

Gabriel Alves Vieira

**FILOGENÔMICA MITOCONDRIAL SEM
CUSTOS: UMA PROVA DE CONCEITO
COM FORMIGAS DA SUBFAMÍLIA
PSEUDOMYRMECINAE
(HYMENOPTERA : FORMICIDAE)**

Brasil

Março de 2019

Gabriel Alves Vieira

**FILOGENÔMICA MITOCONDRIAL SEM CUSTOS:
UMA PROVA DE CONCEITO COM FORMIGAS
DA SUBFAMÍLIA PSEUDOMYRMECINAE
(HYMENOPTERA : FORMICIDAE)**

Dissertação apresentada ao Programa de Pós-Graduação em Química Biológica, Instituto de Bioquímica Médica Leopoldo de Meis, Universidade Federal do Rio de Janeiro, como parte dos requisitos para obtenção do título de Mestre em Química Biológica.

Ministério da Educação
Universidade Federal do Rio de Janeiro
Instituto de Bioquímica Médica Leopoldo de Meis

Orientador: Dr. Francisco Prosdocimi

Brasil
Março de 2019

Gabriel Alves Vieira

FILOGENÔMICA MITOCONDRIAL SEM CUSTOS: UMA PROVA DE CONCEITO COM FORMIGAS DA SUBFAMÍLIA PSEUDOMYRMECINAE (HYMENOPTERA : FORMICIDAE)/ Gabriel Alves Vieira. – Brasil, Março de 2019-

123p. : il. (algumas color.) ; 30 cm.

Orientador: Dr. Francisco Prosdocimi

Tese (Doutorado) – Ministério da Educação

Universidade Federal do Rio de Janeiro

Instituto de Bioquímica Médica Leopoldo de Meis , Março de 2019.

1. Palavra-chave1. 2. Palavra-chave2. 2. Palavra-chave3. I. Orientador. II. Universidade xxx. III. Faculdade de xxx. IV. Título

FILOGENÔMICA MITOCONDRIAL SEM CUSTOS: UMA PROVA DE CONCEITO
COM FORMIGAS DA SUBFAMÍLIA PSEUDOMYRMECINAE (HYMENOPTERA :
FORMICIDAE) Gabriel Alves Vieira

Dissertação apresentada ao Programa de Pós-Graduação em Química Biológica, Instituto de Bioquímica Médica Leopoldo de Meis, Universidade Federal do Rio de Janeiro, como parte dos requisitos para obtenção do título de Mestre em Química Biológica.

Aprovado em: 26/03/2019

BANCA EXAMINADORA

Dr. Francisco Prosdocimi

Prof. Adjunto do Instituto de Bioquímica Médica Leopoldo de Meis da Universidade Federal do Rio de Janeiro – UFRJ

Dr^a. Carla Ribeiro Polycarpo

Prof^a. Associada do Instituto de Bioquímica Médica Leopoldo de Meis da Universidade Federal do Rio de Janeiro – UFRJ

Dr^a Ana Carolina Martins Junqueira

Prof^a. Adjunta do Instituto de Bioquímica Médica Leopoldo de Meis da Universidade Federal do Rio de Janeiro – UFRJ

Dr. Marcus Fernandes de Oliveira

Prof. Adjunto do Instituto de Bioquímica Médica Leopoldo de Meis da Universidade Federal do Rio de Janeiro – UFRJ

Suplente externo: Dr. Marcelo Weksler

Prof. Titular do Programa de Pós-Graduação em Zoologia do Museu Nacional da Universidade Federal do Rio de Janeiro - UFRJ

Revisor: Dr. Fernando Lucas Palhano Soares

Prof. Adjunto do Instituto de Bioquímica Médica Leopoldo de Meis da Universidade Federal do Rio de Janeiro – UFRJ

Agradecimentos

Aos meus pais que, apesar da enorme vontade de me ter ao lado deles, demonstraram seu amor ao darem apoio incondicional a um filho que, tal qual Brás Cubas, foi dominado por uma idéia fixa: ver e aprender mais sobre bioinformática e o mundo.

À Agatinha, minha mafagafa e companheira de todos os momentos, com a qual constituí família e cresci como pessoa. Abro um sorriso cada vez que encontro uma “AGATA” em alguma mitocôndria e me lembro de como sou sortudo. Completamente desprovido de dons artísticos para te escrever músicas ou poemas, o melhor que consigo fazer é dedicar todas essas sequências, junto com meu coração, a ti.

Aos meus filhos: Seth, que vive roubando meu cobertor e pulando no meu colo sem ser convidado; e Bílquis, que só quer saber de me afifar com suas garrinhas afiadas e me trazer proteína na forma de lagartos (vivos, mortos ou em uma estranha mescla dos dois estados). Amo muito ambos e sem eles não seria plenamente feliz.

Ao Exmo. Prof. Dr. Francisco Prosdocimi, por ter lido um email (redigido com pressa e provavelmente contendo múltiplos erros ortográficos) no final de 2016 de um cara meio maluco e afobado que queria estudar bioinformática. Mais do que isso, agradeço por ele ter depositado confiança no dito cujo e aceitado orientá-lo.

A todos os meus colegas do LAMPADA/Laboratório de Genômica e Biodiversidade, em especial à Ana e Deise. Sem a ajuda e apoio de vocês, meu trabalho seria muito menos divertido /empolgante e eu certamente teria traçado arestas menos satisfatórias na tentativa de resolver esse grafo de dois anos que foi meu mestrado.

A todos aqueles que geraram e disponibilizaram os *datasets* utilizados aqui. Se eu não tivesse tido o privilégio de subir nos ombros desses gigantes, esse trabalho não existiria. Embora nós tenhamos explorado perguntas, metodologias e evidências diferentes ao nos aventurar por esses dados, acredito que comungamos de um

objetivo comum: promover o avanço de uma ciência tão aberta e colaborativa quanto possível.

À CAPES e demais agências de fomento, por darem suporte a esse trabalho

“I like the hand we’ve been dealt.”

(Uncharted 4: A Thief’s End - 2016)

Resumo

O advento do Sequenciamento de Nova Geração reduziu os custos de sequenciamento e aumentou o número de projetos genômicos para uma enorme gama de organismos, gerando uma quantidade sem precedentes de conjuntos de dados genômicos publicamente disponíveis. Muitas vezes, apenas uma pequena fração da relevância dos dados produzidos pelos pesquisadores é contemplada em seus trabalhos. Este fato permite que os dados gerados sejam reciclados em outros projetos ao redor do mundo. A montagem de mitogenomas completos é frequentemente negligenciada, embora seja útil para entender as relações evolutivas entre táxons, especialmente aqueles que apresentam baixa amostragem de mtDNA ao nível de gêneros e famílias. Esse é exatamente o caso das formigas (Hymenoptera: Formicidae) e, mais especificamente, da subfamília Pseudomyrmecinae, um grupo de formigas arbóreas com vários casos de coevolução convergente mas sem qualquer sequência mitocondrial completa disponível. Nesta dissertação reunimos, anotamos e realizamos análises genômicas comparativas de 14 novos genomas mitocondriais completos de espécies de Pseudomyrmecinae, usando apenas os conjuntos de dados públicos disponíveis no Sequence Read Archive (SRA). Utilizamos todos os mitogenomas completos disponíveis para formigas para estudar a conservação da ordem gênica e também para gerar duas árvores filogenéticas usando (i) conjunto concatenado de 13 genes mitocondriais e (ii) as seqüências mitocondriais completas. Embora as topologias das árvores tenham divergido sutilmente umas das outras (e de estudos anteriores), nossos resultados confirmam várias relações conhecidas e geram novas evidências para a classificação de grupos irmão dentro de Pseudomyrmecinae. Também realizamos uma análise de sintenia para a família Formicidae e identificamos possíveis sítios nos quais inserções nucleotídicas ocorreram em mitogenomas de formigas do gênero *Pseudomyrmex*. Usando uma abordagem de mineração de dados/bioinformática, a dissertação atual aumentou o número de genomas mitocondriais completos disponíveis para formigas de 15 para 29, demonstrando o potencial único dos bancos de dados públicos para estudos mitogenómicos. As amplas aplicações de mitogenomas na pesquisa e a presença de dados mitocondriais em diferentes tipos de dados públicos tornam a abordagem da “mitogenómica no-budget” ideal para estudos moleculares abrangentes, especialmente para taxóns subamostrados.

Palavras-chave: Pseudomyrmecinae, Mitogenômica, Mineração de dados, Bioinformática, Filogenômica, Biologia evolutiva de formigas, Sequenciamento de nova geração, Dados públicos.

Abstract

The advent of Next Generation Sequencing has reduced sequencing costs and increased genomic projects from a huge amount of organismal taxa, generating an unprecedented amount of genomic datasets publicly available. Often, only a tiny fraction of outstanding relevance of the genome data produced by researchers is used in their works. This fact allows the data generated to be recycled in further projects worldwide. The assembly of complete mitogenomes is frequently overlooked though it is useful to understand evolutionary relationships among taxa, especially those presenting poor mtDNA sampling at the level of genera and families. This is exactly the case for ants (Hymenoptera:Formicidae) and more specifically for the subfamily Pseudomyrmecinae, a group of arboreal ants with several cases of convergent coevolution without any complete mitochondrial sequence available. In this dissertation, we assembled, annotated and performed comparative genomics analyses of 14 new complete mitochondrial genomes from Pseudomyrmecinae species relying solely on public datasets available from the Sequence Read Archive (SRA). We used all complete mitogenomes available for ants to study the gene order conservation and also to generate two phylogenetic trees using both (i) concatenated set of 13 mitochondrial genes and (ii) the whole mitochondrial sequences. Even though the tree topologies diverged subtly from each other (and from previous studies), our results confirm several known relationships and generate new evidences for sister clade classification inside Pseudomyrmecinae clade. We also performed a synteny analysis for Formcidae and identified possible sites in which nucleotidic insertions happened in mitogenomes of pseudomyrmecine ants. Using a data mining/bioinformatics approach, the current dissertation increased the number of complete mitochondrial genomes available for ants from 15 to 29, demonstrating the unique potential of public databases for mitogenomics studies. The wide applications of mitogenomes in research and presence of mitochondrial data in different public dataset types makes the “no budget mitogenomics” approach ideal for comprehensive molecular studies, especially for subsampled taxa.

Keywords: Pseudomyrmecinae, Mitogenomics, Data mining, Bioinformatics, Phy-

logenomics, Ant evolutionary biology, Next Generation Sequencing, Public data.

Listas de ilustrações

Figura 1 – Principais etapas do sequenciamento Illumina	20
Figura 2 – Preparação de bibliotecas <i>paired-end</i> and <i>mate-pair</i>	23
Figura 3 – Conversão de um único spot do formato sra para fastq	27
Figura 4 – Uso de dados pareados para o fechamento de gaps	28
Figura 5 – Anotação de montagens mitocondriais e separação de spots	37
Figura 6 – Esquema representativo das montagens denovo e por referência	38
Figura 7 – Grafos de Bruijn aplicados à montagem de genomas	39
Figura 8 – Mitogenômica coomparativa de <i>Pseudomyrmecinae</i> spp.	50
Figura 9 – Sintenia mitogenômica da família Formicidae	52
Figura 10 – Árvore de concatenação gênica	53
Figura 11 – Árvore de sequência mitocondrial completa	54
Figura S1 – Visualização da cobertura de montagem para todos os 14 mitogenomas de <i>Pseudomyrmecinae</i> fornecidos pelo software TABLET.	76

Lista de tabelas

Tabela 1 – Metadados dos datasets públicos	47
Tabela 2 – Dados da montagem de mitogenomas	48
Tabela S1 – Anotação completa dos 14 genomas mitocondriais descritos nessa dissertação	78
Tabela S2 – Número de acesso, nome da espécie e referência para todos os genomas mitocondriais usados nas analyses de sintenia e filogenéticas.	96

Listas de Abreviaturas e Siglas

3'OH:	Hidroxila presente no carbono 3' de um nucleotídeo
A:	Amina (base púrica)
ATP6:	ATP sintase F0 subunidade 6
ATP8:	ATP sintase F0 subunidade 8
BLAST:	Basic Local Alignment Search Tool
BLASTp:	Protein-protein BLAST
BRIG:	Blast Ring Image Generator
BS:	Suporte de Bootstrap
C:	Citosina (base pirimídica)
COX1:	Citocromo oxidase 1
COX2:	Citocromo oxidase 2
DDBJ:	DNA Data Bank of Japan
DNA:	Ácido desoxirribonucleico
EBI:	European Bioinformatics Institute
EMBL:	European Molecular Biology Laboratory
G:	Guanina (base púrica)
Gpb:	Giga pares de base (1000000000 bp)
GTR+G+I:	Modelo de substituição nucleotídica “General Time Reversible + Gamma distributed + Invariant sites”
INSDC:	International Nucleotide Sequence Database Collaboration

kpb:	Kilo pares de base (1000 pb)
ML:	Máxima Verossimilhança
MMG:	Metagenômica mitocondrial
Mpb:	Mega pares de base (1000000 pb)
mRNA:	RNA mensageiro
mtDNA:	DNA mitocondrial
MYA:	Milhões de anos atrás
N:	Base desconhecida
NCBI:	National Center for Biotechnology Information
NGS:	Sequenciamento de Nova Geração
ORF:	Fase aberta de leitura
pb:	Pares de base
PCG:	Gene codificador de proteína
PCR:	Reação em cadeia da polimerase
RNA:	Ácido ribonucleico
RNA-Seq:	Sequenciamento de mRNA
rRNA:	RNA ribossomal
rrnS:	RNA ribossomal 16S
SBS:	Sequenciamento por síntese
SRA:	Sequence Read Archive
T:	Timina (base pirimídica)
TPA:	Banco de anotação terceirizada do Genbank

- tRNA:** RNA transportador
- trn-X:** RNA transportador relativo ao aminoácido X
- UCE:** Elementos ultra-conservados
- WGS:** Sequenciamento de genoma completo

Sumário

1	INTRODUÇÃO	18
1.1	Bases moleculares do Sequenciamento de Nova Geração II-lumina	19
1.2	O formato sra e o conceito de “spot”	24
1.3	Montagem e anotação de genomas	26
1.4	Mitogenomas: evolução, aplicações e ubiquidade em datasets	30
1.5	Formigas: relevância e disponibilidade de dados moleculares	33
1.6	A subfamília Pseudomyrmecinae: taxonomia, ecologia e evolução	34
2	OBJETIVOS	40
2.1	Objetivos gerais	40
2.2	Objetivos gerais	40
2.3	Objetivos específicos	40
3	METODOLOGIA	41
3.1	Aquisição de dados	41
3.2	Montagem e anotação do genoma mitocondrial	41
3.3	Análises filogenômicas	43
4	RESULTADOS	45
4.1	Montagem e anotação dos mitogenomas de <i>Pseudomyrmecinae</i>	45
4.2	Variação do tamanho de genomas mitocondriais e sítios de inserção no gênero <i>Pseudomyrmex</i>	49
4.3	Ordem gênica em mitogenomas de formigas	49
4.4	Análises filogenéticas de Formicidae usando dados mitogenómicos	51
5	DISCUSSÃO	55

5.1	Cobertura uniforme do mitogenoma e viés de AT	55
5.2	Mitogenômica Comparativa: tamanho do mitogenoma e análises de sintenia	56
5.3	Relações filogenômicas de Formicidae inferidas usando dados de mitogenoma	57
5.4	Mitogenômica no-budget: análises integradas entre datasets e potencial para estudos de larga-escala	64
6	PERSPECTIVAS	67
7	CONCLUSÃO	68
	REFERÊNCIAS	70
	APÊNDICES	74
	APÊNDICE A – FIGURAS E TABELAS ADICIONAIS . . .	75
	APÊNDICE B – ARTIGO PUBLICADO	97

1 Introdução

Mais de uma década após o advento do sequenciamento de nova geração (NGS, do inglês “Next Generation Sequencing”) (MARGULIES et al., 2005), é evidente que essa tecnologia incrível e madura promoveu um aumento sem precedentes na geração de dados genômicos e uma importante redução dos custos de sequenciamento (MARDIS, 2008; GOODWIN; MCPHERSON; MCCOMBIE, 2016). A fim de reunir e democratizar o acesso a dados genômicos, a International Nucleotide Sequence Database Collaboration (INSDC, <http://www.insdc.org/>) foi estabelecida em 1987. Esse esforço contínuo compreende três centros internacionais: (i) o National Center for Biotechnology Information (NCBI), (ii) European Bioinformatics Institute (EBI) e (iii) DNA Data Bank of Japan (DDBJ) (KARSCH-MIZRACHI; TAKAGI; COCHRANE, 2018; COCHRANE et al., 2006). Como parte dessa notável iniciativa, o banco de dados Sequence Read Archive (SRA) foi criado para hospedar reads e metadados de sequências brutas geradas por projetos NGS (KODAMA; SHUMWAY; LEINONEN, 2012). Disponibilizar dados brutos de sequenciamento é fundamental para a reproduzibilidade experimental (STODDEN; SEILER; MA, 2018), um pilar do esforço científico. Além disso, o SRA tem sido utilizado recorrentemente para apoiar novas pesquisas, tais como: a avaliação de polimorfismos e deleções de base única (BORDBARI et al., 2017), o teste de novos programas de bioinformática (SIMPSON et al., 2009; LANGMEAD; SALZBERG, 2012; BOLGER; LOHSE; USADEL, 2014), para avaliar os impactos de alguns procedimentos comuns sobre os dados, como o trimming (Del Fabbro et al., 2013), dentre outros estudos (KAYAL et al., 2015; BERNSTEIN; DOAN; DEWEY, 2017; LINARD et al., 2018).

A disponibilidade de dados públicos está crescendo em conjunto com os potenciais usos de tais bancos de dados para a comunidade científica. Em um período de 2 anos (agosto-2015 a agosto-2017), 3000 trilhões de pares de bases (pb) foram adicionados ao SRA, promovendo um crescimento de 233 % do repositório (Karsch-Mizrachi, Takagi Cochrane, 2017)¹. No entanto, o potencial desses dados

¹ Essa versão da dissertação foi desenvolvida apenas para experimentar as funcionalidades do

está longe de ser totalmente explorado, uma vez que os bancos de dados públicos apresentam recursos que poderiam ser usados para abordar diversos tipos de questões biológicas previamente inexploradas. Neste trabalho, nos concentramos em obter genomas mitocondriais completos usando conjuntos de dados (datasets) genômicos disponíveis publicamente.

1.1 Bases moleculares do Sequenciamento de Nova Geração Illumina

Este trabalho se valeu exclusivamente de conjuntos de dados genômicos paired-end gerados pela plataforma Illumina. Segundo o [site da companhia](#), o processo de sequenciamento gerado por essa tecnologia pode ser dividido em 4 etapas principais:

Preparação de biblioteca ([Figura 1-A](#)): O DNA do organismo de interesse deve ser clivado aleatoriamente (por processos como sonicação ou uso de nucleases inespecíficas) (Knierim et al., 2011). Após isso, fragmentos são selecionados por tamanho através de gel, se obtendo assim sequências com menor variação de tamanho. A preparação das bibliotecas consiste na modificação e amplificação desses fragmentos para que o sequenciamento seja possível. Em um primeiro passo, oligonucleotídeos específicos (chamados adaptadores) se ligam às duas extremidades do fragmento. Na preparação de uma biblioteca, geralmente são usados dois adaptadores: uma para cada extremidade do fragmento (5' e 3'). A sequência dos adaptadores varia de acordo com o modelo de sequenciador utilizado e contêm tanto regiões de ligação a primers, quanto sequências identificadoras (índices) e uma região imobilizadora, que consiste de oligonucleotídeos necessários para a ligação da sequência à flowcell (mais informações na próxima etapa). O tamanho da sequência proveniente do organismo, a qual se encontra entre os adaptadores, é chamado de insert size. Depois da adição dos adaptadores, os fragmentos podem ser

L^AT_EX. Assim sendo, algumas referências que estão no texto - as que não apresentam hyperlink - não serão encontradas na [seção 7](#). O mesmo é válido para a italização de nomes científicos e termos estrangeiros. Para a versão final dessa dissertação, acesse a cópia presente em meu [github](#) ou a disponível no [portal do Departamento de Bioquímica Médica da UFRJ](#) (ano de defesa: 2019)

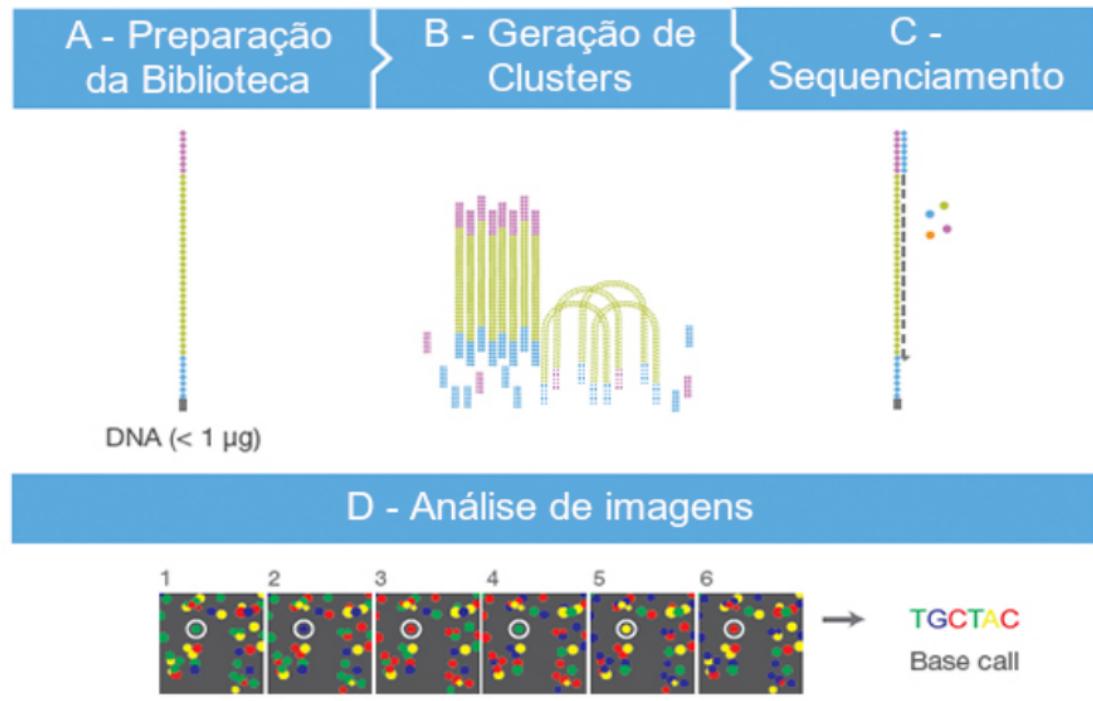


Figura 1 – Quatro principais etapas do sequenciamento Illumina.

A. O DNA obtido a partir do organismo de interesse (seja ele DNA genômico ou cDNA, no caso de RNA-Seq) é modificado de forma a gerar uma biblioteca. **B.** Os fragmentos dessa biblioteca são então ligados à flowcell e amplificados, gerando aglomerações da mesma sequência (clusters). **C.** Uma nova sequência é sintetizada pela adição de bases complementares às sequências dos clusters. Cada uma das quatro bases emite um sinal luminoso de coloração específica que é capturado em imagens. **D.** Essas imagens são analisadas de forma a determinar qual base foi adicionada em cada ciclo de sequenciamento. Adaptado de: <<https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology/2-channel-sbs.html>>

amplificados pela Reação em Cadeia da Polimerase (PCR, do inglês “Polymerase Chain Reaction”). Essa etapa de amplificação (também chamada de “enriquecimento de biblioteca”) nem sempre é realizada, mas é importante por tornar possível a realização de sequenciamentos mesmo a partir de baixas concentrações de DNA.

Ligação e amplificação dos fragmentos na flowcell (Figura 1-B): A biblioteca gerada é então dispersa pela flowcell, que consiste em uma placa

de vidro com um número definido de canais (lanes) densamente populados com oligonucleotídeos. Os adaptadores apresentam sequências que são complementares àquelas presentes nesses canais, hibridizando-se a elas e fixando à flowcell os fragmentos de DNA que serão sequenciados. Essa hibridização se dá de forma aleatória e ocorre com apenas uma das extremidades do fragmento de DNA original, de forma que, em um cenário ideal, é esperado que as sequências se liguem e fiquem posicionadas a uma distância considerável uma das outras, muito embora isso dependa de outros fatores, como a concentração da biblioteca.

Como a próxima etapa do sequenciamento depende da captação de sinal luminoso, há a necessidade de que cada sequência da flowcell gere cópias que estejam próximas a ela para a amplificação desse sinal. Primeiramente, são adicionadas polimerases que gerarão o reverso complementar das sequências de DNA originais. Então, a fita original é lavada e permanecem apenas seus complementares na placa, que passam pelo processo de “amplificação em ponte” (Mayer et al., 2011). Nesse processo, o adaptador do topo do fragmento se hibridiza a uma sequência complementar na superfície da flowcell. Isso fornece uma extremidade livre que permite a amplificação de uma nova cópia de DNA, a princípio resultando em uma dupla fita que é posteriormente desnaturada e dá origem a duas sequências complementares. Em um novo ciclo, os adaptadores dessas sequências novamente se hibridizam aleatoriamente com aqueles presentes na placa, gerando de forma clonal outras sequências. Repetido múltiplas vezes, esse processo acaba por amplificar várias cópias de um fragmento a ser sequenciado, que ficam próximos e consequentemente emitem um sinal luminoso mais forte, perceptível pelo sequenciador. A essas aglomerações de cópias de uma mesma sequência se dá o nome de cluster.

A princípio, os fragmentos em questão podem ser tanto da sequência senso (forward) quanto da antissenso (reverse). Entretanto, para evitar a captação de sinais luminosos conflitantes na próxima etapa, cada etapa de sequenciamento deve ser feita exclusivamente com sequências senso ou antissenso. Isso proporciona dois tipos principais de sequenciamento: (i) single-end, no qual as sequências antissenso são removidas e apenas as sequências senso são sequenciadas; e (ii) paired-end, onde após o sequenciamento das sequências senso, ocorre uma nova série de amplificações em ponte e o processo inverso é realizado, removendo os fragmentos forward e

mantendo os reverse. Com isso, ambas as fitas são sequenciadas e, consequentemente, as duas extremidades do fragmento. Conhecendo as extremidades da sequência e o tamanho dos fragmentos gerados, podemos estimar a distância entre as reads. A obtenção dessa informação é a maior vantagem do sequenciamento paired sobre o single-end, já que saber a distância entre sequências nos permite, dentre outros, fechar gaps reais e estabelecer a ordem e distância entre dois ou mais contigs durante a montagem. O uso dessa informação nas montagens frequentemente aumenta o tamanho dos contigs/scaffolds obtidos e consequentemente gera uma representação mais fidedigna do genoma estudado.

Em trabalhos de sequenciamento de genoma nuclear é comum se usar dados advindos de bibliotecas de diferentes tamanhos médios para se obter uma montagem de maior qualidade (RUBIN et al., 2016; WIRTHLIN et al., 2018). Vale ressaltar que o tamanho de fragmento ou insert size aceito pelo sequenciamento paired-end é limitado (200-800 pb) e para se conseguir ordenar e unir contigs muito afastados geralmente se faz necessário o uso de um terceiro tipo de sequenciamento: o mate-pair, que consegue obter sequências separadas por uma distância muito maior (2000-5000 pb). A única diferença do sequenciamento mate-pair para o paired-end está na preparação de sua biblioteca ([Figura 2](#)): enquanto o paired-end usa os fragmentos gerados na fragmentação diretamente na preparação da biblioteca, no sequenciamento mate-pair moléculas de biotina são ligadas covalentemente às extremidades de DNA (biotinilação), o que leva à circularização dos fragmentos. Posteriormente, os fragmentos são clivados na região de circularização, que é então enriquecida. Assim, as extremidades sequenciadas correspondem às sequências que estão separadas por distâncias muito maiores do que aquelas encontradas em paired-end.

Sequenciamento ([Figura 1-C](#)): O processo é baseado no “Sequenciamento por Síntese” (Sequencing-by-Synthesis ou SBS) (Mardis et al. 2008), metodologia na qual um primer se liga ao adaptador da extremidade livre na flowcell e uma DNA polimerase sintetiza uma nova fita pela incorporação de nucleotídeos marcados com fluorescência. Os quatro tipos de nucleotídeos (A, G, C, T) são adicionados à flowcell simultaneamente e cada um apresenta fluoróforos que emitem um sinal luminoso de comprimento de onda específico. Em outras palavras,

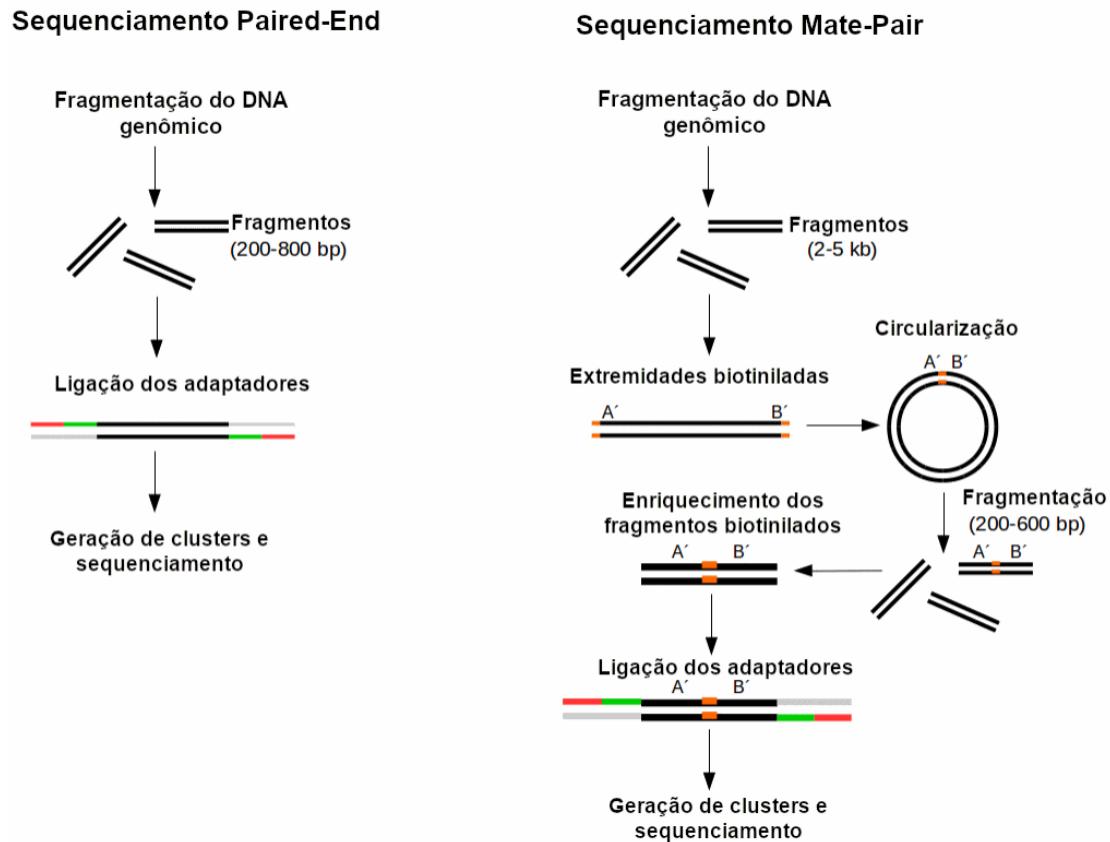


Figura 2 – Comparação entre o processo de preparação de bibliotecas *paired-end* e *mate-pair*

Adaptado de: <<https://www.ecseq.com/support/ngs/what-is-mate-pair-sequencing-useful-for>>

cada nucleotídeo emite luz de uma coloração característica, que é registrada pelo sequenciador e usada posteriormente na identificação da base incorporada. Esses nucleotídeos também são bloqueados de forma reversível em sua extremidade 3'OH, o que permite que apenas um nucleotídeo seja adicionado e identificado por ciclo de síntese. Essa incorporação de um único nucleotídeo por ciclo é necessária para, por exemplo, a determinação correta do comprimento de regiões homopoliméricas.

