapause and development of hymenoptera. **Life Sciences**, v. 8, n. 14, Part 2, p. 767–774, 15 jul. 1969.

HUNT, J. H. **The evolution of social wasps**. Oxford ; New York: Oxford University Press, 2007.

JANDT, J. M.; DORNHAUS, A. Spatial organization and division of labour in the bumblebee Bombus impatiens. **Animal Behaviour**, v. 77, n. 3, p. 641–651, 1 mar. 2009.

JANDT, J. M.; TIBBETTS, E. A.; TOTH, A. L. Polistes paper wasps: a model genus for the study of social dominance hierarchies. **Insectes Sociaux**, v. 61, n. 1, p. 11–27, 1 fev. 2014.

JEANNE, R. L.; FAGEN, R. Polymorphism in *Stelopolybia Areata* (Hymenoptera, Vespidae). **Psyche: A Journal of Entomology**, v. 81, n. 1, p. 155–166, 1 jan. 1974.

JIAN, Y. et al. Protein phosphatase 1 regulatory inhibitor subunit 14C promotes triple-negative breast cancer progression via sustaining inactive glycogen synthase kinase 3 beta. **Clinical and Translational Medicine**, v. 12, n. 1, p. e725, jan. 2022.

JONES, P. et al. InterProScan 5: genome-scale protein function classification. **Bioinformatics**, v. 30, n. 9, p. 1236–1240, 1 maio 2014.

JOSENS, R.; ESCHBACH, C.; GIURFA, M. Differential conditioning and long-term olfactory memory in individual Camponotus fellah ants. **Journal of Experimental Biology**, v. 212, n. 12, p. 1904–1911, 15 jun. 2009.

KAPHEIM, K. M. et al. Genomic signatures of evolutionary transitions from solitary to group living. **Science**, v. 348, n. 6239, p. 1139–1143, 5 jun. 2015.

KAPHEIM, K. M. Nutritional, endocrine, and social influences on reproductive physiology at the origins of social behavior. **Current Opinion in Insect Science**, Vectors and medical and veterinary entomology * Social insects. v. 22, p. 62–70, 1 ago. 2017.

KAPHEIM, K. M.; JOHNSON, M. M. Juvenile hormone, but not nutrition or social cues, affects reproductive maturation in solitary alkali bees (Nomia melanderi). **Journal of Experimental Biology**, v. 220, n. 20, p. 3794–3801, 15 out. 2017.

KARPE, S. D. et al. Computational genome-wide survey of odorant receptors from two solitary bees Dufourea novaeangliae (Hymenoptera: Halictidae) and Habropoda laboriosa (Hymenoptera: Apidae). **Scientific Reports**, v. 7, n. 1, p. 10823, 7 set. 2017.

KASSAMBARA, A. **ggpubr: “ggplot2” Based Publication Ready Plots**. [s.l: s.n.].

KELSTRUP, H. C. et al. Reproductive status, endocrine physiology and chemical signaling in the Neotropical, swarm-founding eusocial wasp Polybia micans. **Journal of Experimental Biology**, v. 217, n. 13, p. 2399–2410, 1 jul. 2014.

KING, R. C.; STANSFIELD, W. D.; MULLIGAN, P. K. **A dictionary of genetics**. 7th ed ed. Oxford ; New York: Oxford University Press, 2006.

KOCHER, S. D.; PAXTON, R. J. Comparative methods offer powerful insights into social evolution in bees. **Apidologie**, v. 45, n. 3, p. 289–305, 1 maio 2014.

KOLKMAN, J. A.; STEMMER, W. P. C. Directed evolution of proteins by exon shuffling. **Nature Biotechnology**, v. 19, n. 5, p. 423–428, maio 2001.

KRONAUER, D. J. C. et al. Non–nest mate discrimination and clonal colony structure in the parthenogenetic ant Cerapachys biroi. **Behavioral Ecology**, v. 24, n. 3, p. 617–622, 1 maio 2013.

KUCHARSKI, R. et al. Nutritional Control of Reproductive Status in Honeybees via DNA Methylation. **Science**, v. 319, n. 5871, p. 1827–1830, 28 mar. 2008.

KVERKOVÁ, K. et al. The evolution of brain neuron numbers in amniotes. **Proceedings of the National Academy of Sciences**, v. 119, n. 11, p. e2121624119, 15 mar. 2022.

LAWSON, S. P.; HELMREICH, S. L.; REHAN, S. M. Effects of nutritional deprivation on development and behavior in the subsocial bee Ceratina calcarata (Hymenoptera: Xylocopinae). **Journal of Experimental Biology**, v. 220, n. 23, p. 4456–4462, 1 dez. 2017.

LESIEUR, V. et al. Phylogeography of the Wheat Stem Sawfly, Cephus cinctus Norton (Hymenoptera: Cephidae): Implications for Pest Management. **PLOS ONE**, v. 11, n. 12, p. e0168370, 13 dez. 2016.

LI, F. et al. Insect genomes: progress and challenges. **Insect Molecular Biology**, v. 28, n. 6, p. 739–758, 2019.

LI, J. et al. Differential Protein Expression in Honeybee (Apis mellifera L.) Larvae: Underlying Caste Differentiation. **PLOS ONE**, v. 5, n. 10, p. e13455, 20 out. 2010.

LI-BYARLAY, H.; CLEARE, X. L. Current trends in the oxidative stress and ageing of social hymenopterans. Em: **Advances in Insect Physiology**. [s.l.] Elsevier, 2020. v. 59p. 43–69.

LIEBERT, A. E.; NONACS, P.; WAYNE, R. K. Solitary nesting and reproductive success in the paper wasp Polistes aurifer. **Behavioral Ecology and Sociobiology**, v. 57, n. 5, p. 445–456, 1 mar. 2005.

LINKSVAYER, T. A.; WADE, M. J. The Evolutionary Origin and Elaboration of Sociality in the Aculeate Hymenoptera: Maternal Effects, Sib-Social Effects, and Heterochrony. **The Quarterly Review of Biology**, v. 80, n. 3, p. 317–336, set. 2005.

LINNEN, C. R.; FARRELL, B. D. Comparison of Methods for Species-Tree Inference in the Sawfly Genus Neodiprion (Hymenoptera: Diprionidae). **Systematic Biology**, v. 57, n. 6, p. 876–890, 1 dez. 2008.

LINNEN, C. R.; SMITH, D. R. Recognition of Two Additional Pine-Feeding Neodiprion Species (Hymenoptera: Diprionidae) in the Eastern United States. **Proceedings of the Entomological Society of Washington**, v. 114, n. 4, p. 492–500, out. 2012.

LIU, L. et al. The Mechanisms of Social Immunity Against Fungal Infections in Eusocial Insects. **Toxins**, v. 11, n. 5, p. 244, maio 2019.

LOBB, B. et al. An assessment of genome annotation coverage across the bacterial tree of life. **Microbial Genomics**, v. 6, n. 3, p. e000341, 3 mar. 2020.

LONG, M. et al. The origin of new genes: glimpses from the young and old. **Nature Reviews Genetics**, v. 4, n. 11, p. 865–875, nov. 2003.

LOUZAO BOADO, Á. Investigating the role of testis-mitochondiral genes in Drosophila lethal (3) malignant brain tumor. 30 abr. 2019.

MA, Y. et al. Lamin B receptor plays a role in stimulating nuclear envelope production

and targeting membrane vesicles to chromatin during nuclear envelope assembly through direct interaction with importin β. **Journal of Cell Science**, v. 120, n. 3, p. 520–530, 1 fev. 2007.

MALM, T.; NYMAN, T. Phylogeny of the symphytan grade of Hymenoptera: new pieces into the old jigsaw(fly) puzzle. **Cladistics**, v. 31, n. 1, p. 1–17, 2015.

MANNI, M. et al. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. **Molecular Biology and Evolution**, n. msab199, 28 jul. 2021.

MARTINS, E. P.; HANSEN, T. F. Phylogenies and the Comparative Method: A General Approach to Incorporating Phylogenetic Information into the Analysis of Interspecific Data. **The American Naturalist**, 1 abr. 1997.

MARTINS, L. C. B. et al. Chemical composition of the intramandibular glands of the ant Neoponera villosa (Fabricius, 1804) (Hymenoptera: Ponerinae). **Chemoecology**, v. 25, n. 1, p. 25–31, 1 fev. 2015.

MCKENZIE, S. K.; KRONAUER, D. J. C. The genomic architecture and molecular evolution of ant odorant receptors. **Genome Research**, v. 28, n. 11, p. 1757–1765, nov. 2018.

MCLENNAN, A. G. The Nudix hydrolase superfamily. **Cellular and Molecular Life Sciences CMLS**, v. 63, n. 2, p. 123–143, 1 jan. 2006.

MERTL, A. L.; TRANIELLO, J. F. A. Behavioral evolution in the major worker subcaste of twig-nesting Pheidole (Hymenoptera: Formicidae): does morphological specialization influence task plasticity? **Behavioral Ecology and Sociobiology**, v. 63, n. 10, p. 1411–1426, 1 ago. 2009.

MICHENER, C. D. Comparative Social Behavior of Bees. **Annual Review of Entomology**, v. 14, n. 1, p. 299–342, jan. 1969.

MICHENER, C. D. **The bees of the world**. 2nd ed ed. Baltimore: Johns Hopkins University Press, 2007.

MICOLINO, R.; CRISTIANO, M. P.; CARDOSO, D. C. Population-Based Cytogenetic Banding Analysis and Phylogenetic Relationships of the Neotropical Fungus-Farming Ant Trachymyrmex holmgreni Wheeler, 1925. **Cytogenetic and Genome Research**, v. 159, n. 3, p. 151–161, 2019.

MILLER, W. et al. COMPARATIVE GENOMICS. **Annual Review of Genomics and Human Genetics**, v. 5, n. 1, p. 15–56, 22 set. 2004.

MISTRY, J. et al. Pfam: The protein families database in 2021. **Nucleic Acids Research**, v. 49, n. D1, p. D412–D419, 8 jan. 2021.

MOURA, A. P. et al. Selectivity evaluation of insecticides used to control tomato pests to Trichogramma pretiosum. **BioControl**, v. 51, n. 6, p. 769, 14 set. 2006.

MÜLLER, U. Inhibition of Nitric Oxide Synthase Impairs a Distinct Form of Long-Term Memory in the Honeybee, Apis mellifera. **Neuron**, v. 16, n. 3, p. 541–549, 1 mar. 1996.

MÜLLER, U. THE NITRIC OXIDE SYSTEM IN INSECTS. **Progress in Neurobiology**, v. 51, n. 3, p. 363–381, 1 fev. 1997.

MUMMERY-WIDMER, J. L. et al. Genome-wide analysis of Notch signalling in Drosophila by transgenic RNAi. **Nature**, v. 458, n. 7241, p. 987–992, abr. 2009.

NAGY, L. G. et al. Novel phylogenetic methods are needed for understanding gene function in the era of mega-scale genome sequencing. **Nucleic Acids Research**, v. 48, n. 5, p. 2209–2219, 18 mar. 2020.

NEELY, G. G. et al. A Genome-wide Drosophila Screen for Heat Nociception Identifies α2δ3 as an Evolutionarily Conserved Pain Gene. **Cell**, v. 143, n. 4, p. 628–638, 12 nov. 2010.

NEUWIRTH, E. **RColorBrewer: ColorBrewer Palettes**. [s.l: s.n.].

NIWA, R.; NIWA, Y. S. Enzymes for ecdysteroid biosynthesis: their biological functions in insects and beyond. **Bioscience, Biotechnology, and Biochemistry**, v. 78, n. 8, p. 1283–1292, 3 ago. 2014.

NOLL, F. B. et al. Marimbondos: systematics, biogeography, and evolution of social behaviour of neotropical swarm-founding wasps (Hymenoptera: Vespidae: Epiponini). **Cladistics**, v. 37, n. 4, p. 423–441, 2021.

NOWAK, M. A.; TARNITA, C. E.; WILSON, E. O. The evolution of eusociality. **Nature**, v. 466, n. 7310, p. 1057–1062, ago. 2010.

O'DONNELL, S.; REICHARDT, M.; FOSTER, R. Individual and colony factors in bumble bee division of labor (Bombus bifarius nearcticus Handl; Hymenoptera, Apidae). **Insectes Sociaux**, v. 47, n. 2, p. 164–170, 1 maio 2000.

OEYEN, J. P. et al. Sawfly Genomes Reveal Evolutionary Acquisitions That Fostered the Mega-Radiation of Parasitoid and Eusocial Hymenoptera. **Genome Biology and Evolution**, v. 12, n. 7, p. 1099–1188, 1 jul. 2020.

OISHI, K. et al. Genetics and biology of the sawfly,Athalia rosae (Hymenoptera). **Genetica**, v. 88, n. 2, p. 119–127, 1 jun. 1993.

OLDROYD, B. P.; YAGOUND, B. The role of epigenetics, particularly DNA methylation, in the evolution of caste in insect societies. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 376, n. 1826, p. 20200115, 7 jun. 2021.

OLZMANN, J. A.; CARVALHO, P. Dynamics and functions of lipid droplets. **Nature reviews. Molecular cell biology**, v. 20, n. 3, p. 137–155, mar. 2019.

ORTIZ-CARREON, F. R. et al. Herbivore-Induced Volatiles from Maize Plants Attract Chelonus insularis, an Egg-Larval Parasitoid of the Fall Armyworm. **Journal of Chemical Ecology**, v. 45, n. 3, p. 326–337, 1 mar. 2019.

PACHER, P.; BECKMAN, J. S.; LIAUDET, L. Nitric Oxide and Peroxynitrite in Health and Disease. **Physiological Reviews**, v. 87, n. 1, p. 315–424, jan. 2007.