Após a captação do sinal luminoso, o bloqueador 3'OH é clivado e um novo nucleotídeo é incorporado. Esse processo ocorre simultaneamente para milhões ou mesmo bilhões de clusters e se repete “n” vezes até gerar uma read de tamanho “n”

por cluster (esse valor pode ir de 50 a 300, dependendo do sequenciador utilizado). A paralelização massiva dessa metodologia, apesar de gerar reads consideravelmente curtas, permite a geração de uma grande quantidade de dados por corrida de sequenciamento. Isso implica uma diminuição do custo de sequenciamento por nucleotídeo quando comparada à tecnologia de sequenciamento Sanger (Mardis, 2008; van Dijk et al., 2014; Goodwin, McPherson McCombie, 2016).

Análise de imagens (Figura 1-D): Cada vez que uma base é adicionada durante o sequenciamento, o sequenciador captura imagens com as fluorescências emitidas por uma grande quantidade de clusters. Essas imagens serão analisadas pelo software do sequenciador, que identifica a posição dos sinais luminosos emitidos por meio de coordenadas X e Y das imagens. As posições emissoras de fluorescência (correspondentes aos clusters) são chamadas de spots. O software do sequenciador analisa essa imagem e, ao avaliar o comprimento de onda e a intensidade do sinal luminoso obtido, realiza a identificação da base (base call) para cada nucleotídeo adicionado a um determinado spot. Ao final desse processo obtemos as sequências de DNA com base na leitura dessas imagens, que por conta disso são chamadas de reads ou leituras. Em sequenciamentos paired-end, as duas extremidades de um mesmo fragmento de DNA pertencerão a um mesmo spot.

O algoritmo utilizado pode apresentar maior ou menor grau de confiança em sua identificação, que é convertido em um valor de qualidade associado a cada base. Essa informação é relevante em diversos procedimentos. Por exemplo, regiões de baixa confiabilidade podem ser identificadas e removidas para que etapas subsequentes (como a montagem) sejam realizada apenas com o “gold standard” dos dados gerados, diminuindo a presença de ruído nas análises e consequentemente propiciando resultados com maior suporte.

1.2 O formato sra e o conceito de “spot”

O Sequence Read Archive disponibiliza seus dados primariamente nos formatos fastq e sra. Os arquivos sra geralmente precisam passar por uma conversão para fastq para serem usados na maior parte dos programas de bioinformática. O programa mais comumente utilizado nessa conversão é o fastq-dump, parte

do pacote de programas SRAtoolkit, distribuído pelo NCBI. Os arquivos sra apresentam duas grandes vantagens sobre os fastq para trabalhos em larga escala: i) ocupam consideravelmente menos espaço do disco rígido, sendo mais eficientes para o armazenamento de cópias locais, especialmente em se tratando de datasets massivos; e ii) o principal programa utilizado para a conversão de sra para fastq (fastq-dump) permite a manipulação de diversos parâmetros para gerar arquivos fastq modificados sem a necessidade de os baixar novamente. Várias opções podem ser utilizadas para se obter exatamente o tipo de dado desejado: tamanho mínimo das reads (-M), qualidade das reads (-R), remoção de sequências adaptadoras (-W), formatação do cabeçalho das reads (-readids e -defline-seq, importantes para que o fastq seja compatível com alguns programas), dentre outros.

Os arquivos sra são divididos em spots, não em reads. Além disso, há tanto uma opção do fastq-dump para imprimir um número fixo de spots (-X) quanto um que permite separar esses spots e armazenar as reads resultantes em um único arquivo (-split-spot) ou em dois arquivos diferentes (-split-files). Embora as documentações oficiais do NCBI não deixem explícito o que é um spot para esses arquivos, a associação com o conceito introduzido no sequenciamento Illumina é muito provável (<https://www.biostars.org/p/12047/>). Conforme supracitado, em um spot são sequenciadas tanto a read forward quanto a reverse relativas às extremidades de um fragmento de DNA. Assim sendo, o mais provável é que o conceito de “spot” em um arquivo sra paired-end seja próximo de algo como “toda a informação gerada por um spot em um sequenciamento”, o que corresponde a dizer que um spot é uma concatenação das reads forward e reverse. Isso é corroborado pelo fato de que, ao se separar os spots, o tamanho das sequências geradas caem pela metade e apresentam os característicos cabeçalhos terminando em “1” ou “2” para indicar os pares de sequências advindas do mesmo fragmento. O dataset paired-end SRR5852657, para o qual mitocôndrias de camundongo (*Mus musculus*) foram isoladas e sequenciadas (<https://www.ncbi.nlm.nih.gov/sra/?term=SRR5852657>), foi utilizado na avaliação da diferença observada entre reads geradas com relação à separação de spots (Figura 3). Esse mesmo dataset foi utilizado para gerar arquivos fastq e montar o genoma mitocondrial do camundongo usando uma referência da mesma espécie (KY018919.1). A anotação dos mitogenomas obtidos com e sem a

separação de spots da [Figura 5](#) evidencia que o uso de spots não partidos acarreta mais erros de montagem.

As sequências concatenadas em um spot, por não corresponderem àquelas encontradas no organismo de estudo, não são recomendadas para realizar montagens. Com isso, concluímos que entender o conceito de spots e utilizar a opção “split-spot” ou “split-files” do fastq-dump é fundamental para se trabalhar com dados públicos resultantes de sequenciamento paired-end.

1.3 Montagem e anotação de genomas

Como o sequenciamento gera sequências curtas, a montagem dessas reads é necessária para se obter sequências maiores ou mesmo genomas completos. O processo de montagem consiste na junção de reads contíguas guiada pela sobreposição dos nucleotídeos presentes em suas extremidades, o que permite gerar sequências maiores (chamadas de contigs). Também há os scaffolds, que consistem em contigs adjacentes cujo afastamento consegue ser estimada pela distância entre as reads geradas por sequenciamentos paired-end ou mate-pair. Como as bases que ligam os contigs não são conhecidas, elas são substituídas por “N”s, que podem ser lidos como “bases desconhecidas” ([página 28](#)).

Figura 3 – Conversão de um único spot do formato sra para fastq-dump

O dataset mitocondrial SRR5852657 foi utilizado para se obter um único spot com cabeçalho adaptado para dados paired-end (opções “-X 1” e “-read-ids”). Em **A**, nenhuma outra opção foi especificada, e o spot corresponde a uma única sequência de 202 nucleotídeos, que corresponde à concatenação das duas reads de 101 pb geradas em **B** (onde a opção “-split-spot” foi utilizada). As reads de **B** também se diferenciam pelo cabeçalho (.1 e .2), indicando que elas vieram do mesmo fragmento sequenciado, ao contrário de **A**.

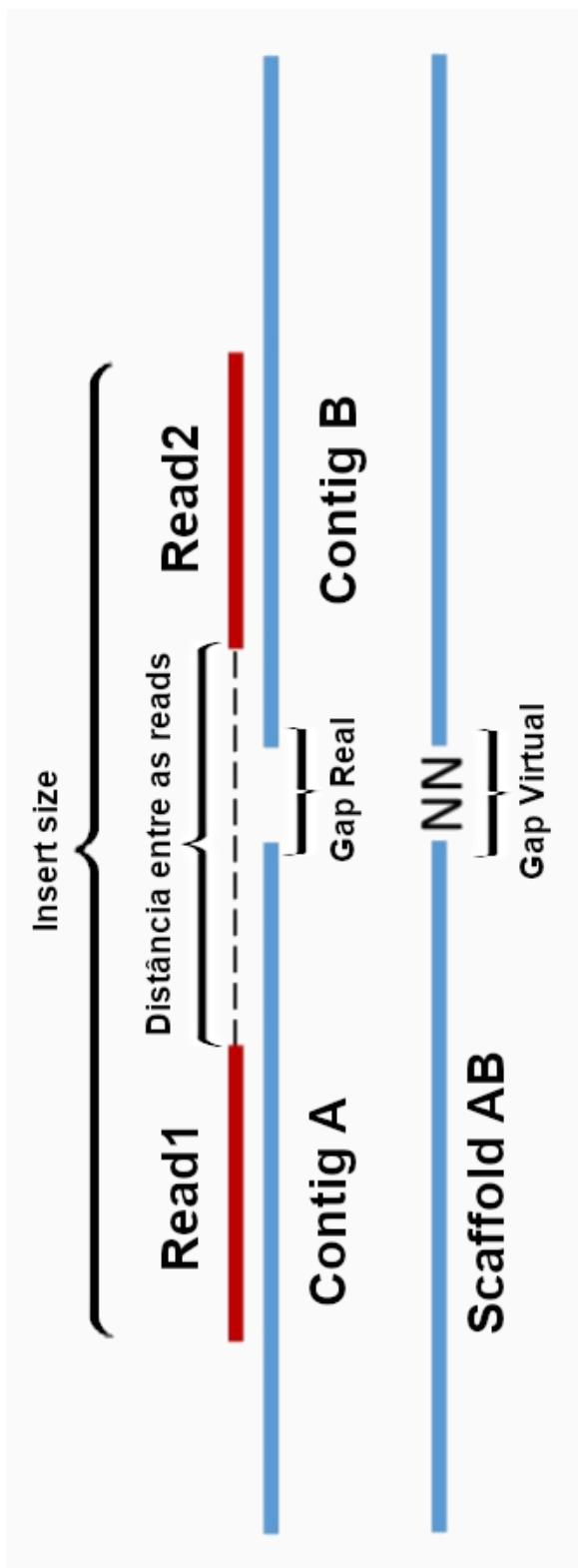


Figura 4 – Uso de dados pareados para o fechamento de gaps

Em dados pareados (paired-end ou mate-pair) é possível estimar a distância entre as reads, já que tanto o tamanho do fragmento sequenciado (Insert size) quanto o das sequências em si são conhecidas. Logo, quando as reads pareadas são mapeadas em contigs diferentes, podemos calcular a distância entre eles e uní-los em uma única sequência (scaffold). No processo de fusão, o que antes era um gap real (no qual tanto o tamanho quanto a sequência entre os contigs é desconhecida) se torna um gap virtual (onde se conhece a extensão do gap, que é preenchida por N's). Adaptado de: <http://madsalbertsen.github.io/multi-metagenome/docs/step10.html>

Há dois tipos principais de montagem (Lesk, 2014):

- (i) a montagem de novo, na qual somente a informação de sobreposição/distância entre as reads é utilizada na montagem de uma sequência (chamada de sequência consenso - [Figura 6-A](#))
- (ii) a montagem por referência, na qual informação exterior (na forma de uma sequência já existente, chamada de referência ou backbone) é utilizada na montagem. As reads são mapeadas a essa referência que idealmente deve pertencer a um organismo evolutivamente próximo do objeto de estudo, e a sobreposição entre as sequências é então utilizada para se obter o consenso ([Figura 6-B](#)).

Entretanto, geralmente priorizamos realizar montagens de novo pois elas não incluem viéses associados a referências, que levam a erros de montagem quando, por exemplo, a referência apresenta ordem gênica (sintenia) diferente daquela encontrada no genoma que está sendo montado. Além disso, ao se trabalhar com organismos não-modelo, referências próximas ao objeto de estudo comumente não estão disponíveis.

Múltiplos algoritmos foram desenvolvidos para a montagem de genomas, mas os mais amplamente utilizados são aqueles que se baseiam nos Grafos de Bruijn (Compeau et al., 2011; Lesk, 2014). Nas implementações desse algoritmo, uma subsequências das reads de tamanho k (chamadas de k-mers) são os vértices e arestas direcionadas são traçadas entre os k-mers nos quais há uma sobreposição de k-1 nucleotídeos, visitando cada nó apenas uma vez ([Figura 7](#)). Ao ser resolvido, o grafo nos dá um caminho entre os nós que indica a sucessão de sequências contíguas sobrepostas, correspondente à sequência montada. O valor de k-mer é geralmente estabelecido por aquele que irá montar o genoma, e a escolha do k-mer mais adequado é pautado pelo balanço entre especificidade e sensibilidade. Por exemplo, ao se escolher k-mers longos, se torna mais fácil estabelecer a direção correta do grafo em uma região repetitiva de DNA ou na ocorrência de erros de sequenciamento (alta especificidade), mas o número de reads sobrepostas acaba sendo muito reduzido (baixa sensibilidade), o que pode levar a uma montagem muito

fragmentada. Por outro lado, k-mers curtos geralmente não conseguem resolver corretamente regiões repetitivas do genoma, mas conseguem recrutar um número consideravelmente maior de reads para a montagem.

Múltiplas sequências podem se sobrepor em uma mesma região, e o número de reads que se sobrepõem confirmando uma determinada base é denominado cobertura (Verli, 2014). Esse conceito pode ser estendido para o cálculo da cobertura média do genoma inteiro que consiste na divisão do total de bases sobrepostas (ou seja, número de reads multiplicado pelo tamanho das mesmas) pelo total de bases da montagem. De forma geral, a cobertura está associada à qualidade da montagem, já que regiões de baixa cobertura apresentam menos dados confirmando sua sequência e consequentemente estão mais sujeitas a estarem incorretas.

A montagem, por melhor que seja, contempla apenas a sequência de nucleotídeos encontrada no objeto de estudo, o que geralmente é insuficiente. Para complementar essa informação e realizar análises adicionais que revelem novos aspectos sobre a biologia do organismo estudado, elementos relevantes dessa sequência nucleotídica (como genes codificadores de proteínas, tRNAs, rRNAs e afins) devem ser identificados e nomeados na montagem. A esse processo de adição de informação funcional às sequências montadas se dá o nome de anotação (Verli, 2014), a qual geralmente é feita em duas etapas: primeiramente, é realizada a anotação automática, que se vale de programas que, geralmente por meio do alinhamento de sequências contra bancos de dados de genes conhecidos (Wyman et al., 2004; Bernt et al., 2013), identificam a localização de vários ou mesmo de todos os genes presentes na sequência fornecida. Entretanto, muitas vezes o estabelecimento dos limites gênicos não é tão preciso, fazendo-se necessária uma segunda etapa de curadoria manual para confirmar ou refinar os resultados obtidos pela anotação automática.

1.4 Mitogenomas: evolução, aplicações e ubiquidade em datasets

Em animais, a mitocôndria é uma organela de origem materna - salvo raras exceções, como em algumas espécies de bivalves (Zouros et al., 1994; Theologidis et

al., 2008) - na qual ocorre o processo de fosforilação oxidativa, indispensável para a obtenção primária de energia em eucariotos. Essa organela está envolvida não só no metabolismo energético (BRAND, 1997), como também na apoptose (Wang, 2001), diferenciação celular (WANET et al., 2015) e várias doenças (CHAN, 2006). Apesar da gênese dessa organela ser tópico de discussão ainda hoje, classicamente sua origem comumente é explicada por meio da teoria endossimbiótica, a qual dita que a mitocôndria evoluiu a partir de uma bactéria de vida livre fagocitada e incorporada por uma célula hospedeira (GRAY, 2017). Algumas características mitocondriais corroboram essa hipótese, como a presença de dupla membrana e ribossomos similares aos encontrados em procariotos, mas nenhuma delas é tão contundente quanto a presença de DNA nessas organelas (GRAY; BURGER; LANG, 1999; KUTSCHERA; NIKLAS, 2005) . O genoma mitocondrial (também chamado de mitogenoma) também apresenta similaridades com um genoma procarioto típico: possui arranjo circular e, no caso do mitogenoma animal, cada fita é transcrita como um único mRNA policistrônico que é então clivado, dando origem a todos os elementos funcionais desse mitogenoma (BOORE, 1999).

O conteúdo gênico mitocondrial varia consideravelmente em diferentes linhagens de eucariotos, o que acredita-se ocorrer em grande parte devido à transferência de vários genes mitocondriais para o núcleo durante a evolução da organela (ADAMS; PALMER, 2003). Em mitogenomas mitocondriais animais tipicamente são encontrados 38 genes (ou features, como são chamados no processo de anotação): 13 genes codificadores de proteínas (Protein Coding Genes ou PCG's); 22 RNAs transportadores e 2 RNAs ribossomais (WOLSTENHOLME, 1992; BOORE, 1999).

Devido aos seus tamanhos pequenos (aproximadamente 16 kpb em animais), alto grau de conservação no conteúdo e ausência de íntrons, os mitogenomas são os cromossomos mais comumente seqüenciados, em especial para metazoários (SMITH, 2016). Os genomas mitocondriais são mal amostrados para muitos táxons e, portanto, nosso conhecimento atual sobre a biologia evolutiva de muitos clados poderia ser incrementado pelo uso de dados públicos. Sendo primariamente de herança materna e não recombinantes, tais sequências são frequentemente usadas para estudar biologia evolutiva (FINSTERMEIER et al., 2013; KRZEMIŃSKA et al., 2018), genética populacional (Pečnerová et al., 2017; Kilinç et al., 2018),

filogeografia (Chang et al., 2017; Fields et al., 2018), sistemática (Lin et al., 2017; Crainey et al., 2018) e conservação (Moritz, 1994; Rubinoff, 2006; Rosel et al., 2017) de vários clados (Avise, 1994), sendo especialmente convenientes para estudos com táxons subamostrados (Gotzek, Clarke Shoemaker, 2010; Duan, Peng Qian, 2016) e organismos não-modelo (Prosdocimi et al., 2012; Tilak et al., 2014; Plese et al., 2018).

Experimentos de sequenciamento de genoma completo (Whole Genome Sequencing or WGS) e projetos de sequenciamento parcial do genoma normalmente produzem reads mitocondriais suficientes para permitir a montagem de mitogenomas completos (Prosdocimi et al., 2012; Smith, 2015). Esses pequenos genomas organelares podem ser frequentemente montados em alta cobertura devido ao alto número de cópias dessa organela presentes nas células (Smith, 2015). Além disso, estudos anteriores indicam que é possível recuperar sequências mitocondriais completas e/ou quase completas a partir de dados de RNA-Seq (Tian Smith, 2016; Rauch et al., 2017, Plese et al., 2018) e dados de estratégias de amplificação que se valem do enriquecimento de regiões específicas do genoma para um sequenciamento direcionado a alguma região de interesse, como o exoma (Picardi Pesole, 2012 ; Guo et al., 2013; Samuels et al., 2013) ou os elementos ultra-conservados (Ultra Conserved Elements ou UCE) ([Do Amaral et al., 2015](#); [MILLER et al., 2016](#)). Apesar do enriquecimento das sequências de interesse, outras regiões são amplificadas aleatoriamente, incluindo as mitocondriais. Logo, o uso de datasets de sequenciamento direcionado para obter mitogenomas completos só é possível devido ao fato dessa metodologia não ser completamente eficiente na amplificação dos sítios desejados. Outra estratégia que já foi realizada com sucesso é montagem de numerosos mitogenomas completos e/ou grandes contigs mitocondriais a partir do sequenciamento de amostras contendo várias espécies (Timmermans et al., 2015; Linard et al., 2018). Essa abordagem é denominada 'mito-metagenômica' (Tang et al., 2014) ou 'metagenômica mitocondrial' (MMG) (Crampton-Platt et al., 2015). Outros trabalhos utilizaram com sucesso dados públicos para montar sequências mitocondriais (Diroma et al., 2014; Kayal et al., 2015; Linard et al., 2018) demonstrando o potencial dos bancos de dados públicos para estudos mitogenômicos. Entretanto, ainda existem várias espécies sem mitogenoma completo disponível

que apresentam um grande número de dados genômicos na base de dados do SRA.

1.5 Formigas: relevância e disponibilidade de dados moleculares

Um exemplo de amostragem deficiente de mitogenomas ocorre na família Formicidae (clado correspondente às formigas). Apesar de ser um grupo onipresente, ecologicamente dominante e hiper-diversificado (Hölldobler Wilson, 1990; Bolton, 1994), com mais de 13.000 espécies descritas (Bolton, 2012), registros de sequências mitocondriais completas são restritos a apenas 15 espécies de formigas no GenBank. Os insetos pertencentes a essa família não só constituem uma parcela significativa da biomassa nos ambientes onde ocorrem - Hölldobler Wilson (1990) estimam que eles podem chegar a constituir mais de 10 % da biomassa faunal -, como também são considerados “engenheiros ecossistêmicos” que influenciam as propriedades físicas, químicas e biológicas do solo (Jones et al, 1994; Folgarait, 1998; Lobry de Bruin, 1999). Além disso, também impactam interações multitróficas (Sanders Veen, 2011) e contribuem para a estabilidade ambiental e subsequente manutenção da prestação de serviços ecossistêmicos como a regulação climática, captura de carbono e purificação de água e ar (Sanford et al., 2009).

As formigas também apresentam grande potencial como bioindicadores em áreas de recuperação ambiental, visto que: (i) elas são abundantes e ubíquas, ocorrendo tanto em habitats intactos quanto em áreas perturbadas (Majer, 1983; Hölldobler Wilson, 1990; Hoffman et al, 2000); (ii) sua coleta é relativamente simples (Majer, 1983; Lopes Vasconcelos, 2008); e (iii) são muito sensíveis a variações ambientais, apresentando respostas claras a essas (Majer, 1983; Hoffman et al., 2000). Estudos usando bioindicadores para analisar distúrbios ambientais comumente são realizados através da comparação da riqueza de espécies em regiões perturbadas e não-perturbadas (Pearson Carroll, 1998; Hoffman et al., 2000). Portanto, a eficácia do uso desses organismos como bioindicadores está diretamente associada à precisão da identificação das espécies. Por último, é necessário ressaltar que as formigas também podem participarativamente do processo de regeneração de áreas degradadas (Lobry de Bruin, 1999; Gallegos et al., 2014).

Várias formigas têm sua notoriedade e importância indissociavelmente ligadas às suas interações com plantas, em especial àquelas que detêm status de pragas agrícolas. De fato, há espécies que são pragas de culturas economicamente importantes, como as formigas cortadeiras (popularmente conhecidas como saúvas), que cortam folhas para cultivar fungos (que constituem a fonte de alimento dessas formigas) e geram prejuízos enormes em várias monoculturas, em especial as de eucalipto (DellaLucia et al., 2014). Entretanto, mesmo as poucas espécies de saúvas que causam dano econômico também impactam positivamente seu ambiente - reduzindo o tempo de ciclagem de nutrientes e aumentando a aeração do solo, por exemplo - o que coloca em xeque sua classificação como “praga” (Fowler et al., 1989; Jones, 1994). Também há espécies de formigas que são utilizadas como agentes de controle biológico no manejo de pragas (Way Khoo, 1992; Philpott Armbrecht, 2006) e a maioria das relações entre formigas e plantas são na verdade mutualísticas (Cannicci et al., 2008), não predatórias. Logo, informações sobre esse grupo é relevante no estudo e preservação desses e de seus ecossistemas.

Evidência molecular, em especial dados de sequenciamento de UCE, já foi usada para estudar a filogenia das formigas nos últimos anos (Blaimer et al., 2015; Ward Branstetter, 2017; Brasnsteller et al., 2017). Entretanto, são raras as tentativas que visam recuperar sequências mitocondriais com base nesses dados gerados para outros propósitos (Ströher et al., 2017) e usar essas informações para entender melhor as relações evolutivas para o clado. Estudos com esse escopo também adicionam evidências moleculares úteis à identificação de espécies, o que no caso das formigas pode potencializar seu uso como bioindicadoras.

1.6 A subfamília Pseudomyrmecinae: taxonomia, ecologia e evolução

Um grupo particular de formigas que sofre de má amostragem de mitogenômica é a subfamília Pseudomyrmecinae, que contém três gêneros: (i) *Pseudomyrmex*, encontrado no Novo Mundo e que possui ≈ 137 espécies, a maioria das quais pode ser classificada em um dos dez grupos de espécies, delimitados com base em caracteres morfológicos (Ward 1989, 1993, 1999, 2017); (ii) *Tetraponera*, com ≈ 93 espécies

e de distribuição paleotropical; e (iii) Myrcidris, gênero sul-americano que possui apenas uma espécie descrita, *Myrcidris epicharis* (Ward Downie, 2005; Bolton, 2012; Ward, 2017).

De acordo com [Janzen \(1966\)](#) e [Ward \(1991\)](#), existem dois grupos ecológicos conhecidos de Pseudomyrmecinae:

Grupo 1 Composto de espécies arbóreas generalistas. Nidificam em galhos mortos de vários tipos de plantas e são geralmente passivas em relação a objetos externos

Grupo 2 Formigas especializadas em colonização de plantas. habitantes obrigatórias de cavidades ocas em tecidos vivos de plantas. Essas cavidades são estruturas vegetais especializadas (chamadas de domatia) que provêm abrigo e proteção às formigas, que são freqüentemente agressivas em relação a outros insetos ou plantas.

As formigas do Grupo 2 fornecem proteção contra herbivoria e competição para sua planta hospedeira em um caso típico de mutualismo ([JANZEN, 1966](#); [WARD, 1991](#)). Estudos anteriores usando dados morfológicos e moleculares (Ward, 1991; Ward Downie, 2005) sugerem que esse tipo de mutualismo das espécies do Grupo 2 evoluiu independentemente pelo menos 12 vezes na subfamília Pseudomyrmecinae. Por exemplo, o trabalho de Chomicki e colaboradores (2015) chama a atenção para o fato de formigas do gênero *Pseudomyrmex* desenvolverem comportamentos similares por convergência, apesar de evoluírem com diferentes hospedeiros vegetais. Esse mesmo trabalho elenca as plantas comumente associadas a essas formigas: (i) leguminosas (Fabaceae) dos gêneros *Vachellia* (anteriormente parte do gênero *Acacia*, cujo mutualismo com as *Pseudomyrmex* levou essas formigas a serem conhecidas como “acacia-ants”), *Platymiscium* e *Tachigali*; e (ii) as poligonáceas dos gêneros *Triplaris* e *Ruprechtia*.

Casos de evolução convergente são freqüentemente caracterizados usando abordagens filogenéticas (Ward Branstetter, 2017). Análises evolutivas de sequências mitocondriais geralmente permitem um melhor entendimento sobre a história de clados superiores ao nível de ordem (Mao,Gibson Dowton, 2015) e família (Miya et al., 2003; Kayal et al. 2015). A mitogenômica já foi usada para resolver relações

evolutivas em clados superiores de insetos (subfilo Hexapoda) (Mao, Gibson Dowton, 2015; Bourguignon et al., 2016). Assim sendo, a subfamília Pseudomyrmecinae é uma ótima candidata a ser estudada usando filogenômica mitocondrial, já que é um clado superior de Formicidae e apresenta diversos casos de coevolução.

Diversos estudos moleculares foram realizados no gênero *Pseudomyrmex*. Estes geralmente abordam questões coevolutivas, como o impacto de associações mutualistas na taxa de evolução do genoma (Rubin Moreau, 2016), ou a caracterização de associações entre formigas e plantas através do estudo de relações filogenéticas e biogeografia (Chomicki, Ward Renner, 2015; Ward Branstetter, 2017). No entanto, análises completas de mitogenomas nunca foram realizadas para Pseudomyrmecinae devido à ausência de sequências mitocondrais completas para o clado. Na atual abordagem de “mitogenômica *no-budget*” (definida aqui como o uso de dados públicos de NGS para montar grandes sequências mitocondriais indisponíveis em bancos de dados), usamos dados genômicos publicamente disponíveis gerados por outros trabalhos (Tabela 1) para montar e analisar a sequência mitocondrial completa para 14 representantes da subfamília Pseudomyrmecinae: 12 espécies do gênero *Pseudomyrmex* e duas espécies de *Tetraponera*. O tempo de divergência estimado entre *Pseudomyrmex* e *Tetraponera* é estimado em \approx 95.8 MYA (Million Years Ago), de acordo com Gómez-Acevedo et al. (2010), apoiando um evento de vicariância durante a separação da América do Sul da África. Assim sendo, este trabalho também testa a eficiência da abordagem mitogenômica em resolver relações filogenéticas entre clados que divergiram em um passado relativamente distante.

Apresentamos as primeiras sequências mitocondriais completas para esta subfamília e realizamos análises evolutivas delas em conjunto com todos os outros mitogenomas de Formicidae disponíveis, tentando entender melhor as relações de grupos-irmãos dentro deste clado altamente diverso. A dissertação atual apresenta novos genomas mitocondriais para espécies de formigas que cobrem cinco dos 10 grupos de espécies de *Pseudomyrmex* e quase duplica o número de genomas mitocondriais disponíveis para formigas, aumentando este número de 15 para 29.

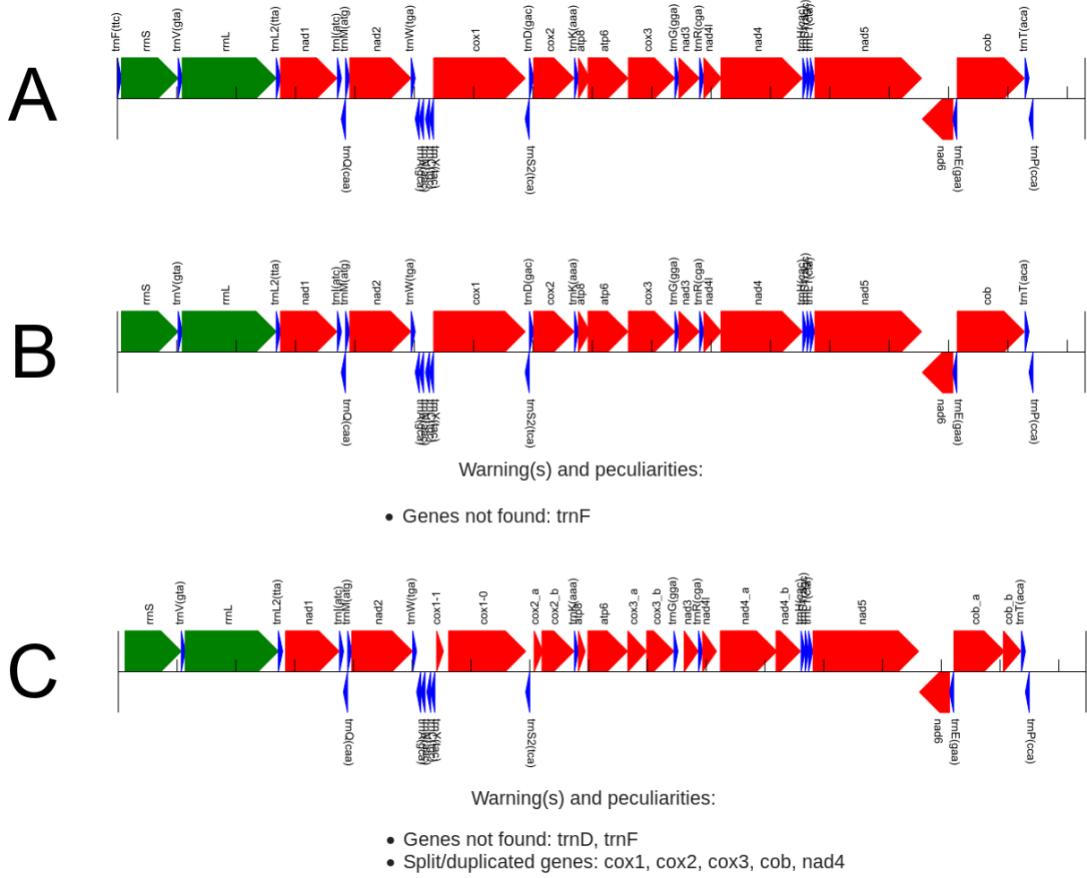


Figura 5 – Anotação de montagens mitocondriais com e sem a separação de spots

Por meio do software MIRA (Chevreux et al., 1999), realizamos montagens mitocondriais por referência para averiguar o impacto das separação dos spots nas montagens. A referência utilizada e os mitogenomas gerados foram anotados utilizando o MITOS Web Server (BERNT et al., 2013). Podemos observar que a montagem **A** (obtida através do Genbank – Accession number [KY018919.1](#)) não apresenta nenhum problema em sua anotação. As duas outras montagens foram realizadas utilizando a sequência de **A** como backbone. Na montagem **B** (obtida usando spots partidos), o tRNA da fenilalanina (trn-F) não pôde ser identificado. Por último, temos que a anotação da montagem **C** (spots inteiros) apresenta mais problemas: além da ausência de trn-D e trn-F, há vários genes duplicados. De maneira geral, a anotação de **B** é muito mais similar à da referência, o que sugere que a separação de spots é importante para a obtenção de sequências de maior qualidade.

A) MONTAGEM DE NOVO

```

ACCGGATTTCAGGTTACCACCG
GCGATTTCAGGTTACCACCGCG
GATTCAAGGTTACCACCGCGTA
TTCAGGTTACCACCGCGTAGC
CAGGTACCCACCGTAGCGC
GGTTACCAACCGCGTAGCGCAT
TTACACCGCGTAGCGCATTACA
ACCACCGCGTAGCGCATTACAA
CACCGCGTAGCGCATTACACAGA
CGCGTAGCGCATTACACAGATT
CTAGCGCATTACACAGATTAG
TAGCGCATTACACAGATTAG

```

Consenso

Reads sobrepostas

B) MONTAGEM POR REFERÊNCIA

```

...ACGTACGGTTACACAAACCGTTGCACGTACGTAAACCGTTGTGACG...
TTACACAAATCCCCTTCGCA
TACACAAATCCCCTTCGCA
ACACAAATCCCCTTCGCA
CACAAACCCGTTTCGCA
CAATCCCCTTCGCA
AAATCCCCTTCGCA
ATCCCCTTCGCA

```

Referência

Reads sobrepostas

Consenso

Figura 6 – Esquema representativo das montagens denovo e por referência

A. A montagem de novo se vale exclusivamente da informação contida no sequenciamento. Adaptado de <<https://contig.files.wordpress.com/2010/02/alignment1.jpg>>. **B.** A montagem por referência utiliza informação externa na forma de uma sequência previamente sequenciada. Adaptado de Wajid Serpedin (2014). Ambas geram a sequência final com base no consenso das sobreposições.

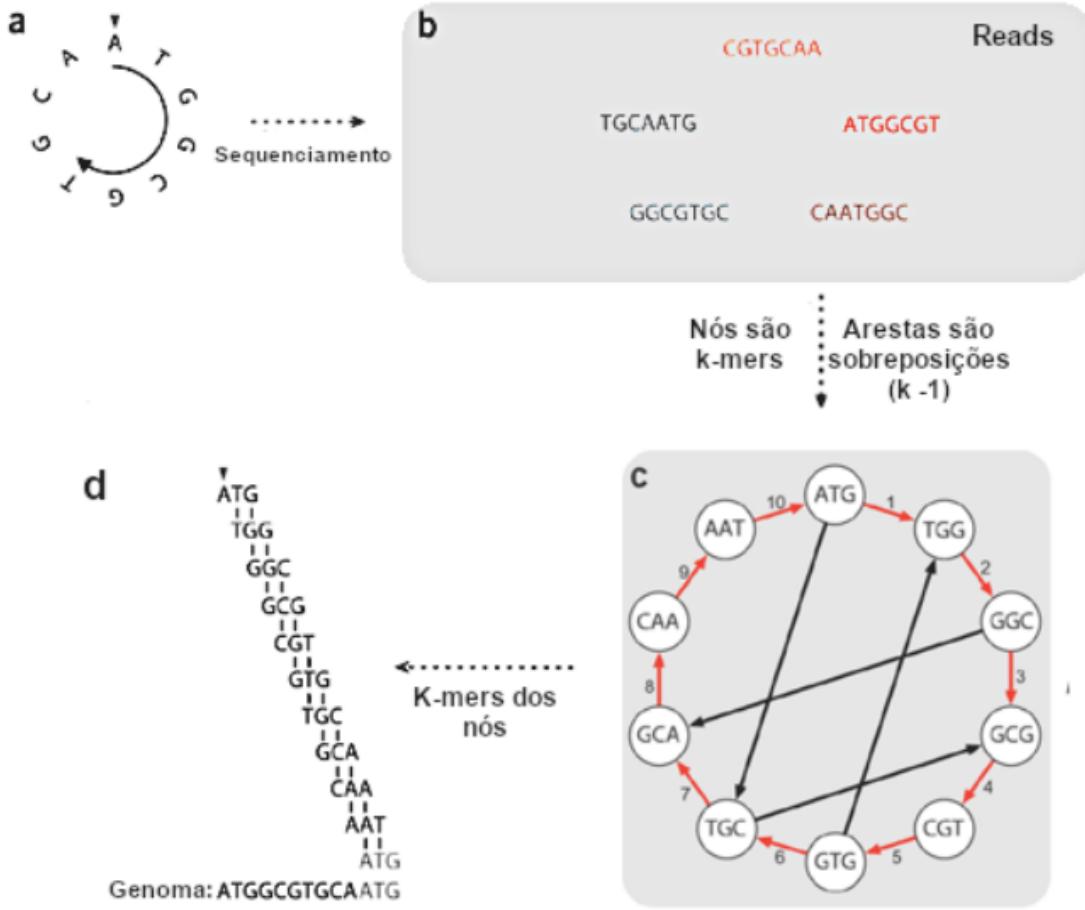


Figura 7 – Grafos de Bruijn aplicados à montagem de genomas

A. Um pequeno genoma circular é sequenciado. **B.** As reads são geradas (cada uma com sete nucleotídeos). **C.** Com $k=3$, um grafo de Bruijn é construído, onde os nós são sequências de três nucleotídeos (3-mers) e as arestas são traçadas ao se encontrar sobreposições de dois ($k-1$) nucleotídeos entre as sequências. Repare que o baixo valor de k -mer utilizado facilita a sobreposição entre vários nós, o que faz com que múltiplos caminhos sejam possíveis nesse grafo (arestas pretas) e dificultam a montagem. **D.** Ao final, o caminho que visita cada nó apenas uma vez (arestas vermelhas) nos dá a sequência montada. Adaptado de Compeau et al. (2011).

2 Objetivos

2.1 Objetivos gerais

2.2 Objetivos gerais

- Obter genomas mitocondriais completos usando dados públicos presentes no Sequence Read Archive
- Estudar a subfamília de formigas Pseudomyrmecinae por meio da análise de sequências mitocondriais

2.3 Objetivos específicos

- Montar, anotar e disponibilizar mitogenomas da subfamília Pseudomyrmecinae
- Realizar estudos de genômica comparativa que englobem os mitogenomas completos de formigas presentes no Genbank e as sequências geradas nessa dissertação
- Construir árvores filogenômicas para estudar a evolução da família Formicidae, com ênfase na subfamília Pseudomyrmecinae

3 Metodologia

3.1 Aquisição de dados

Quatorze datasets paired-end obtidos por sequenciamento Illumina foram baixados do EMBL Nucleotide Archive (<https://www.ebi.ac.uk/ena>) no formato de arquivo SRA (consulte a Tabela 1). Esses conjuntos de dados contêm tanto reads mitocondriais quanto nucleares que foram convertidas para FASTQ usando o software fastq-dump (com parâmetros `-readids` e `-split-files`) que integra o pacote SRAtoolkit.2.8.2.

3.2 Montagem e anotação do genoma mitocondrial

Cada dataset completo foi usados como entrada para a montagem de novo usando NOVOPlasty2.6.3 (Dierckxsens et al., 2016) com os valores padrão dos parâmetros. Já que NOVOPlasty foi nosso montador principal e [Dierckxsens, Mardulyn e Smits \(2016\)](#) recomendam o uso de dados não trimados nesse software, optamos por usar os datasets brutos como entrada para a montagem. A única exceção foi o dataset de *Tetraponera rufonigra*, que teve de ser ajustado com o software Trimmomatic v.0.36 (Bolger, Lohse & Usadel, 2014) para produzir sequências com o mesmo comprimento em pares de base. Este controle dos dados foi realizado ao ajustar o parâmetro MINLEN do Trimmomatic para 125, que é o comprimento das maiores reads encontradas no dataset. Com isso, descartamos sequências menores e mantemos apenas aquelas de tamanho máximo para serem usadas como entrada na montagem inicial. Montagens NOVOPlasty precisam de uma sequência (denominada seed ou semente) que é utilizada para identificar uma read mitocondrial do dataset, a qual por sua vez será usada para iniciar a montagem (Dierckxsens et al., 2016). As seeds foram selecionadas utilizando sequências de COX1 (Citocromo Oxidase I) da mesma espécie (quando disponíveis) ou utilizando regiões de COX1 de espécies proximamente relacionadas. As montagens de mitogenoma preliminares realizadas pelo NOVOPlasty foram utilizadas como

referência para uma segunda etapa de montagem utilizando o software MIRA v.4.0.2 com parâmetros padrão (Chevreux et al., 1999). NOVOPlasty não gera um arquivo de alinhamento mostrando as sequências mapeadas à montagem, então MIRA foi utilizado para mapear reads brutos à montagem preliminar e permitir a análise de cobertura da sequência mitocondrial consenso nos próximos passos. Quando a primeira montagem não gerou o mitogenoma completo, nós usamos o MITObim v.1.9 (Hahn et al., 2013) sem alterar seus parâmetros. Este programa realiza montagens MIRA sucessivas para estender o(s) contig(s) mitocondrial(ais) e fechar pequenas lacunas da montagem, gerando a versão final e circularizada do genoma mitocondrial.

O software Tablet versão 1.17.08.17 (Milne et al., 2012) foi usado com valores paramétricos padrão para verificar a cobertura de reads e a circularização dos mitogenomas completos. O processo de anotação automática foi realizado usando MITOSWebServer (Bernt et al., 2013) sem a alteração de parâmetros. Em seguida foi realizada uma etapa de curadoria manual com o software Artemis v.17.0.1 (Carver et al., 2012) usando a tabela de código genético número cinco (correspondente à mitocôndria dos invertebrados) para identificar os limites das fases abertas de leitura (Open Reading Frames ou ORF's).

Já que o genoma mitocondrial dá origem a um grande mRNA policistrônico que então é clivado (Boore, 1999), sobreposições gênicas poderiam incorrer na formação de proteínas, rRNAs ou tRNAs não funcionais. É então razoável pensarmos que mitogenomas nos quais as features não se sobreponem são energeticamente mais econômicas para a célula e, consequentemente, foram selecionados ao longo da evolução. Partindo dessa premissa, tentamos ao máximo evitar sobreposições durante a anotação das sequências mitocondriais. Assim sendo, os limites gênicos dos tRNAs e rRNAs foram mantidos de acordo com os resultados do MITOS Web Server, salvo quando encontrada sobreposição entre duas features (gene codificador de proteína - PCG, tRNA ou rRNA) na mesma fita. Nesse caso, os nucleotídeos sobrepostos foram retirados de uma das features para que a sobreposição fosse completamente removida. O D-loop não foi explicitamente anotado, já que seu caráter hipervariável e de baixa complexidade (Moritz, 1994; Vanecek et al., 2004; Zhang & Hewitt, 1997) torna difícil estabelecer limites precisos para essa região,

em especial quando não se tem uma referência próxima. Entretanto, com base na análise comparativa de sintenia em Formicidae e no fato do D-loop ser geralmente a maior região intergênica do genoma mitocondrial (Liu et al., 2015; Zhang et al., 2016; Huang et al., 2017), identificamos que essa região variável se encontra entre o rrnS e trn-M.

Para os PCGs, em vários casos se fez necessário expandir a anotação fornecida pelo MITOS Web Server de forma a englobar a maior ORF que não apresente sobreposição com outras features na mesma fita. Então, essa ORF foi utilizada como entrada na versão online do BLASTp (Altschul et al., 1997), sendo alinhada contra o banco de sequências pertencentes a família Formicidae do GenBank. As informações obtidas por esse alinhamento contra sequências mitocondriais de outras formigas nos permitiu considerar a conservação de sequências entre as espécies e determinar o tamanho mais provável da proteína, refinando a anotação. Seguindo esse procedimento, nós alcançamos uma decisão racional, com base em genômica comparativa, sobre os limites gênicos. O conteúdo de AT para (i) o genoma mitocondrial completo; e (ii) a região intergênica que contém o D-loop foram calculados usando o programa online OligoCalc (Kibbe, 2007) com os valores padrão para seus parâmetros.

3.3 Análises filogenômicas

As relações filogenéticas de Formicidae foram reconstruídas usando (i) os 14 mitogenomas completos por nós produzidos juntamente com (ii) todos os outros 15 genomas mitocondriais completos atualmente disponíveis para o clado; e (iii) dois mitogenomas de abelhas (família Apidae) utilizados como grupos externos. Duas árvores filogenéticas foram construídas usando (i) toda a sequência mitocondrial e (ii) o conjunto de genes concatenados de todos os 13 genes codificadores de proteínas (PCGs). Para o primeiro, editamos manualmente as sequências para iniciar no gene COX1 quando necessário e alinhamos os mitogenomas inteiros usando o software ClustalW v.2.1 usando os parâmetros padrão (Thompson, Gibson & Higgins, 2003). Para o segundo, alinhamos e concatenamos os nucleotídeos de todos os PCGs usando o programa Phylomito ([<https://github.com/igorrcosta/phylomito>](https://github.com/igorrcosta/phylomito)) sem

alteração de seus parâmetros. Modeltest (Posada & Crandall, 1998) foi executado através do software MEGA7 (Kumar, Stecher & Tamura, 2016) para os dois conjuntos de dados e identificou o modelo GTR + G + I como o modelo de substituição de nucleotídeos que melhor explica a variação das sequências. As sequências alinhadas foram usadas como entrada para uma análise de Máxima Verossimilhança (Maximum Likelihood ou ML) usando o MEGA7. A reamostragem foi realizada por bootstrap usando 1000 réplicas. O software BRIG (Blast Ring Image Generator) v.0.95 (Alikhan et al. 2011) foi utilizado com valores paramétricos padrão para comparar e visualizar todos os mitogenomas de Pseudomyrmecinae produzidos aqui.

4 Resultados

4.1 Montagem e anotação dos mitogenomas de *Pseudomyrmecinae*

Os 14 conjuntos de dados genômicos usados para montar o mitogenoma completo das formigas pertencentes à subfamília Pseudomyrmecinae foram baixados do banco de dados SRA (Tabela 1). Dois tipos de datasets diferentes foram usados: (i) Sequenciamento do Genoma Completo (WGS), que frequentemente continha uma quantidade maior de dados de sequenciamento, totalizando 212,7 Giga pares de base (Gpb) para seis espécies (de acordo com a informação fornecida pelo SRA); uma média de 35,45 Gpb por espécie (Rubin & Moreau, 2016); e (ii) experimentos de UCE, para os quais realizamos o download de 5,94 Gpb para oito espécies; uma média de 742,5 Mpb por espécie (Branstetter et al., 2017; Ward & Bristetter, 2017).

O dataset completo baixado para cada espécie foi usado como entrada para uma montagem de novo usando o montador NOVOPlasty. Após essa primeira etapa de montagem do genoma, usamos um subconjunto contendo dois ou quatro milhões de reads de sequenciamento como entrada para uma segunda etapa de montagem do genoma usando o software MIRA. Este procedimento foi realizado para mapear as reads na montagem preliminar e melhorar a qualidade do mitogenoma. Para alguns mitogenomas, o MIRA não conseguiu produzir o genoma mitocondrial completo e circularizado. Nesse caso, uma terceira etapa de montagem foi necessária, na qual o maior contig gerado pelo MIRA foi usado como backbone para concluir a montagem usando o MITObim (Tabela 2). Esta metodologia foi capaz de montar a mitocôndria completa de todas as espécies de Pseudomyrmecinae analisadas, com exceção da *T. aethiops*, para a qual tivemos que usar o dataset completo como entrada para MIRA e MITObim ao invés de filtrar o subconjunto de reads na segunda etapa. O uso de múltiplas estratégias para montar as sequências mitocondriais completas era esperado, já que dados de NGS são variáveis entre

diferente espécies e corridas de sequenciamento. Além disso, os datasets usados aqui vieram de trabalhos com abordagens experimentais diferentes, o que provavelmente potencializou a variabilidade de conjuntos de dados já muito díspares entre si. Os 14 genomas mitocondriais construídos aqui foram verificados quanto à circularidade e confirmaram apresentar, como esperado para metazoários, 13 genes codificadores de proteínas, 22 tRNAs, dois rRNAs e uma região de controle (Wolstenholme, 1992; Boore, 1999). A anotação do genoma para todos os mitogenomas completos é apresentada na Tabela S1. Todos os genomas mitocondriais produzidos aqui foram submetidos ao GenBank sob o banco de dados de anotação terceirizada (Third Party Annotation ou TPA) ([COCHRANE et al., 2006](#)) que forneceu números de acesso para cada genoma mitocondrial, permitindo a visualização e download das sequências (Tabela 1).

De acordo com as estimativas fornecidas pelo software TABLET, a cobertura média de reads para os mitogenomas variou entre 85x e 292x para as sequências mitocondriais nas quais um subconjunto dos dados foi usado. Para *T. aethiops*, a cobertura foi maior, dado que o dataset inteiro foi utilizado (712x). Observou-se uma distribuição uniforme da cobertura ao longo dos mitogenomas (Figura S1), exceto em casos nos quais segmentos ricos em AT da região de controle apresentaram baixa cobertura, geralmente próximos a seqüências poli-T.

Tabela 1 – Informação acerca dos 14 datasets genômicos baixados do Sequence Read Archive para a montagem de mitogenomas completos de formigas da subfamília Pseudomyrmecinae

Species name	Bioproject	Experiment	Biosample	SRA Run number	Dataset type	# Sequencing Reads	# Downloaded Bases	Reference
Pseudomyrmex concolor	PRJNA268384	SRX831102	SAMN03275516	SRR1742927	WGS	359,475,424	35.9 Gpb	Rubin & Moreau, 2016
Pseudomyrmex dendroicus	PRJNA268384	SRX831097	SAMN03275515	SRR1742922	WGS	366,341,280	36.6 Gpb	Rubin & Moreau, 2016
Pseudomyrmex elongatus	PRJNA268384	SRX831106	SAMN03275518	SRR1742975	WGS	409,687,406	41 Gpb	Rubin & Moreau, 2016
Pseudomyrmex feralis	PRJNA357470	SRX2424867	SAMN06141944	SRR5112519	UCE	4,552,328	569 MpB	Ward & Branstetter, 2017
Pseudomyrmex ferrugineus	PRJNA357470	SRX2424886	SAMN06141956	SRR5112538	UCE	5,274,142	659.3 MpB	Ward & Branstetter, 2017
Pseudomyrmex flavicornis	PRJNA268384	SRX831107	SAMN03275519	SRR1742976	WGS	290,503,558	29.1 Gpb	Rubin & Moreau, 2016
Pseudomyrmex gracilis	PRJNA268384	SRX831110	SAMN03219222	SRR1742979	WGS	358,526,654	35.9 Gpb	Rubin & Moreau, 2016
Pseudomyrmex janiensi	PRJNA357470	SRX2424860	SAMN06141954	SRR5112512	UCE	3,720,456	465.1 MpB	Ward & Branstetter, 2017
Pseudomyrmex pallidus	PRJNA268384	SRX831105	SAMN03275517	SRR1742947	WGS	342,184,040	34.2 Gpb	Rubin & Moreau, 2016
Pseudomyrmex particeps	PRJNA357470	SRX2424875	SAMN06141966	SRR5112527	UCE	7,821,658	977.7 MpB	Ward & Branstetter, 2017
Pseudomyrmex peperi	PRJNA357470	SRX2424871	SAMN06141946	SRR5112523	UCE	4,383,700	548 MpB	Ward & Branstetter, 2017

Tabela 2 – Informação sobre a montagem dos genomas mitocondriais das 14 espécies de formigas da subfamília Pseudomyrmecinae

Pseudomyrmecinae Species	Species group	Mitogenome TPA accession number	NOVOPlasty seed	MITObim third assembly round needed	Mitogenome Coverage	Low coverage Region	Mitogenome Size (pb)	AT content: mitogene-	AT content: D-loop region (%)
<i>P. concolor</i>	<i>P. viidus</i>	BK010475	KU985552.1	No	193.2x	No	15906	75	91
<i>P. deroicinus</i>	<i>P. viidus</i>	BK010473	KP271186.1	Yes	123.9x	No	17362	81	94
<i>P. pallidus</i>	<i>P. pallidus</i>	BK010383	KU985552.1	No	91.9x	No	17117	74	84
<i>P. elongatus</i>	<i>P. oculatus</i>	BK010474	KP271181.1	No	115.4x	No	17304	78	93
<i>P. gracilis</i>	<i>P. gracilis</i>	BK010472	FJ436821.1	No	165.5x	13761-13928	15704	77	93
<i>P. feralis</i>	<i>P. ferrugineus</i>	BK010379	FJ436819.1	No	128.0x	No	18835	78	92
<i>P. ferrugineus</i>	<i>P. ferrugineus</i>	BK010380	FJ436819.1	Yes	87.0x	No	18480	77	90
<i>P. janzeni</i>	<i>P. ferrugineus</i>	BK010382	FJ436819.1	No	125.8x	15848-15867	18380	77	89
<i>P. particeps</i>	<i>P. ferrugineus</i>	BK010384	FJ436819.1	No	126.8x	15799-15820	18524	80	90
<i>P. peperi</i>	<i>P. ferrugineus</i>	BK010385	FJ436819.1	Yes	87.4x	16006-16023	18709	78	91
<i>P. veneficus</i>	<i>P. ferrugineus</i>	BK010386	FJ436819.1	No	155.4x	15889-15928	18410	79	91
T. rufonigra	NE	BK010387	KX398231.1	No	292.2x	13889-13982	15907	74	91

4.2 Variação do tamanho de genomas mitocondriais e sítios de inserção no gênero *Pseudomyrmex*

Mitogenomas de *Pseudomyrmex* mostraram variação significativa de tamanho, indo de 15704 a 18835 pb (Tabela 2). Observamos três faixas distintas de tamanho de mitogenoma para o gênero. O tamanho do genoma mitocondrial variou de: (i) menos de 16 kpb em *P. gracilis* e *P. concolor*; (ii) entre 17 kpb e 18 kpb em *P. pallidus* e *P. dendroicus*; e (iii) maior que 18 kpb em outras espécies, pertencentes ao grupo de espécies *P. ferrugineus*. Uma análise de genômica comparativa usando o software BRIG identificou quatro regiões variáveis como segmentos putativos de inserção (Figura 8). Após a anotação do genoma, identificamos que essas supostas inserções possivelmente estão localizadas entre (i) COX2 e trn-K; (ii) ATP8 e ATP6; (iii) trn-N e trn-F; e (iv) trn-W e COX1.

4.3 Ordem gênica em mitogenomas de formigas

Apesar da amostragem limitada de genomas mitocondriais completos disponível para formigas, cinco rearranjos de sintenia ligeiramente diferentes (Figura 9) foram observados na família Formicidae. Todos os mitogenomas de *Pseudomyrmecinae* e *Dolichoderinae* analisados mostraram uma única sintenia, conservada para todas as suas espécies e compartilhada pela maioria das espécies de *Formicinae*. Essa conservação de arranjo gênico foi levado em conta para determinar a região na qual o D-loop se encontra nas espécies de *Pseudomyrmecinae*. Também observamos que os clados *Formicinae* e *Myrmicinae* apresentam um arranjo modal de sintenia sugerindo uma possível ordem gênica ancestral para cada grupo. Uma única espécie de *Formicinae* (*Camponotus atrox*) apresenta inversões entre trn-M, I e Q que diferem de outros mitogenomas desta subfamília, possivelmente representando uma variação derivada. A subfamília *Myrmicinae* também apresenta dois outros rearranjos únicos restritos a uma única espécie cada, sugerindo sintenias derivadas: (i) *P. punctatus* tem uma inversão entre trn-K e D; e (ii) *W. auropunctata* apresenta uma inversão entre trn-V e D-loop e trnY na fita oposta quando comparada com as outras.

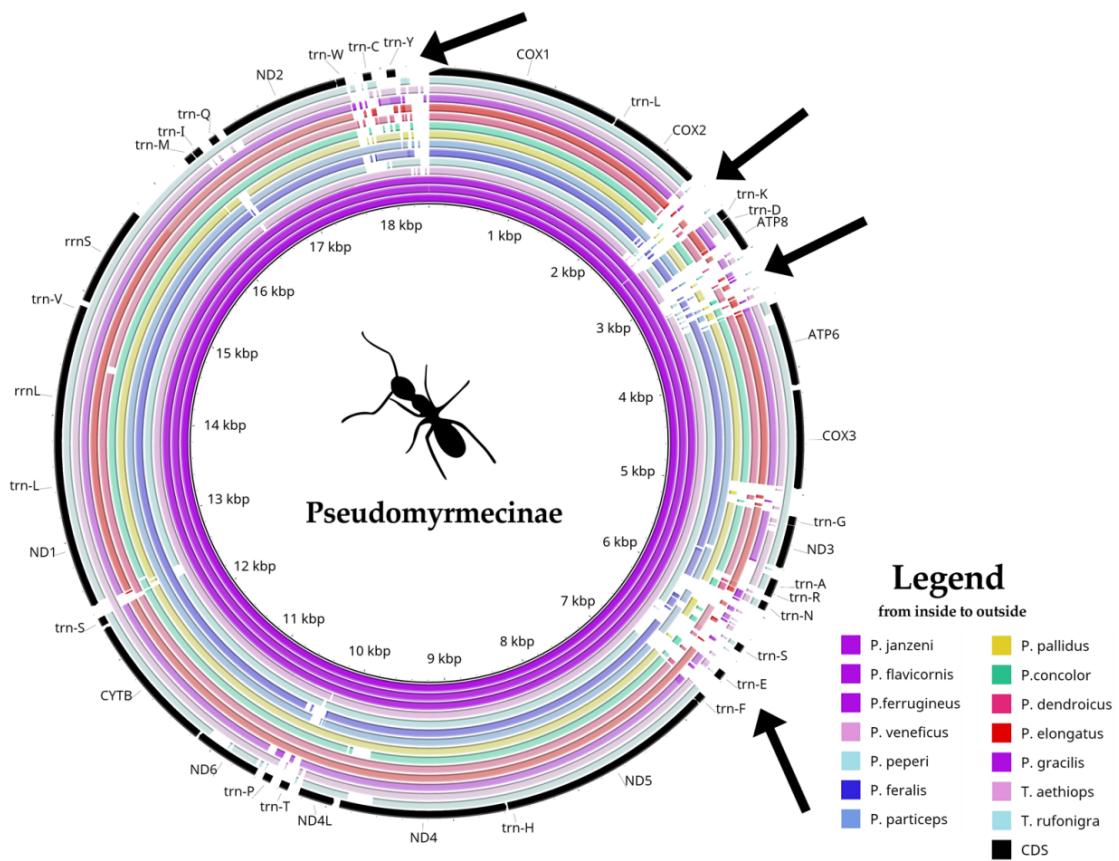


Figura 8 – Análise de genômica comparativa de todas as 14 formigas da subfamília Pseudomyrmecinae

Comparação por BLAST de todos os genomas mitocondriais de Pseudomyrmecinae contra uma referência (*Pseudomyrmex janzeni*) gerada pelo Blast Ring Image Generator (BRIG). As lacunas presentes nos anéis correspondem a regiões com menos de 50 % de identidade com a seqüência de referência. A maioria das características mitocondriais é conservada dentro do clado, embora ATP8 e alguns tRNAs (trn-S, trn-E e trn-T) tenham apresentado maior variabilidade. Quatro regiões (identificadas por setas) apresentam variações de tamanho de nucleotídeos e são encontradas entre (i) COX2 e trn-K; (ii) ATP8 e ATP6; (iii) trn-N e trn-F e; (iv) trn-W e COX1.

4.4 Análises filogenéticas de Formicidae usando dados mito- genômicos

Para avaliar a filogenia do grupo, duas árvores de Máxima Verossimilhança foram produzidas usando dados de entrada ligeiramente diferentes: (i) as sequências alinhadas e concatenadas para todos os 13 PCGs mitocondriais (Figura 10); e (ii) os genomas mitocondriais completos (Figura 11). Analisamos todas as espécies de formigas que apresentam mitogenomas completos disponíveis no Genbank (Gotzek et al., 2010; Hasegawa et al., 2011; Berman et al., 2014; Babbucci et al., 2014; Kim et al., 2015; Duan et al., 2016; Liu et al., 2016; Yang et al., 2016) e duas abelhas da família Apidae como outgroups (Crozier & Crozier, 1993; Cha et al., 2007) (ver números de acesso e referências para todas as sequências na Tabela S2. As árvores reconstruídas a partir de dados mitocondriais corroboraram a maioria das relações filogenéticas conhecidas para formigas com vários clados observados como monofiléticos com alta confiança (bootstrap = 100). Ambas as árvores apresentaram resultados semelhantes, embora diferenças possam ser observadas em vários nós quanto à topologia da árvore e/ou suporte estatístico. A principal diferença observada é que a árvore de genes concatenados exibiu todas as subfamílias como monofiléticas, enquanto Myrmicinae foi recuperada como parafilética na árvore baseada em mitogenomas completos.

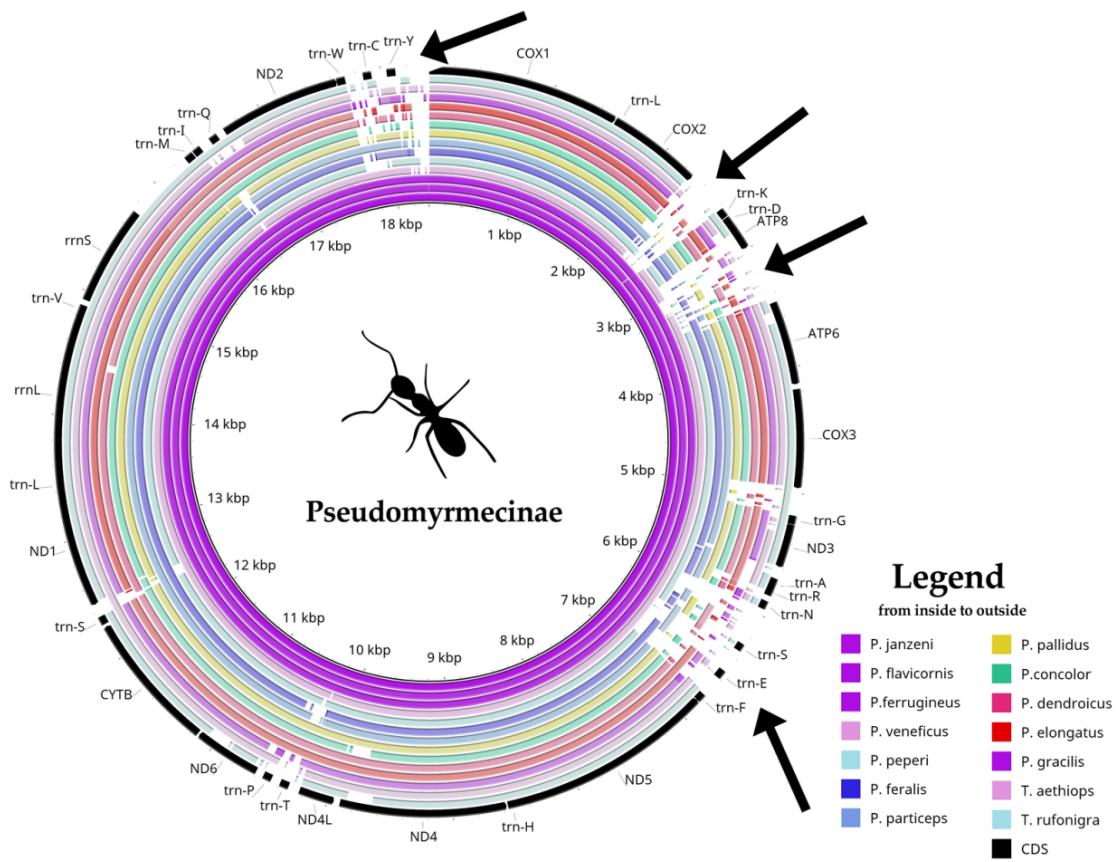


Figura 9 – Cinco sintenias observadas em mitogenomas completos da família Formicidae disponíveis no Genbank

Os dois arranjos gênicos modais estão representados dentro do retângulo horizontal e foram observados em 26 das 29 espécies analizadas: todas as Pseudomyrmecinae (14 espécies); todas as Dolichoderinae (duas espécies: *L. pallens* e *L. humile*); três das quatro Formicinae (*F. fusca*, *F. selysi* e *P. dives*) e em sete das nove Myrmicinae (*A. texana*; *C. obscurior*; *M. scabrinodis*; *S. richteri*; *S. geminata*; *S. invicta*; *V. emeryi*). Nós sugerimos que essas sintenias podem representar arranjos ancestrais para esses clados. As sintenias fora do retângulo horizontal correspondem às ordens gênicas cuja ocorrência é limitada a uma única espécie. Os retângulos verticais e linhas indicam regiões nas quais mudanças de sintenia ocorreram e tanto o asterisco (*) quanto a linha vertical no trn-Y de *W. auropunctata* indicam que essa é a única feature em uma mitocôndria de formiga que mudou sua fita codificante ao longo de sua evolução.