PARADIS, E.; SCHLIEP, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. **Bioinformatics**, v. 35, n. 3, p. 526–528, 1 fev. 2019.

PARK, J.-H. et al. Critical roles of T-LAK cell-originated protein kinase in cytokinesis. **Cancer Science**, v. 101, n. 2, p. 403–411, 2010.

PATEL, A. et al. The Making of a Queen: TOR Pathway Is a Key Player in Diphenic Caste Development. **PLOS ONE**, v. 2, n. 6, p. e509, 6 jun. 2007.

PATTHY, L. Genome evolution and the evolution of exon-shuffling — a review. **Gene**, v. 238, n. 1, p. 103–114, 30 set. 1999.

PAULA JUNIOR, D. E. DE; BARCHUK, A. R. **Aspectos morfológicos e moleculares do desenvolvimento cerebral casta-específico em Apis mellifera**. Alfenas, Minas Gerais, Brasil: Universidade Federal de Alfenas, 30 out. 2019.

PENNELL, M. W. et al. geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. **Bioinformatics (Oxford, England)**, v. 30, n. 15, p. 2216–2218, 1 ago. 2014.

PÉREZ-LACHAUD, G.; BATCHELOR, T. P.; HARDY, I. C. W. Wasp eat wasp: facultative hyperparasitism and intra-guild predation by bethylid wasps. **Biological Control**, v. 30, n. 2, p. 149–155, 1 jun. 2004.

PETERS, R. S. et al. Evolutionary History of the Hymenoptera. **Current Biology**, v. 27, n. 7, p. 1013–1018, 3 abr. 2017.

PETERSON, R. K. D. et al. Parasitism and the demography of wheat stem sawfly larvae, Cephus cinctus. **BioControl**, v. 56, n. 6, p. 831–839, 1 dez. 2011.

PIEKARSKI, P. K. et al. Phylogenomic Evidence Overturns Current Conceptions of Social Evolution in Wasps (Vespidae). **Molecular Biology and Evolution**, v. 35, n. 9, p. 2097–2109, 1 set. 2018.

PINHEIRO, J.; BATES, D.; R CORE TEAM. **nlme: Linear and Nonlinear Mixed Effects Models**. [s.l: s.n.].

PIQUERET, B.; SANDOZ, J.-C.; D'ETTORRE, P. Ants learn fast and do not forget: associative olfactory learning, memory and extinction in Formica fusca. **Royal Society Open Science**, v. 6, n. 6, p. 190778, 19 jun. 2019.

PROVECHO, Y.; JOSENS, R. Olfactory memory established during trophallaxis

affects food search behaviour in ants. **Journal of Experimental Biology**, v. 212, n. 20, p. 3221–3227, 15 out. 2009.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, AustriaR Foundation for Statistical Computing, , 2021. Disponível em: <https://www.R-project.org/>

RAMSAY, C.; LASKO, P.; ABOUHEIF, E. Evo-Devo Lessons from the Reproductive Division of Labor in Eusocial Hymenoptera. Em: NUÑO DE LA ROSA, L.; MÜLLER, G. B. (Eds.). **Evolutionary Developmental Biology: A Reference Guide**. Cham: Springer International Publishing, 2021. p. 791–804.

RAVARY, F.; JAISSON, P. The reproductive cycle of thelytokous colonies of Cerapachys biroi Forel (Formicidae, Cerapachyinae). **Insectes Sociaux**, v. 49, n. 2, p. 114–119, 1 maio 2002.

RAVARY, F.; JAISSON, P. Absence of individual sterility in thelytokous colonies of the ant Cerapachys biroi Forel (Formicidae, Cerapachyinae). **Insectes Sociaux**, v. 51, n. 1, p. 67–73, 1 fev. 2004.

REHAN, S. M.; RICHARDS, M. H. Nesting biology and subsociality in Ceratina calcarata (Hymenoptera: Apidae). **The Canadian Entomologist**, v. 142, n. 1, p. 65–74, fev. 2010.

REVELL, L. J. phytools: an R package for phylogenetic comparative biology (and other things): *phytools: R package*. **Methods in Ecology and Evolution**, v. 3, n. 2, p. 217–223, abr. 2012.

RICHARDS, O. W. The Biology of the Social Wasps (hymenoptera, Vespidae). **Biological Reviews**, v. 46, n. 4, p. 483–528, 1971.

ROBERTSON, H. M.; GADAU, J.; WANNER, K. W. The insect chemoreceptor superfamily of the parasitoid jewel wasp Nasonia vitripennis. **Insect Molecular Biology**, v. 19, n. s1, p. 121–136, 2010.

ROBINSON, G. E. Chapter 77 - Division of Labor in Insect Societies. Em: RESH, V. H.; CARDÉ, R. T. (Eds.). **Encyclopedia of Insects (Second Edition)**. San Diego: Academic Press, 2009. p. 297–299.

ROMERO, P. R. et al. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. **Proceedings of the National Academy of Sciences**, v. 103, n. 22, p. 8390–8395, 30 maio 2006.

ROMIGUIER, J. et al. Ant phylogenomics reveals a natural selection hotspot preceding the origin of complex eusociality. **Current Biology**, v. 32, n. 13, p. 2942-2947.e4, 11 jul. 2022.

ROUSSE, P.; HARRIS, E.; QUILICI, S. Fopius arisanus, an egg-pupal parasitoid of Tephritidae. Overview. **Biocontrol News and Information**, v. 26, n. 2, p. 59–69, 2005.

RUNYON, J. B. et al. Parasitism of the Wheat Stem Sawfly (Hymenoptera: Cephidae) by Bracon cephi and B. lissogaster (Hymenoptera: Braconidae) in Wheat Fields Bordering Tilled and Untilled Fallow in Montana. **Journal of Economic Entomology**, v. 95, n. 6, p. 1130–1134, 1 dez. 2002.

SANTOS, B. F. et al. Phylogeny and historical biogeography of the paper wasp genus Polistes (Hymenoptera: Vespidae): implications for the overwintering hypothesis of social evolution. **Cladistics**, v. 31, n. 5, p. 535–549, 2015.

SANTOS JÚNIOR, J. E. et al. Biogeography and Diversification of Bumblebees (Hymenoptera: Apidae), with Emphasis on Neotropical Species. **Diversity**, v. 14, n. 4, p. 238, abr. 2022.

SAVERSCHEK, N. et al. Avoiding plants unsuitable for the symbiotic fungus: learning and long-term memory in leaf-cutting ants. **Animal Behaviour**, v. 79, n. 3, p. 689–698, 1 mar. 2010.

SCHMIDT, C. Molecular phylogenetics of ponerine ants (Hymenoptera: Formicidae: Ponerinae). **Zootaxa**, v. 3647, n. 2, p. 201, 9 maio 2013.

SCHMIDT, S. et al. Chemical detoxification vs mechanical removal of host plant toxins in Eucalyptus feeding sawfly larvae (Hymenoptera: Pergidae). **Journal of Insect Physiology**, v. 56, n. 12, p. 1770–1776, 1 dez. 2010.

SCHNORRER, F. et al. Systematic genetic analysis of muscle morphogenesis and function in Drosophila. **Nature**, v. 464, n. 7286, p. 287–291, mar. 2010.

SCHRICK, K. et al. FACKEL is a sterol C-14 reductase required for organized cell division and expansion in Arabidopsis embryogenesis. **Genes & Development**, v. 14, n. 12, p. 1471–1484, 15 jun. 2000.

SEEHUUS, S.-C.; KREKLING, T.; AMDAM, G. V. Cellular senescence in honey bee brain is largely independent of chronological age. **Experimental Gerontology**, v. 41, n. 11, p. 1117–1125, 1 nov. 2006.

SHARANOWSKI, B. J. et al. Phylogenomics of Ichneumonoidea (Hymenoptera) and implications for evolution of mode of parasitism and viral endogenization. **Molecular Phylogenetics and Evolution**, v. 156, p. 107023, 1 mar. 2021.

SHARKEY, M. J. Phylogeny and Classification of Hymenoptera. **Zootaxa**, v. 1668, n. 1, p. 521–548, 21 dez. 2007.

SHELL, W. A. et al. Sociality sculpts similar patterns of molecular evolution in two independently evolved lineages of eusocial bees. **Communications Biology**, v. 4, n. 1, p. 1–9, 26 fev. 2021.

SHELL, W. A.; REHAN, S. M. Behavioral and genetic mechanisms of social evolution: insights from incipiently and facultatively social bees. **Apidologie**, v. 49, n. 1, p. 13–30, 1 fev. 2018.

SHI, M.; CHEN, X. X.; VAN ACHTERBERG, C. Phylogenetic relationships among the Braconidae (Hymenoptera: Ichneumonoidea) inferred from partial 16S rDNA, 28S rDNA D2, 18S rDNA gene sequences and morphological characters. **Molecular Phylogenetics and Evolution**, v. 37, n. 1, p. 104–116, 1 out. 2005.

SIMOLA, D. F. et al. Epigenetic (re)programming of caste-specific behavior in the ant Camponotus floridanus. **Science**, v. 351, n. 6268, p. aac6633, 1 jan. 2016.

SIMPSON, J. C. et al. Genome-wide RNAi screening identifies human proteins with a regulatory function in the early secretory pathway. **Nature Cell Biology**, v. 14, n. 7, p. 764–774, jul. 2012.

SMITH, A. R. et al. Socially induced brain development in a facultatively eusocial sweat bee Megalopta genalis (Halictidae). **Proceedings of the Royal Society B: Biological Sciences**, v. 277, n. 1691, p. 2157–2163, 22 jul. 2010.

SNODGRASS, R. E. **Anatomy and physiology of the honeybee**. [s.l.] McGraw-Hill Book Company, Incorporated, 1925. v. 20

SOMAVILLA, A. et al. Total-Evidence Phylogeny of the New World Polistes Lepeletier, 1836, Paper Wasps (Vespidae, Polistinae, Polistini). **American Museum Novitates**, v. 2021, n. 3973, p. 1–42, jul. 2021.

SOPKO, R. et al. Combining Genetic Perturbations and Proteomics to Examine Kinase-Phosphatase Networks in Drosophila Embryos. **Developmental Cell**, v. 31, n. 1, p. 114–127, 13 out. 2014.

STAMP, N. E.; BOWERS, M. D. Direct and indirect effects of predatory wasps (Polistes sp.: Vespidae) on gregarious caterpillars (Hemileuca lucina: Saturniidae). **Oecologia**, v. 75, n. 4, p. 619–624, 1 maio 1988.

STELINSKI, L. L.; LIBURD, O. E. Behavioral evidence for host fidelity among populations of the parasitic wasp, Diachasma alloeum (Muesebeck). **Naturwissenschaften**, v. 92, n. 2, p. 65–68, 1 fev. 2005.

SUBRAMANIAN, G. et al. Lamin B Receptor Regulates the Growth and Maturation of Myeloid Progenitors via its Sterol Reductase Domain: Implications for Cholesterol Biosynthesis in Regulating Myelopoiesis. **The Journal of Immunology**, v. 188, n. 1, p. 85–102, 1 jan. 2012.

SUH, J. M. et al. An RGS-Containing Sorting Nexin Controls Drosophila Lifespan. **PLOS ONE**, v. 3, n. 5, p. e2152, 14 maio 2008.

SUHONEN, J. et al. Brood parasitism in eusocial insects (Hymenoptera): role of host geographical range size and phylogeny. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 374, n. 1769, p. 20180203, abr. 2019.

SUZZONI, J. P.; PASSERA, L.; STRAMBI, A. Ecdysteroid titre and caste determination in the ant,Pheidole pallidula (Nyl.) (Hymenoptera: Formicidae).

Experientia, v. 36, n. 10, p. 1228–1229, 1 out. 1980.

SYMONDS, M. R. E.; BLOMBERG, S. P. A Primer on Phylogenetic Generalised Least Squares. Em: GARAMSZEGI, L. Z. (Ed.). **Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology**. Berlin, Heidelberg: Springer, 2014. p. 105–130.

SZABÓ, C.; ISCHIROPOULOS, H.; RADI, R. Peroxynitrite: biochemistry, pathophysiology and development of therapeutics. **Nature Reviews Drug Discovery**, v. 6, n. 8, p. 662–680, ago. 2007.

TATUSOV, R. L.; KOONIN, E. V.; LIPMAN, D. J. A Genomic Perspective on Protein Families. **Science**, v. 278, n. 5338, p. 631–637, 24 out. 1997.

TERBOT II, J. **THE SOCIAL BEHAVIOR OF PINE SAWFLIES IN THE GENUS NEODIPRION**. [s.l.] University of Kentucky Libraries, 2021.

THE HONEYBEE GENOME SEQUENCING CONSORTIUM. Insights into social insects from the genome of the honeybee Apis mellifera. **Nature**, v. 443, n. 7114, p. 931–949, 26 out. 2006.

THE INTERPRO CONSORTIUM et al. InterPro: An integrated documentation resource for protein families, domains and functional sites. **Briefings in Bioinformatics**, v. 3, n. 3, p. 225–235, 1 set. 2002.

THE PANDAS DEVELOPMENT TEAM. **pandas-dev/pandas: Pandas**. Zenodo, , 19 set. 2022. Disponível em: <https://zenodo.org/record/3509134>. Acesso em: 10 out. 2022

TORRES, V. DE O.; GIANOTTI, E.; ANTONIALLI-JR, W. F. Temporal Polyethism and Life Expectancy of Workers in the Eusocial Wasp Polistes canadensis canadensis Linnaeus (Hymenoptera: Vespidae). **Sociobiology**, v. 60, n. 1, p. 107–113, 27 mar. 2013.

TOTH, A. L.; ROBINSON, G. E. Evo-devo and the evolution of social behavior. **Trends in Genetics**, v. 23, n. 7, p. 334–341, 1 jul. 2007.