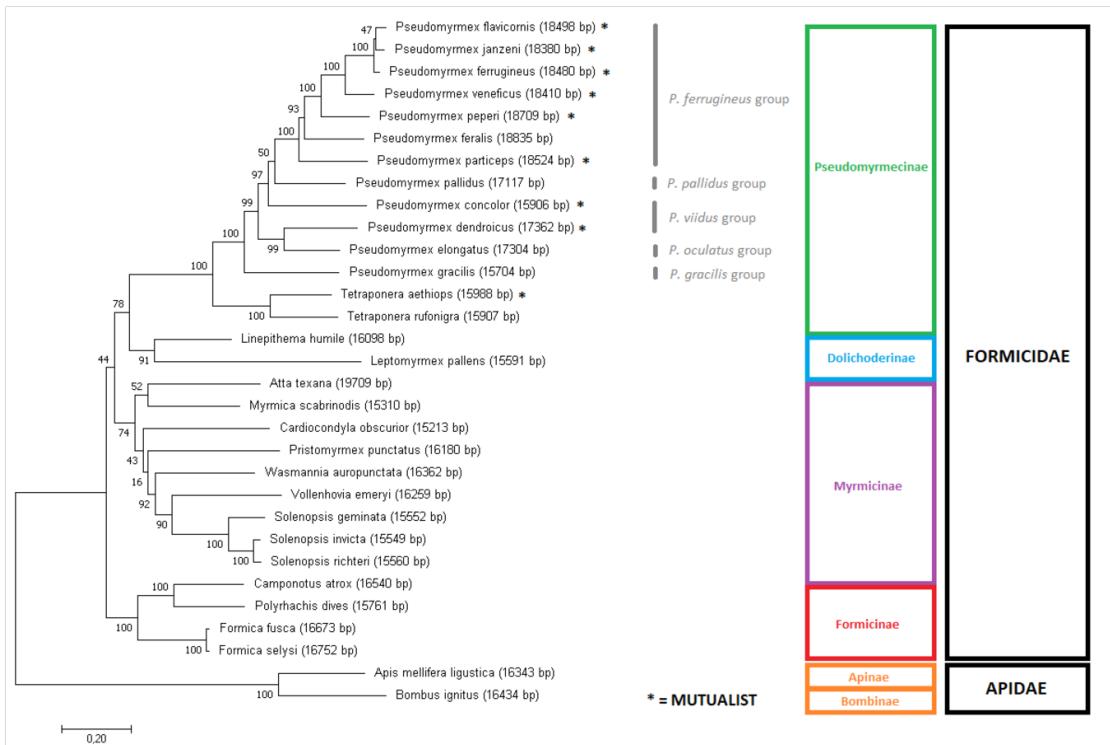


Figura 10 – Árvore filogenômica de concatenação gênica para todos os mitogenomas completos de Formicidae disponíveis no Genbank

A árvore foi construída usando as sequências nucleotídicas alinhadas e concatenadas para todos os 13 genes mitocondriais codificadores de proteínas. Modeltest identificou o GTR + G + I como o modelo de substituição mais adequado e a filogenia foi reconstruída por Maximum Likelihood usando o software MEGA7, com 1000 replicatas geradas pelo método de bootstrap. Abelhas da família Apidae foram utilizadas como grupo externo. Grupos de espécies do gênero Pseudomyrmex são evidenciados e espécies de Pseudomyrmecinae que apresentam características mutualistas são indicadas pela presença de um asterisco “*”.

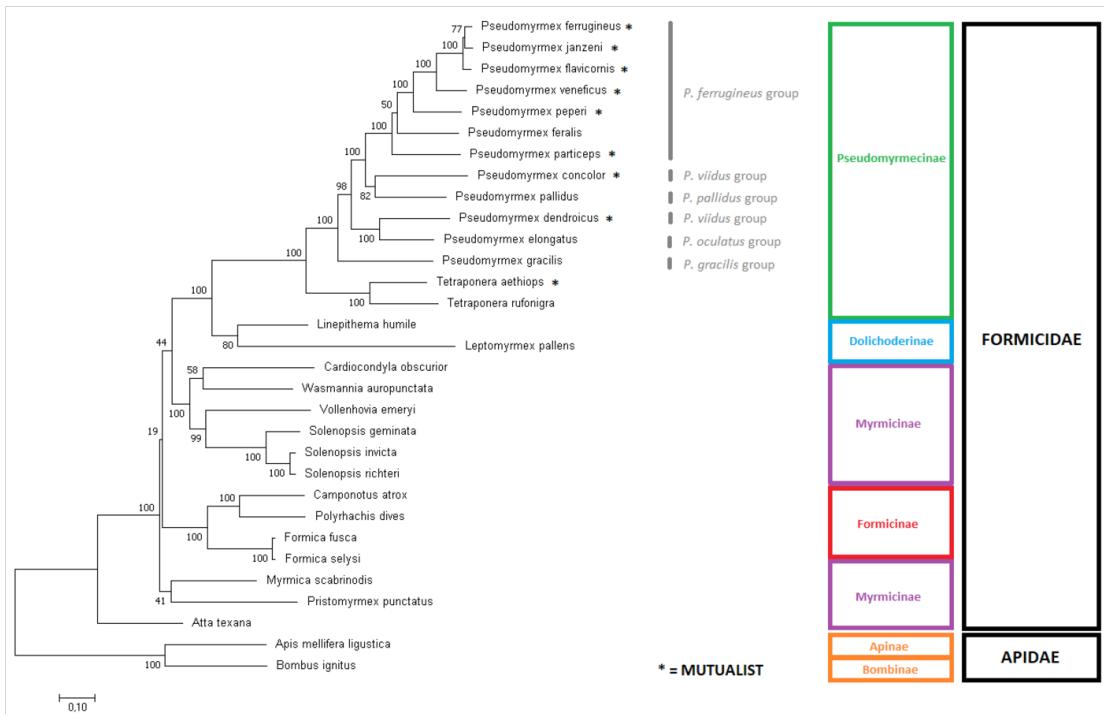


Figura 11 – Árvore filogenética usando a sequência mitocondrial completa de todos os mitogenomas de formigas disponíveis no Genbank.

“GTR + G + I” foi escolhido como modelo de substituição, conforme sugerido pelo software Modeltest. A árvore foi construída usando MEGA7 pelo método de Maximum Likelihood com 1000 replicatas de bootstrap. Mitogenomas de abelhas foram usados como grupos externos. Grupos de espécies de Pseudomyrmecinae e espécies mutualistas de Pseudomyrmecinae são evidenciados.

5 Discussão

Neste estudo, usamos dados públicos para montar, anotar, comparar e fornecer análises evolutivas de 14 sequências completas de genoma mitocondrial da subfamília Pseudomyrmecinae e outros 15 mitogenomas de formigas baixados do GenBank.

5.1 Cobertura uniforme do mitogenoma e viés de AT

Mesmo que segmentos do genoma mitocondrial possam ser copiados para o núcleo formando NuMTs (Sequências Nucleares Mitocondriais), a cobertura genômica obtida para as montagens frequentemente apresentou distribuições uniformes (Figura S1), mesmo para *Pseudomyrmex gracilis* em que os NuMTs foram previamente identificados (Rubin & Moreau, 2016). A montagem correta dos genomas mitocondriais foi possível porque o número de reads mitocondriais é provavelmente muito maior do que o número de sequências provenientes de NuMTs.

A baixa cobertura em segmentos com uma tendência a AT pronunciada deve ser esperada pois as regiões ricas em AT são conhecidas por ter amplificação reduzida em protocolos de preparação de bibliotecas Illumina (Dohm et al., 2008; Aird et al., 2011; Oyola et al., 2012). Nas espécies analisadas, a região intergênica na qual o D-loop se encontra está localizada entre o RNA ribossomal 12S (rrns) e o RNA transportador da metionina (trn-M) e varia em tamanho de 527 pb (*P. concolor*) a 697 pb (*P. elongatus*) (Tabela S1). Vários estudos demonstraram que formigas apresentam viés de AT em seus mitogenomas, em especial na região de controle, que pode exceder 90 % (Berman et al., 2014; Liu et al., 2016). Nossos dados corroboram isso, já que a menor porcentagem de AT dentre as 14 espécies de Pseudomyrmecinae analisadas corresponde a 74 % e 12 dessas espécies apresentam valor de conteúdo de AT igual ou superior a 90 % na região intergênica que contém o D-loop (Tabela 1). Além disso, a região de controle já se mostrou particularmente difícil de sequenciar em himenópteros (Castro & Dowton, 2005; Dowton et al., 2009;

Rodovalho et al., 2014). Tendo isso em vista, o fato de que segmentos de baixa cobertura em nossas montagens sempre ocorrerem no D-loop provavelmente está associado ao viés de AT pronunciado dessa região.

O viés de AT dos mitogenomas de formigas, associado às limitações do sequenciamento para regiões ricas nesses nucleotídeos podem dificultar a obtenção de mitogenomas completos de formigas. Essa dificuldade provavelmente é parte do motivo pelo qual há tão poucos genomas mitocondriais completos disponíveis para esse grupo apesar da ampla disponibilidade de dados públicos. Ao mesmo tempo, devemos considerar que o advento de novas ferramentas podem tornar a montagem de mitogenomas mais acessível. Por exemplo, novos montadores como o NOVOPlasty superam programas clássicos (Dierckxsens, Mardulyn & Smits, 2016; Plese et al., 2018) e facilitam a produção de mitogenomas completos. Assim sendo, os incessantes avanços técnicos no ramo da bioinformática prenunciam perspectivas favoráveis para o fechamento de lacunas filogenéticas em Formicidae, especialmente se os dados públicos disponíveis para o clado forem utilizados para esse fim.

5.2 Mitogenômica Comparativa: tamanho do mitogenoma e análises de sintenia

Além da identificação de quatro sítios putativos de inserção que podem explicar as diferenças observadas no tamanho do mitogenoma (apontado pelas setas na Figura 8), também observamos que todos os sete mitogenomas incluídos no grupo de espécies *P. ferrugineus* têm aproximadamente o mesmo tamanho de sequência em pb, sugerindo que esse grupo é monofilético. Por outro lado, existe uma diferença significativa entre o tamanho do mitogenoma de *P. concolor* (15906 pb) e *P. dendroicus* (17362 pb), ambas pertencentes ao grupo *P. viidus*. Isto corrobora trabalhos anteriores que apontam este grupo de espécies como parafilético (Ward, 1989; Ward & Downie, 2005).

Há uma correlação positiva entre as múltiplas sintenias encontradas nos clados Myrmicinae e Formicinae e a notável biodiversidade observada para estas duas subfamílias: Myrmicinae, que apresentou o maior número de rearranjos gênicos

(três), é a maior subfamília de formigas em termos de riqueza de espécies, com mais de 6.600 espécies descritas, quase metade de toda a biodiversidade documentada para formigas; e Formicidae, que apresentou duas sintenias distintas, é a segunda mais biodiversa, com mais de 3.100 espécies. As outras subfamílias de Formicidae analisadas neste estudo (ambas com um único arranjo gênico) são menos diversas: Dolichoderinae exibe 713 espécies enquanto Pseudomyrmecinae apresenta 231 espécies documentadas (Bolton, 2012). Como o arranjo de genes ancestrais para Formicinae é idêntico ao observado em Pseudomyrmecinae e Dolichoderinae, a análise de sintenia indica que Formicinae está mais próximo filogeneticamente a este grupo do que a Myrmicinae. Um número maior de mitogenomas associado a uma cobertura taxonômica mais ampla melhorarão a avaliação da correlação entre a ordem dos genes mitocondriais e a biodiversidade da subfamília, permitindo um melhor entendimento da evolução mitocondrial da sintenia em Formicidae.

5.3 Relações filogenômicas de Formicidae inferidas usando dados de mitogenoma

As árvores filogenômicas geradas forneceram topologias ligeiramente diferentes devido a informação adicional presente na análise do mitogenoma completo. Enquanto a árvore de concatenação gênica utiliza apenas a informação contida nos genes codificadores de proteínas, a árvore construída com base na sequência mitocondrial completa usa, além das PCG's, informação proveniente das regiões intergênicas e dos tRNAs, rRNAs e D-loop para a inferência filogenética. Ademais, o DNA mitocondrial apresenta uma taxa de substituição relativamente elevada em regiões não codificantes (Vanecek, Vorel & Sip, 2004; DeSalle, 2017) e a adição dessa variabilidade às análises também justifica as diferenças topológicas observadas.

Em geral, nas árvores filogenômicas geradas para todas as formigas apresentando mitogenoma completo, a filogenia da subfamília Pseudomyrmecinae foi fortemente recuperada como monofilética, e as posições filogenéticas da maioria dos clados foram bem resolvidas. A monofilia para a subfamília Pseudomyrmecinae e também para os gêneros *Pseudomyrmex* e *Tetraponera* foi recuperada com 100 % de suporte de bootstrap (BS) em ambas as árvores. O gênero *Pseudomyrmex*

apresentou poucos nós não suportados, mas *Tetraponera* foi completamente resolvida em ambas as árvores ($BS = 100$). Com relação aos grupos de espécies em *Pseudomyrmex*, em ambas as árvores o status monofilético do grupo *P. flavidornis* e o estado parafilético do grupo *P. viidus* confirmam (i) observações prévias baseadas exclusivamente em morfologia (Ward, 1989), (ii) filogenias usando caracteres morfológicos em conjunto com marcadores nucleares (Ward & Downie, 2005), e (iii) nossas próprias observações em relação ao tamanho do mitogenoma. Embora a divisão morfológica em grupos de espécies não tenha sido formalizada ou regulada nomenclaturalmente (Ward, 2017), o trabalho usando uma abordagem híbrida morfológica/molecular de Ward & Downie, 2005 mostra que apenas dois dos nove grupos definidos na época eram parafiléticos: *P. pallens* e *P. viidus*. A corroboração de estudos morfológicos pela análise de dados mitocondriais confirma a relevância do uso de caracteres morfológicos na determinação das relações entre clados. Ao mesmo tempo, nossos resultados reforçam que evidências moleculares podem esclarecer e complementar tais estudos, refinando e melhorando o suporte geral das filogenias reconstruídas. Neste trabalho, geramos sequências mitocondriais completas para formigas classificadas em cinco dos 10 grupos descritos para espécies de *Pseudomyrmex*, cobrindo pelo menos metade da diversidade genética do gênero e adicionando uma nova fonte de evidência molecular para estudos posteriores sobre o clado.

Ambas as árvores sugerem fortemente que os mutualismos de formigas são parafiléticos em *Pseudomyrmecinae* (por favor, verifique os asteriscos presentes nas Figuras 10 e 11), adicionando também evidências às suposições prévias de comportamento generalista como um traço basal do gênero *Pseudomyrmex* (Ward & Branstetter, 2017). Isso sugere que a relação de co-evolução entre plantas e essas formigas se desenvolveu mais tarde (e independentemente) várias vezes no clado. Espécies mutualistas são mais comuns no grupo de espécies *P. ferrugineus*, reforçando a hipótese do mutualismo ser uma característica derivada. No gênero *Pseudomyrmex*, o grupo *P. ferrugineus* possivelmente apresenta duas linhagens independentes de formigas mutualísticas (já que *P. feralis* é frequentemente considerada como exibindo comportamento generalista; $BS = 50$), enquanto outras duas linhagens mutualistas independentes podem ser observadas dentro do gênero ao se

analisar o posicionamento filogenético de *P. concolor* e *P. dendroicus* nas árvores. Considerando o gênero *Tetraponera*, *T. aethiops* e *T. rufonigra* são espécies intimamente relacionadas e apenas *T. aethiops* apresenta comportamento mutualístico (Ward & Downie, 2005), mostrando-nos que a diferenciação de traços ecológicos pode ser observado mesmo entre espécies aparentadas. Infelizmente não há estudos que estimem o tempo de divergência para espécies do gênero *Tetraponera*, mas essa diferenciação ecológica observada em espécies próximas (para as quais é esperado um passado evolutivo comum relativamente recente) pode ser indicativa de que a evolução de traços mutualistas em *Pseudomyrmecinae* pode ocorrer em períodos de tempo consideravelmente curtos. Considerando o número limitado de espécies amostradas aqui, fomos capazes de identificar 5 das 12 vezes em que associações mutualísticas desenvolvidas independentemente foram relatadas no clado (Ward, 1991; Ward & Downie, 2005). Com uma melhor cobertura taxonômica, esse número pode ser aumentado e novas análises realizadas, gerando resultados mais robustos e elucidando cada vez mais esses eventos coevolutivos.

Há diversas relações filogenéticas bem resolvidas para várias espécies de *Pseudomyrmecinae* (como *P. peperi*, *P. veneficus*, *P. particeps*, *P. gracilis*, *T. aethiops* e *T. rufonigra*) que corroboram tanto os resultados de Ward & Downie (2005) quanto a árvore de Máxima Verossimilhança gerada usando dados de elementos ultra-conservados de Ward & Branstetter (2017). A relação do grupo irmão entre *P. dendroicus* e *P. elongatus* também é bem suportada (BS = 100 na árvore de mitocôndria completa; e BS = 99 na árvore de concatenação gênica), indo ao encontro de um trabalho recente utilizando scaffolds de WGS concatenados como entrada para a construção de árvores de ML (Rubin & Moreau, 2016).

No entanto, diferenças sutis foram observadas entre nossos resultados e as relações filogenéticas inferidas com base em elementos ultra-conservados de Ward & Branstetter, 2017. Usando dados de UCE, *P. janzeni* foi observado como grupo irmão de *P. ferrugineus*. Neste trabalho, a árvore, usando a sequência mitogenômica completa, recapturou essa mesma relação com um valor replicado de bootstrap de 77. Por outro lado, na árvore de genes concatenados, além da relação de grupo irmão ter sido observada entre *P. janzeni* e *P. flavigaster*, ela mostrou um suporte inferior (BS = 47). No geral, essa relação pareceu ser melhor reconstruída pela análise

da sequência mitocondrial completa, que corrobora as análises de UCE. Dentro de Pseudomyrmecinae, observamos duas espécies cujas posições filogenéticas não foram bem resolvidas pelas análises mitocondriais atuais e, portanto, sua relação pode ser vista como inconclusiva: (i) *P. feralis* em ambas as árvores filogenéticas; e (ii) *P. pallidus* na árvore de genes concatenados. Essas posições apresentam um valor de suporte de bootstrap de 50.

Ambas as árvores apresentaram a subfamília Dolichoderinae como monofilética, embora este resultado não tenha sido recuperado em todas as replicatas. Dolichoderinae é uma subfamília altamente diversificada e contém mais de 700 espécies, mas foi aqui representada por apenas duas espécies. Assim, acreditamos que uma maior cobertura de espécies irá melhorar a robustez das análises filogenéticas para o clado.

Trabalhos anteriores com caracteres morfológicos e/ou genes nucleares apresentam evidências de relação de grupo irmão entre Pseudomyrmecinae e Myrmeciinae (Ward & Downie, 2005; Brady et al., 2006). Nós esperaríamos que Myrmeciinae fosse o grupo irmão de Pseudomyrmecinae de acordo com os dados mitocondriais, mas como genomas mitocondriais completos não estão disponíveis para a subfamília Myrmeciinae, nós não pudemos testar essa hipótese. A ausência de mitogenomas para essa e outras subfamílias podem estar associadas ao fato de que sua biodiversidade e importância econômica não são tão expressivas quando comparadas à das subfamílias de formigas mais estudadas. Por exemplo, Myrmeciinae não apresenta nenhuma espécie de pronunciada relevância econômica e conta com apenas 94 espécies descritas, enquanto Myrmicinae engloba as saúvas (tribo Attini), notoriamente conhecidas como importantes pragas agrícolas e possui mais de 6600 espécies (Bolton, 2012).

Na ausência de Myrmeciinae, espera-se que Dolichoderinae seja o grupo mais próximo de Pseudomyrmecinae em nossas árvores. Isso foi confirmado em ambas as árvores, nas quais as duas subfamílias aparecem como grupos irmãos entre si, corroborando filogenias moleculares de larga escala usando poucos genes nucleares (Brady et al., 2006) e dados de UCE (Branstetter et al., 2017). A sintenia compartilhada entre todas as espécies de Pseudomyrmecinae e Dolichoderinae também suporta a relação do grupo irmão observada. Nossos resultados sugerem

que Myrmicinae é o táxon mais próximo de um clado contendo Pseudomyrmecinae e Dolichoderinae, enquanto Formicinae foi observado como um grupo mais basal na família Formicidae. Esta posição basal de Formicinae é altamente suportada na árvore de concatenação de genes, mas não na árvore usando mitogenomas completos, ao contrário do que é mostrado por outros trabalhos usando dados nucleares que apontam para uma relação de grupo irmão entre Myrmicinae e Formicinae (Brady et al., 2006; Branstetter et al., 2017).

A monofilia da subfamília Formicinae e todos os seus nós mostram suporte máximo em ambas as árvores ($BS = 100$). Nossos resultados também corroboram o caráter monofilético do gênero *Formica* e apresentam os gêneros *Camponotus* e *Polyrhachis* como intimamente relacionados entre si, conforme observado no trabalho de Blaimer e colaboradores (2015), que utilizaram locus de UCE para inferência filogenética. O único problema com relação a essa subfamília diz respeito à posição mal suportada de Formicinae em relação às outras subfamílias. Os dados de mitogenoma forneceram com sucesso relações filogenéticas robustas, mesmo para *Camponotus atrox*, uma espécie que mostrou sintenia única, mas teve sua posição bem resolvida em ambas as inferências, inclusive na árvore mitocondrial completa, que pode estar propensa a ser afetada por alterações de sintenia. Esta questão confirma a robustez das sequências mitocondriais para inferir filogenias de formigas.

No geral, os resultados mais controversos obtidos aqui estão relacionados à posição da subfamília Myrmicinae. Para esse clado, a árvore de concatenação de genes foi capaz de indicar monofilia ($BS = 74$), mas dados de mitogenoma total produziram parafilia. Neste último caso, as espécies *Atta texana*, *Myrmica scabrinodis* e *Pristomyrmex punctatus* divergiram antes das outras formigas. Por outro lado, ambas as árvores recapturaram com sucesso o caráter monofilético do gênero *Solenopsis* e as relações entre suas espécies (*S. geminata* como grupo irmão do clado consistindo de *S. invicta* e *S. richteri*) com 100 % de suporte de bootstrap. A relação do grupo irmão entre *Solenopsis* spp. e *Vollenhovia emeryi* também é recuperada. Estes resultados corroboram aqueles obtidos pelo uso de sequências de aminoácidos concatenados de todos os PCGs mitocondriais para inferência de árvores (Duan et al., 2016). Entretanto, nossa avaliação da posição

de *V. emeryi* foi melhor suportada ($BS = 90$ na árvore de concatenação gênica e $BS = 99$ na árvore mitocondrial completa) do que a deste trabalho anterior ($BS = 75$). Considerando que Duan e colaboradores (2016) utilizaram uma abordagem semelhante à nossa (concatenação gênica seguida por construção de árvore usando Máxima Verossimilhança), podemos concluir que esses resultados indicam que os dados nucleotídicos apresentam informações mais confiáveis para a inferência filogenômica desses clados do que os dados aminoacídicos. Isto é consistente com pesquisa anterior na qual a inferência filogenética utilizando nucleotídeos obteve resultados melhor suportados do que as análises ao nível de aminoácidos ou codons (Holder, Zwickl & Dessimoz, 2008). Além disso, os valores de bootstrap obtidos através de dados nucleotídicos já foram relatados como geralmente maiores do que aqueles provenientes de seus correspondentes em aminoácidos (Regier et al., 2010). Essas observações são ao menos parcialmente explicadas pelas diferenças na quantidade de sinal filogenético considerados por esses dois métodos. Sinal adicional presente em sequências de nucleotídeos é perdido na tradução para aminoácidos. Isso é particularmente importante em se tratando de aminoácidos hexacodônicos como a serina, que é codificada tanto por TCN quanto por AGY (Regier et al., 2010; Zwick, Regier & Zwickl, 2012).

As relações filogenéticas de outras espécies da subfamília Myrmicinae na nossa árvore de concatenação gênica não estão bem resolvidas, como a posição de *Myrmica scabrinodis* ($BS = 52$), *Wasmannia auropunctata* ($BS = 43$) e *Pristomyrmex punctatus* ($BS = 16$). No entanto, a posição dessas espécies na árvore de aminoácidos de Duan e colaboradores (2016) também é inconclusiva e difere daqui, agrupando *W. auropunctata* e *M. scabrinodis* em uma relação suportada apenas em 35 % das replicatas de bootstrap. Este clado é colocado como grupo irmão de *Solenopsis* spp. e *V. emeryi* com suporte ainda menor ($BS = 21$) e *P. punctatus* assume uma posição mais basal na árvore em 46 % das repetições. No entanto, *Atta laevigata* aparece na base de todas as Myrmecinae com suporte de bootstrap de 100 % na árvore de aminoácidos. Como o mitogenoma de *A. laevigata* disponível não está completo, ele não foi usado como entrada para a concatenação de nucleotídeos das PCGs aqui realizada, ao contrário de sua congenérica *Atta texana*, cujo mitogenoma completo foi analisado aqui. *A. texana* também aparece

na base da subfamília Myrmicinae, mas sob uma relação de grupo irmão com *M. scabrinodis*, mesmo que com baixa resolução (BS = 52). Este clado é irmão de todas as outras espécies de Myrmicinae (BS = 74). Por fim, a posição de *Cardiocondyla obscurior* também não foi bem suportada (BS = 43), mas como esse é um mitogenoma recentemente publicado, não foi utilizado no trabalho de Duan e colaboradores.

Em ambos os trabalhos, as análises mitogenômicas não foram totalmente capazes de resolver importantes nós do ramo das Myrmicinae e vários fatores podem estar associados a esses resultados insatisfatórios. É necessário destacar que Myrmicinae é a subfamília mais biodiversa (Bolton, 2012) e é conhecida por apresentar vários grupos monofiléticos duvidosos (Brady et al., 2006; Ward, 2011; Ward et al., 2015). Essa diversidade é evidenciada pelo fato de que, apesar de apenas nove mitogenomas estarem disponíveis para o grupo, três arranjos diferentes de genes mitocondriais podem ser observados, sugerindo uma alta taxa de evolução mitocondrial nessa subfamília.

Além disso, houveram divergências no ramo das Myrmicinae em estudos filogenéticos moleculares anteriores que tentaram estudar a família Formicidae como um todo (Brady et al., 2006; Moreau et al., 2006). Por outro lado, [Ward et al. \(2015\)](#) se foca no estudo dessa subfamília ao reconstruir uma filogenia em grande escala usando 11 marcadores nucleares de 251 espécies amostradas em todas as 25 tribos de Myrmicinae, a maioria delas parafiléticas. Utilizando uma grande quantidade de dados que cobre uma extensa parcela da diversidade de espécies dessa subfamília, eles conseguiram propor uma nova classificação de Myrmicinae composta exclusivamente por tribos monofiléticas, o que também reduziu o número de gêneros parafiléticos.

Assim, a natureza hiperdiversa deste clado, associada à subamostragem ou mesmo ausência de mitogenomas para vários táxons da subfamília e uma possível alta taxa de evolução do genoma mitocondrial são fatores que podem ter contribuído para os resultados inconclusivos das análises mitocondriais. Além disso, apesar de algumas relações não terem sido elucidadas pelo uso exclusivo da filogenômica mitocondrial, a informação fornecida pelo mitogenoma é classicamente considerada como útil no estudo das relações evolutivas para diversos táxons, seja

confirmando (Prosdocimi et al., 2012; Finstermeier et al. al., 2013) ou refutando hipóteses filogenéticas anteriores ([KAYAL et al., 2015](#); [ULIANO-SILVA et al., 2016](#)). Assim, ainda recomendamos o uso de dados mitocondriais, de preferência ao lado de outros marcadores (por exemplo, genes nucleares), para aumentar o sinal filogenético e recapturar filogenias mais robustas. Entretanto, graças à taxa de substituição do mtDNA, árvores geradas a partir de dados mitocondriais apresentam uma maior probabilidade de resolver ramos curtos corretamente (DeSalle, 2017). Portanto, também acreditamos que o uso de dados mitocondriais para inferência filogenômica, mesmo sem outros marcadores, produzirá resultados mais satisfatórios se trabalharmos no sentido de mitigar o problema da escassez de mitogenomas disponíveis para esse clado e melhorarmos a cobertura mitocondrial de seus táxons. Essa afirmativa não só é válida para a subfamília Myrmicinae, como também para a família Formicidae como um todo e para qualquer outro grupo com escassez de mitogenomas conhecidos e necessidade de elucidação sobre suas relações filogenéticas. Nesse sentido, os resultados aqui apresentados são extremamente relevantes para mostrar que as informações já disponíveis em bancos de dados públicos devem ser usadas para obter genomas mitocondriais completos e fomentar novas pesquisas que gerarão conhecimento sem incorrer em custos adicionais de sequenciamento.

5.4 Mitogenômica no-budget: análises integradas entre datasets e potencial para estudos de larga-escala

Os resultados aqui apresentados confirmam que os dados de UCE e WGS publicamente disponíveis podem ser usados para montar genomas mitocondriais completos com alta cobertura (Tabela 2), o que pode ser explicado pelo alto número de cópias de reads mitocondriais que pode alcançar algo entre 0,25 % a 0,5 % do número total de bases geradas ([PROSDOCIMI et al., 2012](#)) e chegar a 2 % do total de reads mapeando ao mtDNA ([EKBLOM; WOLF, 2014](#)). Também confirmamos o potencial dos dados de UCE como uma alternativa de baixo custo para sequenciar mitogenomas completos com alta cobertura, conforme descrito por Raposo do Amaral et al. (2015). Dados de mitogenoma são usados em várias análises e seqüências mitocondriais são encontradas em vários tipos de datasets,

que geralmente fornecem informação suficiente para montar toda a sequência mitocondrial. Essa versatilidade e onipresença de sequências mitocondriais deve ser usada em favor dos estudos de biodiversidade, especialmente considerando que os datasets públicos estão disponíveis para um número cada vez maior de espécies.

O potencial dessas seqüências na elucidação de filogenias não deve ser menos-prezado, especialmente se considerarmos que existem diferentes tipos de conjuntos de dados disponíveis para diferentes espécies (WGS, RNA-Seq, enriquecimento de UCE, dentre outros). Esses diferentes recursos dificultam a obtenção de árvores filogenéticas/filogenômicas que integrem esses diferentes dados públicos, já que muitas vezes as análises dependem da ortologia das sequências comparadas ([KUZNIAR et al., 2008](#)). Assim, o uso de diferentes tipos de dados para montar os mitogenomas completos ou quase completos para espécies com dados publicamente disponíveis apresenta uma solução para este problema com o genoma mitocondrial agindo como uma “sequência normalizadora” que permite a comparação de diferentes conjuntos de dados. Por exemplo, neste trabalho algumas espécies tinham apenas dados de UCE disponíveis publicamente, enquanto outros apresentavam datasets padrão de WGS. No entanto, anotando e analisando o mitogenoma completo para essas espécies, conseguimos ampliar nosso escopo e estudar todas elas juntas. Assim, sugerimos que o uso de mitogenomas obtidos a partir de dados públicos tem o potencial de se tornar uma importante fonte de informação filogenética. Além disso, o estudo das sequências mitocondriais pode ser uma das rotas mais rápidas para a obtenção de árvores abrangentes de alta qualidade para táxons hiperdiversos, como os insetos. Progresso se tem feito nesse sentido, como pode ser visto no trabalho recente de Linard et al. (2018), onde a mineração de dados do Genbank e montagem usando datasets metagenômicos forneceram contigs mitocondriais ($> 3\text{kpb}$) para quase 16.000 espécies de coleópteros. Essa enorme quantidade de dados mitogenômicos foi usada para gerar a maior árvore filogenética já vista para o clado.

Estudos que tentam montar mitogenomas completos usando dados públicos ainda são escassos, ao passo que o tamanho e a amplitude dos bancos de dados públicos estão em crescimento, juntamente com seu potencial para responder questões filogenéticas, dentre tantas outras. A mitogenômica no-budget é uma

oportunidade sem precedentes de reconstruir e analisar filogenias em larga escala para vários grupos em diferentes níveis taxonômicos, o que por sua vez pode subsidiar estudos evolutivos e de biologia da conservação e incrementar nosso conhecimento sobre espécies não-modelo e sua diversidade.

6 Perspectivas

Usando um protocolo ligeiramente modificado para a realização de montagens mitogenômicas em computadores domésticos, pretendemos montar e anotar pelo menos 100 sequências mitocondriais de insetos usando dados públicos (chamado “projeto 100 MITO”). Trata-se de um projeto de escopo bem maior se comparado àquele descrito nessa dissertação, que consequentemente envolve grande parte dos integrantes do Laboratório de Genômica e Biodiversidade. Até o momento, possuímos 40 mitogenomas completos ou quase completos (aqui definidos como sequências que, apesar de não circularizadas, apresentam todas as 37 features mitocondriais). Destes, 34 pertencem à espécies de um gênero de formigas hiperdiverso (*Temnothorax* spp.), montadas a partir dos datasets de UCE gerados pelo trabalho de Prebus (2017). Pretendemos continuar usando os dados públicos disponíveis no Sequence Read Archive para aumentar a cobertura mitogenômica e desvendar relações evolutivas não só de formigas, como também de outros grupos de insetos.

7 Conclusão

Neste trabalho, que foi publicado pela revista PeerJ em Janeiro de 2019 ([VI-EIRA; PROSDOCIMI, 2019](#)), montamos e anotamos os primeiros 14 mitogenomas para a subfamília Pseudomyrmecinae. Para tal, utilizamos uma metodologia que se baseia no uso de dados públicos de diferentes fontes e tipos, em conjunto com a aplicação de softwares gratuitos de bioinformática para a manipulação desses dados. As sequências obtidas foram utilizadas para estudar a sintenia, genómica comparativa e relações filogenéticas desses organismos, fornecendo informações valiosas sobre a filogenia e evolução de Pseudomyrmecinae, como:

- (i) identificação de quatro regiões putativas de inserção nucleotídica em mitogenomas do gênero *Pseudomyrmex*;
- (ii) corroboração de que as associações mutualísticas com plantas encontradas na subfamília são parafiléticas, tendo ocorrido independentemente múltiplas vezes no clado;
- (iii) indicação de que o grupo de espécies *P. ferrugineus* é monofilético, enquanto o grupo *P. viidus* é parafilético; e
- (iv) corroboração da monofilia dos gêneros *Pseudomyrmex* e *Tetraponera*.

Dados mitocondriais em outros clados de formigas, mesmo que limitados, nos permitiram ampliar nosso escopo e estudar a família Formicidae como um todo. Isso nos possibilitou elucidar relações de grupo irmão para a família, como a descrita entre Pseudomyrmecinae e Dolichoderinae, assim como o caráter monofilético de todas as subfamílias analisadas. Entretanto, uma definição mais precisa das relações entre os diferentes grupos de formigas idealmente devem se valer de grandes datasets genómicos e concatâmeros de centenas a milhares de genes, atualmente indisponíveis. Os baixos valores de bootstrap observados em alguns nós indicam que os dados mitocondriais disponíveis no momento não apresentam variabilidade o bastante para elucidar algumas relações, muito embora isso possa

mudar com a ampliação tanto da quantidade quanto da cobertura taxonômica de genomas mitocondriais. As sequências mitocondriais montadas abarcam uma porção considerável da biodiversidade de Pseudomyrmecinae e serão úteis em novos estudos sobre a evolução e conservação desse grupo.

Este trabalho praticamente dobra o número de mitogenomas de formiga completos disponíveis sem custos adicionais de sequenciamento. Uma vez que há poucos grupos de formigas com genomas mitocondriais completos disponíveis, o aprimoramento da cobertura mitogenômica é necessário para uma melhor resolução e robustez de filogenias em larga escala para o clado. A metodologia apresentada aqui também pode ser usada para estudar os já citados clados de monofilia duvidosa da subfamília Myrmicinae (Brady et al., 2006; Ward, 2011; Ward et al., 2015) ou grupos notoriamente parafiléticos, como o gênero *Camponotus* da subfamília Formicinae (Blaimer et al., 2015). Com base nesses resultados, enfatizamos que a cobertura filogenética cada vez maior dos bancos de dados públicos, associada à presença de sequências mitocondriais em diferentes tipos de dados de sequenciamento, torna a mitogenômica no-budget a abordagem ideal para o estudo da diversidade de espécies. Trabalhos que utilizem dados públicos para a montagem e análise de mitogenomas, como este e o projeto em andamento “100 MITO”, são possivelmente o caminho mais rápido para se obter amplas árvores filogenéticas que representem a história evolutiva das espécies da forma mais fidedigna possível.

Referências

- ADAMS, K. L.; PALMER, J. D. Evolution of mitochondrial gene content: Gene loss and transfer to the nucleus. *Molecular Phylogenetics and Evolution*, v. 29, n. 3, p. 380–395, 2003. ISSN 10557903. Citado na página 31.
- BERNSTEIN, M. N.; DOAN, A.; DEWEY, C. N. MetaSRA: Normalized human sample-specific metadata for the Sequence Read Archive. *Bioinformatics*, v. 33, n. 18, p. 2914–2923, 2017. ISSN 14602059. Citado na página 18.
- BERNT, M. et al. MITOS: Improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution*, Elsevier Inc., v. 69, n. 2, p. 313–319, 2013. ISSN 10557903. Disponível em: <<http://dx.doi.org/10.1016/j.ympev.2012.08.023>>. Citado na página 37.
- BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, v. 30, n. 15, p. 2114–2120, 2014. ISSN 14602059. Citado na página 18.
- BOORE, J. L. Animal mitochondrial genomes. *Nucleic Acids Research*, v. 27, n. 8, p. 1767–1780, 1999. ISSN 03051048. Citado na página 31.
- BORDBARI, M. H. et al. Deletion of 2.7 kb near HOXD3 in an Arabian horse with occipitoatlantoaxial malformation. *Animal Genetics*, v. 48, n. 3, p. 287–294, 2017. ISSN 13652052. Citado na página 18.
- BRAND, M. D. Regulation analysis of energy metabolism. *The Journal of experimental biology*, v. 200, n. Pt 2, p. 193–202, 1997. ISSN 0022-0949. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/9050227>>. Citado na página 31.
- CHAN, D. C. Mitochondria: Dynamic Organelles in Disease, Aging, and Development. *Cell*, v. 125, n. 7, p. 1241–1252, 2006. ISSN 00928674. Citado na página 31.
- COCHRANE, G. et al. Evidence Standards in Experimental and Inferential INSDC Third Party Annotation Data. *OMICS A Journal of Integrative Biology*, v. 10, n. 2, 2006. Citado 2 vezes nas páginas 18 e 46.
- Del Fabbro, C. et al. An extensive evaluation of read trimming effects on illumina NGS data analysis. *PLoS ONE*, 2013. ISSN 19326203. Citado na página 18.

- DIERCKXSENS, N.; MARDULYN, P.; SMITS, G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, v. 45, n. 4, p. gkw955, 2016. ISSN 0305-1048. Disponível em: <<https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw955>>. Citado na página 41.
- DO AMARAL, F. R. et al. Ultraconserved elements sequencing as a low-cost source of complete mitochondrial genomes and microsatellite markers in non-model amniotes. *PLoS ONE*, v. 10, n. 9, p. 1–9, 2015. ISSN 19326203. Citado na página 32.
- EKBLOM, R.; WOLF, J. B. W. *A field guide to whole-genome sequencing, assembly and annotation*. 2014. 1026–1042 p. Citado na página 64.
- FINSTERMEIER, K. et al. A Mitogenomic Phylogeny of Living Primates. *PLoS ONE*, 2013. ISSN 19326203. Citado na página 31.
- GOODWIN, S.; MCPHERSON, J. D.; MCCOMBIE, W. R. Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, Nature Publishing Group, v. 17, n. 6, p. 333–351, 2016. ISSN 14710064. Disponível em: <<http://dx.doi.org/10.1038/nrg.2016.49>>. Citado na página 18.
- GRAY, M. W. Lynn Margulis and the endosymbiont hypothesis: 50 years later. *Molecular Biology of the Cell*, v. 28, n. 10, p. 1285–1287, 2017. ISSN 1059-1524. Disponível em: <<http://www.molbiolcell.org/lookup/doi/10.1091/mbc.E16-07-0509>>. Citado na página 31.
- GRAY, M. W.; BURGER, G.; LANG, B. F. *Mitochondrial evolution*. 1999. Citado na página 31.
- JANZEN, D. H. Coevolution of Mutualism Between Ants and Acacias in Central America. *Evolution*, 1966. ISSN 00143820. Citado na página 35.
- KARSCH-MIZRACHI, I.; TAKAGI, T.; COCHRANE, G. The international nucleotide sequence database collaboration. *Nucleic Acids Research*, v. 46, n. D1, p. D48–D51, 2018. ISSN 13624962. Citado na página 18.
- KAYAL, E. et al. Phylogenetic analysis of higher-level relationships within Hydrodololina (Cnidaria: Hydrozoa) using mitochondrial genome data and insight into their mitochondrial transcription. *PeerJ*, v. 3, p. e1403, 2015. ISSN 2167-8359. Disponível em: <<https://peerj.com/articles/1403>>. Citado 2 vezes nas páginas 18 e 64.
- KODAMA, Y.; SHUMWAY, M.; LEINONEN, R. The sequence read archive: Explosive growth of sequencing data. *Nucleic Acids Research*, v. 40, n. D1, p. 2011–2013, 2012. ISSN 03051048. Citado na página 18.

- KRZEMIŃSKA, U. et al. Population mitogenomics provides insights into evolutionary history, source of invasions and diversifying selection in the House Crow (*Corvus splendens*). *Heredity*, 2018. ISSN 13652540. Citado na página 31.
- KUTSCHERA, U.; NIKLAS, K. J. Endosymbiosis, cell evolution, and speciation. *Theory in Biosciences*, v. 124, n. 1, p. 1–24, 2005. ISSN 14317613. Citado na página 31.
- KUZNIAR, A. et al. The quest for orthologs: finding the corresponding gene across genomes. *Trends in Genetics*, v. 24, n. 11, p. 539–551, 2008. ISSN 01689525. Citado na página 65.
- LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, v. 9, n. 4, p. 357–359, 2012. ISSN 15487091. Citado na página 18.
- LINARD, B. et al. The contribution of mitochondrial metagenomics to largescale data mining and phylogenetic analysis of Coleoptera. *bioRxiv*, n. March, p. 280792, 2018. Disponível em: <<https://www.biorxiv.org/content/early/2018/03/12/280792.figures-only>>. Citado na página 18.
- MARDIS, E. R. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, v. 24, n. 3, p. 133–141, 2008. ISSN 01689525. Citado na página 18.
- MARGULIES, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 2005. ISSN 00280836. Citado na página 18.
- MILLER, M. J. et al. Complete mitochondrial genomes of the New World jacanas: Jacana spinosa and Jacana jacana. *Mitochondrial DNA*, v. 27, n. 1, p. 764–765, 2016. ISSN 19401744. Citado na página 32.
- PROSDOCIMI, F. et al. The complete mitochondrial genome of two recently derived species of the fish genus *Nannoperca* (Perciformes, Percichthyidae). *Molecular Biology Reports*, v. 39, n. 3, p. 2767–2772, 2012. ISSN 03014851. Citado na página 64.
- RUBIN, B. E. R. et al. Comparative genomics reveals convergent rates of evolution in ant–plant mutualisms. *Nature Communications*, Nature Publishing Group, v. 7, p. 12679, 2016. ISSN 2041-1723. Disponível em: <<http://www.nature.com/doifinder/10.1038/ncomms12679>>. Citado na página 22.
- SIMPSON, J. T. et al. ABYSS : A parallel assembler for short read sequence data ABYSS : A parallel assembler for short read sequence data. p. 1117–1123, 2009. Citado na página 18.

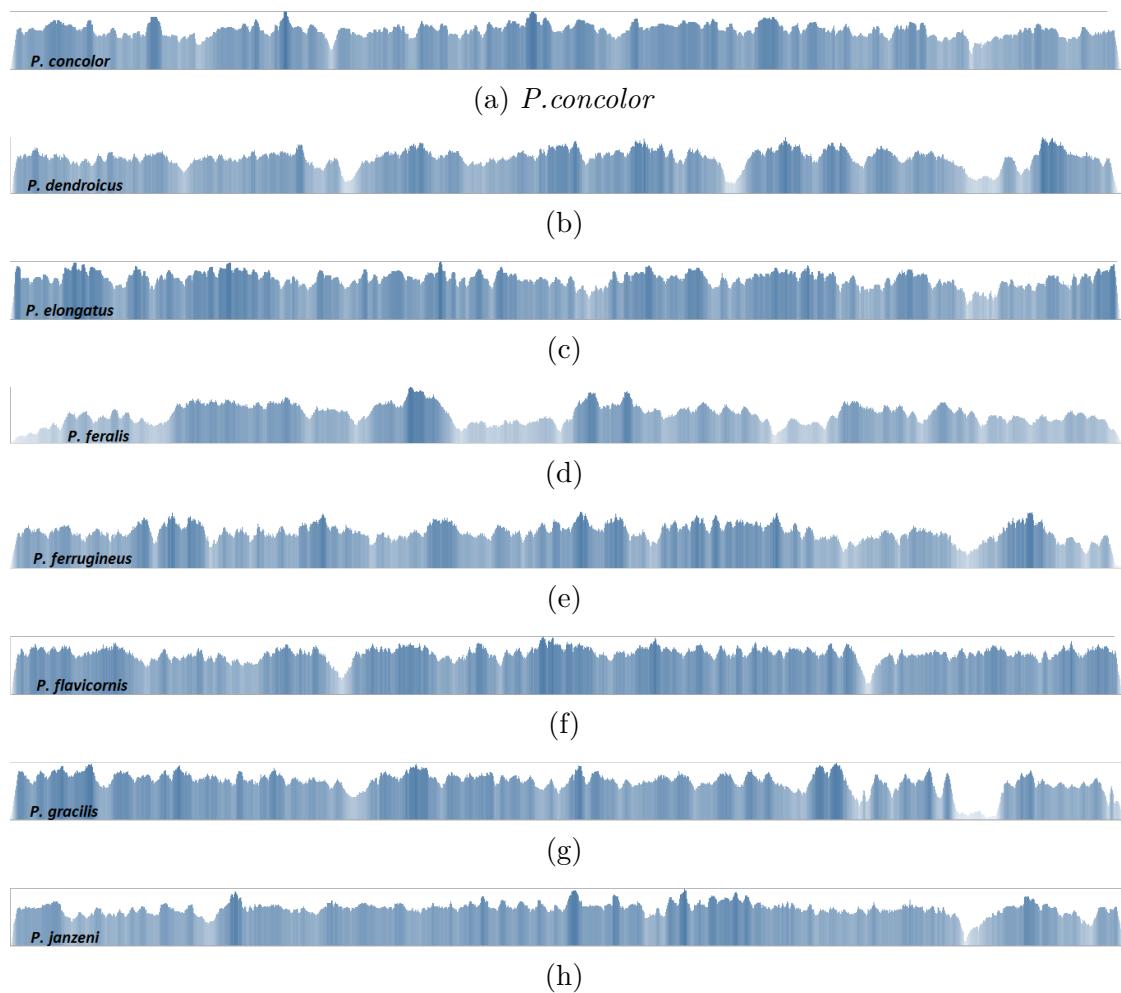
- SMITH, D. R. Goodbye genome paper, hello genome report: the increasing popularity of 'genome announcements' and their impact on science. *Briefings in Functional Genomics*, v. 16, n. 3, p. 156–162, 2016. Disponível em: <www.arrogantgenome.com>. Citado na página 31.
- STODDEN, V.; SEILER, J.; MA, Z. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, v. 115, n. 11, p. 2584–2589, 2018. ISSN 0027-8424. Disponível em: <<http://www.pnas.org/lookup/doi/10.1073/pnas.1708290115>>. Citado na página 18.
- ULIANO-SILVA, M. et al. The complete mitochondrial genome of the golden mussel *Limnoperna fortunei* and comparative mitogenomics of Mytilidae. *Gene*, Elsevier B.V., v. 577, n. 2, p. 202–208, 2016. ISSN 18790038. Disponível em: <<http://dx.doi.org/10.1016/j.gene.2015.11.043>>. Citado na página 64.
- VIEIRA, G. A.; PROSDOCIMI, F. Accessible molecular phylogenomics at no cost: obtaining 14 new mitogenomes for the ant subfamily Pseudomyrmecinae from public data. *PeerJ*, v. 7, p. e6271, jan 2019. ISSN 2167-8359. Disponível em: <<https://peerj.com/articles/6271>>. Citado na página 68.
- WANET, A. et al. Connecting Mitochondria, Metabolism, and Stem Cell Fate. *Stem Cells and Development*, v. 24, n. 17, p. 1957–1971, 2015. ISSN 1547-3287. Disponível em: <<http://online.liebertpub.com/doi/10.1089/scd.2015.0117>>. Citado na página 31.
- WARD, P. S. *Phylogenetic analysis of pseudomyrmecine ants associated with domatia-bearing plants*. 1991. Citado na página 35.
- 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. ISSN 13653113. Citado na página 63.
- WIRTHLIN, M. et al. Parrot Genomes and the Evolution of Heightened Longevity and Cognition. *Current Biology*, p. 1–8, 2018. ISSN 09609822. Disponível em: <<https://linkinghub.elsevier.com/retrieve/pii/S0960982218314179>>. Citado na página 22.
- WOLSTENHOLME, D. R. Animal Mitochondrial DNA: Structure and Evolution. *International Review of Cytology*, v. 141, n. C, p. 173–216, 1992. ISSN 00747696. Citado na página 31.

Apêndices

APÊNDICE A – Figuras e tabelas adicionais

Figura S1 – Visualização da cobertura de montagem para todos os 14 mitogenomas de Pseudomyrmecinae fornecidos pelo software TABLET.

O eixo X representa a posição dos nucleotídeos no mitogenoma, enquanto o Y corresponde à cobertura de reads. As espécies estão na seguinte ordem: A) *P. concolor*; B) *P. dendroicus*; C) *P. elongatus*; D) *P. feralis* E) *P. ferrugineus*; F) *P. flavidornis*; G) *P. gracilis*; H) *P. janzeni*; I) *P. pallidus*; J) *P. particeps*; K) *P. peperi*; L) *P. veneficus*. Picos de alta cobertura não são observados em nenhuma espécie e regiões de baixa cobertura estão presentes em G, H, J, K, L, M e N, coincidindo com regiões ricas em AT.



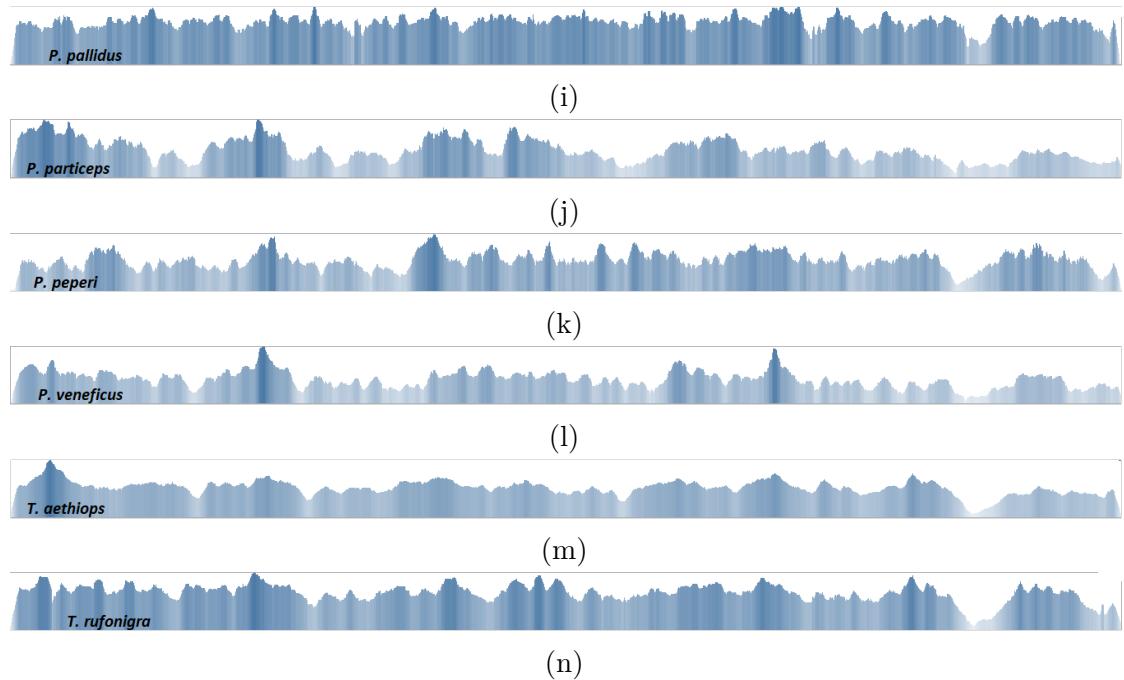


Tabela S1 – Anotação completa dos 14 genomas mitocondriais descritos nessa dissertação

Gene	Pseudomyrmex gracilis						
	Position		Size		Codon	Intergenic	
	From	To	Nucleotide	Aminoacid	Start	Stop	nucleotide
COX1	1	1533	1533	510	ATG	TAA	0
tRNA(Leu)	1534	1599	66				0
COX2	1600	2271	672	223	ATC	TAA	16
tRNA(Lys)	2288	2359	72				0
tRNA(Asp)	2360	2422	63				0
ATP8	2423	2584	162	53	ATT	TAA	21
ATP6	2606	3272	667	222	ATG	T-	0
COX3	3273	4052	780	259	ATG	TAA	3
tRNA(Gly)	4056	4125	70				3
ND3	4129	4479	351	116	ATG	TAA	6
tRNA(Ala)	4486	4547	62				0
tRNA(Arg)	4548	4615	68				2
tRNA(Asn)	4618	4684	67				28
tRNA(Ser)	4713	4772	60				5
tRNA(Glu)	4778	4850	73				10
tRNA(Phe)	4861	4929	69				39
ND5	4969	6633	1665	554	ATT	TAA	0
tRNA(His)	6634	6698	65				0
ND4	6699	8022	1324	441	ATG	T-	0
ND4L	8023	8305	283	94	ATA	T-	27
tRNA(Thr)	8333	8398	66				1
tRNA(Pro)	8400	8466	67				20
ND6	8487	9018	532	177	ATG	T-	0
CYTB	9019	10139	1120	373	ATG	TA-	0
tRNA(Ser)	10140	10208	69				11
ND1	10220	11170	951	316	ATT	TAA	0
tRNA(Leu)	11171	11238	68				11
16S rRNA	11250	12527	1278				20

tRNA(Val)	12548	12615	68				10
12S rRNA	12626	13401	776				628
tRNA(Met)	14030	14098	69				0
tRNA(Ile)	14099	14165	67				45
tRNA(Gln)	14211	14280	70				61
ND2	14342	15303	961	320	ATT	TA-	0
tRNA(Trp)	15304	15369	66				1
tRNA(Cys)	15371	15434	64				38
tRNA(Tyr)	15473	15538	66				166

2 Pseudomyrmex
concolor

Gene	Position		Size		Codon		Intergenic nucleotide
	From	To	Nucleotide	Aminoacid	Start	Stop	
COX1	1	1533	1533	510	ATG	TAA	2
tRNA(Leu)	1536	1607	72				0
COX2	1608	2285	678	225	ATA	TAA	60
tRNA(Lys)	2346	2415	70				0
tRNA(Asp)	2416	2482	67				0
ATP8	2483	2662	180	59	ATT	TAA	35
ATP6	2698	3361	664	221	ATG	T-	0
COX3	3362	4141	780	259	ATG	TAA	26
tRNA(Gly)	4168	4245	78				4
ND3	4250	4597	348	115	ATG	TAA	71
tRNA(Ala)	4669	4737	69				0
tRNA(Arg)	4738	4805	68				7
tRNA(Asn)	4813	4879	67				15
tRNA(Ser)	4895	4952	58				19
tRNA(Glu)	4972	5044	73				106
tRNA(Phe)	5151	5220	70				3
ND5	5224	6885	1662	553	GTG	TAA	0
tRNA(His)	6886	6954	69				6
ND4	6961	8283	1323	440	ATG	TAA	0
ND4L	8284	8566	283	94	ATA	T-	97

tRNA(Thr)	8664	8730	67				26
tRNA(Pro)	8757	8823	67				17
ND6	8841	9372	532	177	ATG	T-	0
CYTB	9373	10493	1121	373	ATG	TA-	0
tRNA(Ser)	10494	10563	70				3
ND1	10567	11514	948	315	ATT	TAG	0
tRNA(Leu)	11515	11582	68				7
16S rRNA	11590	12880	1291				0
tRNA(Val)	12881	12971	91				2
12S rRNA	12974	13760	787				527
tRNA(Met)	14288	14357	70				5
tRNA(Ile)	14363	14431	69				40
tRNA(Gln)	14472	14546	75				56
ND2	14603	15571	969	322	ATA	TAA	8
tRNA(Trp)	15580	15650	71				20
tRNA(Cys)	15671	15733	63				48
tRNA(Tyr)	15782	15850	69				56

3 Pseudomyrmex
elongatus

Gene	Position		Size		Codon		Intergenic nucleotide
	From	To	Nucleotide	Aminoacid	Start	Stop	
COX1	1	1531	1531	510	ATG	T-	0
tRNA(Leu)	1532	1599	68				0
COX2	1600	2277	678	225	ATT	TAA	94
tRNA(Lys)	2372	2444	73				0
tRNA(Asp)	2445	2513	69				0
ATP8	2514	2690	177	58	ATT	TAA	165
ATP6	2856	3522	667	222	ATG	T-	0
COX3	3523	4302	780	259	ATG	TAA	15
tRNA(Gly)	4318	4385	68				5
ND3	4391	4738	348	115	ATG	TAA	62
tRNA(Ala)	4801	4869	69				0
tRNA(Arg)	4870	4937	68				0

tRNA(Asn)	4938	5008	71				9
tRNA(Ser)	5018	5079	62				35
tRNA(Glu)	5115	5184	70				159
tRNA(Phe)	5344	5412	69				12
ND5	5425	7092	1668	555	ATG	TAG	0
tRNA(His)	7093	7159	67				12
ND4	7172	8503	1332	443	ATG	TAA	0
ND4L	8504	8786	283	94	ATA	T-	129
tRNA(Thr)	8916	8984	69				131
tRNA(Pro)	9116	9180	65				136
ND6	9317	9848	532	177	ATG	T-	0
CYTB	9849	10964	1116	371	ATG	TAG	14
tRNA(Ser)	10979	11050	72				604
ND1	11655	12605	951	316	ATG	TAA	0
tRNA(Leu)	12606	12672	67				9
16S rRNA	12682	14008	1327				34
tRNA(Val)	14043	14097	55				23
12S rRNA	14121	14906	786				697
tRNA(Met)	15604	15673	70				6
tRNA(Ile)	15680	15747	68				37
tRNA(Gln)	15785	15854	70				72
ND2	15927	16892	966	321	ATA	TAA	2
tRNA(Trp)	16895	16964	70				40
tRNA(Cys)	17005	17075	71				92
tRNA(Tyr)	17168	17236	69				68

4	Pseudomyrmex dendroicus					
Gene	Position		Size		Codon	Intergenic
	From	To	Nucleotide	Aminoacid	Start	Stop
COX1	1	1533	1533	510	ATG	TAA
						nucleotide
						6

tRNA(Leu)	1540	1608	69				0
COX2	1609	2286	678	225	ATT	TAA	88
tRNA(Lys)	2375	2446	72				0
tRNA(Asp)	2447	2512	66				0
ATP8	2513	2686	174	57	ATT	TAA	147
ATP6	2834	3500	667	222	ATG	T-	0
COX3	3501	4280	780	259	ATG	TAA	7
tRNA(Gly)	4288	4359	72				2
ND3	4362	4709	348	115	ATG	TAA	71
tRNA(Ala)	4781	4848	68				0
tRNA(Arg)	4849	4919	71				6
tRNA(Asn)	4926	4994	69				10
tRNA(Ser)	5005	5068	64				40
tRNA(Glu)	5109	5178	70				266
tRNA(Phe)	5445	5514	70				10
ND5	5525	7192	1668	555	ATA	TAA	0
tRNA(His)	7193	7262	70				25
ND4	7288	8625	1338	445	ATG	TAA	2
ND4L	8628	8908	281	94	ATT	T-	125
tRNA(Thr)	9034	9098	65				112
tRNA(Pro)	9211	9278	68				52
ND6	9331	9862	532	177	ATG	T-	0
CYTB	9863	10987	1125	374	ATG	TAA	7
tRNA(Ser)	10995	11062	68				629
ND1	11692	12639	948	315	GTA	TAA	0
tRNA(Leu)	12640	12708	69				12
16S rRNA	12721	14079	1359				24
tRNA(Val)	14104	14174	71				6
12S rRNA	14181	14966	786				658
tRNA(Met)	15625	15693	69				6
tRNA(Ile)	15700	15765	66				118
tRNA(Gln)	15884	15956	73				57
ND2	16014	16979	966	321	ATA	TAA	9

tRNA(Trp)	16989	17058	70			17	
tRNA(Cys)	17076	17148	73			72	
tRNA(Tyr)	17221	17288	68			74	
				Pseudomyrmex			
5				feralis			
Gene	Position		Size		Codon	Intergenic	
	From	To	Nucleotide	Aminoacid	Start	Stop	
COX1	1	1533	1533	510	ATG	TAA	1
tRNA(Leu)	1535	1605	71				0
COX2	1606	2283	678	225	ATT	TAA	166
tRNA(Lys)	2450	2523	74				14
tRNA(Asp)	2538	2605	68				0
ATP8	2606	2836	231	76	ATA	TAA	426
ATP6	3263	3931	669	222	ATG	TAA	27
COX3	3959	4744	786	261	ATG	TAA	175
tRNA(Gly)	4920	4992	73				3
ND3	4996	5343	348	115	ATG	TAA	302
tRNA(Ala)	5646	5709	64				0
tRNA(Arg)	5710	5776	67				72
tRNA(Asn)	5849	5921	73				119
tRNA(Ser)	6041	6101	61				225
tRNA(Glu)	6327	6400	74				105
tRNA(Phe)	6506	6573	68				79
ND5	6653	8314	1662	553	ATT	TAA	0
tRNA(His)	8315	8383	69				11
ND4	8395	9717	1323	440	ATG	TAG	52
ND4L	9770	10054	285	94	ATA	TAG	294
tRNA(Thr)	10349	10423	75				361
tRNA(Pro)	10785	10853	69				87
ND6	10941	11474	534	177	ATG	TAA	7
CYTB	11482	12600	1119	372	ATG	TAA	13
tRNA(Ser)	12614	12682	69				139
ND1	12822	13766	945	314	ATT	TAA	0

tRNA(Leu)	13767	13834	68				0
16S rRNA	13835	15174	1340				0
tRNA(Val)	15175	15246	72				7
12S rRNA	15254	16049	796				566
tRNA(Met)	16616	16686	71				10
tRNA(Ile)	16697	16767	71				161
tRNA(Gln)	16929	17000	72				69
ND2	17070	18038	969	322		ATA TAA	2
tRNA(Trp)	18041	18111	71				229
tRNA(Cys)	18341	18412	72				137
tRNA(Tyr)	18550	18618	69				217

6
Pseudomyrmex
ferrugineus

Gene	Position		Size	Aminoacid	Codon	Intergenic	Strand
	From	To	Nucleotide		Start	Stop	nucleotide
COX1	1	1533	1533	510	ATG	TAA	7
tRNA(Leu)	1541	1611	71			0	+
COX2	1612	2289	678	225	ATT	TAA	399
tRNA(Lys)	2689	2760	72			6	+
tRNA(Asp)	2767	2833	67			0	+
ATP8	2834	3049	216	71	ATA	TAA	481
ATP6	3531	4199	669	222	ATG	TAA	46
COX3	4246	5028	783	260	ATG	TAA	274
tRNA(Gly)	5303	5371	69			3	+
ND3	5375	5722	348	115	ATG	TAA	92
tRNA(Ala)	5815	5889	75			0	+
tRNA(Arg)	5890	5954	65			53	+
tRNA(Asn)	6008	6075	68			316	+
tRNA(Ser)	6392	6453	62			202	+
tRNA(Glu)	6656	6727	72			166	+
tRNA(Phe)	6894	6963	70			0	-
ND5	6964	8627	1664	554	ATA	TA-	0
tRNA(His)	8628	8696	69			9	-