TRUJILLO, M.; FERRER-SUETA, G.; RADI, R. Peroxynitrite Detoxification and Its Biologic Implications. **Antioxidants & Redox Signaling**, v. 10, n. 9, p. 1607–1620, set. 2008.

TSAI, P.-L. et al. The Lamin B receptor is essential for cholesterol synthesis and perturbed by disease-causing mutations. **eLife**, v. 5, p. e16011, 1 jun. 2016.

TSUJI, K.; YAMAUCHI, K. Production of females by parthenogenesis in the ant,Cerapachys biroi. **Insectes Sociaux**, v. 42, n. 3, p. 333–336, 1 set. 1995.

URBANEK, S.; HORNER, J. **Cairo: R Graphics Device using Cairo Graphics Library for Creating High-Quality Bitmap (PNG, JPEG, TIFF), Vector (PDF, SVG, PostScript) and Display (X11 and Win32) Output**. [s.l: s.n.].

VAN EECKHOVEN, J.; DUNCAN, E. J. Mating status and the evolution of eusociality: Oogenesis is independent of mating status in the solitary bee Osmia bicornis. **Journal of Insect Physiology**, v. 121, p. 104003, 1 fev. 2020.

VIZEACOUMAR, F. J. et al. A negative genetic interaction map in isogenic cancer cell lines reveals cancer cell vulnerabilities. **Molecular Systems Biology**, v. 9, p. 696, 8 out. 2013.

VOGEL, C.; CHOTHIA, C. Protein Family Expansions and Biological Complexity. **PLOS Computational Biology**, v. 2, n. 5, p. e48, 26 maio 2006.

WANG, A. R. et al. Comparative description of mitochondrial genomes of the honey bee Apis (Hymenoptera: Apidae): four new genome sequences and Apis phylogeny using whole genomes and individual genes. **Journal of Apicultural Research**, v. 57, n. 4, p. 484–503, 8 ago. 2018.

WANG, L.-G. et al. Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. **Molecular Biology and Evolution**, v. 37, n. 2, p. 599–603, 1 fev. 2020.

WARD, P. S. et al. The evolution of myrmicine ants: phylogeny and biogeography of a hyperdiverse ant clade (Hymenoptera: Formicidae). **Systematic Entomology**, v. 40, n. 1, p. 61–81, 2015.

WARNER, M. R. et al. Convergent eusocial evolution is based on a shared reproductive groundplan plus lineage-specific plastic genes. **Nature Communications**, v. 10, n. 1, p. 2651, 14 jun. 2019.

WARNER, N. et al. A Genome-wide Small Interfering RNA (siRNA) Screen Reveals Nuclear Factor-κB (NF-κB)-independent Regulators of NOD2-induced Interleukin-8 (IL-8) Secretion. **The Journal of Biological Chemistry**, v. 289, n. 41, p. 28213–28224, 10 out. 2014.

WATERHAM, H. R. Defects of cholesterol biosynthesis. **FEBS Letters**, Lipidome and Disease. v. 580, n. 23, p. 5442–5449, 9 out. 2006.

WEIBLEN, G. D.; BUSH, G. L. Speciation in fig pollinators and parasites. **Molecular Ecology**, v. 11, n. 8, p. 1573–1578, 2002.

WELTE, M. A.; GOULD, A. P. Lipid droplet functions beyond energy storage. **Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids**, Recent Advances in Lipid Droplet Biology. v. 1862, n. 10, Part B, p. 1260–1272, 1 out. 2017.

WEST-EBERHARD, M. J. The Evolution of Social Behavior by Kin Selection. **The Quarterly Review of Biology**, v. 50, n. 1, p. 1–33, mar. 1975.

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. [s.l.] Springer-Verlag New York, 2016.

WICKHAM, H. et al. **dplyr: A Grammar of Data Manipulation**. [s.l: s.n.].

WICKHAM, H. **forcats: Tools for Working with Categorical Variables (Factors)**. [s.l: s.n.].

WIEMER, A. P. et al. Functional morphology and wasp pollination of two South American asclepiads (Asclepiadoideae–Apocynaceae). **Annals of Botany**, v. 109, n. 1, p. 77–93, 1 jan. 2012.

WILKE, C. O. **cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2"**. [s.l: s.n.].

WILSON, E. O. **The insect societies**. Cambridge, Mass: Belknap Press of Harvard University Press, 1971.

YANDELL, M.; ENCE, D. A beginner's guide to eukaryotic genome annotation. **Nature Reviews Genetics**, v. 13, n. 5, p. 329–342, maio 2012.

YU, G. et al. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. **Methods in Ecology and Evolution**, v. 8, n. 1, p. 28–36, 2017.

ZHAO, Y.-H. et al. Nesting biology of Colletes gigas Cockerell (Hymenoptera: Colletidae). **Acta Entomologica Sinica**, v. 53, n. 11, p. 1287, 2010.

APÊNDICE A - Tabela contendo as espécies de Hymenoptera com genomas anotados recuperados da plataforma *Genome*.

Espécie	**Código**	**Acesso *Genome***	**Grande grupo**	**Alta qualidade**
Acromyrmex charruanus	Acha	GCA_017607545	Formigas	Não
Acromyrmex echinatior	Aech	GCF_000204515	Formigas	Sim
Acromyrmex heyeri	Ahey	GCA_017607565	Formigas	Não
Acromyrmex insinuator	Ains	GCA_017607455	Formigas	Não
Ampulex compressa	Acom	GCA_019049445	Vespas	Sim
Aphidius gifuensis	Agif	GCF_014905175	Vespas	Não
Apis cerana	Acer	GCF_001442555	Abelhas	Não
Apis dorsata	Ador	GCF_000469605	Abelhas	Não
Apis florea	Aflo	GCF_000184785	Abelhas	Sim
Apis laboriosa	Alab	GCF_014066325	Abelhas	Sim
Apis mellifera	Amel	GCF_003254395	Abelhas	Sim
Athalia rosae	Aros	GCF_917208135	Moscas-serra	Sim
Atta cephalotes	Acep	GCF_000143395	Formigas	Sim
Atta colombica	Acol	GCF_001594045	Formigas	Sim
Belonocnema kinseyi	Bkin	GCF_010883055	Vespas	Sim
Bombus bifarius	Bbif	GCF_011952205	Abelhas	Sim
Bombus impatiens	Bimp	GCF_000188095	Abelhas	Sim
Bombus pyrosoma	Bpyr	GCF_014825855	Abelhas	Sim
Bombus terrestris	Bter	GCF_910591885	Abelhas	Sim
Bombus vancouverensis nearticus	Bvan	GCF_011952275	Abelhas	Sim
Bombus vosnesenskii	Bvos	GCF_011952255	Abelhas	Sim
Camponotus floridanus	Cflo	GCF_003227725	Formigas	Sim

Espécie	**Código**	**Acesso *Genome***	**Grande grupo**	**Alta qualidade**
Cephus cinctus	Ccin	GCF_000341935	Moscas-serra	Sim
Ceratina calcarata	Ccal	GCF_001652005	Abelhas	Sim
Ceratosolen solmsi marchali	Csol	GCF_000503995	Vespas	Não
Chelonus insularis	Cins	GCF_013357705	Vespas	Sim
Colletes gigas	Cgig	GCF_013123115	Abelhas	Sim
Copidosoma floridanum	Cofl	GCF_000648655	Vespas	Não
Cotesia congregata	Ccon	GCA_905319865	Vespas	Não
Cotesia glomerata	Cglo	GCF_020080835	Vespas	Sim
Cyphomyrmex costatus	Ccos	GCF_001594065	Formigas	Sim
Diachasma alloeum	Dall	GCF_001412515	Vespas	Sim
Dinoponera quadriceps	Dqua	GCF_001313825	Formigas	Sim
Diprion similis	Dsim	GCF_021155765	Moscas-serra	Sim
Dufourea novaeangliae	Dnov	GCF_001272555	Abelhas	Sim
Eciton burchellii	Ebur	GCA_020341155	Formigas	Sim
Eufriesea mexicana	Emex	GCF_001483705	Abelhas	Sim
Fopius arisanus	Fari	GCF_000806365	Vespas	Sim
Formica exsecta	Fexs	GCF_003651465	Formigas	Sim
Frieseomelitta varia	Fvar	GCF_011392965	Abelhas	Sim
Habropoda laboriosa	Hlab	GCF_001263275	Abelhas	Não
Harpegnathos saltator	Hsal	GCF_003227715	Formigas	Sim
Heterotrigona itama	Hita	GCA_903986555	Abelhas	Não
Lasius niger	Lnig	GCA_001045655	Formigas	Não
Linepithema humile	Lhum	GCF_000217595	Formigas	Sim
Megachile rotundata	Mrot	GCF_000220905	Abelhas	Sim

Espécie	**Código**	**Acesso *Genome***	**Grande grupo**	**Alta qualidade**
Megalopta genalis	Mgen	GCF_011865705	Abelhas	Sim
Melipona quadrifasciata	Mqua	GCA_001276565	Abelhas	Não
Microplitis demolitor	Mdem	GCF_000572035	Vespas	Sim
Mischocyttarus mexicanus	Mmex	GCA_023678845	Vespas	Não
Monomorium pharaonis	Mpha	GCF_013373865	Formigas	Sim
Nasonia vitripennis	Nvit	GCF_009193385	Vespas	Sim
Neodiprion fabricii	Nfab	GCF_021155785	Moscas-serra	Sim
Neodiprion lecontei	Nlec	GCF_021901455	Moscas-serra	Sim
Neodiprion pinetum	Npin	GCF_021155775	Moscas-serra	Sim
Neodiprion virginianus	Nvir	GCF_021901495	Moscas-serra	Sim
Nomia melanderi	Nmel	GCF_003710045	Abelhas	Sim
Nylanderia fulva	Nful	GCF_005281655	Formigas	Sim
Odontomachus brunneus	Obru	GCF_010583005	Formigas	Sim
Ooceraea biroi	Obir	GCF_003672135	Formigas	Sim
Orussus abietinus	Oabi	GCF_000612105	Moscas-serra	Sim
Osmia bicornis bicornis	Obic	GCF_907164935	Abelhas	Sim
Pogonomyrmex barbatus	Pbar	GCF_000187915	Formigas	Sim
Polistes canadensis	Pcan	GCF_001313835	Vespas	Sim
Polistes dominula	Pdom	GCF_001465965	Vespas	Sim
Polistes exclamans	Pexc	GCA_023678865	Vespas	Sim
Polistes fuscatus	Pfus	GCF_010416935	Vespas	Sim
Pseudoatta argentina	Parg	GCA_017607525	Formigas	Não

Espécie	**Código**	**Acesso *Genome***	**Grande grupo**	**Alta qualidade**
Pseudomyrmex gracilis	Pgra	GCF_002006095	Formigas	Sim
Solenopsis invicta	Sinv	GCF_016802725	Formigas	Sim
Temnothorax curvispinosus	Tcur	GCF_003070985	Formigas	Não
Temnothorax longispinosus	Tlon	GCA_004794745	Formigas	Não
Trachymyrmex cornetzi	Tcor	GCF_001594075	Formigas	Sim
Trachymyrmex septentrionalis	Tsep	GCF_001594115	Formigas	Sim
Trachymyrmex zeteki	Tzet	GCF_001594055	Formigas	Sim
Trichogramma pretiosum	Tpre	GCF_000599845	Vespas	Sim
Trichomalopsis sarcophagae	Tsar	GCA_002249905	Vespas	Não
Venturia canescens	Vcan	GCF_019457755	Vespas	Sim
Vespa crabro	Vcra	GCF_910589235	Vespas	Sim
Vespa velutina	Vvel	GCF_912470025	Vespas	Sim
Vespula germanica	Vger	GCA_014466195	Vespas	Não
Vesupula pensylvanica	Vpen	GCF_014466175	Vespas	Sim
Vespula vulgaris	Vvul	GCA_014466185	Vespas	Não
Vollenhovia emeryi	Veme	GCF_000949405	Formigas	Sim
Wasmannia auropunctata	Waur	GCF_000956235	Formigas	Sim

Todas as espécies tiveram a completude de anotação avaliada por BUSCO. O nome de cada espécie foi abreviado para um código não ambíguo de 4 letras. Apenas espécies com alta qualidade de anotação (“Sim”) foram utilizadas nas etapas seguintes do trabalho.

APÊNDICE B - Tabela contendo a classificação de cada espécie de Hymenoptera de acordo com seu nível de socialidade

Código	Socialidade	Observação	Referência
Aech	Eussocial	-	Hölldobler e Wilson, 1990
Acom	Solitário	Parasitóide de baratas	Fox, Bressan-Nascimento e Eizemberg, 2009
Aflo	Eussocial	-	Kapheim et al., 2015
Alab	Eussocial	-	Michener, 2007
Amel	Eussocial	-	Kapheim et al., 2015
Aros	Solitário	-	Hunt, 2007
Acep	Eussocial	-	Hölldobler e Wilson, 1990
Acol	Eussocial	-	Hölldobler e Wilson, 1990
Bkin	Solitário	Formadora de galhas em plantas	Forbes et al., 2016
Bbif	Eussocial	Eussocial “simples”	Foster et al., 2004
Bimp	Eussocial	Eussocial “simples”	Jandt e Dornhaus, 2009
Bpyr	Eussocial	Eussocial “simples”	Michener, 2007
Bter	Eussocial	Eussocial “simples”	Kapheim et al., 2015
Bvan	Eussocial	Eussocial “simples”	O’Donnell, Reichardt e Foster, 2000
Bvos	Eussocial	Eussocial “simples”	Heinrich, 1974
Cflo	Eussocial	-	Hölldobler e Wilson, 1990
Ccin	Solitário	Ciclo de desenvolvimento muito similar ao de parasitóides	Peterson et al., 2011
Ccal	Intermediário	-	Rehan e Richards, 2010

Código	Socialidade	Observação	Referência
Cins	Solitário	Parasitóide de Lepidoptera	Ortiz-Carreon et al., 2019
Cgig	Solitário	-	Zhao et al., 2010
Cglo	Solitário	Parasitóide de Lepidoptera	Geervilet et al., 1998
Ccos	Eussocial	-	Hölldobler e Wilson, 1990
Dall	Solitário	Parasitóide de Tephritidae	Stelinski e Liburd, 2005
Dqua	Eussocial	-	Hölldobler e Wilson, 1990
Dsim	Solitário	-	Hunt, 2007
Dnov	Solitário	-	Kapheim et al., 2015
Ebur	Eussocial	Formiga de correição	Hölldobler e Wilson, 1990
Emex	Intermediário	-	Kapheim et al., 2015
Fari	Solitário	Parasitóide de Tephritidae	Rousse, Harris e Quilici, 2005
Fexs	Eussocial	-	Hölldobler e Wilson, 1990
Fvar	Eussocial	-	De Paula Freitas et al., 2020
Hsal	Eussocial	-	Hölldobler e Wilson, 1990
Lhum	Eussocial	-	Hölldobler e Wilson, 1990
Mrot	Solitário	-	Kapheim et al., 2015
Mgen	Eussocial	Abelha eussocial facultativa	Smith et al., 2010
Mdem	Solitário	Parasitóide de Noctuidae (Lepidoptera)	Harvey et al., 2004
Mpha	Eussocial	-	Hölldobler e Wilson, 1990

Código	Socialidade	Observação	Referência
Nvit	Solitário	Parasitóide de Cyclorrhapha	Robertson, Gaudau e Wanner, 2010
Nfab	Solitário	-	Hunt, 2007
Nlec	Solitário	-	Hunt, 2007
Npin	Solitário	-	Hunt, 2007
Nvir	Solitário	-	Hunt, 2007
Nmel	Intermediário	-	Kapheim e Johnson, 2017
Nful	Eussocial	-	Hölldobler e Wilson, 1990
Obru	Eussocial	-	Hölldobler e Wilson, 1990
Obir	Eussocial	Formiga de correição	Hölldobler e Wilson, 1990
Oabi	Solitário	Ectoparasitóide de larvas xilófagas	Oeyen et al., 2020
Obic	Solitário	-	Van Eeckhoven e Duncan, 2020
Pbar	Eussocial	-	Hölldobler e Wilson, 1990
Pcan	Eussocial	Eussocial “simples”	Hunt, 2007
Pdom	Eussocial	Eussocial “simples”	Hunt, 2007
Pexc	Eussocial	Eussocial “simples”	Hunt, 2007
Pfus	Eussocial	Eussocial “simples”	Hunt, 2007
Pgra	Eussocial	-	Hölldobler e Wilson, 1990
Sinv	Eussocial	-	Hölldobler e Wilson, 1990
Tcor	Eussocial	-	Hölldobler e Wilson, 1990
Tsep	Eussocial	-	Hölldobler e Wilson, 1990

Código	Socialidade	Observação	Referência
Tzet	Eussocial	-	Hölldobler e Wilson, 1990
Tpre	Solitário	Parasitóide de Lepidoptera	Moura et al., 2006
Vcan	Solitário	Parasitóide de Lepidoptera	Harvey, Harvey e Thompson, 2001
Vcra	Eussocial	-	Hunt, 2007
Vvel	Eussocial	-	Hunt, 2007
Vpen	Eussocial	-	Hunt, 2007
Veme	Eussocial	-	Hölldobler e Wilson, 1990
Waur	Eussocial	-	Hölldobler e Wilson, 1990

Classificação baseada em extensa revisão literária. Observações pontuam espécies com traços característicos notáveis relacionados a seu nível de socialidade. Apenas espécies que puderam ser classificadas como “Eussocial” ou “Solitário” foram utilizadas nas etapas seguintes do trabalho.

APÊNDICE C - Tabela relacionando espécies presentes na árvore filogenética final

Código	Ponto de inserção	Obervações	Referências
Aech	*Acromyrmex echinatior*	Presente na árvore original	Peters et al., 2017
Acom	*Ampulex compressa*	Presente na árvore original	Peters et al., 2017
Aflo	*Apis mellifera*	-	Bossert et al., 2019; Chhakchhuak et al., 2016; Wang et al., 2018
Alab	*Apis mellifera*	-	Chhakchhuak et al., 2016; Wang et al., 2018
Amel	*Apis mellifera*	Presente na árvore original	Peters et al., 2017
Aros	*Tenthredo koehleri, Nematus ribesii*	-	Branstetter et al., 2017; Peters et al., 2017

Código	Ponto de inserção	Obervações	Referências
Acep	*Acromyrmex echinatior*	-	Branstetter et al., 2017
Acol	*Acromyrmex echinatior*	Inferido pela posição do gênero *Atta*	Branstetter et al., 2017
Bkin	*Andricus quercuscalicis*	Resolvido a nível da tribo Cynipini	Peters et al., 2017
Bbfi	*Bombus rupestris*	-	Santos Júnior et al., 2022
Bimp	*Bombus rupestris*	-	Bossert et al., 2019; Branstetter et al., 2017; Santos Júnior et al., 2022
Bpyr	*Bombus rupestris*	-	Santos Júnior et al., 2022
Bter	*Bombus rupestris*	-	Bossert et al., 2019; Branstetter et al., 2017; Santos Júnior et al., 2022
Bvan	*Bombus rupestris*	Sinônimo de *Bombus bifarius nearticus* e *Bombus nearticus*	Ghisbain et al., 2020
Bvos	*Bombus rupestris*	-	Santos Júnior et al., 2022
Cflo	*Camponotus floridanus*	Presente na árvore original	Peters et al., 2017
Ccin	*Cephus spinipes*	-	Branstetter et al., 2017; Peters et al., 2017
Ccal	*Ceratina chalybea*	-	Bossert et al., 2019; Da Silva, 2021; Peters et al., 2017

Código	Ponto de inserção	Obervações	Referências
Cins	*Cotesia vestalis, Macrocentrus marginator, Diaeretus essingellae, Aphidius colemani, Aleoides testaceus, Dacnusa sibirica*	-	Shi, Chen e Van Achterberg, 2005
Cgig	*Colletes cunicularius*	-	Peters et al., 2017
Cglo	*Cotesia vestalis*	-	Branstetter et al., 2017; Peters et al., 2017; Shi, Chen e Van Achterberg, 2005
Ccos	*Acromyrmex echinatior*	Resolvido a nível do gênero *Cyphomyrmex*	Micolino, Cristiano e Cardoso, 2019; Ward et al., 2015
Dall	*Cotesia vestalis, Macrocentrus marginator, Diaeretus essingellae, Aphidius colemani, Aleoides testaceus, Dacnusa sibirica*	-	Sharanowski et al., 2021
Dqua	*Harpegnathos saltator*	Resolvido a nível do gênero *Dinoponera*	Schmidt, 2013
Dsim	*Diprion pini*	-	Malm e Nyman, 2015; Peters et al., 2017
Dmel	*Gyrinus marinus, Pseudomallada prasinus*	*Drosophila melanogaster*, inserida para enraizamento apenas	Bewick et al., 2017
Dnov	*Dufourea dentiventris*	-	Branstetter et al., 2017; Da Silva, 2021; Peters et al., 2017
Ebur	*Acromyrmex echinatior, Harpegnathos saltator, Camponotus floridanus*	Resolvido a nível da subfamília Dorylinae	Romiguier et al., 2022

Código	Ponto de inserção	Obervações	Referências
Emex	*Euglossa dilemma*	-	Bossert et al., 2019; Branstetter et al., 2017; Peters et al., 2017
Fari	*Cotesia vestalis, Macrocentrus marginator, Diaeretus essingellae, Aphidius colemani, Aleoides testaceus, Dacnusa sibirica*	-	Branstetter et al., 2017; Shi, Chen e Van Achterberg, 2005
Fexs	*Camponotus floridanus*	-	Romiguier et al., 2022
Fvar	*Tetragonula carbonaria*	Resolvido a nível da tribo Meliponini	Peters et al., 2017
Hsal	*Harpegnathos saltator*	Presente na árvore original	Peters et al., 2017
Lhum	*Acromyrmex echinatior, Harpegnathos saltator, Camponotus floridanus*	-	Branstetter et al., 2017
Mrot	*Megachile willughbiella*	-	Bossert et al., 2019; Branstetter et al., 2017; Peters et al., 2017
Mgen	*Halictus quadricinctus, Lasioglossum xanthopus, Sphecodes albilabris, Nomioides sp., Nomia diversipes, Systropha curvicornis, Dufourea dentiventris*	-	Da Silva, 2021
Mdem	*Cotesia vestalis*	-	Branstetter et al., 2017; Shi, Chen e Van Achterberg, 2005
Mpha	*Acromyrmex echinatior*	-	Branstetter et al., 2017
Nvit	*Nasonia vitripennis*	Presente na árvore original	Peters et al., 2017

Código	Ponto de inserção	Obervações	Referências
Nfab	*Diprion pini*	-	Linnen e Smith, 2012; Malm e Nyman, 2015; Terbot II, 2021
Nlec	*Diprion pini*	-	Branstetter et al., 2017; Linnen e Farrell, 2008; Malm e Nyman, 2015; Terbot II, 2021
Npin	*Diprion pini*	-	Linnen e Farrell, 2008; Malm e Nyman, 2015; Terbot II, 2021
Nvir	*Diprion pini*	-	Linnen e Farrell, 2008; Linnen e Smith, 2012; Malm e Nyman, 2015; Terbot II, 2021
Nful	*Camponotus floridanus*	-	Romiguier et al., 2022
Obru	*Harpegnathos saltator*	Resolvido a nível do gênero *Odontomachus*	Romiguier et al., 2022; Schmidt, 2013
Obir	*Acromyrmex echinatior, Harpegnathos saltator, Camponotus floridanus*	Resolvido a nível da subfamília Dorylinae	Romiguier et al., 2022
Oabi	*Orussus abietinus*	Presente na árvore original	Peters et al., 2017
Obic	*Osmia cornuta*	-	Peters et al., 2017
Pbar	*Acromyrmex echinatior*	-	Branstetter et al., 2017
Pcan	*Polistes dominula*	-	Santos et al., 2015; Somavilla et al., 2021
Pdom	*Polistes dominula*	Presente na árvore original	Peters et al., 2017

Código	Ponto de inserção	Obervações	Referências
Pexc	*Polistes dominula*	-	Da Silva, 2021; Santos et al., 2015; Somavilla et al., 2021
Pfus	*Polistes dominula*	-	Da Silva, 2021; Santos et al., 2015; Somavilla et al., 2021
Pgra	*Acromyrmex echinatior, Harpegnathos saltator, Camponotus floridanus*	Resolvido a nível do gênero *Pseudomyrmex*	Romiguier et al., 2022
Sinv	*Acromyrmex echinatior*	-	Branstetter et al., 2017
Tcor	*Acromyrmex echinatior*	-	Micolino, Cristiano e Cardoso, 2019
Tsep	*Acromyrmex echinatior*	-	Micolino, Cristiano e Cardoso, 2019
Tzet	*Acromyrmex echinatior*	-	Micolino, Cristiano e Cardoso, 2019
Tcas	*Tribolium castaneum*	Presente na árvore original, mantido apenas para o enraizamento	Peters et al., 2017
Tpre	*Gonatocerus morrilli, Nasonia vitripennis, Torymus bedeguaris, Brachymeria minuta, Orasema simulatrix, Diglyphus isaea, Cosmocoidea morrilli*	-	Branstetter et al., 2017
Vcan	*Hyposoter didymator*	-	Peters et al., 2017; Shi, Chen e Van Achterberg, 2005
Vcra	*Vespa crabro*	Presente na árvore original	Peters et al., 2017
Vvel	*Vespa crabro*	-	Peters et al., 2017; Suhonen et al., 2019

Código	Ponto de inserção	Obervações	Referências
Vpen	*Vespula germanica*	-	Peters et al., 2017; Suhonen et al., 2019
Veme	*Acromyrmex echinatior*	-	Branstetter et al., 2017
Waur	*Acromyrmex echinatior*	-	Branstetter et al., 2017

Estão indicados os taxa presentes em Peters et al. (2017) utilizados como pontos de referência para inserção e as referências literárias utilizadas para resolver posições de taxa não presentes em Peters et al. (2017).