ND4	8706	10031	1326	441	ATG	TAA	47	-
ND4L	10079	10363	285	94	ATA	TAA	99	-
tRNA(Thr)	10463	10534	72				83	+
tRNA(Pro)	10618	10686	69				63	-
ND6	10750	11283	534	177	ATG	TAA	12	+
CYTB	11296	12411	1116	371	ATG	TAA	18	+
tRNA(Ser)	12430	12498	69				101	+
ND1	12600	13544	945	314	GTT	TAA	0	-
tRNA(Leu)	13545	13616	72				0	-
16S rRNA	13617	14956	1340				0	-
tRNA(Val)	14957	15027	71				46	-
12S rRNA	15074	15866	793				572	-
tRNA(Met)	16439	16511	73				8	+
tRNA(Ile)	16520	16587	68				94	+
tRNA(Gln)	16682	16750	69				65	-
ND2	16816	17784	969	322	ATA	TAA	0	+
tRNA(Trp)	17785	17853	69				140	+
tRNA(Cys)	17994	18060	67				106	-
tRNA(Tyr)	18167	18236	70				244	-

7 Pseudomyrmex
flavicornis

Gene	Position		Size		Codon	Start	Stop	Intergenic nucleotide	Strand
	From	To	Nucleotide	Aminoacid					
COX1	1	1533	1533	510	ATG	TAA	7		+
tRNA(Leu)	1541	1611	71				0		+
COX2	1612	2289	678	225	ATT	TAA	412		+
tRNA(Lys)	2702	2773	72				6		+
tRNA(Asp)	2780	2848	69				0		+
ATP8	2849	3064	216	71	ATT	TAA	474		+
ATP6	3539	4207	669	222	ATG	TAA	46		+
COX3	4254	5036	783	260	ATG	TAG	272		+
tRNA(Gly)	5309	5377	69				3		+
ND3	5381	5728	348	115	ATG	TAA	94		+

tRNA(Ala)	5823	5901	79			-2	+
tRNA(Arg)	5900	5968	69			51	+
tRNA(Asn)	6020	6087	68			306	+
tRNA(Ser)	6394	6455	62			209	+
tRNA(Glu)	6665	6736	72			165	+
tRNA(Phe)	6902	6968	67			0	-
ND5	6969	8632	1664	554	ATA	TA-	0
tRNA(His)	8633	8701	69			9	-
ND4	8711	10036	1326	441	ATG	TAA	47
ND4L	10084	10368	285	94	ATA	AAT	98
tRNA(Thr)	10467	10538	72			80	+
tRNA(Pro)	10619	10687	69			70	-
ND6	10758	11291	534	177	ATG	TAA	12
CYTB	11304	12419	1116	371	ATG	TAA	20
tRNA(Ser)	12440	12510	71			102	+
ND1	12613	13557	945	314	GTT	TAA	0
tRNA(Leu)	13558	13629	72			0	-
16S rRNA	13630	14968	1339			0	-
tRNA(Val)	14969	15039	71			44	-
12S rRNA	15084	15879	796			570	-
tRNA(Met)	16450	16522	73			8	+
tRNA(Ile)	16531	16598	68			88	+
tRNA(Gln)	16687	16755	69			67	-
ND2	16823	17791	969	322	ATA	TAA	0
tRNA(Trp)	17792	17860	69			142	+
tRNA(Cys)	18003	18069	67			111	-
tRNA(Tyr)	18181	18250	70			248	-

8 Pseudomyrmex
janzeni

Gene	Position		Size		Codon	Intergenic	Strand
	From	To	Nucleotide	Aminoacid	Start	Stop	nucleotide
COX1	1	1533	1533	510	ATG	TAA	7
tRNA(Leu)	1541	1611	71				0

COX2	1612	2289	678	225	ATT	TAA	337	+
tRNA(Lys)	2627	2698	72				6	+
tRNA(Asp)	2705	2773	69				0	+
ATP8	2774	2989	216	71	ATC	TAA	481	+
ATP6	3471	4139	669	222	ATG	TAA	38	+
COX3	4178	4960	783	260	ATG	TAA	217	+
tRNA(Gly)	5178	5246	69				3	+
ND3	5250	5597	348	115	ATG	TAA	95	+
tRNA(Ala)	5693	5771	79				-2	+
tRNA(Arg)	5770	5836	67				46	+
tRNA(Asn)	5883	5960	78				307	+
tRNA(Ser)	6268	6329	62				194	+
tRNA(Glu)	6524	6595	72				172	+
tRNA(Phe)	6768	6834	67				0	-
ND5	6835	8498	1664	554	ATA	TA-	0	-
tRNA(His)	8499	8565	67				10	-
ND4	8576	9901	1326	441	ATG	TAA	48	-
ND4L	9950	10234	285	94	ATA	TAA	91	-
tRNA(Thr)	10326	10397	72				75	+
tRNA(Pro)	10473	10541	69				65	-
ND6	10607	11140	534	177	ATG	TAA	14	+
CYTB	11155	12270	1116	371	ATG	TAA	20	+
tRNA(Ser)	12291	12359	69				102	+
ND1	12462	13406	945	314		TAA	0	-
tRNA(Leu)	13407	13478	72				0	-
16S rRNA	13479	14816	1338				0	-
tRNA(Val)	14817	14887	71				44	-
12S rRNA	14932	15726	795				571	-
tRNA(Met)	16298	16370	73				8	+
tRNA(Ile)	16379	16446	68				93	+
tRNA(Gln)	16540	16608	69				67	-
ND2	16676	17644	969	322	ATA	TAA	0	+
tRNA(Trp)	17645	17713	69				143	+

tRNA(Cys)	17857	17923	67			121	-	
tRNA(Tyr)	18045	18114	70			266	-	
9				Pseudomyrmex				
				palidus				
Gene	Position		Size		Codon	Intergenic	Strand	
	From	To	Nucleotide	Aminoacid	Start	Stop	nucleotide	
COX1	1	1533	1533	510	ATG	TAA	47	+
tRNA(Leu)	1581	1649	69				0	+
COX2	1650	2330	681	226	ATT	TAA	84	+
tRNA(Lys)	2415	2485	71				31	+
tRNA(Asp)	2517	2590	74				0	+
ATP8	2591	2767	177	58	ATA	TAA	116	+
ATP6	2884	3552	669	222	ATG	TAA	87	+
COX3	3640	4419	780	259	ATG	TAG	117	+
tRNA(Gly)	4537	4609	73				5	+
ND3	4615	4962	348	115	ATG	TAA	70	+
tRNA(Ala)	5033	5099	67				0	+
tRNA(Arg)	5100	5164	65				25	+
tRNA(Asn)	5190	5256	67				191	+
tRNA(Ser)	5448	5508	61				108	+
tRNA(Glu)	5617	5689	73				107	+
tRNA(Phe)	5797	5865	69				6	-
ND5	5872	7545	1674	557	ATT	TAG	0	-
tRNA(His)	7546	7610	65				8	-
ND4	7619	8953	1335	444	ATG	TAA	0	-
ND4L	8954	9236	283	94	ATT	T-	155	-
tRNA(Thr)	9392	9461	70				21	+
tRNA(Pro)	9483	9551	69				41	-
ND6	9593	10121	529	176	ATG	T-	0	+
CYTB	10122	11240	1119	372	ATG	TAA	11	+
tRNA(Ser)	11252	11320	69				184	+
ND1	11505	12449	945	314	ATA	TAA	0	-
tRNA(Leu)	12450	12517	68				0	-

16S rRNA	12518	13853	1336			0	-
tRNA(Val)	13854	13924	71			7	-
12S rRNA	13932	14726	795			650	-
tRNA(Met)	15377	15444	68			2	+
tRNA(Ile)	15447	15512	66			9	+
tRNA(Gln)	15522	15593	72			53	-
ND2	15647	16615	969	322	ATA	TAA	17
tRNA(Trp)	16633	16703	71				25
tRNA(Cys)	16729	16795	67				99
tRNA(Tyr)	16895	16959	65				158

10 Pseudomyrmex
particeps

Gene	Position		Size		Codon	Intergenic	Strand
	From	To	Nucleotide	Aminoacid	Start	Stop	nucleotide
COX1	1	1533	1533	510	ATG	TAA	5
tRNA(Leu)	1539	1610	72				0
COX2	1611	2291	681	226	ATT	TAA	219
tRNA(Lys)	2511	2580	70				4
tRNA(Asp)	2585	2651	67				0
ATP8	2652	2849	198	65	ATA	TAG	363
ATP6	3213	3881	669	222	ATG	TAA	10
COX3	3892	4671	780	259	ATG	TAA	177
tRNA(Gly)	4849	4915	67				0
ND3	4916	5263	348	115	ATA	TAA	468
tRNA(Ala)	5732	5801	70				-2
tRNA(Arg)	5800	5866	67				33
tRNA(Asn)	5900	5970	71				169
tRNA(Ser)	6140	6199	60				151
tRNA(Glu)	6351	6425	75				36
tRNA(Phe)	6462	6527	66				4
ND5	6532	8196	1665	554	ATA	TAA	0
tRNA(His)	8197	8268	72				11
ND4	8280	9602	1323	440	ATG	TAG	23

ND4L	9626	9910	285	94	ATA	TAA	189	-
tRNA(Thr)	10100	10166	67				353	+
tRNA(Pro)	10520	10591	72				77	-
ND6	10669	11205	537	178	ATG	TAA	7	+
CYTB	11213	12331	1119	372	ATG	TAA	30	+
tRNA(Ser)	12362	12430	69				48	+
ND1	12479	13423	945	314	ATT	TAA	0	-
tRNA(Leu)	13424	13498	75				0	-
16S rRNA	13499	14823	1325				0	-
tRNA(Val)	14824	14892	69				27	-
12S rRNA	14920	15692	773				562	-
tRNA(Met)	16255	16323	69				10	+
tRNA(Ile)	16334	16398	65				264	+
tRNA(Gln)	16663	16735	73				67	-
ND2	16803	17769	967	322	ATA	T-	0	+
tRNA(Trp)	17770	17837	68				192	+
tRNA(Cys)	18030	18098	69				106	-
tRNA(Tyr)	18205	18275	71				249	-

11 Pseudomyrmex
peperi

Gene	Position		Size	Aminoacid	Codon	Intergenic	Strand	
	From	To	Nucleotide		Start	Stop	nucleotide	
COX1	1	1533	1533	510	ATG	TAA	5	+
tRNA(Leu)	1539	1608	70			0		+
COX2	1609	2280	672	223	ATT	TAA	415	+
tRNA(Lys)	2696	2764	69				9	+
tRNA(Asp)	2774	2845	72				3	+
ATP8	2849	3058	210	69	ATA	TAG	282	+
ATP6	3341	4009	669	222	ATG	TAA	71	+
COX3	4081	4863	783	260	ATG	TAA	254	+
tRNA(Gly)	5118	5186	69				3	+
ND3	5190	5537	348	115	ATG	TAA	116	+
tRNA(Ala)	5654	5720	67				0	+

tRNA(Arg)	5721	5787	67			62	+
tRNA(Asn)	5850	5920	71			308	+
tRNA(Ser)	6229	6290	62			232	+
tRNA(Glu)	6523	6594	72			95	+
tRNA(Phe)	6690	6757	68			2	-
ND5	6760	8421	1662	553	ATA	TAA	0
tRNA(His)	8422	8491	70				10
ND4	8502	9824	1323	440	ATG	TAA	45
ND4L	9870	10154	285	94	ATA	TAA	203
tRNA(Thr)	10358	10424	67				82
tRNA(Pro)	10507	10574	68				98
ND6	10673	11206	534	177	ATG	TAA	17
CYTB	11224	12348	1125	374	ATG	TAA	21
tRNA(Ser)	12370	12439	70				203
ND1	12643	13590	948	315	ATA	TAG	0
tRNA(Leu)	13591	13661	71				0
16S rRNA	13662	15015	1354				0
tRNA(Val)	15016	15087	72				10
12S rRNA	15098	15886	789				562
tRNA(Met)	16449	16517	69				4
tRNA(Ile)	16522	16589	68				109
tRNA(Gln)	16699	16772	74				63
ND2	16836	17806	971	323	ATA	TA-	0
tRNA(Trp)	17807	17875	69				199
tRNA(Cys)	18075	18146	72				160
tRNA(Tyr)	18307	18372	66				-

12 Pseudomyrmex
veneficus

Gene	Position		Size	Aminoacid	Codon	Intergenic	Strand
	From	To	Nucleotide		Start	Stop	nucleotide
COX1	1	1533	1533	510	ATG	TAA	8
tRNA(Leu)	1542	1611	70				0
COX2	1612	2289	678	225	ATT	TAA	199

tRNA(Lys)	2489	2560	72			8	+	
tRNA(Asp)	2569	2641	73			0	+	
ATP8	2642	2857	216	71	ATT	TAA	339	+
ATP6	3197	3865	669	222	ATG	TAA	57	+
COX3	3923	4702	780	259	ATG	TAA	187	+
tRNA(Gly)	4890	4958	69			3	+	
ND3	4962	5309	348	115	ATG	TAA	106	+
tRNA(Ala)	5416	5490	75			0	+	
tRNA(Arg)	5491	5556	66			52	+	
tRNA(Asn)	5609	5683	75			328	+	
tRNA(Ser)	6012	6073	62			220	+	
tRNA(Glu)	6294	6365	72			220	+	
tRNA(Phe)	6586	6654	69			2	-	
ND5	6657	8324	1668	555	ATA	TAA	0	-
tRNA(His)	8325	8388	64			18	-	
ND4	8407	9738	1332	443	ATG	TAA	46	-
ND4L	9785	10069	285	94	ATA	TAG	244	-
tRNA(Thr)	10314	10383	70			105	+	
tRNA(Pro)	10489	10554	66			79	-	
ND6	10634	11167	534	177	ATG	TAA	39	+
CYTB	11207	12325	1119	372	ATG	TAA	27	+
tRNA(Ser)	12353	12421	69			80	+	
ND1	12502	13446	945	314	ATT	TAA	0	-
tRNA(Leu)	13447	13518	72			0	-	
16S rRNA	13519	14877	1359			0	-	
tRNA(Val)	14878	14952	75			40	-	
12S rRNA	14993	15774	782			566	-	
tRNA(Met)	16341	16413	73			12	+	
tRNA(Ile)	16426	16492	67			159	+	
tRNA(Gln)	16652	16720	69			68	-	
ND2	16789	17757	969	322	ATA	TAA	3	+
tRNA(Trp)	17761	17829	69			139	+	
tRNA(Cys)	17969	18034	66			102	-	

tRNA(Tyr)	18137	18205	69			205	-	
13				Tetraponera aethiops				
Gene	Position		Size		Codon		Intergenic	
	From	To	Nucleotide	Aminoacid	Start	Stop	nucleotide	
COX1	1	1533	1533	510	ATG	TAA	22	+
tRNA(Leu)	1556	1625	70				0	+
COX2	1626	2300	675	224	ATT	TAG	3	+
tRNA(Lys)	2304	2373	70				0	+
tRNA(Asp)	2374	2440	67				0	+
ATP8	2441	2614	174	57	ATC	TAA	153	+
ATP6	2768	3436	669	222	ATG	TAA	25	+
COX3	3462	4241	780	259	ATG	TAA	7	+
tRNA(Gly)	4249	4314	66				144	+
ND3	4459	4806	348	115	ATT	TAA	57	+
tRNA(Ala)	4864	4934	71				1	+
tRNA(Arg)	4936	5001	66				5	+
tRNA(Asn)	5007	5074	68				1	+
tRNA(Ser)	5076	5135	60				16	+
tRNA(Glu)	5152	5223	72				8	+
tRNA(Phe)	5232	5298	67				0	-
ND5	5299	6979	1681	560	ATT	T-	0	-
tRNA(His)	6980	7047	68				11	-
ND4	7059	8384	1326	441	ATG	TAG	0	-
ND4L	8385	8667	283	94	ATT	T-	2	-
tRNA(Thr)	8670	8740	71				7	+
tRNA(Pro)	8748	8818	71				31	-
ND6	8850	9380	531	176	ATA	TAA	38	+
CYTB	9419	10531	1113	370	ATG	TAA	43	+
tRNA(Ser)	10575	10643	69				10	+
ND1	10654	11604	951	316		TAA	0	-
tRNA(Leu)	11605	11671	67				13	-
16S rRNA	11685	12979	1295				0	-

tRNA(Val)	12980	13043	64			3	-
12S rRNA	13047	13811	765			548	-
tRNA(Met)	14360	14429	70			9	+
tRNA(Ile)	14439	14507	69			28	+
tRNA(Gln)	14536	14605	70			58	-
ND2	14664	15629	966	321	ATA	TAA	37
tRNA(Trp)	15667	15739	73			-8	+
tRNA(Cys)	15732	15801	70			73	-
tRNA(Tyr)	15875	15941	67			47	-

14 Tetraponera
rufonigra

Gene	Position		Size		Codon	Start	Stop	Intergenic nucleotide	Strand
	From	To	Nucleotide	Aminoacid					
COX1	1	1533	1533	510	ATG	TAA	14		+
tRNA(Leu)	1548	1615	68				0		+
COX2	1616	2290	675	224	ATT	TAA	3		+
tRNA(Lys)	2294	2362	69				0		+
tRNA(Asp)	2363	2432	70				0		+
ATP8	2433	2606	174	57	ATC	TAA	149		+
ATP6	2756	3424	669	222	ATG	TAA	21		+
COX3	3446	4225	780	259	ATG	TAA	13		+
tRNA(Gly)	4239	4308	70				162		+
ND3	4471	4818	348	115	ATT	TAA	33		+
tRNA(Ala)	4852	4916	65				1		+
tRNA(Arg)	4918	4981	64				3		+
tRNA(Asn)	4985	5051	67				1		+
tRNA(Ser)	5053	5109	57				13		+
tRNA(Glu)	5123	5193	71				8		+
tRNA(Phe)	5202	5267	66				0		-
ND5	5268	6942	1675	558	ATT	T-	0		-
tRNA(His)	6943	7012	70				22		-
ND4	7035	8354	1320	439	ATG	TAG	0		-
ND4L	8355	8637	283	94	ATT	T-	2		-

tRNA(Thr)	8640	8711	72			7	+	
tRNA(Pro)	8719	8788	70			46	-	
ND6	8835	9359	525	174	ATG	TAA	23	+
CYTB	9383	10495	1113	370	ATG	TAA	17	+
tRNA(Ser)	10513	10583	71				31	+
ND1	10615	11565	951	316	ATT	TAG	0	-
tRNA(Leu)	11566	11631	66				0	-
16S rRNA	11632	12915	1284				0	-
tRNA(Val)	12916	12980	65				3	-
12S rRNA	12984	13744	761				568	-
tRNA(Met)	14313	14380	68				14	+
tRNA(Ile)	14395	14463	69				43	+
tRNA(Gln)	14507	14575	69				58	-
ND2	14634	15599	966	321	ATA	TAA	18	+
tRNA(Trp)	15618	15685	68				-8	+
tRNA(Cys)	15678	15740	63				78	-
tRNA(Tyr)	15819	15883	65				24	-

Tabela S2 – Número de acesso, nome da espécie e referência para todos os genomas mitocondriais usados nas analyses de sintenia e filogenéticas.

Accession Number	Species name	Reference
NC_028534	<i>Linepithema humile</i>	Bi et al., Unpublished
NC_023093	<i>Leptomyrmex pallens</i>	Berman, Austin & Miller, 2014
MF417380	<i>Atta texana</i>	Almeida,C.S., Unpublished
NC_026133	<i>Myrmica scabrinodis</i>	Babbucci et al., 2014
KX951753	<i>Cardiocondyla obscurior</i>	Liu & Qian, Unpublished
NC_015075	<i>Pristomyrmex punctatus</i>	Hasegawa et al., 2011
NC_030541	<i>Wasmannia auropunctata</i>	Duan, Peng & Qian, 2016
NC_030176	<i>Vollenhovia emeryi</i>	Liu et al., 2016
NC_014669	<i>Solenopsis geminata</i>	Gotzek, Clarke & Shoemaker, 2010
NC_014672	<i>Solenopsis invicta</i>	Gotzek, Clarke & Shoemaker, 2010
NC_014677	<i>Solenopsis richteri</i>	Gotzek, Clarke & Shoemaker, 2010
NC_029357	<i>Camponotus atrox</i>	Berman, Austin & Miller, 2014
NC_030790	<i>Polyrhachis dives</i>	Song et al., Unpublished
NC_026132	<i>Formica fusca</i>	Babbucci et al., 2014
NC_026711	<i>Formica selysi</i>	Yang et al., 2016
NC_001566	<i>Apis mellifera ligustica</i>	Crozier & Crozier, 1993
NC_010967	<i>Bombus ignitus</i>	Cha et al., 2007
BK010475	<i>P. concolor</i>	This work
BK010473	<i>P. dendroicus</i>	This work
BK010474	<i>P. elongatus</i>	This work
BK010379	<i>P. feralis</i>	This work
BK010380	<i>P. ferrugineus</i>	This work
BK010381	<i>P. flavicornis</i>	This work
BK010472	<i>P. gracilis</i>	This work
BK010382	<i>P. janzeni</i>	This work
BK010383	<i>P. pallidus</i>	This work
BK010384	<i>P. particeps</i>	This work
BK010385	<i>P. peperi</i>	This work
BK010386	<i>P. veneficus</i>	This work
BK010476	<i>T. aethiops</i>	This work
BK010387	<i>T. rufonigra</i>	This work

APÊNDICE B – Artigo Publicado

Accessible molecular phylogenomics at no cost: obtaining 14 new mitogenomes for the ant subfamily Pseudomyrmecinae from public data

Gabriel A. Vieira and Francisco Prosdocimi

Instituto de Bioquímica Médica Leopoldo de Meis, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

ABSTRACT

The advent of Next Generation Sequencing has reduced sequencing costs and increased genomic projects from a huge amount of organismal taxa, generating an unprecedented amount of genomic datasets publicly available. Often, only a tiny fraction of outstanding relevance of the genomic data produced by researchers is used in their works. This fact allows the data generated to be recycled in further projects worldwide. The assembly of complete mitogenomes is frequently overlooked though it is useful to understand evolutionary relationships among taxa, especially those presenting poor mtDNA sampling at the level of genera and families. This is exactly the case for ants (Hymenoptera:Formicidae) and more specifically for the subfamily Pseudomyrmecinae, a group of arboreal ants with several cases of convergent coevolution without any complete mitochondrial sequence available. In this work, we assembled, annotated and performed comparative genomics analyses of 14 new complete mitochondria from Pseudomyrmecinae species relying solely on public datasets available from the Sequence Read Archive (SRA). We used all complete mitogenomes available for ants to study the gene order conservation and also to generate two phylogenetic trees using both (i) concatenated set of 13 mitochondrial genes and (ii) the whole mitochondrial sequences. Even though the tree topologies diverged subtly from each other (and from previous studies), our results confirm several known relationships and generate new evidences for sister clade classification inside Pseudomyrmecinae clade. We also performed a synteny analysis for Formicidae and identified possible sites in which nucleotidic insertions happened in mitogenomes of pseudomyrmecine ants. Using a data mining/bioinformatics approach, the current work increased the number of complete mitochondrial genomes available for ants from 15 to 29, demonstrating the unique potential of public databases for mitogenomics studies. The wide applications of mitogenomes in research and presence of mitochondrial data in different public dataset types makes the “no budget mitogenomics” approach ideal for comprehensive molecular studies, especially for subsampled taxa.

Submitted 7 November 2018

Accepted 10 December 2018

Published 24 January 2019

Corresponding authors

Gabriel A. Vieira,
gabriel.vieira@bioqmed.ufrj.br,
fpdocsocimi@gmail.com
Francisco Prosdocimi,
prosdocimi@bioqmed.ufrj.br

Academic editor

Kimberly Bishop-Lilly

Additional Information and
Declarations can be found on
page 17

DOI 10.7717/peerj.6271

© Copyright
2019 Vieira and Prosdocimi

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

Subjects Biodiversity, Bioinformatics, Entomology, Evolutionary Studies, Genomics

Keywords Pseudomyrmecinae, Mitogenomics, Data mining, Bioinformatics, Phylogenomics, Ant evolutionary biology, Next Generation Sequencing, Public data

INTRODUCTION

More than one decade after the advent of next-generation sequencing (NGS) (*Margulies et al., 2005*), it is now clear that this mature technology fostered an unprecedented increase in the generation of genomic data together with an important reduction in sequencing costs (*Mardis, 2008; Van Dijk et al., 2014; Goodwin, McPherson & McCombie, 2016*). In order to gather and democratize the access to genomic data, the International Nucleotide Sequence Database Collaboration (INSDC, <http://www.insdc.org/>) has been established in 1987. This continuous effort comprises three international centers: (i) the National Center for Biotechnology Information (NCBI), (ii) the European Bioinformatics Institute (EBI) and (iii) the DNA Data Bank of Japan (DDBJ) (*Karsch-Mizrachi et al., 2017*). As part of this initiative, the Sequence Read Archive (SRA) was created to host raw sequence reads and metadata generated by NGS projects (*Kodama, Shumway & Leinonen, 2012*). Making raw sequence data available is key for the experimental reproducibility (*Stoddan, Seiler & Ma, 2018*), a pillar of scientific endeavor. SRA database has been recurrently used to support new research, such as: the evaluation of single nucleotide polymorphisms and deletions (*Bordbari et al., 2017*), the test of new bioinformatics pieces of software (*Simpson et al., 2009; Langmead & Salzberg, 2012; Bolger, Lohse & Usadel, 2014*), and also to evaluate the impacts of common procedures on data, such as trimming (*Del Fabbro et al., 2013*), among other studies (*Kayal et al., 2015; Bernstein, Doan & Dewey, 2017; Linard et al., 2018*).

The availability of public data is continuously growing together with the potential uses of such databases to the scientific community. In a 2-year period (August-2015 to August-2017), 3,000 trillion base pairs have been added to SRA, promoting a 233% growth of the repository (*Karsch-Mizrachi et al., 2017*). However, potential uses of these data are far from being fully exploited, once public databases present resources that could be used to address diverse ranges of unexplored biological questions, as pointed out in the previous paragraph. Here we focus in searching for the presence of complete mitochondrial genomes in public genomics datasets.

Mitogenomes: ubiquity in datasets and relevance in scientific research

Whole-Genome Sequencing (WGS) experiments and partial genome sequencing projects normally yield enough sequencing reads from mitochondria to allow the assembly of complete mitogenomes (*Prosdocimi et al., 2012; Smith, 2015*). These small organellar genomes can be often assembled in high coverage due to the high copy number of these organelles (*Smith, 2015*). Also, previous studies indicate that it is possible to recover complete and/or nearly complete mitochondrial sequences from RNA-Seq data (*Tian & Smith, 2016; Rauch et al., 2017; Plese et al., 2018*) and targeted sequencing strategies as exome (*Picardi & Pesole, 2012; Guo et al., 2013; Samuels et al., 2013*) and UCE (Ultra Conserved Elements) off-target data (*Raposo do Amaral et al., 2015; Miller et al., 2016*). The assembly of numerous complete mitogenomes and/or large mitochondrial contigs from the sequencing of pooled multi-species samples has also been performed

successfully ([Timmermans et al., 2016](#); [Linard et al., 2018](#)) under an approach named ‘mito-metagenomics’ ([Tang et al., 2014](#)) or ‘mitochondrial metagenomics’ (MMG) ([Crampton-Platt et al., 2015](#)).

Some works have demonstrated the potential of public data to mitogenomic studies by successfully using public data to assemble mitochondrial sequences ([Diroma et al., 2014](#); [Kayal et al., 2015](#); [Linard et al., 2018](#)). However, a large number of species that have genomic data available in the SRA database are still lacking works describing their complete mitochondrial sequences.

Due to their small sizes, high conservation and the absence of introns, mitogenomes are the most commonly sequenced chromosomes, especially for metazoans ([Smith, 2015](#)). Mitochondrial genomes are poorly sampled for many taxa and therefore our current knowledge about evolutionary biology of many clades could be improved with the use of public data. Being primarily maternally inherited and non-recombinant, such sequences are often used to study evolutionary biology ([Finstermeier et al., 2013](#); [Krzemińska et al., 2017](#)), population genetics ([Pečnerová et al., 2017](#); [Kılınç et al., 2018](#)), phylogeography ([Chang et al., 2017](#); [Fields et al., 2018](#)), systematics ([Lin et al., 2017](#); [Crainey et al., 2018](#)) and conservation ([Moritz, 1994](#); [Rubinoff, 2006](#); [Rosel et al., 2017](#)) of various clades ([Avise, 1994](#)), specially from subsampled taxa ([Gotzek, Clarke & Shoemaker, 2010](#); [Duan, Peng & Qian, 2016](#)) and non-model organisms ([Prosdocimi et al., 2012](#); [Tilak et al., 2014](#); [Plese et al., 2018](#)).

Mitogenome sampling in the Formicidae family

An example of poor mitogenome taxon sampling occurs in ants (Hymenoptera: Formicidae). Despite being an ubiquitous, ecologically dominant and hyper diverse group ([Holldobler & Wilson, 1990](#)) with over 13,000 species described ([Bolton, 2012](#)), complete mitogenome records are available for mere 15 species in GenBank.

UCE sequencing data have been previously used to study ant phylogeny ([Blaimer et al., 2015](#); [Ward & Branstetter, 2017](#); [Branstetter et al., 2017](#)), but attempts to recover mitochondrial sequences from the off-target data generated are limited ([Ströher et al., 2017](#)). Thus, these pieces of information have not been used to further understand evolutionary relationships for the clade.

The Pseudomyrmecinae subfamily: taxonomy, ecology and evolution

One particular ant group that suffers from poor mitogenome sampling is the ant subfamily Pseudomyrmecinae that contains three genera: (i) the New-World genus *Pseudomyrmex*, consisting of ~137 species, most of which can be classified in one of the ten morphological species groups described ([Ward, 1989](#); [Ward, 1993](#); [Ward, 1999](#); [Ward, 2017](#)); (ii) the Paleotropical *Tetraponera*, with ~93 species; and (iii) the South American *Myrcidris*, that has only one species described, *Myrcidris epicharis* ([Ward & Downie, 2004](#); [Bolton, 2012](#); [Ward, 2017](#)).

According to [Janzen \(1966\)](#) and [Ward \(1991\)](#), there are two known Pseudomyrmecinae ecological groups: (i) one composed of generalist arboreal species, and another consisting of (ii) ants specialized in plant colonization. While ants from the Group one nest in dead

sticks of various types of plants and are generally passive in relation to external objects, ants from the Group two are obligate inhabitants of hollow cavities in live tissues (domatia) of plants and are often aggressive towards other insects or plants. Also, ants from Group two provide protection from herbivory and competition to its host plant in a relationship commonly associated with coevolved mutualism (Janzen, 1966; Ward, 1991).

Previous works using morphological and molecular data (Ward, 1991; Ward & Downie, 2004) suggest that this kind of mutualism has evolved independently at least 12 times in the Pseudomyrmecinae subfamily. For instance, *Pseudomyrmex* ants evolved similar behaviors by convergence, despite coevolving with different plant hosts (Ward & Downie, 2004; Chomicki, Ward & Renner, 2015).

Cases of convergent evolution are frequently characterized using phylogenetics approaches (Ward & Branstetter, 2017). Evolutionary analyses of mitochondrial sequences often allow a better understanding about the history of taxonomic clades in the level of family (Miya et al., 2003; Kayal et al., 2015), including inside the subphylum Hexapoda (Mao, Gibson & Dowton, 2015; Bourguignon et al., 2016). Thus, analyses of mitochondrial genes should be taken on account to study Pseudomyrmecinae, a subfamily that presents several cases of coevolution.