APÊNDICE D - Código em Python utilizado para a filtragem de nucleotídeos e aminoácidos não canônicos presentes nas sequências dos genomas anotados de Hymenoptera baixados do NCBI

```
# Carregamento das bibliotecas necessárias
import argparse
import glob
import os

from Bio import SeqIO

# Listas de nucleotídeos e aminoácidos canônicos
CANNONICAL_NUC: list = ["A", "C", "T", "G"]
CANNONICAL_AA: list = [
    "A", ”R", "N", "D", "C", "Q", "E", "G", "H", "I", "L", "K", "M", "F",
"P", "S", "T", "W", "Y", "V",
]


# Função para recebimento de argumentos da linha de comando
def get_args() -> argparse.Namespace:
    parser = argparse.ArgumentParser(
        description="Filters FASTA files for nucleotide or protein
sequences containing non-cannonical nucleotides or aminoacids."
    )
    parser.add_argument(
        "input_directory",
        metavar="DIR",
        type=str,
        help="the directory containing the sequences to be filtered",
    )
    parser.add_argument(
        "-o",
```

```
        "--out",
        type=str,
        help="the directory where the result files should be written to
(default: the current working directory)",
        default="./",
    )
    parser.add_argument(
       "-t",
       "--type",
       type=str,
       help="the type of sequence to be expected from the input files,
which must all be of the same type: 'protein' or 'p' for aminoacids or
'nucleotide' or 'n' for nucleotides (default: n)",
       choices=["protein", "nucleotide", "p", "n"],
       default="n",
       required=False,
    )

    return parser.parse_args()


# Função que confere se uma sequência de nucleotídeos ou aminoácidos inicia
com
# códon de iniciação e termina com um códon de término
def is_sane(record: SeqIO.SeqRecord, seq_type: str) -> bool:
    if seq_type in ["n", "nucleotide"]:
        return record.seq.upper().startswith("ATG") and
record.seq.upper().endswith(("TAA", "TAG", "TGA"))
    else:
        return record.seq.upper().startswith("M") and
record.seq.upper().endswith("*")


# Função que remove sequências de um arquivo FASTA que possuam nucleotídeos
ou
# aminoácidos não canônicos e que não possuam códon de início e término
def filter_sequences(input_path: str, outdir: str, seq_type: str):
    reference = CANNONICAL_NUC if seq_type in ["n", "nucleotide"] else
CANNONICAL_AA

    for file in glob.glob(os.path.join(input_path, "*")):
        genome_acc = os.path.basename(file)

        print(f"File {genome_acc}:")

        records = SeqIO.parse(file, "fasta")

        with open(os.path.join(outdir, genome_acc), "w") as outfile:
            for record in records:
                if not any(
                    [char not in reference for char in record.upper()]
```

```
                ) and is_sane(record, seq_type):
                    outfile.write(record.format("fasta"))
                else:
                    print(f"\tRecord {record.id} removed")


# Função principal que invoca as demais funcionalidades do script
def main():
    args = get_args()
    filter_sequences(args.input_directory, args.out, args.type)
    print("Done!")


if __name__ == "__main__":
    main()
```

APÊNDICE E - Código em Python utilizado para a sumarização por lócus, filtragem e tradução das sequências nucleotídicas dos genomas anotados de Hymenoptera baixados do NCBI

```
# Carregamento das bibliotecas necessárias
import argparse
import glob
import os
import re
from typing import List

from Bio import SeqIO
import pandas as pd


# Função para recebimento de argumentos da linha de comando
def get_args() -> argparse.Namespace:
    parser = argparse.ArgumentParser(
        description="Filters FASTA files for nucleotide or protein
sequences containing non-cannonical nucleotides or aminoacids."
    )
    parser.add_argument(
        "input_directory",
        metavar="DIR",
        type=str,
        help="the directory containing the sequences to be filtered",
    )
    parser.add_argument(
        "-o",
        "--out",
        type=str,
        help="the directory where the result files should be written to
(default: the current working directory)",
        default="./",
```

```
    )
    parser.add_argument(
        "-t",
        "--no-translate",
        action="store_true",
        help="tell the program to not output the translated sequences (it
still translates to check for internal stop codons)",
    )

    return parser.parse_args()


# Função que utiliza os dados de anotação das sequências para obter apenas
a
# maior isoforma por locus
def get_longest_seq(records: List[SeqIO.SeqRecord]) -> List[str]:
    all_loci = []

    for record in records:
        each_locus = {}

        record_desc = record.description
        record_len = len(record)

        # Sequências que não possuam algum dos tipos de anotação a seguir
são
        # desconsideradas pois não possuem informações sobre o locus a que
        # pertencem ou a proteína que codificam
        found_locus = re.search(r"\[gene=.*?\]", record_desc)
        found_locus_tag = re.search(r"\[locus_tag=.*?\]", record_desc)
        found_gene_id = re.search(r"\[db_xref=GeneID:.*?\]", record_desc)
        found_protein = re.search(r"\[protein_id=.*?\]", record_desc)

        if not found_locus and not found_locus_tag and not found_gene_id:
            print(f"\tCouldn't find locus tag for record {record.id}")
            continue
        elif not found_protein:
            print(f"\tCouldn't find protein id for record {record.id}")
            continue

        locus_id = ""

        if found_locus:
            locus_id = found_locus.group(0).replace("[gene=",
"").replace("]", "")
        elif found_locus_tag:
            locus_id = (found_locus_tag.group(0).replace("[locus_tag=",
"").replace("]", "")
        )
        elif found_gene_id:
            locus_id = (
```

```
            found_gene_id.group(0).replace("[db_xref=GeneID:",
"").replace("]", "")
        )

        protein_id = found_protein.group(0).replace("[protein_id=",
"").replace("]", "")

        each_locus["locus"] = locus_id
        each_locus["protein_id"] = protein_id
        each_locus["length"] = record_len

        all_loci.append(each_locus)

        if len(all_loci) == 0:
            return []

        loci_df = pd.DataFrame(all_loci)
        longest_isoforms = loci_df.loc[loci_df.groupby(["locus"])
["length"].idxmax()]["protein_id"].tolist()

        return longest_isoforms


# Função que lê um arquivo FASTA, sumariza todas as sequências por locus,
desde
# que possuam a informação necessária para tal, traduz as sequências,
remove as
# sequências que não tenham tamanho múltiplo de 3, remove o códon de
término,
# seguindo o padrão de tradução do NCBI, e remove as sequências com códon
de
# término prematuro
def summarize_and_translate(input_path: str, outdir: str, no_translate:
bool):
    outfile_extension = ".fna" if no_translate else ".faa"

    for file in glob.glob(os.path.join(input_path, "*")):
        genome_acc = os.path.basename(file)

        print(f"File {genome_acc}")

        records: List[SeqIO.SeqRecord] = list(SeqIO.parse(file, "fasta"))
        longest_seqs = get_longest_seq(records)

        if len(longest_seqs) == 0:
            print(f"File {genome_acc} has no valid records, skipping...")
            continue

        print(f"\tProtIDs: {', '.join(longest_seqs)}")
```

```
        with open(os.path.join(outdir, genome_acc.replace(".fna",
outfile_extension)), "w") as outfile:
            for record in records:
                # Checagem adicional para confirmar que o locus possui uma
                # proteína anotada
                record_desc = record.description
                found_protein = re.search(r"\[protein_id=.*?\]",
record_desc)

                if not found_protein:
                    print(f"\tCouldn't find protein id for record
{record.id}")
                    continue
                elif len(record) % 3 != 0:
                    print(f"\tLength of record {record.id} not multiple of
3")
                    continue

                protein_id =
(found_protein.group(0).replace("[protein_id=", "").replace("]", ""))

                if protein_id not in longest_seqs:
                    continue

                translated: SeqIO.SeqRecord = record.translate(
                    id=True, name=True, description=True
                )
                translated.seq = translated.seq.rstrip("*")

                if "*" in translated:
                    print(f"\tRecord {record.id} contains premature stop
codon")
                    continue

                if no_translate:
                    outfile.write(record.format("fasta"))
                else:
                    outfile.write(translated.format("fasta"))


def main():
    args = get_args()
    summarize_and_translate(args.input_directory, args.out,
args.no_translate)
    print("Done!")


if __name__ == "__main__":
    main()
```

APÊNDICE F - Código em R utilizado para a construção da árvore filogenética ultramétrica e construção das respectivas figuras (Figura 8-11).

```
# Carregamento das bibliotecas necessárias
library(treeio)
library(ggtree)
library(RRphylo)
library(ape)
library(phytools)
library(Cairo)
library(ggplot2)
library(dplyr)

# Limpeza da área de trabalho
rm(list = ls())

# Diretório onde se encontra este arquivo. Alterar para o caminho
apropriado
setwd(".")

print("Iniciando construção da árovre filogenética ultramétrica...")

# Árvore "backbone", retirada de Peters et al. 2017
backbone_treedata <-
read.mcmctree("../Dados/dated_tree_aa_inde_2_used_in_Fig1.tre")

# `force.ultrametric` é usado aqui para resolver problemas de aproximação
numérica
# A árvore original é ultramétrica
backbone_tree <- force.ultrametric(backbone_treedata@phylo)

# Árvore de Santos Junior et al. 2022
bombus_treedata <-
read.beast("../Dados/bombus_UCLN_BEAST_MODEL_last_fossils.tre")
bombus_tree <- bombus_treedata@phylo

# Dados utilizados para mapear as espécies de Santos Junior et al. 2022 na
árvore "backbone"
bombus_data <- data.frame(
  bind = c("Bombus_impatiens_voucher_SC060",
"Bombus_bifarius_voucher_SC208",
            "Bombus_vancouverensis_nearticus",
"Bombus_vosnesenskii_voucher_SC112",
            "Bombus_terrestris_voucher_SC003", "Bombus_pyrosoma_B04"),
  reference = c("Bombus_rupestris", "Bombus_impatiens_voucher_SC060",
                  "Bombus_bifarius_voucher_SC208",
                  "Bombus_bifarius_voucher_SC208-
Bombus_vancouverensis_nearticus",
```

```
                  "Bombus_impatiens_voucher_SC060-
Bombus_vancouverensis_nearticus",
                  "Bombus_terrestris_voucher_SC003-
Bombus_vancouverensis_nearticus"),
  poly = rep(FALSE, times = 6)
)

# Árvore incluindo as espécies de Santos Junior et al. 2022
backbone_part_1 <- tree.merger(
  backbone = backbone_tree,
  source.tree = bombus_tree,
  data = bombus_data,
  plot = FALSE
)

# Árvore de Bossert et al. 2019
apidae_tree <- read.newick("../Dados/Phylobayes_80p.tre")

# Dados utilizados para mapear as espécies de Bossert et al. 2019 na árvore
"backbone"
apidae_data <- data.frame(
  bind = c("Apis_laboriosa", "GEN_apis_florea",
           "GEN_ceratina_calcarata", "GEN_eufriesea_mexicana",
           "GEN_megachile_rotundata"),
  reference = c("Apis_mellifera", "Apis_mellifera-Apis_laboriosa",
                "Ceratina_chalybea", "Euglossa_dilemma",
                "Megachile_willughbiella"),
  poly = rep(FALSE, times = 5)
)

# Árvore incluindo as espécies de Bossert et al. 2019
backbone_part_2 <- tree.merger(
  backbone = backbone_part_1,
  source.tree = apidae_tree,
  data = apidae_data,
  plot = FALSE
)

# Árore de Branstetter et al. 2017
branstetter_treedata <- read.beast("../Dados/hym-187t-f75-beast-
rand50.tre")

# `force.ultrametric` é usado aqui para resolver problemas de aproximação
numérica
# A árvore original é ultramétrica
branstetter_tree <- force.ultrametric(branstetter_treedata@phylo)

# Renomeando os nós terminais para remover o caractere "-" dos nomes das
espécies
# O caractere interfere com a funcionalidade da funçao `tree.merger`
```

```
branstetter_tree$tip.label <- sapply(branstetter_tree$tip.label,
function(x) sub("-", "_", x))

# Dados utilizados para mapear as espécies de Branstetter et al. 2017 na
árvore "backbone"
branstetter_data <- data.frame(
  bind = c(
      "TENTHREDINIDAE_Athalia_rosae_Genome",
"FORMICIDAE_Atta_cephalotes_Genome",
      "Atta_colombica", "FORMICIDAE_Wasmannia_auropunctata_Genome",
      "FORMICIDAE_Monomorium_pharaonis_Genome",
"FORMICIDAE_Solenopsis_invicta_Genome",
      "FORMICIDAE_Vollenhovia_emeryi_Genome",
"FORMICIDAE_Pogonomyrmex_barbatus_Genome",
      "CEPHIDAE_Cephus_cinctus_Genome",
"HALICTIDAE_Dufourea_novaeangliae_Genome",
      "BRACONIDAE_Microplitis_demolitor_Genome",
      "BRACONIDAE_Fopius_arisanus_Genome",
"FORMICIDAE_Linepithema_humile_Genome",
      "Neodiprion_fabricii", "Neodiprion_virginianus",
      "DIPRIONIDAE_Neodiprion_lecontei_Genome", "Neodiprion_pinetum",
      "TRICHOGRAMMATIDAE_Trichogramma_pretiosum_Genome"
  ),
  reference = c(
      "Tenthredo_koehleri-Nematus_ribesii", "Acromyrmex_echinatior",
      "FORMICIDAE_Atta_cephalotes_Genome", "Acromyrmex_echinatior-
FORMICIDAE_Atta_cephalotes_Genome",
"FORMICIDAE_Wasmannia_auropunctata_Genome-
FORMICIDAE_Atta_cephalotes_Genome",
      "FORMICIDAE_Monomorium_pharaonis_Genome",
      "FORMICIDAE_Solenopsis_invicta_Genome-
FORMICIDAE_Atta_cephalotes_Genome",
      "FORMICIDAE_Vollenhovia_emeryi_Genome-
FORMICIDAE_Atta_cephalotes_Genome",
      "Cephus_spinipes", "Dufourea_dentiventris",
      "Cotesia_vestalis-Dacnusa_sibirica",
      "BRACONIDAE_Microplitis_demolitor_Genome-Dacnusa_sibirica",
      "Camponotus_floridanus-FORMICIDAE_Atta_cephalotes_Genome",
      "Diprion_pini", "Neodiprion_fabricii",
      "Neodiprion_virginianus", "DIPRIONIDAE_Neodiprion_lecontei_Genome",
      "Gonatocerus_morrilli-Nasonia_vitripennis"
  ),
  poly = rep(FALSE, times = 18)
)

# Árvore incluindo as espécies de Branstetter et al. 2017
backbone_part_3 <- tree.merger(
  backbone = backbone_part_2,
  source.tree = branstetter_tree,
  data = branstetter_data,
  plot = FALSE
```

```
)

# Árvore de Vespidae de Silva 2021
silva_vespidae_tree <- read.newick("../Dados/Vespidae all data 2.tre")

# Dados utilizados para mapear as espécies de Vespidae de Silva 2021 na
árvore "backbone"
silva_vespidae_data <- data.frame(
  bind = c(
      "Polistes_fuscatus",
      "Polistes_exclamans",
      "Polistes_canadensis"
  ),
  reference = c(
      "Polistes_dominula",
      "Polistes_fuscatus",
      "Polistes_exclamans"
  ),
  poly = rep(FALSE, times = 3)
)

# Árvore incluindo as espécies de Vespidae de Silva 2021
backbone_part_4 <- tree.merger(
  backbone = backbone_part_3,
  source.tree = silva_vespidae_tree,
  data = silva_vespidae_data,
  plot = FALSE
)

# Árvore de Halictidae de Silva 2021
silva_halictidae_tree <- read.newick("../Dados/Halictidae all data.tre")

# Dados utilizados para mapear as espécies de Halictidae de Silva 2021 na
árvore "backbone"
silva_halictidae_data <- data.frame(
  bind = c(
      "Nomia_melanderi",
      "Megalopta_genalis"
  ),
  reference = c(
      "Nomia_diversipes",
      "Nomia_diversipes-Lasioglossum_xanthopus"
  ),
  poly = rep(FALSE, times = 2)
)

# Árvore incluindo as espécies de Halictidae de Silva 2021
backbone_part_5 <- tree.merger(
  backbone = backbone_part_4,
  source.tree = silva_halictidae_tree,
  data = silva_halictidae_data,
```

```
  plot = FALSE
)

# Árvore de Romiguier et al. 2022
romiguier_treedata <- read.mcmctree("../Dados/FigTree.tre")

# `force.ultrametric` é usado aqui para resolver problemas de aproximação
numérica
# A árvore original é ultramétrica
romiguier_tree <- force.ultrametric(romiguier_treedata@phylo)

# Dados utilizados para mapear as espécies de Romiguier et al. 2022 na
árvore "backbone"
romiguier_data <- data.frame(
  bind = c(
      "Formica_sanguinea",
      "Formica_exsecta",
      "Nylanderia_fulva",
      "Odontomachus_haematodus",
      "Odontomachus_brunneus",
      "Pseudomyrmex_pallidus",
      "Pseudomyrmex_gracilis",
      "Ooceraea_biroi",
      "Eciton_burchellii"
  ),
  reference = c(
      "Camponotus_floridanus",
      "Formica_sanguinea",
      "Formica_exsecta-Camponotus_floridanus",
      "Harpegnathos_saltator",
      "Odontomachus_haematodus",
      "FORMICIDAE_Linepithema_humile_Genome",
      "Pseudomyrmex_pallidus",
      "Pseudomyrmex_gracilis-FORMICIDAE_Atta_cephalotes_Genome",
      "Ooceraea_biroi"
  ),
  poly = rep(FALSE, times = 9)
)

# Árvore incluindo as espécies de Romiguier et al. 2022
backbone_part_6 <- tree.merger(
  backbone = backbone_part_5,
  source.tree = romiguier_tree,
  data = romiguier_data,
  plot = FALSE
)

# Remoção de espécies utilizadas apenas como pontos de ancoragem
backbone_part_6 <- drop.tip(backbone_part_6, c(
  "Formica_sanguinea",
  "Odontomachus_haematodus",
```

```
  "Pseudomyrmex_pallidus")
)

# A seguir, espécies que foram manualmente adicionadas baseadas em
referências da literatura
# que não disponibilizaram arquivos das árvores
manual_data <- data.frame(
  bind = c(
      "Belonocnema_kinseiy",
      "Chelonus_insularis", "Cotesia_glomerata",
      "Venturia_canescens", "Colletes_gigas",
      "Frieseomelitta_varia", "Osmia_bicornis_bicornis",
      "Vespa_velutina", "Vespula_pensylvanica",
      "Diprion_similis", "Diachasma_alloeum",
      "Trachymyrmex_septentrionalis", "Trachymyrmex_cornetzi",
      "Trachymyrmex_zeteki", "Cyphomyrmex_costatus",
      "Dinoponera_quadriceps", "Drosophila_melanogaster"
  ),
  reference = c(
      "Andricus_quercuscalicis",
      "BRACONIDAE_Microplitis_demolitor_Genome-Dacnusa_sibirica",
"Cotesia_vestalis",
      "Hyposoter_didymator", "Colletes_cunicularius",
      "Tetragonula_carbonaria", "Osmia_cornuta",
      "Vespa_crabro", "Vespula_germanica",
      "Diprion_pini", "BRACONIDAE_Fopius_arisanus_Genome",
      "Acromyrmex_echinatior-FORMICIDAE_Atta_cephalotes_Genome",
"Trachymyrmex_septentrionalis-FORMICIDAE_Atta_cephalotes_Genome",
      "Trachymyrmex_cornetzi-FORMICIDAE_Atta_cephalotes_Genome",
"Trachymyrmex_zeteki-FORMICIDAE_Atta_cephalotes_Genome",
      "Harpegnathos_saltator-Odontomachus_brunneus",
"Pseudomallada_prasinus-Xanthostigma_xanthostigma"
  ),
  poly = rep(FALSE, times = 17)
)

# Árvore final
final_tree <- tree.merger(
  backbone = backbone_part_6,
  data = manual_data,
  plot = FALSE
)

# Renomeação das espécies na árovre final para o código de 4 letras
utilizado no trabalho
names_map <- c(
  # Espécies presentes originalmente na árvore "backbone" de Peters et al.
  "Acromyrmex_echinatior" = "Aech",
  "Ampulex_compressa" = "Acom",
  "Apis_mellifera" = "Amel",
  "Camponotus_floridanus" = "Cflo",
```

```
"Harpegnathos_saltator" = "Hsal",
"Nasonia_vitripennis" = "Nvit",
"Orussus_abietinus" = "Oabi",
"Polistes_dominula" = "Pdom",
"Tribolium_castaneum" = "Tcas",
"Vespa_crabro" = "Vcra",
# Demais espécies
"Bombus_bifarius_voucher_SC208" = "Bbif",
"Bombus_impatiens_voucher_SC060" = "Bimp",
"Bombus_pyrosoma_B04" = "Bpyr",
"Bombus_terrestris_voucher_SC003" = "Bter",
"Bombus_vosnesenskii_voucher_SC112" = "Bvos",
"Bombus_vancouverensis_nearticus" = "Bvan",
"GEN_apis_florea" = "Aflo",
"GEN_ceratina_calcarata" = "Ccal",
"GEN_eufriesea_mexicana" = "Emex",
"GEN_megachile_rotundata" = "Mrot",
"Apis_laboriosa" = "Alab",
"TENTHREDINIDAE_Athalia_rosae_Genome" = "Aros",
"FORMICIDAE_Atta_cephalotes_Genome" = "Acep",
"FORMICIDAE_Wasmannia_auropunctata_Genome" = "Waur",
"FORMICIDAE_Monomorium_pharaonis_Genome" = "Mpha",
"FORMICIDAE_Solenopsis_invicta_Genome" = "Sinv",
"FORMICIDAE_Vollenhovia_emeryi_Genome" = "Veme",
"FORMICIDAE_Pogonomyrmex_barbatus_Genome" = "Pbar",
"CEPHIDAE_Cephus_cinctus_Genome" = "Ccin",
"HALICTIDAE_Dufourea_novaeangliae_Genome" = "Dnov",
"BRACONIDAE_Microplitis_demolitor_Genome" = "Mdem",
"BRACONIDAE_Fopius_arisanus_Genome" = "Fari",
"FORMICIDAE_Linepithema_humile_Genome" = "Lhum",
"DIPRIONIDAE_Neodiprion_lecontei_Genome" = "Nlec",
"TRICHOGRAMMATIDAE_Trichogramma_pretiosum_Genome" = "Tpre",
"Atta_colombica" = "Acol",
"Neodiprion_virginianus" = "Nvir",
"Neodiprion_fabricii" = "Nfab",
"Neodiprion_pinetum" = "Npin",
"Polistes_fuscatus" = "Pfus",
"Polistes_exclamans" = "Pexc",
"Nomia_melanderi" = "Nmel",
"Megalopta_genalis" = "Mgen",
"Polistes_canadensis" = "Pcan",
"Formica_exsecta" = "Fexs",
"Nylanderia_fulva" = "Nful",
"Odontomachus_brunneus" = "Obru",
"Pseudomyrmex_gracilis" = "Pgra",
"Ooceraea_biroi" = "Obir",
"Eciton_burchellii" = "Ebur",
"Belonocnema_kinseiy" = "Bkin",
"Chelonus_insularis" = "Cins",
"Cotesia_glomerata" = "Cglo",
"Venturia_canescens" = "Vcan",
```

```
  "Colletes_gigas" = "Cgig",
  "Frieseomelitta_varia" = "Fvar",
  "Osmia_bicornis_bicornis" = "Obic",
  "Vespa_velutina" = "Vvel",
  "Vespula_pensylvanica" = "Vpen",
  "Diprion_similis" = "Dsim",
  "Diachasma_alloeum" = "Dall",
  "Trachymyrmex_septentrionalis" = "Tsep",
  "Trachymyrmex_cornetzi" = "Tcor",
  "Trachymyrmex_zeteki" = "Tzet",
  "Cyphomyrmex_costatus" = "Ccos",
  "Dinoponera_quadriceps" = "Dqua",
  "Drosophila_melanogaster" = "Dmel"
)

# Renomeação dos ramos terminais e remoção daqueles não utilizados na
árvore final
for (tip_name in final_tree$tip.label) {
  if (tip_name %in% names(names_map)) {
      final_tree$tip.label[which(final_tree$tip.label == tip_name)] <-
names_map[[tip_name]]
  } else {
      final_tree <- drop.tip(final_tree, tip_name)
  }
}

# Salvando a árvore final
write.tree(final_tree,
"../Dados/hymenoptera_65_dmel_tcas_stitched_ultrametric.nwk")

print("Árvore construída com sucesso!")

###### FIGURAS
print("Gerando figuras das árvores...")

## ÁRVORE "BACKBONE"
# Renomeando os ramos terminais para ficar de acordo com nossa convenção
for (tip_name in backbone_tree$tip.label) {
  split_name <- strsplit(tip_name, "_")
  if (split_name[[1]][2] == "sp") {
      fixed_name <- paste0(substr(split_name[[1]][1], 1, 2),
substr(split_name[[1]][2], 1, 2))
  }
  else {
      fixed_name <- paste0(substr(split_name[[1]][1], 1, 1),
substr(split_name[[1]][2], 1, 3))
  }
  backbone_tree$tip.label[which(backbone_tree$tip.label == tip_name)] <-
fixed_name
}
```

```
# Re-enraizamento da árvore "backbone"
backbone_tree <- reroot(backbone_tree, 343)

tree_font_size <- 6
tree_font_family <- "Consolas"
tree_pdf_width <- 20
tree_pdf_height <- 46
tree_font_size_legend <- 18
tree_font_size_legend_title <- 20
tree_legend_margin <- margin(0, 0, 0, 0)
tree_legend_box_margin <- margin(r = 10)

# Classificação dos nós removidos ou mantidos como pontos de ancoragem na
árvore "backbone"
backbone_removed_nodes <- data.frame(
  node = c(
      304, 119, 312, 118, 116, 115, 295, 111, 110, 141, 144, 325, 327, 109,
282,
      272, 252, 70, 69, 246, 233, 49, 53, 48, 221, 154, 216, 212, 211, 31,
200, 11,
      10, 13, 14, 15, 9, 196, 197, 7, 186, 8, 2, 331, 169, 339, 162, 161,
335, 120,
      309, 117, 114, 319, 142, 143, 329, 230, 30, 203, 12, 6, 1, 168, 342,
344
  ),
  status = c(rep("removed", times = 49), rep("anchor", times = 17))
)

# Geração da figura
cairo_pdf(
  "../Figuras/Figura_6.pdf",
  width = tree_pdf_width,
  height = tree_pdf_height
)
ggtree(backbone_tree, size = 2, branch.length = "none") +
  geom_tiplab(size = tree_font_size, family = tree_font_family) +
  geom_highlight(data = backbone_removed_nodes, mapping = aes(node = node,
fill = status, extend = 1.2)) +
  geom_rootedge(rootedge = 0.5, size = 2) +
  scale_fill_manual(
      name = "",
      breaks = c("removed", "anchor"),
      labels = c("Removido", "Referência"),
      values = c("red", "blue")) +
  theme(legend.text = element_text(size = tree_font_size_legend),
      legend.title = element_text(size = tree_font_size_legend_title),
      legend.margin = tree_legend_margin,
      legend.box.margin = tree_legend_box_margin) +
  scale_y_continuous(limits = c(-1, NA))
dev.off()
```

```
## ÁRVORE INTERMEDIÁRIA
# Nós terminais não presentes na árvore final e não utilizados como pontos
de ancoragem
intermediate_to_remove <- c(
  126, 125, 124, 123, 122, 121, 119, 134, 133, 132, 131, 135, 118, 116,
115, 113,
  112, 111, 110, 141, 144, 147, 146, 145, 150, 149, 148, 109, 108, 107,
106, 105,
  103, 102, 104, 99, 98, 97, 96, 95, 94, 101, 100, 93, 92, 91, 90, 89, 88,
87,
  86, 85, 84, 83, 82, 81, 79, 78, 80, 77, 76, 75, 74, 73, 72, 70, 69, 68,
67, 64,
  63, 62, 66, 65, 61, 60, 59, 58, 56, 55, 54, 57, 49, 53, 48, 45, 44, 43,
42, 47,
  46, 154, 41, 40, 39, 38, 37, 36, 35, 34, 33, 32, 31, 22, 21, 23, 11, 10,
13,
  14, 15, 9, 17, 16, 19, 18, 20, 7, 5, 4, 8, 2, 1, 157, 156, 155, 158, 169,
165,
  164, 163, 162, 161, 160, 159
)

# Remoção dos nós não presentes na árvore final e não utilizados como
pontos de ancoragem
intermediate_tree <- drop.tip(backbone_tree, intermediate_to_remove)

# Nós usados como ancoragem marcados para posterior remoção
intermediate_removed_nodes <- data.frame(
  node = c(
      64, 63, 18, 17, 59, 12, 46, 2, 72, 77, 35, 39
  ),
  status = c(rep("removed", times = 12))
)

# Geração da figura
cairo_pdf(
  "../Figuras/Figura_7.pdf",
  width = tree_pdf_width,
  height = 16
)
ggtree(intermediate_tree, size = 2, branch.length = "none") +
  geom_tiplab(size = tree_font_size, family = tree_font_family) +
  geom_highlight(data = intermediate_removed_nodes, mapping = aes(node =
node, fill = status, extend = 0.6)) +
  geom_rootedge(rootedge = 0.5, size = 2) +
  scale_fill_manual(
      name = "",
      breaks = c("removed"),
      labels = c("Removido"),
      values = c("red")) +
  theme(legend.text = element_text(size = tree_font_size_legend),
      legend.title = element_text(size = tree_font_size_legend_title),
```

```
      legend.margin = tree_legend_margin,
      legend.box.margin = tree_legend_box_margin) +
  scale_y_continuous(limits = c(-1, NA))
dev.off()

## ÀRVORE FINAL
# Re-enraizamento da árvore usando o grupo externo (Tcas + Dmel)
final_tree <- reroot(final_tree, 132)

# Classificação dos nós adicionados não presentes em Peters et al. 2017
# de acordo com o grande grupo
highlighted_nodes <- data.frame(
  node = c(
      # formigas
      124, 47, 48, 49, 50, 51, 125, 54, 55, 42, 43, 126, 127, 39, 40,
      # abelhas
      98, 25, 26, 23, 22, 92, 104,
      # grupo externo
      133,
      # moscas-serra
      1, 128,
      # vespas
      84, 18, 16, 76, 5, 3
  ),
  group = c(
      rep("ants", times = 15),
      rep("bees", times = 7),
      "outgroup",
      rep("sawflies", times = 2),
      rep("wasps", times = 6)
  )
)

# Geração da figura
cairo_pdf(
  "../Figuras/Figura_8.pdf",
  width = tree_pdf_width,
  height = 22
)
ggtree(final_tree, size = 2, branch.length = "none") +
  geom_tiplab(size = tree_font_size, family = tree_font_family) +
  geom_highlight(data = highlighted_nodes, mapping = aes(node = node, fill
= group, extend = 0.95)) +
  geom_rootedge(rootedge = 0.5, size = 2) +
  scale_fill_manual(
      name = "Grande grupo",
      breaks = c("bees", "ants", "outgroup", "sawflies", "wasps"),
      labels = c("Abelhas", "Formigas", "Grupo externo", "Moscas-serra",
"Vespas"),
      values = c("#66C2A5", "#8DA0CB", "#FFD92F", "#FC8D62", "#E78AC3")) +
  theme(legend.text = element_text(size = tree_font_size_legend),
```

```
      legend.title = element_text(size = tree_font_size_legend_title),
      legend.margin = tree_legend_margin,
      legend.box.margin = tree_legend_box_margin) +
  scale_y_continuous(limits = c(-1, NA))
dev.off()

## ÁRVORE PRINCIPAL
# Remoção dos grupos externos e das espécies com socialidade intermediária
main_tree <- drop.tip(final_tree, c("Dmel", "Tcas", "Ccal", "Nmel",
"Emex"))

# Leitura dos metadados sobre eussocialidade
eusociality_data <- read.csv(
  "../Dados/plot_data/species_metadata.tsv",
  header = T,
  sep = "\t"
) %>% select(ABBREV, SOCIALITY_LEVEL, GROUP) %>%
  dplyr::rename(
      species = ABBREV,
      level = SOCIALITY_LEVEL,
      major_group = GROUP
) %>% filter(!species %in% c("Dmel", "Tcas"))

# Geração da árvore
cairo_pdf(
  "../Figuras/Figura_9.pdf",
  width = tree_pdf_width,
  height = 22
)
ggtree(main_tree, size = 2) %<+% eusociality_data +
  geom_tiplab(size = tree_font_size, family = tree_font_family) +
  geom_tippoint(aes(color = as.factor(level)), position = position_nudge(x
= .22), size = 5) +
  geom_cladelab(
      node = 121,
      label = "\"Symphyta\"",
      image = "../pics/sawfly.png",
      geom = "image",
      imagecolor = "black",
      alpha = 1,
      offset = 0.25,
      extend = 2.3,
      barsize = 2,
      offset.text = 0.25
  ) +
  geom_cladelab(
      node = 121,
      label = "\"Symphyta\"",
      color = "black",
      angle = 90,
      offset = 0.25,
```

```
    barsize = 0,
    fontsize = tree_font_size,
    family = tree_font_family,
    hjust = 0.5,
    offset.text = 0.05
) +
geom_cladelab(
    node = 67,
    label = "Parasitoida",
    image = "../pics/parasitoida.png",
    geom = "image",
    imagecolor = "black",
    alpha = 1,
    offset = 0.25,
    extend = 0.3,
    barsize = 2,
    offset.text = 0.25
) +
geom_cladelab(
    node = 67,
    label = "Parasitoida",
    color = "black",
    angle = 90,
    offset = 0.25,
    barsize = 0,
    fontsize = tree_font_size,
    family = tree_font_family,
    hjust = 0.5,
    offset.text = 0.05
) +
geom_cladelab(
    node = 76,
    label = "Vespoidea",
    image = "../pics/vespoidea.png",
    geom = "image",
    imagecolor = "black",
    alpha = 1,
    offset = 0.25,
    extend = 0.3,
    barsize = 2,
    offset.text = 0.25
) +
geom_cladelab(
    node = 76,
    label = "Vespoidea",
    color = "black",
    angle = 90,
    offset = 0.25,
    barsize = 0,
    fontsize = tree_font_size,
    family = tree_font_family,
```

```
    hjust = 0.5,
    offset.text = 0.05
) +
geom_cladelab(
    node = 83,
    label = "Apoidea",
    image = "../pics/apoidea.png",
    geom = "image",
    imagecolor = "black",
    alpha = 1,
    offset = 0.25,
    extend = 0.3,
    barsize = 2,
    offset.text = 0.25
) +
geom_cladelab(
    node = 83,
    label = "Apoidea",
    color = "black",
    angle = 90,
    offset = 0.25,
    barsize = 0,
    fontsize = tree_font_size,
    family = tree_font_family,
    hjust = 0.5,
    offset.text = 0.05
) +
geom_cladelab(
    node = 98,
    label = "Formicoidea",
    image = "../pics/formicoidea.png",
    geom = "image",
    imagecolor = "black",
    alpha = 1,
    offset = 0.25,
    extend = 0.3,
    barsize = 2,
    offset.text = 0.25,
    imagesize = 0.035
) +
geom_cladelab(
    node = 98,
    label = "Formicoidea",
    color = "black",
    angle = 90,
    offset = 0.25,
    barsize = 0,
    fontsize = tree_font_size,
    family = tree_font_family,
    hjust = 0.5,
    offset.text = 0.05
```

```
  ) +
  geom_rootedge(rootedge = 0.2, size = 2) +
  geom_treescale(fontsize = tree_font_size, family = tree_font_family, y =
-1) +
  scale_color_manual(
      name = "Nível de socialidade",
      breaks = c(1, 3),
      labels = c("Solitário", "Eussocial"),
      values = c("black", "red")
  ) +
  guides(color = guide_legend(override.aes = list(size = 5))) +
  theme(legend.text = element_text(size = tree_font_size_legend),
      legend.title = element_text(size = tree_font_size_legend_title),
      legend.margin = tree_legend_margin,
      legend.box.margin = tree_legend_box_margin) +
  scale_y_continuous(limits = c(-1, NA))
dev.off()

print("Execução concluída com sucesso!")
```