Several molecular studies have been described on *Pseudomyrmex*, generally addressing co-evolutionary questions, such as the impact of mutualistic associations in the rate of genome evolution (Rubin & Moreau, 2016) or characterization of ant-plant associations through the study of phylogenetic relationships and biogeography (Chomicki, Ward & Renner, 2015; Ward & Branstetter, 2017). However, complete mitogenomes analyses have never been performed for Pseudomyrmecinae due to absence of these data. In the current “no budget mitogenomics” approach (defined here as the usage of public raw data to assemble large mitochondrial sequences unavailable at public databases), we used publicly available genomic data generated elsewhere (Table S1) to assemble and analyze the complete mitochondrial sequence for 12 *Pseudomyrmex* and two *Tetraponera* species from Pseudomyrmecinae subfamily. Thus, we present the first dozen of mitogenomes for this subfamily and performed evolutionary analyses on them and all other available Formicidae mitogenomes, trying to better understand the sister clade relationships inside this highly diverse clade. Given the “no budget” nature of this work, the choice of Pseudomyrmecinae species analyzed took advantage of the availability of public data for this clade. The current study presents new complete mitochondrial sequences for ant species that cover five out of 10 *Pseudomyrmex* species groups and almost duplicates the number of mitochondrial genomes available for ants, increasing this number from 15 to 29.

METHODS

Data acquisition

Fourteen Illumina paired-end datasets were downloaded from EMBL Nucleotide Archive (<https://www.ebi.ac.uk/ena>) in SRA file format (see Table S1). The datasets containing both mitochondrial and nuclear data were converted to FASTQ using fastq-dump (with –readids and –split-files parameters) from the SRAtoolkit.2.8.2.

Mitochondrial genome assembly and annotation

The complete datasets with different number of sequencing reads were used as input for *de novo* assembly using NOVOPlasty2.6.3 ([Dierckxsens, Mardulyn & Smits, 2016](#)) with default parameters. Since NOVOPlasty was our primary assembler and [Dierckxsens, Mardulyn & Smits \(2016\)](#) recommend the use of untrimmed data for this software, we decided to use all datasets without trimming. The only exception was the dataset for *Tetraponera rufonigra*, that had to be trimmed with Trimmomatic v.0.36 ([Bolger, Lohse & Usadel, 2014](#)) to produce sequences with the same length. This trimming has been performed by setting the parameter MINLEN to match the longest (and modal) read size, therefore discarding shorter reads. NOVOPlasty assemblies needs a seed sequence to start the assembly. Seeds were selected using COX1 (Cytochrome Oxidase I) sequences from the same species (when available) or using COX1 regions from closely-related species. Preliminary mitogenome assemblies by NOVOPlasty were used as reference to a second round of genome assembly using MIRA v.4.0.2 with default parameters ([Chevreux, Wetter & Suhai, 1999](#)). NOVOPlasty does not generate an alignment file showing the reads mapped to the assembly, so MIRA has been used to map raw sequencing reads to the consensus mitochondrial sequence for the next steps. When the first assembly did not generate the complete mitogenome, we used the largest NOVOPlasty contig as reference for a first mapping assembly using MIRA. The results of this first mapping step were then used as input to MITObim v.1.9 ([Hahn, Bachmann & Chevreux, 2013](#)) with default parameters, that performed successive iterations to elongate the mitochondrial contig and assemble a circularized version of the mitochondrial genome.

Tablet software version 1.17.08.17 ([Milne et al., 2013](#)) was used with default parameters to check read coverage and circularization of complete mitogenomes. Automatic annotation performed using MITOS Web Server ([Bernt et al., 2013](#)) with default parameters and was followed by manual curation using Artemis ([Carver et al., 2012](#)). The annotation of tRNAs and rRNAs were used in accordance to MITOS Web Server data, except for removing few bases overlapping features in the same strand, when encountered. For the protein-coding genes (PCGs), in many cases, we needed to expand the annotation provided by MITOS Web Server to the closest start codon in order to match the largest Open Reading Frame (ORF) that did not overlap other features in the same strand. Then, we used the online version of BLASTp ([Altschul et al., 1997](#)) against a database of ants (clade Formicidae) to consider sequence conservation and have information available to decide about the most likely size of the protein, fine-tuning our annotation. Following this procedure, we reached a rational decision on gene boundaries. AT content for (i) the complete mitochondrial genome sequence; and (ii) the intergenic region that contains the D-loop were calculated using the OligoCalc web application ([Kibbe, 2007](#)).

Phylogenomics analyses

Formicidae phylogenetics relationships were reconstructed using (i) the 14 complete mitogenomes produced by us together with (ii) all other 15 complete mitochondrial genomes currently available for the clade; and (iii) two mitogenomes of bees (Apidae family) used as outgroups. Two phylogenetics trees have been built using (i) the whole

mitochondrial sequence; and (ii) the concatenated gene set of all 13 protein-coding genes. For the former, we manually edited the sequences to start at the COX1 gene when necessary and aligned the whole mitogenomes using ClustalW version 2.1 using default parameters (*Thompson, Gibson & Higgins, 2003*). For the latter, we aligned and concatenated the nucleotides for all protein-coding genes (PCGs) using the Phylomito script (<https://github.com/igorrcosta/phylomito>) with default parameters. Modeltest (*Posada & Crandall, 1998*) was run through MEGA7 (*Kumar, Stecher & Tamura, 2016*) with the two datasets and identified the model GTR+G+I as the nucleotide substitution model that better explained sequence variation. Aligned sequences were used as input to a Maximum Likelihood (ML) analysis in MEGA7. Resampling was conducted by bootstrap using 1000 replicates. Blast Ring Image Generator (BRIG) software v.0.95 was run with default parameters (*Alikhan et al., 2011*) to compare and visualize all mitogenomes of Pseudomyrmecinae produced here.

RESULTS

Mitogenome assembly and annotation of Pseudomyrmecinae

The 14 genomic datasets used to assemble complete mitogenome sequences for pseudomyrmecine ants were downloaded from SRA database (Table S1). Two different dataset types were used: (i) Whole Genome Sequencing (WGS), that often contained a higher amount of sequencing data totaling 212.7 Gbp (according to SRA information) for six species; an average of 35.45 Gbp per species (*Rubin & Moreau, 2016*); and (ii) UCE experiments, on which we have downloaded 5.94 Gbp for eight species; an average of 742.5 Mbp per species (*Branstetter et al., 2017*; *Ward & Bristetter, 2017*).

The complete dataset downloaded for each species was used as input for a *de novo* sequence assembly using NOVOPlasty. After this first round of genome assembly, we used a subset containing either two or four million sequencing reads as input for a second round of genome assembly using MIRA software. This procedure was performed to both map the sequencing reads into the preliminary assembly and improve the mitogenome quality. For some mitogenomes MIRA could not produce the complete, circularized mitochondrial genome; and a third round of assembly was needed. In that case, the largest contig generated by MIRA has been used as backbone to finish the assembly using MITObim (Table 1). This pipeline was capable to assemble the whole mitochondria of all Pseudomyrmecinae except for *T. aethiops*, on which we have had to use the entire sequencing read dataset for MIRA and MITObim instead of filtering the subset of reads on round 2. The use of multiple strategies to assemble the complete mitochondrial sequences for these species was expected once NGS data is variable amongst different species and sequencing runs; also, the datasets used here came from both different sources and experimental approaches probably potentializing the data variability. The 14 mitochondrial genomes built here were checked for circularity and confirmed to present, as expected for metazoans, 13 protein-coding genes, 22 tRNAs, two rRNAs and a variable control region or D-loop (*Wolstenholme, 1992*; *Prosdocimi et al., 2012*). The genome annotation for all complete mitogenomes is presented (Table S2). All mitochondrial genomes produced here were submitted to GenBank under

Table 1 Information about mitochondrial genome assemblies of Pseudomyrmecinae.

Pseudomyrmecinae species	Species ubrk group	Mitogenome TPA accession number	NOVOPlasty seed	MITObim third assembly round needed	Mitogenome coverage	Low coverage region	Mitogenome size (bp)	AT content: mitogenome (%)	AT content: D-loop region (%)
<i>P. concolor</i>	<i>P. viidus</i>	BK010475	KU985552.1	No	193.2×	No	15,906	75	91
<i>P. dendroicus</i>	<i>P. viidus</i>	BK010473	KP271186.1	Yes	123.9×	No	17,362	81	94
<i>P. pallidus</i>	<i>P. pallidus</i>	BK010383	KU985552.1	No	91.9×	No	17,117	74	84
<i>P. elongatus</i>	<i>P. oculatus</i>	BK010474	KP271181.1	No	115.4×	No	17,304	78	93
<i>P. gracilis</i>	<i>P. gracilis</i>	BK010472	FJ436821.1	No	165.5×	13,761–13,928	15,704	77	93
<i>P. feralis</i>	<i>P. ferrugineus</i>	BK010379	FJ436819.1	No	128.0×	No	18,835	78	92
<i>P. ferrugineus</i>	<i>P. ferrugineus</i>	BK010380	FJ436819.1	Yes	87.0×	No	18,480	77	90
<i>P. flavigaster</i>	<i>P. ferrugineus</i>	BK010381	FJ436819.1	Yes	152.7×	No	18,498	77	90
<i>P. janzeni</i>	<i>P. ferrugineus</i>	BK010382	FJ436819.1	No	125.8×	15,848–15,867	18,380	77	89
<i>P. particeps</i>	<i>P. ferrugineus</i>	BK010384	FJ436819.1	No	126.8×	15,799–15,820	18,524	80	90
<i>P. peperi</i>	<i>P. ferrugineus</i>	BK010385	FJ436819.1	Yes	87.4×	16,006–16,023	18,709	78	91
<i>P. veneficus</i>	<i>P. ferrugineus</i>	BK010386	FJ436819.1	No	155.4×	15,889–15,928	18,410	79	91
<i>T. aethiops</i>	NE	BK010476	KX398231.1	Yes	712.9×	13,934–13,982	15,988	79	93
<i>T. rufonigra</i>	NE	BK010387	KX398231.1	No	292.2×	13,889–13,982	15,907	74	91

the Third Party Annotation (TPA) database ([Cochrane et al., 2006](#)) that provided accession numbers allowing sequence retrieval ([Table 1](#)).

According to the average coverage estimate provided by TABLET software, the sequencing read coverage for mitogenomes ranged between 85x and 292x for mitogenomes on which a subset of reads was used. For *T. aethiops*, the coverage was higher as the entire dataset was used (712x). Assembly coverage was observed to be evenly distributed ([Fig. S1](#)), except in cases of AT-rich regions that presented low coverage, generally close to poly-T sequences.

Mitogenome size variation and putative insertion sites in the *Pseudomyrmex* genus

Pseudomyrmex mitogenomes have shown significant variation in size, ranging from 15,704 to 18,835 bp ([Table 1](#)). We observed three distinct mitogenome size ranges for the clade ([Table 1](#)). Mitogenome size in the genus varied from: (i) less than 16 kb in *P. gracilis* and *P. concolor*; (ii) between 17 kb and 18 kb in *P. pallidus* and *P. dendroicus*; and (iii) higher than 18 kb in other species, that belong to *P. ferrugineus* group. A comparative genomics analysis using BRIG software identified four variable regions as putative insertion segments ([Fig. 1](#)). After genome annotation, we identified these presumed insertions to be located between (i) COX2 and *trn-K*; (ii) ATP8 and ATP6; (iii) *trn-N* and *trn-F*; and (iv) *trn-W* and COX1.

Gene order arrangements in ant mitogenomes

Regardless of the limited sample of complete mitochondrial genomes analyzed for ants, in general, five slightly different synteny rearrangements ([Fig. 2](#)) could be observed in Formicidae family ([Duan, Peng & Qian, 2016](#)). All Pseudomyrmecinae and Dolichoderinae mitogenomes analyzed showed a single conserved gene arrangement for all species that is also shared by most of Formicinae species. We also observed that Formicinae and Myrmicinae clades present a modal synteny arrangement suggesting a possible ancestral gene arrangement for each group. One single species of Formicinae (*Camponotus atrox*) presents inversions between *trn-M*, *I* and *Q* that differ from other mitogenomes from this subfamily, possibly representing a derived variation. Myrmicinae also present two other unique rearrangements restricted to a single species each, suggesting derived syntenies: (i) *P. punctatus* has an inversion between *trn-K* and *D*; and (ii) *W. auropunctata* presents both an inversion between *trn-V* and D-loop and a feature (*trnY*) on the opposite strand when compared to the others.

Phylogenetic analyses of Formicidae using mitogenome data

In order to assess the phylogeny of the group, two Maximum Likelihood trees were produced using slightly different input data: (i) the aligned and concatenated sequences for all 13 mitochondrial PCG's ([Fig. 3](#)); and (ii) the complete mitochondrial genomes ([Fig. 4](#)). We analyzed all ant species presenting complete mitogenomes available on Genbank ([Gotzek, Clarke & Shoemaker, 2010](#); [Hasegawa et al., 2011](#); [Berman, Austin & Miller, 2014](#); [Babbucci et al., 2014](#); [Kim, Hong & Kim, 2016](#); [Duan, Peng & Qian, 2016](#); [Liu et al., 2016](#)); [Yang et al., 2015](#)) and two Apidae bees as outgroups ([Crozier & Crozier,](#)

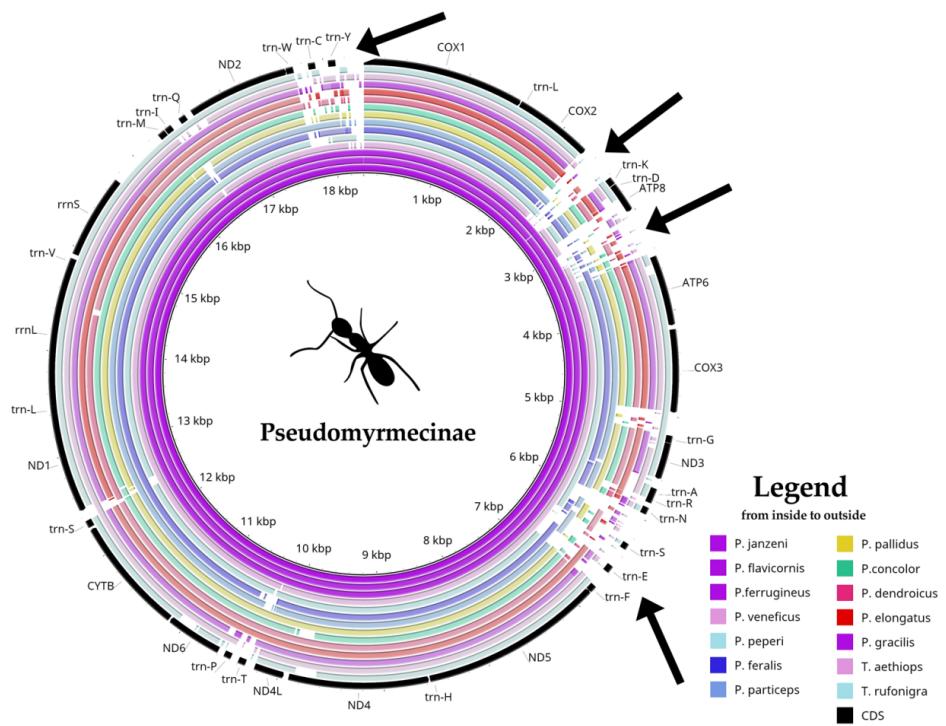


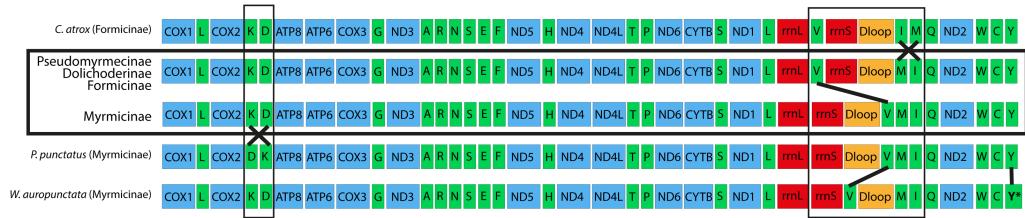
Figure 1 Comparative genomics analysis of all 14 Pseudomyrmecinae ants. BLAST comparison of all Pseudomyrmecinae mitochondrial genomes against a reference (*Pseudomyrmex janzeni*) generated by Blast Ring Image Generator (BRIG). Gaps in rings correspond to regions with less than 50% identity to the reference sequence. Most mitochondrial features are conserved within the clade, even though ATP8 and some tRNAs (trn-S, trn-E and trn-T) were observed to be less conserved. Four regions (identified by arrows) present nucleotide size variations between (i) COII and trn-K; (ii) ATP8 and ATP6; (iii) trn-N and trn-F and; (iv) trn-W and COI.

Full-size DOI: 10.7717/peerj.6271/fig-1

1993; Cha et al., 2007) (see accession numbers and references for all sequences on Table S3). The trees reconstructed from mitochondrial data corroborated most of the phylogenetic relationships known for ants, with several clades observed as monophyletic with high confidence (bootstrap = 100). Both trees showed similar results, though differences can be observed in several nodes regarding tree topology and/or statistical support. The major difference observed is that the gene-concatenation tree displayed all subfamilies as monophyletic, while Myrmicinae was recovered as paraphyletic in the tree based on complete mitogenomes.

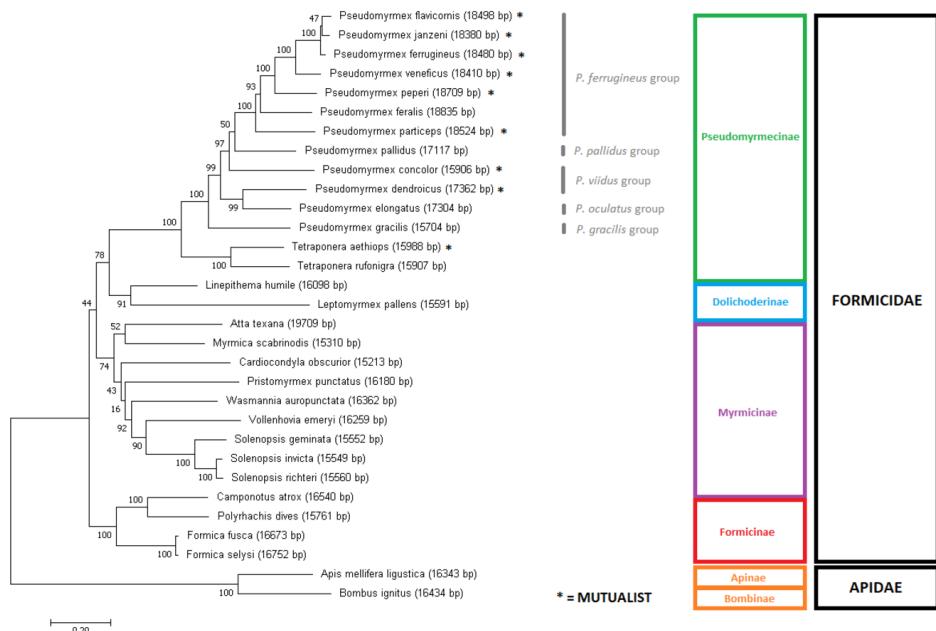
DISCUSSION

In this study, we used public data to assemble, annotate, compare and provide evolutionary analyses of 14 complete mitochondrial genome sequences from the ant subfamily Pseudomyrmecinae plus 15 other ant mitogenomes downloaded from GenBank.

**Figure 2** Five different syntenies observed in complete Formicidae mitogenomes available on

Genbank. The two modal gene arrangements are represented inside the horizontal rectangle and were observed in 26 out of 29 species analyzed: all Pseudomyrmecinae (14 species); all Dolichoderinae (two species: *L. pallens* and *L. humile*); three out of four Formicinae (*F. fusca*, *F. selysi* and *P. dives*) and seven out of nine Myrmicinae (*A. texana*; *C. obscurior*; *M. scabrinodis*; *S. richteri*; *S. geminata*; *S. invicta*; *V. emeryi*). We suggest that these may represent the ancestral arrangements for their clades. The syntenies outside the rectangle correspond to unique gene orders encountered in single species. Vertical rectangles and lines indicate regions on which synteny changes occurred, and both the asterisk (*) and the vertical line in the *trn-Y* of *W. auropunctata* indicates that it is the only feature in Formicinae mitochondria that changed its coding strand.

Full-size DOI: 10.7717/peerj.6271/fig-2

**Figure 3** Gene-concatenation phylogenetic tree for all Formicidae complete mitogenomes available on Genbank. The tree was built using the aligned and concatenated nucleotidic sequences for all 13 protein-coding mitochondrial genes. Modeltest identified 'GTR + G + I' as the most adequate substitution model and phylogeny was reconstructed by Maximum Likelihood using MEGA7 software, with 1000 bootstrap replicates. Bees from the Apidae family were used as outgroup. *Pseudomyrmex* species groups are described and mutualistic pseudomyrmecines are evidenced by the presence of an asterisk (*).

Full-size DOI: 10.7717/peerj.6271/fig-3

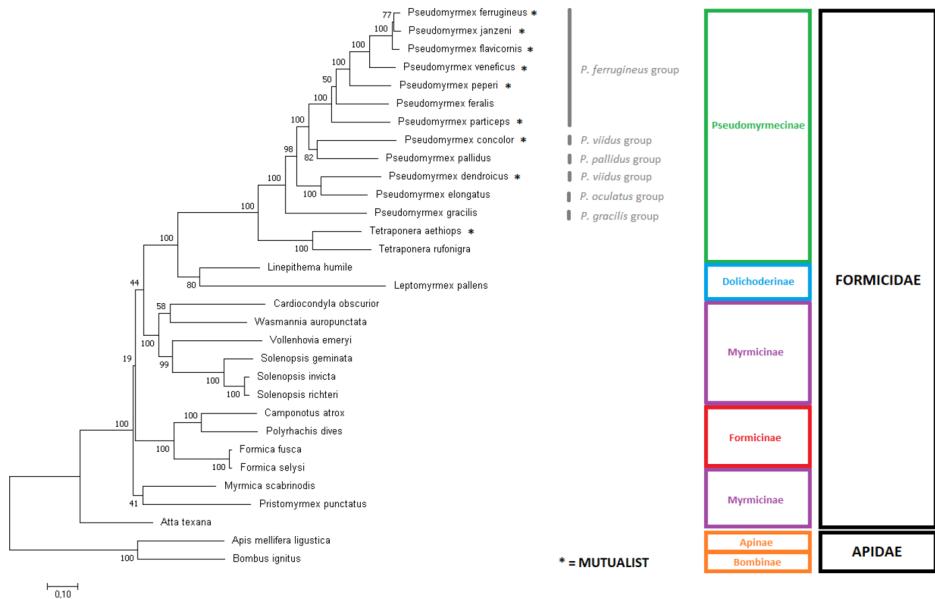


Figure 4 Phylogenetic tree using the complete mitochondrial sequence of all complete ant mitogenomes available on Genbank. 'GTR + G+I' was chosen as substitution model as suggested by Modeltest. The tree was built with MEGA7 using Maximum Likelihood with 1,000 bootstrap replicates. Mitogenomes from bees were used as outgroups. *Pseudomyrmex* species groups and mutualistic pseudomyrmecines are evidenced.

Full-size DOI: 10.7717/peerj.6271/fig-4

Uniform genome coverage and expected AT-bias

Even though pieces of the mitochondrion genome may be copied to the nucleus forming NuMTs (Nuclear Mitochondrion Sequences), the genome coverage obtained for the assemblies often presented uniform distributions (Fig. S1), even for *Pseudomyrmex gracilis* on which NuMTs have been previously identified (Rubin & Moreau, 2016). The correct assembly of mitochondrial genomes was possible because the number of mitochondrial reads is probably much higher than the number of reads coming from NuMTs.

The low coverage in segments with a pronounced AT-bias should be expected because AT-rich regions are known to have reduced amplification in Illumina library preparation protocols (Dohm et al., 2008; Aird et al., 2011; Oyola et al., 2012). It has been shown that ant mitogenomes have a remarkable AT bias in the D-loop that exceeds 90% (Berman, Austin & Miller, 2014; Liu et al., 2016). Our data also corroborates this, as 12 out of 14 *Pseudomyrmecinae* species presents an AT content value that is equal or superior to 90% in the intergenic region between rrnS and trn-M that contains the D-loop (Table 1). Also, this region has already been proved to be particularly difficult to sequence in hymenopterans (Castro & Dowton, 2005; Dowton et al., 2009; Rodovalho, 2014). The difficulties in obtaining the complete mitogenomes for ants could be the reason why there are so few mitochondrial genomes available for this group despite the availability of public data. At the same time, new assemblers like NOVOPlasty outperform classic programs (Dierckxsens, Mardulyn & Smits, 2016; Plese et al., 2018), facilitating the production of complete mitogenomes. These

advances provide favorable prospects for the closure of the mitogenome phylogenetic gaps in Formicidae, especially if public data is employed for that end.

Comparative mitogenomics: mitogenome size and synteny analyses

Aside from the identification of four putative insertion sites that could explain the differences observed in mitogenome size (pointed by arrows in Fig. 1), we also observed that all seven mitogenomes included in *P. ferrugineus* group have approximately the same genome size in bp, suggesting that this group is monophyletic. On the other hand, there is a significant difference in mitogenome size between *P. concolor* (15906 bp) and *P. dendroicus* (17362 bp), both belonging to *P. viidus* species group. This corroborates previous works indicating that this species group is paraphyletic ([Ward, 1989](#); [Ward & Downie, 2004](#)).

There is a positive correlation between the multiple syntenies encountered within Myrmicinae and Formicinae clades and the remarkable biodiversity observed for these subfamilies: Myrmicinae is the largest ant subfamily in species richness, with over 6,600 species described, almost half of all biodiversity documented for ants; and Formicidae is the second most biodiverse, featuring over 3,100 species. Other Formicidae subfamilies in this study are not nearly as diverse: Dolichoderinae has ~713 species while Pseudomyrmecinae presents ~231 species documented ([Bolton, 2012](#)). Ancestral gene arrangement for Formicinae is identical to the one observed in Pseudomyrmecinae and Dolichoderinae, signaling that Formicinae is closely related to this group than to Myrmicinae.

A higher number of mitogenomes and broader taxon coverage will improve the assessment of correlation between mitochondrial gene order and subfamily biodiversity, allowing a better understanding of synteny evolution in ant mitochondria.

Phylogenomic relationships of Formicidae inferred using mitogenome data

The phylogenomic trees generated provided slightly different topologies due to the additional data present in intergenic regions, such as tRNAs, rRNAs and D-loop ([Wolstenholme, 1992](#); [Prosdocimi et al., 2012](#)). It is also known that mitochondrial DNA presents a relatively high substitution rate in non-coding regions ([Vanecek, Vorel & Sip, 2004](#); [Desalle, 2017](#)).

Overall, in the phylogenomic trees generated for the whole Formicidae family, the phylogeny of the subfamily Pseudomyrmecinae was strongly retained as monophyletic, and the phylogenetic positions of most clades were well resolved. The monophyly for the Pseudomyrmecinae subfamily and also for *Pseudomyrmex* and *Tetraponera* genera were recovered with 100% bootstrap support (BS) in both trees. The genus *Pseudomyrmex* presented few unsupported nodes, but *Tetraponera* was completely resolved on both trees (BS = 100). In both trees, both the monophyletic status of the *P. flavidornis* group and the paraphyletic status of the *P. viidus* group confirms (i) previous observations based exclusively on morphology ([Ward, 1989](#)), (ii) phylogenies using both morphological characters and few nuclear markers ([Ward & Downie, 2004](#)), and (iii) our own observations regarding mitogenome size. Although the morphological division in species groups has not been formalized or regulated under nomenclatures ([Ward, 2017](#)), the work using a

hybrid morphological/molecular approach of [Ward & Downie \(2004\)](#) shows that only two out of nine groups defined at the time were paraphyletic: *P. pallens* and *P. viidus* groups. The corroboration of morphological studies by mitochondrial data analysis confirms the relevance of using morphological characters in determining relationships between clades, but also reinforces that molecular evidence can clarify and complement such studies, refining and improving the overall support of the phylogenies reconstructed. Under this work, we generated complete mitochondrial sequences for ants representing five out of the 10 groups described for *Pseudomyrmex* species, covering at least half of *Pseudomyrmex* genetic diversity and adding a new source of molecular evidence for further studies on the clade.

Both trees suggest strongly that ant-plant mutualisms are paraphyletic in *Pseudomyrmecinae* (please check asterisks in [Figs. 3](#) and [4](#)), also adding evidence to previous assumptions of generalist behavior as a basal trait in the *Pseudomyrmex* genus ([Ward & Branstetter, 2017](#)). This suggests that ant-plant coevolution developed later (and independently) several times in the clade. Mutualistic species are more common in the *P. ferrugineus* species group, strengthening the hypothesis of mutualism being a derived trait in *Pseudomyrmecinae*. In the *Pseudomyrmex* genus, the *P. ferrugineus* group features may present two independent lineages of mutualistic ants (considering that *P. feralis* is often considered to display generalist behavior; BS = 50), while other two independent mutualistic lineages can be observed by the phylogenetic placement of the species *P. concolor* and *P. dendroicus*. Considering the *Tetraponera* genus, *T. aethiops* and *T. rufonigra* are closely related species and only *T. aethiops* presents exclusive ant-plant mutualistic behavior. This shows that evolution of mutualistic traits in *Pseudomyrmecinae* may have occurred in close related species. So, considering the limited number of species sampled here, we were able to identify five out of the 12 times that mutualistic associations have been reported to appear in the clade ([Ward, 1991](#); [Ward & Downie, 2004](#)). With a better taxonomic coverage, this number can be increased and new analyses performed, further improving our understanding about these coevolutionary events.

Well resolved relationships for several *Pseudomyrmecinae* species (such as *P. peperi*, *P. veneficus*, *P. particeps*, *P. gracilis*, *T. aethiops* and *T. rufonigra*) corroborate both the results of [Ward & Downie \(2004\)](#) and the ML tree generated using UCE data from [Ward & Branstetter \(2017\)](#). The sister group relationship between *P. dendroicus* and *P. elongatus* is also well supported (BS = 100 in complete mitochondria tree; and BS = 99 in gene-concatenation tree), in line with a recent work using concatenated WGS scaffolds as input for ML tree reconstruction ([Rubin & Moreau, 2016](#)).

However, subtle differences were observed between our results and the inferred UCE multiloci phylogenetic relationships ([Ward & Branstetter, 2017](#)). Using UCE data, *P. janzeni* was observed as sister group to *P. ferrugineus*. Here, the complete mitogenome tree recaptured this same relationship with a bootstrap replicate value of 77. On the other hand, in the concatenated gene set, the sister group relationship observed between *P. janzeni* and *P. flavigaster* showed a lower support (BS = 47). Overall, this relationship seemed to be better resolved by the analysis of the complete mitochondrial sequence, also corroborating the UCE analyses.

Both trees showed Dolichoderinae subfamily as monophyletic, even though this result was not recovered in all replicates. Dolichoderinae is a highly diverse subfamily and contains over 700 species, but it has been represented here by merely two species. Thus, we believe that a higher coverage of species will improve the robustness of the phylogenetic analyses.

Previous work with morphological characters and/or nuclear genes presents evidence of sister group relationship between Pseudomyrmecinae and Myrmeciinae ([Ward & Downie, 2004](#); [Brady et al., 2006](#)). We also expected Myrmeciinae to be sister group to Pseudomyrmecinae according to mitochondrial data, but once complete mitochondrial genomes are not available for the subfamily Myrmeciinae we could not test this hypothesis. The complete absence of mitogenomes for this and other subfamilies might lie in the fact that their diversity is not expressive in comparison to the species richness of the most diverse, sampled and studied ant subfamilies. While Myrmeciinae has only 94 described species, Myrmicinae has over 6,600 species ([Bolton, 2012](#)). In the absence of Myrmeciinae, Dolichoderinae is expected to be the closest relative to Pseudomyrmecinae in our trees, which occurs in, both trees corroborates large-scale molecular phylogenies using few nuclear genes ([Brady et al., 2006](#)) and UCE data ([Branstetter et al., 2017](#)). Shared synteny between all Pseudomyrmecinae and Dolichoderinae sampled also supports the sister group relationship observed. Our results support the evidence that Myrmicinae as sister taxa to a clade containing both Pseudomyrmecinae and Dolichoderinae, while Formicinae has been observed as a more basal group in the Formicidae family. This position for Formicinae is highly supported in the gene-concatenation tree but not in the tree using complete mitogenomes. This position is not supported by other works using nuclear data that supports the evidence of a sister group relationship between Myrmicinae and Formicinae ([Brady et al., 2006](#); [Branstetter et al., 2017](#)).