APÊNDICE G - Código em R utilizado para realização dos testes de ANOVA filogenético e construção da figura dos resultados (Figura 12).

```
# Carregamento das bibliotecas necessárias
library(geiger)
library(phytools)
library(CALANGO)
library(ggplot2)
library(ggpubr)
library(dplyr)
library(cowplot)
library(nlme)
library(dendextend)
library(ComplexHeatmap)

# Limpeza da área de trabalho
rm(list = ls())

# Definição de função para negar pertencimento ("not in")
`%!in%` <- Negate(`%in%`)

# Diretório onde se encontra este arquivo. Alterar para o caminho
apropriado
setwd("")

# Valor de corte de valor q para os resultados do teste ANOVA
q_value_cutoff = 0.05
```

```
# Carregamento dos dados gerados numa rodada do pacote CALANGO, que gera
uma
# matriz de contagem de IPR por espécie
load("../Dados/IPR_CALANGO.RData")

# Metadados de eussocialidade
metadata <- read.table("../Dados/species_metadata.tsv", sep = "\t", header
= TRUE)

# Dados de grande grupo excluindo espécies intermediárias e grupos externos
major_group <- c(
  "Ccin" = "Moscas-serra", "Oabi" = "Moscas-serra", "Bkin" = "Moscas-
serra",
  "Nvit" = "Vespas", "Tpre" = "Vespas", "Cglo" = "Vespas", "Mdem" =
"Vespas",
  "Cins" = "Vespas", "Fari" = "Vespas", "Dall" = "Vespas", "Vcan" =
"Vespas",
  "Pdom" = "Vespas", "Pfus" = "Vespas", "Pexc" = "Vespas", "Pcan" =
"Vespas",
  "Vpen" = "Vespas", "Vcra" = "Vespas", "Vvel" = "Vespas", "Acom" =
"Vespas",
  "Mrot" = "Abelhas", "Obic" = "Abelhas", "Amel" = "Abelhas", "Alab" =
"Abelhas",
  "Aflo" = "Abelhas", "Bimp" = "Abelhas", "Bbif" = "Abelhas", "Bvan" =
"Abelhas",
  "Bvos" = "Abelhas", "Bter" = "Abelhas", "Bpyr" = "Abelhas", "Fvar" =
"Abelhas",
  "Mgen" = "Abelhas", "Dnov" = "Abelhas", "Cgig" = "Abelhas", "Hsal" =
"Formigas",
  "Obru" = "Formigas", "Dqua" = "Formigas", "Cflo" = "Formigas", "Fexs" =
"Formigas",
  "Nful" = "Formigas", "Aech" = "Formigas", "Acep" = "Formigas", "Acol" =
"Formigas",
  "Tsep" = "Formigas", "Tcor" = "Formigas", "Tzet" = "Formigas", "Ccos" =
"Formigas",
  "Waur" = "Formigas", "Mpha" = "Formigas", "Sinv" = "Formigas", "Veme" =
"Formigas",
  "Pbar" = "Formigas", "Lhum" = "Formigas", "Pgra" = "Formigas", "Obir" =
"Formigas",
  "Ebur" = "Formigas", "Aros" = "Moscas-serra", "Dsim" = "Moscas-serra",
"Nfab" = "Moscas-serra",
  "Nvir" = "Moscas-serra", "Nlec" = "Moscas-serra", "Npin" = "Moscas-serra"
)

# Lista de espécies intermediárias e grupos externos a serem removidos da
árvore
spp2remove = c("Dmel", "Tcas", metadata$ABBREV[metadata$SOCIALITY_LEVEL ==
2])

# Lista de IPR redundantes a serem removidos, que foram identificados a
partir
```

```
# de uma execução prévia do código
ipr_to_ignore <- c("IPR036658")

# Árvore ultramétrica incluindo todas as espécies de Hymenoptera e os
grupos externos
tree <-
read.tree("../Dados/hymenoptera_65_dmel_tcas_stitched_ultrametric.nwk")

# Remoção das espécies não incluídas nas análises da árvore
pruned.tree <- drop.tip(tree, tree$tip.label[match(spp2remove,
tree$tip.label)])
pruned.tree$tip.label <- metadata$GENOME[match(pruned.tree$tip.label,
metadata$ABBREV)]

# Conversão das classes (eussocial ou solitário) em fatores para construção
dos gráficos
classes <- as.factor(metadata$IS_EUSOCIAL[match(pruned.tree$tip.label,
metadata$GENOME)])
names(classes) <- metadata$GENOME[match(pruned.tree$tip.label,
metadata$GENOME)]

### Construção de um dataframe associando espécies e IPR a serem analisados
# Identificadores de espécies
spp_IDs <- pruned.tree$tip.label

# Todos os IPR associados às respectivas espécies a serem avaliadas
df_IPR <- IPR$y[spp_IDs, colnames(IPR$y)]

# Lista para armazenamento dos valores p
list <- vector()

# Apenas para impressão do status de processamento
i = 1

print("Computando ANOVA filogenético...")

# ANOVA com correção filogenética
for (IPR_ID in unique(colnames(df_IPR)))  {
  if (IPR_ID %!in% ipr_to_ignore) {
      print(paste(IPR_ID, " - ", i))
      i <- i + 1

      # Dados de contagem por IPR
      data_tmp <- as.vector(df_IPR[,IPR_ID])
      names(data_tmp) <- metadata$GENOME[match(rownames(df_IPR),
metadata$GENOME)]

      # Computando coeficiente de variação para remover dados sem variação
      cv <- sd(data_tmp)/mean(data_tmp)

      if ((sum(data_tmp) > 0) && (cv > 0)) {
```

```
      table <- cbind.data.frame(data_tmp, classes)
      # No caso de haver apenas duas categorias de variação, apenas levar
em conta
      # os IPR que possuam pelo menos 3 espécies em cada categoria
      if (length(unique(data_tmp)) == 2) {
      first_group_count <- table(table[,1])[1] >= 3
      second_group_count <- table(table[,1])[2] >= 3
      if (isTRUE(first_group_count) && isTRUE(second_group_count)) {
            spp <- rownames(table)
            # Correlação com movimento Browniano
            corBM <- corBrownian(phy=pruned.tree,form = ~spp)
            # Parâmetros estimados por mínimos quadrados ordinários
            ancova <- gls(data_tmp~classes, data = table, correlation =
corBM)
            # ANOVA
            tmp <- anova(ancova)
            list[IPR_ID] <- tmp$`p-value`[2]
      }
      }
      # No caso de haver apenas três categorias de variação, apenas levar
em conta
      # os IPR que possuam pelo menos 3 espécies em cada categoria
      else if (length(unique(data_tmp)) == 3) {
      first_group_count <- table(table[,1])[1] >= 3
      second_group_count <- table(table[,1])[2] >= 3
      third_group_count <- table(table[,1])[3] >= 3
      if (isTRUE(first_group_count) && isTRUE(second_group_count) &
isTRUE(third_group_count)) {
            spp <- rownames(table)
            # Correlação com movimento Browniano
            corBM <- corBrownian(phy=pruned.tree,form = ~spp)
            # Parâmetros estimados por mínimos quadrados ordinários
            ancova <- gls(data_tmp~classes, data = table, correlation =
corBM)
            # ANOVA
            tmp <- anova(ancova)
            list[IPR_ID] <- tmp$`p-value`[2]
      }
      }
      # No caso de existirem mais de três categorias, proceder normalmente
com
      # as análises
      else {
      spp <- rownames(table)
      # Correlação com movimento Browniano
      corBM <- corBrownian(phy=pruned.tree,form = ~spp)
      # Parâmetros estimados por mínimos quadrados ordinários
      ancova <- gls(data_tmp~classes, data = table, correlation = corBM)
      # ANOVA
      tmp <- anova(ancova)
      list[IPR_ID] <- tmp$`p-value`[2]
```

```
      }
      }
  }
}

# Correção para o teste de múltiplas hipóteses (FDR)
print("Realizando correção para o teste de múltiplas hipóteses (FDR)...")
q_values_all  <- p.adjust(list, method = "BH")

# Lista de IPR associados significativamente com o fenótipo
(eussocialidade)
significant_IDs <- names(q_values_all[q_values_all < q_value_cutoff])

# Filtragem do dataframe incluindo apenas os IPR significativamente
associados
df_IPR <- IPR$y[spp_IDs, significant_IDs]

print("ANOVA filogenético computado com sucesso!")
print("Iniciando construção da figura final...")

# Identificadores para os heatmaps
names <- data.frame(metadata$GENOME, metadata$IS_EUSOCIAL, metadata$ABBREV)
colnames(names) = c("genomeID", "group", "shortName")
# IPR incluindo descrição
colnames_df <- paste0(significant_IDs, " - ",
IPR$annotation.contrasts[significant_IDs])

rownames(df_IPR) <- names$shortName[match(rownames(df_IPR),
names$genomeID)]
colnames(df_IPR) <- colnames_df

# Confirmação de que a árvore não contém espécies não analisadas
pruned.tree$tip.label <- names$shortName[match(pruned.tree$tip.label,
names$genomeID)]

# Transformação da árvore para visualização nos heatmaps
tree2 <- as.dendrogram(as.hclust.phylo(pruned.tree))

# Paleta de cores para as contagens de IPR
my_palette <- colorRampPalette(c("white","blue"))(n = 51)

# Paleta de cores para os níveis de eussocialidade
jColors <- data.frame(
  LABEL = levels(as.factor(names$group)),
  COLOR = I(c("black", "red"))
)
col_letters <- jColors$COLOR
names(col_letters) <- jColors$LABEL

# Paleta de cores para os grandes grupos
colors_groups <- c(
```

```
  "Abelhas" = "#66C2A5", "Formigas" = "#8DA0CB", "Moscas-serra" =
"#FC8D62",
  "Vespas" = "#E78AC3"
)

# Aplicação das definições de cores para o fenótipo (eussocialidade) nos
dados
species2color <- list()
species2color$species <- rownames(df_IPR)
species2color$group <- names$group[match(species2color$species,
names$shortName)]
species2color$group <- jColors$LABEL[match(species2color$group,
jColors$LABEL)]
le <- species2color$group
names(le) <- species2color$species

# Clusterização dos IPR baseda na abundância relativa
distance2 = dist(as.matrix(t(df_IPR)), method = "euclidean")
cluster2_final = set(as.dendrogram(hclust(distance2, method =
c("ward.D2"))), "branches_lwd", 2)
tree_final <- set(tree2, "branches_lwd", 2)

# Configurações adicionais de parâmetros visuais das legendas dos heatmaps
hm1_legend_params = list(
  "at" = c(0, 1, 2), "labels" = c("0", "1", "2"),
  "color_bar" = "discrete", "border" = "black"
)
hm2_legend_params = list(
  "at" = c("FALSE", "TRUE"),
  "labels" = c("Solitário", "Eussocial")
)

### Heatmaps utilizando a filogenia para agrupar espécies (genomas)
# Heatmap com as contagens de termos IPR significativamente associados
ht1 = Heatmap(
  as.matrix(df_IPR),
  cluster_rows = as.hclust(tree_final),
  name = "Contagem do IPR",
  col = my_palette,
  column_title = "Contagem do IPR",
  column_names_rot = 60,
  row_dend_width = unit(5, "cm"),
  row_dend_gp = gpar(lwd = 1.2),
  heatmap_legend_param = hm1_legend_params,
  row_names_side = "left",
  row_names_gp = gpar(fontfamily = "Consolas"),
  # rect_gp = (gpar(col = "black"))
)

# Heatmap nível de socialidade por espécie
ht2 = Heatmap(
```

```
  t(rbind(letters = le)),
  cluster_rows = as.hclust(tree_final),
  name = "Nível de socialidade",
  col = col_letters,
  column_dend_height = unit(2, "cm"),
  row_dend_width = unit(5, "cm"),
  column_dend_gp = gpar(lwd = 1.2),
  column_labels = "",
  height = 1,
  heatmap_legend_param = hm2_legend_params,
  row_names_side = "left",
  row_names_gp = gpar(fontfamily = "Consolas"),
  row_title = "Espécie"
)

# Heatmap associando grande grupo por espécie
ht3 = Heatmap(
  t(rbind(letters = major_group)),
  cluster_rows = as.hclust(tree_final),
  name = "Grande grupo",
  col = colors_groups,
  column_dend_height = unit(2, "cm"),
  row_dend_width = unit(5, "cm"),
  column_dend_gp = gpar(lwd = 1.2),
  column_labels = "",
  height = 1,
  row_names_side = "left",
  row_names_gp = gpar(fontfamily = "Consolas"),
  row_title = "Espécie"
)

# Construção da figura final incluindo os dados dos três heatmaps
ht_list = ht3 + ht2 + ht1

# Salvando a figura final
cairo_pdf(
  "../Figuras/Figura_10.pdf",
  width = 1.2 * ncol(df_IPR),
  height = 15
)
draw(ht_list, padding = unit(c(50, 2, 2, 2), "mm"))
dev.off()

print("Execução concluída com sucesso!")
```

APÊNDICE H - Código em R utilizado para construção da figura com os resultados da análise de completude com BUSCO incluindo apenas os valores de completude de cópia simples (Figura 6).

```
# Carregamento das bibliotecas necessárias
library(ggplot2)
library("grid")
library(dplyr)
library(ggpubr)
library(forcats)
library(RColorBrewer)
library(Cairo)

# Limpeza da área de trabalho
rm(list = ls())

# Diretório onde se encontra este arquivo. Alterar para o caminho
apropriado
setwd(".")

# Resultados sumarizados das análises de completude com BUSCO
data <- read.csv("../Dados/busco_data.csv")

# Definições pertinentes à figura final
my_output_single <- "../Figuras/Figura_5.pdf"
my_width_single <- 10
my_height_single <- 18
my_family <- "Consolas"
colors <- c("#66C2A5", "#8DA0CB", "#FC8D62", "#E78AC3")
breaks <- c("bees", "ants", "sawflies", "wasps")
labels <- c("Abelhas", "Formigas", "Moscas-serra", "Vespas")
y_breaks <- c(0, 25, 50, 75, 100)
y_limits <- c(0, 100)
y_labels <- c("0", "25", "50", "75", "100")
font_size_legend <- 14
title_bottom_margin <- 10
x_title_size <- 14
line_color <- "black"
line_size <- 1
line_type <- 2

# Geração da figura
print("Gerando a figura ...")

# Completude total
complete <- data %>%
  arrange(C) %>%
  mutate(species = factor(species, levels = species)) %>%
  ggplot() +
  geom_bar(aes(y = C, x = species, fill = group), stat = "identity") +
  xlab(NULL) +
  ylab("Completo (%)") +
  scale_y_continuous(
      position = "right",
      breaks = y_breaks,
```

```
      limits = y_limits,
      labels = y_labels
  ) +
  coord_flip() +
  labs(fill = NULL) +
  scale_fill_manual(
      breaks = breaks,
      labels = labels,
      values = colors
  ) +
  theme_light() +
  theme(
      axis.text.y = element_text(family = my_family, color =
ifelse(arrange(data, C)$S < 90, "red", "black")),
      axis.title.x = element_text(size = x_title_size),
      axis.title.x.top = element_text(margin = margin(b =
title_bottom_margin)),
      legend.text = element_text(size = font_size_legend),
  ) +
  geom_hline(
      yintercept = 90,
      color = line_color,
      size = line_size,
      linetype = line_type
  )

# Figura 5, incluindo apenas os valores de completude de cópia simples
cairo_pdf(my_output_single, my_width_single, my_height_single)
print(
  single +
      ylab("Completude de cópia simples (%)") +
      guides(fill = guide_legend("Grande grupo")) +
      theme(
      axis.text.y = element_text(size = 14),
      axis.text.x = element_text(size = 14),
      axis.title.x = element_text(size = 20),
      legend.text = element_text(size = 18),
      legend.title = element_text(size = 20)
      )
  )
dev.off()

print("Execução concluída com sucesso!")
```

APÊNDICE I - Código em R utilizado para a avaliação do tamanho dos proteomas e cobertura de anotação dos proteomas não redundantes de alta qualidade de Hymenoptera e construção das respectivas figuras (Figuras 5 e 7).

```
# Carregando bibliotecas necessárias
library(dplyr)
library(ggplot2)
library(Cairo)

# Limpeza do ambiente de trabalho
rm(list = ls())

# Diretório onde se encontra este arquivo. Alterar para o caminho
apropriado
setwd("")

# Dados de número de proteínas e proteína anotadas por proteoma, com
informação
# de grande grupo
count_data <- read.csv("../Dados/count_data.csv", header = TRUE) %>%
  select(abbrev, major_group, proteins, annotated)

# Reordenação alfabética em português dos grandes grupos
count_data$major_group <- factor(
  count_data$major_group,
  levels = c("bees", "ants", "sawflies", "wasps")
)

# Semente para reprodução de resultados baseados em aleatoriedade
seed <- 666
set.seed(seed)

# Variáveis para controlar a aparência dos gráficos
colors <- c("#66C2A5", "#8DA0CB", "#FC8D62", "#E78AC3")
jitter_width <- 0.3
point_alpha <- 0.25
point_size <- 3
text_size_plot <- 18
text_size_legend <- 16

# Variáveis para controlar o tamnho das figuras finais
pdf_width <- 12
pdf_height <- 10
pdf_height_annotation <- 7

print("Gerando as figuras...")
```

```
# Tamanho do proteoma
cairo_pdf(
  "../Figuras/Figura_Boxplot_Tamanho.pdf",
  width = pdf_width,
  height = pdf_height
)
ggplot(
      count_data,
      mapping = aes(x = major_group, y = proteins, fill = major_group)
  ) +
  geom_boxplot(outlier.shape = NA) +
  geom_point(
      position = position_jitterdodge(jitter.width = jitter_width, seed =
seed),
      show.legend = FALSE,
      alpha = point_alpha,
      size = point_size
  ) +
  scale_fill_manual(
      name = "Grande grupo",
      values = colors,
      labels = c("Abelhas", "Formigas", "Moscas-serra", "Vespas")
  ) +
  scale_x_discrete(
      labels = c("Abelhas", "Formigas", "Moscas-serra", "Vespas")
  ) +
  xlab("Grande grupo") +
  ylab("Tamanho do proteoma (# de proteínas)") +
  theme_bw() +
  theme(
      text = element_text(size = text_size_plot),
      legend.text = element_text(size = text_size_legend)
  )
dev.off()

# Total anotado do proteoma
cairo_pdf(
  "../Figuras/Figura_Boxplot_Anotado.pdf",
  width = pdf_width,
  height = pdf_height_annotation
)
ggplot(
      count_data,
      mapping = aes(x = major_group, y = annotated / proteins, fill =
major_group)
  ) +
  geom_boxplot(outlier.shape = NA) +
  geom_point(
      position = position_jitterdodge(jitter.width = jitter_width, seed =
seed),
      show.legend = FALSE,
```

```
      alpha = point_alpha,
      size = point_size
  ) +
  scale_fill_manual(
      name = "Grande grupo",
      values = colors,
      labels = c("Abelhas", "Formigas", "Moscas-serra", "Vespas")
  ) +
  scale_x_discrete(
      labels = c("Abelhas", "Formigas", "Moscas-serra", "Vespas")
  ) +
  scale_y_continuous(
      limits = c(0.8, 1),
      breaks = c(0.8, 0.9, 1),
      labels = c("80", "90", "100")
  ) +
  xlab("Grande grupo") +
  ylab("Anotação do proteoma (%)") +
  theme_bw() +
  theme(
      text = element_text(size = text_size_plot),
      legend.text = element_text(size = text_size_legend)
  )
dev.off()

print("Execução concluída com sucesso!")
```