The monophyly of the subfamily Formicinae and all its nodes show maximum support on both trees (BS = 100). These trees also confirm the monophyly for the genus *Formica* and show genera *Camponotus* and *Polyrhachis* as closely related to each other, as observed in the work of Blaimer and collaborators ([2017](#)) that used UCE loci for tree inference. The only issue in this subfamily concerns the unsupported phylogenetic placement of Formicinae in relation to other subfamilies. Mitogenome data successfully delivered sound phylogenetic relationships even for *Camponotus atrox* that showed a unique synteny but have had its position well resolved in both trees, including in the complete mitochondrial tree, that may be prone to suffer from synteny changes. This issue confirms the robustness of mitochondrial sequences to infer ant phylogenies.

Overall, the most controversial results obtained here are related to the position of the subfamily Myrmicinae. For that clade, the gene concatenation tree was capable to indicate monophyly (BS = 74) but whole mitogenome data produced paraphyly. In the latter case, the myrmicine ants *Atta texana*, *Myrmica scabrirostris* and *Pristomyrmex punctatus* derived earlier than the other ants. On the other hand, some relationships were recovered with 100% bootstrap support, such as the monophyly of the genus *Solenopsis*. Our results corroborate those obtained by the use of concatenated amino acid sequences of all mitochondrial PCGs for tree inference ([Duan, Peng & Qian, 2016](#)). However, our assessment of the position of *V. emeryi* was better supported (BS = 90 on gene-concatenation tree and BS = 99 on

complete mitochondria tree) than that of this previous work ($BS = 75$). Considering that Duan and collaborators (2016) used a similar approach to ours (gene concatenation under a Maximum Likelihood method), we may conclude that these better results indicate that nucleotide data presents more reliable information for these clades than amino acid data. This is consistent with previous comparative work in which tree inference at nucleotide level has outperformed amino acid and codon analyses (Holder, Zwickl & Dessimoz, 2008). Moreover, bootstrap values obtained from nucleotide data have been reported to be often higher than their amino acid correspondents (Regier et al., 2010). This observation is at least partly explained by the differences in the amount of phylogenetic signal considered by these two methods. Additional signal present in nucleotide data is missed when they are translated into amino acids. This is particularly important when hexadecadic amino acids are considered; serine, for instance, is encoded by TCN and AGY (Regier et al., 2010; Zwick, Regier & Zwickl, 2012).

In both works, mitogenome analyses were not fully capable of resolving important nodes of the myrmicine branch and several factors may be involved in these unsatisfactory results. It is necessary to highlight that Myrmicinae is the most biodiverse ant subfamily (Bolton, 2012) and it is known to feature several dubious monophyletic groups (Brady et al., 2006; Ward, 2011; Ward et al., 2015). This diversity is evidenced by the fact that, despite only nine mitogenomes are available for the group, three different mitochondrial gene arrangements can be observed, suggesting a high rate of mitochondrial evolution in this subfamily.

Also, there have been divergences in the Myrmicinae branch of previous molecular phylogenetic studies attempting to study the Formicidae family (Brady et al., 2006; Moreau et al., 2006). On the other hand, Ward et al. (2015) focuses on the subfamily by reconstructing a large-scale phylogeny using 11 nuclear markers from 251 species sampled across all 25 myrmicine tribes, most of them nonmonophyletic. By using such huge amounts of data covering a great part of Myrmicinae species diversity, they managed to propose a new classification of Myrmicinae consisting of exclusively monophyletic tribes, which also reduced the number of genera that are not monophyletic.

Thus, the hyperdiverse nature of this clade, associated to poor taxon sampling and a possible high rate of mitochondrial genome evolution may have contributed to produce inconclusive results in mitochondrial analyses. Also, even though some relationships were not elucidated by mitochondrial phylogenomics alone, the information provided by the mitogenome has been proven several times to be useful in the study of evolutionary relationships for several taxa, either confirming (Prosdocimi et al., 2012; Finstermeier et al., 2013) or refuting previous phylogenetic hypotheses (Kayal et al., 2015; Uliano-Silva et al., 2016). Therefore, we still recommend the use of mitochondrial data, preferably alongside other markers (i.e., nuclear genes), to increase phylogenetic signal and recapture phylogenies. However, due to the substitution rate of mtDNA, trees generated from mitochondrial data have higher probability of resolving short internodes correctly (Desalle, 2017). Thus, we also believe that mitochondrial data alone will yield better results for this and other branches if we address the shortage of mitogenomes available for this clade by improving

mitochondrial taxon coverage and reducing tree internodes. In that sense, results present here are extremely relevant to show that information already available in public databases should be used to obtain such sequences at no additional sequencing costs.

No budget mitogenomics: integrative analyses between datasets and potential for large-scale studies

The results presented here confirm that both UCE and WGS data publicly available can be used to assemble complete mitochondrial genomes with high coverage (Table 1), which can be explained by the high copy number of mitochondrial genome reads compared to nuclear genomes sequencing reads that may reach something between 0.25% to 0.5% of the total number of bases generated ([Prosdocimi et al., 2012](#)), sometimes reaching percentages as high as 2% of reads mapping to mtDNA ([Ekblom, Smeds & Ellegren, 2014](#)). We also confirm the potential of UCE data as a low-cost alternative to sequence complete mitogenomes with high coverage as described by [Raposo do Amaral et al. \(2015\)](#). Mitogenome data is used in various types of analyses and mitochondrial sequences are encountered in several types of datasets, normally providing enough information to assemble the entire mitochondrial sequence. This versatility and ubiquity of mitogenome information should be used in favor of biodiversity studies, especially considering the increasingly available public datasets for a great number of species.

The potential of these sequences in unveiling phylogenies must not be overlooked, especially if we consider that there are different dataset types available for different species (WGS, RNA-Seq, UCE enrichment, among others). These different resources makes it difficult to achieve an integrated phylogenetic/phylogenomic analyses using the public data, that often depends on sequence orthology to be performed ([Kuzniar et al., 2008](#)). Thus, the use of different types of data to assemble the complete or nearly complete mitogenomes for species with publicly available data presents a solution to this problem, with the mitochondrial genome acting as a “normalizing sequence” that allows the comparison of different datasets. For instance, in this work some species had only UCE data publicly available, while others presented standard WGS datasets. Yet, by assembling, annotating and analyzing the complete mitogenome for these species, we were able to broaden our scope and study all of them together. Thus, we suggest that the use of mitogenomes obtained from public data has the potential to become an important source of phylogenetic information. Besides, the study of mitochondrial sequences may be one of the fastest routes towards a high-quality comprehensive species-level tree for hyperdiverse taxa such as insects. Steps have been taken that way, as it can be seen on recent work by [Linard et al. \(2018\)](#), where data mining from Genbank and assembly of metagenomic datasets provided mitochondrial contigs (>3 kbp) for almost 16,000 coleopteran species. This huge amount of data was used to generate the largest phylogenetic tree for the clade.

Studies that attempt to assemble complete mitogenomes using public data are yet scarce whereas the size and breadth of public databases is ever growing, along with its potential to answer phylogenetic questions. No budget mitogenomics represents an unprecedented opportunity to reconstruct and analyze large-scale phylogenies for various groups at

different taxa levels, which in turn may help other evolutionary and conservation biology studies and promote an overall increase on our knowledge about non-model species and their diversity.

CONCLUSION

Here we assembled and annotated the first 14 mitogenomes for the ant subfamily Pseudomyrmecinae using a pipeline that relies solely on public data from different sources and types, making profit of bioinformatics software publicly available. These sequences were used to study synteny, comparative genomics and phylogenomic analyses providing valuable information regarding Pseudomyrmecinae phylogeny and evolution, such as: (i) identification of four putative mitochondrial indel sites in *Pseudomyrmex*; (ii) corroboration that mutualistic associations independently arose in the clade many times; (iii) indication that *P. ferrugineus* group is monophyletic and *P. viidus* species groups is paraphyletic; and (iv) corroboration of monophyletism for the *Pseudomyrmex* and *Tetraponera* genera. Mitochondrial data on other ant clades, though limited, were useful in both synteny and phylogenomic analyses to broaden our scope and allow the study other ant groups. This allowed us to unveil sister group relationships throughout the family, such as the one between Pseudomyrmecinae and Dolichoderinae; as well as the monophyletic status of all subfamilies analyzed. However, a more precise definition about the relationship among groups should make use of large genome datasets and gene concatamers built with hundreds to thousands of genes; currently unavailable. Besides, low bootstrap values observed at some nodes, indicate that mitochondrial data available do not present enough variability to elucidate some relationships. The mitochondrial sequences assembled cover a considerable portion of Pseudomyrmecinae biodiversity and will be useful for further evolutionary and conservational studies. This work practically doubles the number of complete ant mitogenomes available at no additional sequencing costs. Since mitogenome taxon coverage is still lacking for Formicidae, its improvement is desirable for better resolution and robustness of large scale phylogenies in the group. This pipeline can also be used to study the aforementioned dubious monophyletic clades in Myrmicinae ([Brady et al., 2006](#); [Ward, 2011](#); [Ward et al., 2015](#)) or paraphyletic groups, such as the *Camponotus* genus from the Formicinae subfamily ([Blaimer et al., 2015](#)). Based on these results, we emphasize that the ever-increasing breadth of public databases, associated to the possibility of obtaining mitochondrial sequences from different types of sequencing data makes no budget mitogenomics the ideal approach for the study of species diversity and, possibly, the fastest route toward species-level phylogenetic trees.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by CAPES and Fundação de Amparo a Pesquisa do Estado do Rio de Janeiro (FAPERJ) (grant numbers 202.810/2015 and 202.780/2018). The funders had

no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
CAPES.

Fundação de Amparo a Pesquisa do Estado do Rio de Janeiro (FAPERJ): 202.810/2015,
202.780/2018.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Gabriel A. Vieira and Francisco Prosdocimi conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:
The following information was supplied regarding the deposition of DNA sequences:

GenBank accession numbers [BK010472–BK010476](#) and [BK010379–BK010387](#).

The complete sequence for all assembled mitogenomes is provided in [File S1](#).

Data Availability

The following information was supplied regarding data availability:

The direct links for download of all 14 Pseudomyrmecinae genomic public datasets (in SRA file format) used for mitogenome assembly are provided in [Table S1](#).

Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj.6271#supplemental-information>.

REFERENCES

- Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* 12:Article R18 DOI [10.1186/gb-2011-12-2-r18](https://doi.org/10.1186/gb-2011-12-2-r18).
- Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA. 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12:402 DOI [10.1186/1471-2164-12-402](https://doi.org/10.1186/1471-2164-12-402).
- Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, Miller W, Lipman D. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25:3389–3402 DOI [10.1093/nar/25.17.3389](https://doi.org/10.1093/nar/25.17.3389).

- Avise JC.** 1994. *Molecular markers, natural history and evolution*. Boston: Springer US
DOI 10.1007/978-1-4615-2381-9.
- Babbucci M, Basso A, Scupola A, Patarnello T, Negrisolo E.** 2014. Is it an ant or a butterfly? convergent evolution in the mitochondrial gene order of hymenoptera and lepidoptera. *Genome Biology and Evolution* **6**:3326–3343 DOI 10.1093/gbe/evu265.
- Berman M, Austin CM, Miller AD.** 2014. Characterisation of the complete mitochondrial genome and 13 microsatellite loci through next-generation sequencing for the New Caledonian spider-ant Leptomyrmex pallens. *Molecular Biology Reports* **41**:1179–1187 DOI 10.1007/s11033-013-2657-5.
- Bernstein M, Doan A, Dewey C.** 2017. MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive. *Bioinformatics* **33**:2914–2923.
- Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, Pütz J, Midendorf M, Stadler PF.** 2013. MITOS: improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution* **69**:313–319 DOI 10.1016/j.ympev.2012.08.023.
- Blaimer BB, Brady SG, Schultz TR, Lloyd MW, Fisher BL, Ward PS.** 2015. Phylogenomic methods outperform traditional multi-locus approaches in resolving deep evolutionary history: a case study of formicine ants. *BMC Evolutionary Biology* **15**:271 DOI 10.1186/s12862-015-0552-5.
- Bolger AM, Lohse M, Usadel B.** 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**:2114–2120 DOI 10.1093/bioinformatics/btu170.
- Bolton B.** 2012. AntCat. An online catalog of the ants of the world. Available at <http://www.antcat.org/> (accessed on 22 June 2018).
- Bordbari MH, Penedo MCT, Aleman M, Valberg SJ, Mickelson J, Finno CJ.** 2017. Deletion of 2.7 kb near HOXD3 in an Arabian horse with occipitoatlantoaxial malformation. *Animal Genetics* **48**:287–294 DOI 10.1111/age.12531.
- Bourguignon T, Lo N, Šobotník J, Ho SYW, Iqbal N, Coissac E, Lee M, Jendryka MM, Sillam-Dussès D, Křížková B, Roisin Y, Evans TA.** 2016. Mitochondrial phylogenomics resolves the global spread of higher termites, ecosystem engineers of the Tropics. *Molecular Biology and Evolution* **34**(3):589–597 DOI 10.1093/molbev/msw253.
- Brady SG, Schultz TR, Fisher BL, Ward PS.** 2006. Evaluating alternative hypotheses for the early evolution and diversification of ants. *Proceedings of the National Academy of Sciences of the United States of America* **103**:18172–18177 DOI 10.1073/pnas.0605858103.
- Branstetter MG, Longino JT, Ward PS, Faircloth BC.** 2017. Enriching the ant tree of life: enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. *Methods in Ecology and Evolution* **8**:768–776 DOI 10.1111/2041-210X.12742.
- Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA.** 2012. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28**:464–469 DOI 10.1093/bioinformatics/btr703.

- Castro LR, Downton M.** 2005. The position of the Hymenoptera within the Holometabola as inferred from the mitochondrial genome of *Perga condei* (Hymenoptera: Symphyta: Pergidae). *Molecular Phylogenetics and Evolution* **34**:469–479 DOI [10.1016/j.ympev.2004.11.005](https://doi.org/10.1016/j.ympev.2004.11.005).
- Cha SY, Yoon HJ, Lee EM, Yoon MH, Hwang JS, Jin BR, Han YS, Kim I.** 2007. The complete nucleotide sequence and gene organization of the mitochondrial genome of the bumblebee, *Bombus ignitus* (Hymenoptera: Apidae). *Gene* **392**:206–220 DOI [10.1016/j.gene.2006.12.031](https://doi.org/10.1016/j.gene.2006.12.031).
- Chang D, Knapp M, Enk J, Lippold S, Kircher M, Lister A, Macphee RDE, Widga C, Czechowski P, Sommer R, Hodges E, Stümpel N, Barnes I, Dalén L, Derevianko A, Germonpré M, Hillebrand-Voiculescu A, Constantin S, Kuznetsova T, Mol D, Rathgeber T, Rosendahl W, Tikhonov AN, Willerslev E, Hannon G, Lalueza-Fox C, Joger U, Poinar H, Hofreiter M, Shapiro B.** 2017. The evolutionary and phylogeographic history of woolly mammoths: a comprehensive mitogenomic analysis. *Scientific Reports* **7**:44585 DOI [10.1038/srep44585](https://doi.org/10.1038/srep44585).
- Chevreux B, Wetter T, Suhai S.** 1999. October. Genome sequence assembly using trace signals and additional sequence information. In *German conference on bioinformatics* **99**(1):45–56.
- Chomicki G, Ward PS, Renner SS.** 2015. Macroevolutionary assembly of ant/plant symbioses: *Pseudomyrmex* ants and their ant-housing plants in the Neotropics. *Proceedings of the Royal Society B: Biological Sciences* **282**:Article 20152200 DOI [10.1098/rspb.2015.2200](https://doi.org/10.1098/rspb.2015.2200).
- Cochrane G, Bates K, Apweiler R, Tateno Y, Mashima J, Kosuge T, Mizrahi IK, Schafer S, Fetchko M.** 2006. Evidence standards in experimental and inferential INSDC third party annotation data. *OMICS: A Journal of Integrative Biology* **10**:105–113 DOI [10.1089/omi.2006.10.105](https://doi.org/10.1089/omi.2006.10.105).
- Crainey JL, Marín MA, Silva TRRD, Medeiros JFD, Pessoa FAC, Santos YV, Vicente ACP, Luz SLB.** 2018. *Mansonella ozzardi* mitogenome and pseudogene characterisation provides new perspectives on filarial parasite systematics and CO-1 barcoding. *Scientific Reports* **8**:6158 DOI [10.1038/s41598-018-24382-3](https://doi.org/10.1038/s41598-018-24382-3).
- Crampton-Platt A, Timmermans MJTN, Gimmel ML, Kutty SN, Cockerill TD, Vun Khen C, Vogler AP.** 2015. Soup to tree: the phylogeny of beetles inferred by mitochondrial metagenomics of a bornean rainforest sample. *Molecular Biology and Evolution* **32**:2302–2316 DOI [10.1093/molbev/msv111](https://doi.org/10.1093/molbev/msv111).
- Crozier RH, Crozier YC.** 1993. The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organization. *Genetics* **133**:97–117.
- Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM.** 2013. An extensive evaluation of read trimming effects on illumina NGS data analysis. *PLOS ONE* **8**:e85024 DOI [10.1371/journal.pone.0085024](https://doi.org/10.1371/journal.pone.0085024).
- Desalle R.** 2017. MtDNA The small workhorse of evolutionary studies. *Frontiers in Bioscience* **22**:873–887 DOI [10.2741/4522](https://doi.org/10.2741/4522).

- Dierckxsens N, Mardulyn P, Smits G.** 2016. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research* **45**:e18 DOI [10.1093/nar/gkw955](https://doi.org/10.1093/nar/gkw955).
- Diroma MA, Calabrese C, Simone D, Santorsola M, Calabrese FM, Gasparre G, Attimonelli M.** 2014. Extraction and annotation of human mitochondrial genomes from 1000 Genomes Whole Exome Sequencing data. *BMC Genomics* **15**:S2 DOI [10.1186/1471-2164-15-S3-S2](https://doi.org/10.1186/1471-2164-15-S3-S2).
- Dohm JC, Lottaz C, Borodina T, Himmelbauer H.** 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research* **36**:e105–e105 DOI [10.1093/nar/gkn425](https://doi.org/10.1093/nar/gkn425).
- Downton M, Cameron SL, Dowavic JI, Austin AD, Whiting MF.** 2009. Characterization of 67 Mitochondrial tRNA Gene rearrangements in the hymenoptera suggests that mitochondrial tRNA gene position is selectively neutral. *Molecular Biology and Evolution* **26**:1607–1617 DOI [10.1093/molbev/msp072](https://doi.org/10.1093/molbev/msp072).
- Duan X-Y, Peng X-Y, Qian Z-Q.** 2016. The complete mitochondrial genomes of two globally invasive ants, the Argentine ant *Linepithema humile* and the little fire ant *Wasmannia auropunctata*. *Conservation Genetics Resources* **8**:275–277 DOI [10.1007/s12686-016-0555-6](https://doi.org/10.1007/s12686-016-0555-6).
- Ekblom R, Smeds L, Ellegren H.** 2014. Patterns of sequencing coverage bias revealed by ultra-deep sequencing of vertebrate mitochondria. *BMC Genomics* **15**:467 DOI [10.1186/1471-2164-15-467](https://doi.org/10.1186/1471-2164-15-467).
- Fields PD, Obbard DJ, McTaggart SJ, Galimov Y, Little TJ, Ebert D.** 2018. Mitogenome phylogeographic analysis of a planktonic crustacean. *Molecular Phylogenetics and Evolution* **129**:138–148 DOI [10.1016/j.ympev.2018.06.028](https://doi.org/10.1016/j.ympev.2018.06.028).
- Finstermeier K, Zinner D, Bräuer M, Meyer M, Kreuz E, Hofreiter M, Roos C.** 2013. A mitogenomic phylogeny of living primates. *PLOS ONE* **8**:e69504 DOI [10.1371/journal.pone.0069504](https://doi.org/10.1371/journal.pone.0069504).
- Goodwin S, McPherson JD, McCombie WR.** 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17**:333–351 DOI [10.1038/nrg.2016.49](https://doi.org/10.1038/nrg.2016.49).
- Gotzek D, Clarke J, Shoemaker D.** 2010. Mitochondrial genome evolution in fire ants (Hymenoptera: Formicidae). *BMC Evolutionary Biology* **10**:300 DOI [10.1186/1471-2148-10-300](https://doi.org/10.1186/1471-2148-10-300).
- Guo Y, Li J, Li C, Shyr Y, Samuels D.** 2013. MitoSeek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis. *Bioinformatics* **29**:1210–1211.
- Hahn C, Bachmann L, Chevreux B.** 2013. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Research* **41**:e129–e129 DOI [10.1093/nar/gkt371](https://doi.org/10.1093/nar/gkt371).
- Hasegawa E, Kobayashi K, Yagi N, Tsuji K.** 2011. Complete mitochondrial genomes of normal and cheater morphs in the parthenogenetic ant *Pristomyrmex punctatus* (Hymenoptera: Formicidae). *Myrmecol News* **15**(85):85–90.

- Holder MT, Zwickl DJ, Dessimoz C.** 2008. Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363**:4013–4021 DOI [10.1098/rstb.2008.0162](https://doi.org/10.1098/rstb.2008.0162).
- Holldobler B, Wilson E.** 1990. *The ants*. Harvard University Press.
- Janzen DH.** 1966. Coevolution of mutualism between ants and acacias in central America. *Evolution* **20**:249–275 DOI [10.2307/2406628](https://doi.org/10.2307/2406628).
- Karsch-Mizrachi I, Takagi T, Cochrane G, and International Nucleotide Sequence Database Collaboration.** 2017. The international nucleotide sequence database collaboration. *Nucleic Acids Research* **46**(D1):D48–D51 DOI [10.1093/nar/gkx1097](https://doi.org/10.1093/nar/gkx1097).
- Kayal E, Bentlage B, Cartwright P, Yanagihara AA, Lindsay DJ, Hopcroft RR, Collins AG.** 2015. Phylogenetic analysis of higher-level relationships within Hydroidolina (Cnidaria: Hydrozoa) using mitochondrial genome data and insight into their mitochondrial transcription. *PeerJ* **3**:e1403 DOI [10.7717/peerj.1403](https://doi.org/10.7717/peerj.1403).
- Kibbe WA.** 2007. OligoCalc: an online oligonucleotide properties calculator. *Nucleic Acids Research* **35**:W43–W46 DOI [10.1093/nar/gkm234](https://doi.org/10.1093/nar/gkm234).
- Kılınç GM, Kashuba N, Yaka R, Sümer AP, Yüncü E, Shergin D, Ivanov GL, Kichigin D, Pestereva K, Volkov D, Mandryka P, Kharinskii A, Tishkin A, Ineshin E, Kovychev E, Stepanov A, Alekseev A, Fedoseeva SA, Somel M, Jakobsson M, Krzewińska M, Storå J, Götherström A.** 2018. Investigating Holocene human population history in North Asia using ancient mitogenomes. *Scientific Reports* **8**:8969 DOI [10.1038/s41598-018-27325-0](https://doi.org/10.1038/s41598-018-27325-0).
- Kim MJ, Hong EJ, Kim I.** 2016. Complete mitochondrial genome of Camponotus atrox (Hymenoptera: Formicidae): a new tRNA arrangement in Hymenoptera. *Genome* **59**:59–74 DOI [10.1139/gen-2015-0080](https://doi.org/10.1139/gen-2015-0080).
- Kodama Y, Shumway M, Leinonen R, on behalf of the International Nucleotide Sequence Database Collaboration.** 2012. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Research* **40**:D54–D56 DOI [10.1093/nar/gkr854](https://doi.org/10.1093/nar/gkr854).
- Krzemińska U, Morales HE, Greening C, Nyári ÁS, Wilson R, Song BK, Austin CM, Sunnucks P, Pavlova A, Rahman S.** 2017. Population mitogenomics provides insights into evolutionary history, source of invasions and diversifying selection in the House Crow (*Corvus splendens*). *Heredity* **120**:296–309 DOI [10.1038/s41437-017-0020-7](https://doi.org/10.1038/s41437-017-0020-7).
- Kumar S, Stecher G, Tamura K.** 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* **33**:1870–1874 DOI [10.1093/molbev/msw054](https://doi.org/10.1093/molbev/msw054).
- Kuzniar A, Van Ham RCHJ, Pongor S, Leunissen JAM.** 2008. The quest for orthologs: finding the corresponding gene across genomes. *Trends in Genetics* **24**:539–551 DOI [10.1016/j.tig.2008.08.009](https://doi.org/10.1016/j.tig.2008.08.009).
- Langmead B, Salzberg SL.** 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**:357–359 DOI [10.1038/nmeth.1923](https://doi.org/10.1038/nmeth.1923).

- Lin Z-Q, Song F, Li T, Wu Y-Y, Wan X.** 2017. New mitogenomes of two chinese stag beetles (Coleoptera, Lucanidae) and their implications for systematics. *Journal of Insect Science* 17(2):63 DOI [10.1093/jisesa/ixw041](https://doi.org/10.1093/jisesa/ixw041).
- Linard B, Crampton-Platt A, Morinier J, Timmermans MJTN, Andújar C, Arribas P, Miller KE, Lipecki J, Favreau E, Hunter A, Gómez-Rodríguez C, Barton C, Nie R, Gillett CPDT, Breeschoten T, Bocak L, Vogler AP.** 2018. The contribution of mitochondrial metagenomics to large-scale data mining and phylogenetic analysis of Coleoptera. *Molecular Phylogenetics and Evolution* 128:1–11 DOI [10.1016/j.ympev.2018.07.008](https://doi.org/10.1016/j.ympev.2018.07.008).
- Liu N, Duan X-Y, Qian Z-Q, Wang X-Y, Li X-L, Ding M-Y.** 2016. Characterization of the complete mitochondrial genome of the myrmicine ant Vollenhovia emeryi (Insecta: Hymenoptera: Formicidae). *Conservation Genetics Resources* 8:211–214 DOI [10.1007/s12686-016-0535-x](https://doi.org/10.1007/s12686-016-0535-x).
- Mao M, Gibson T, Dowton M.** 2015. Higher-level phylogeny of the Hymenoptera inferred from mitochondrial genomes. *Molecular Phylogenetics and Evolution* 84:34–43 DOI [10.1016/j.ympev.2014.12.009](https://doi.org/10.1016/j.ympev.2014.12.009).
- Mardis ER.** 2008. The impact of next-generation sequencing technology on genetics. *Trends in Genetics* 24:133–141 DOI [10.1016/j.tig.2007.12.007](https://doi.org/10.1016/j.tig.2007.12.007).
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380 DOI [10.1038/nature03959](https://doi.org/10.1038/nature03959).
- Miller MJ, Aguilar C, De León LF, Loaiza JR, McMillan WO.** 2016. Complete mitochondrial genomes of the New World jacanas: Jacana spinosa and Jacana jacana. *Mitochondrial DNA* 27:764–765 DOI [10.3109/19401736.2014.915530](https://doi.org/10.3109/19401736.2014.915530).
- Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD, Marshall D.** 2013. Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics* 14:193–202 DOI [10.1093/bib/bbs012](https://doi.org/10.1093/bib/bbs012).
- Miya M, Takeshima H, Endo H, Ishiguro NB, Inoue JG, Mukai T, Satoh TP, Yamaguchi M, Kawaguchi A, Mabuchi K, Shirai SM, Nishida M.** 2003. Major patterns of higher teleostean phylogenies: a new perspective based on 100 complete mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution* 26:121–138 DOI [10.1016/s1055-7903\(02\)00332-9](https://doi.org/10.1016/s1055-7903(02)00332-9).
- Moreau C, Bell C, Vila R, Archibald S, Pierce N.** 2006. Phylogeny of the Ants: diversification in the Age of Angiosperms. *Science* 312:101–104 DOI [10.1126/science.1124891](https://doi.org/10.1126/science.1124891).
- Moritz C.** 1994. Applications of mitochondrial DNA analysis in conservation: a critical review. *Molecular Ecology* 3:401–411 DOI [10.1111/j.1365-294X.1994.tb00080.x](https://doi.org/10.1111/j.1365-294X.1994.tb00080.x).

- Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, Campino S, Turner DJ, MacInnis B, Kwiatkowski DP, Swerdlow HP, Quail MA. 2012. Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes. *BMC Genomics* 13:1 DOI [10.1186/1471-2164-13-1](https://doi.org/10.1186/1471-2164-13-1).
- Picardi E, Pesole G. 2012. Mitochondrial genomes gleaned from human whole-exome sequencing. *Nature Methods* 9:523–524 DOI [10.1038/nmeth.2029](https://doi.org/10.1038/nmeth.2029).
- Plese B, Rossi ME, Kenny N, Taboada S, Koutsouveli V, Riesgo A. 2018. Trimitomics: an efficient pipeline for mitochondrial assembly from transcriptomic reads in non-model species. *bioRxiv* 2018:Article 413138 DOI [10.1101/413138](https://doi.org/10.1101/413138).
- Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14:817–818 DOI [10.1093/bioinformatics/14.9.817](https://doi.org/10.1093/bioinformatics/14.9.817).
- Prosdocimi F, De Carvalho DC, De Almeida RN, Beheregaray LB. 2012. The complete mitochondrial genome of two recently derived species of the fish genus *Nannoperca* (Perciformes, Percichthyidae). *Molecular Biology Reports* 39:2767–2772 DOI [10.1007/s11033-011-1034-5](https://doi.org/10.1007/s11033-011-1034-5).
- Raposo do Amaral F, Neves LG, Resende MFR, Mobili F, Miyaki CY, Pellegrino KCM, Biondo C. 2015. Ultraconserved elements sequencing as a low-cost source of complete mitochondrial genomes and microsatellite markers in non-model amniotes. *PLOS ONE* 10:e0138446 DOI [10.1371/journal.pone.0138446](https://doi.org/10.1371/journal.pone.0138446).
- Rauch C, Christa G, de Vries J, Woehle C, Gould SB. 2017. Mitochondrial genome assemblies of *elysia timida* and *elysia cornigera* and the response of mitochondrion-associated metabolism during starvation. *Genome Biology and Evolution* 9:1873–1879 DOI [10.1093/gbe/evx129](https://doi.org/10.1093/gbe/evx129).
- Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* 463:1079–1083 DOI [10.1038/nature08742](https://doi.org/10.1038/nature08742).
- Rodovalho C de M, Lyra ML, Ferro M, Bacci M. 2014. The mitochondrial genome of the leaf-cutter ant atta laevigata: a mitogenome with a large number of intergenic spacers. *PLOS ONE* 9:e97117.
- Rosel PE, Hancock-Hanser BL, Archer FI, Robertson KM, Martien KK, Leslie MS, Berta A, Cipriano F, Viricel A, Viaud-Martinez KA, Taylor BL. 2017. Examining metrics and magnitudes of molecular genetic differentiation used to delimit cetacean subspecies based on mitochondrial DNA control region sequences. *Marine Mammal Science* 33:76–100.
- Rubin BER, Moreau CS. 2016. Comparative genomics reveals convergent rates of evolution in ant–plant mutualisms. *Nature Communications* 7:Article 12679 DOI [10.1038/ncomms12679](https://doi.org/10.1038/ncomms12679).
- Rubinoff D. 2006. Essays: utility of Mitochondrial DNA barcodes in species conservation. *Conservation Biology* 20:1026–1033 DOI [10.1111/j.1523-1739.2006.00372.x](https://doi.org/10.1111/j.1523-1739.2006.00372.x).
- Samuels D, Han L, Li J, Quanghu S, Clark T, Shyr Y, Guo Y. 2013. Finding the lost treasures in exome sequencing data. *Trends in Genetics* 29:593–599.

- Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I.** 2009. ABySS: a parallel assembler for short read sequence data. *Genome Research* **19**:1117–1123 DOI [10.1101/gr.089532.108](https://doi.org/10.1101/gr.089532.108).
- Smith DR.** 2015. The past, present and future of mitochondrial genomics: have we sequenced enough mtDNAs? *Briefings in Functional Genomics* **15**(1):47–54 DOI [10.1093/bfgp/elv027](https://doi.org/10.1093/bfgp/elv027).
- Stodden V, Seiler J, Ma Z.** 2018. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences of the United States of America* **115**:2584–2589 DOI [10.1073/pnas.1708290115](https://doi.org/10.1073/pnas.1708290115).
- Ströher PR, Zarza E, Tsai WLE, McCormack JE, Feitosa RM, Pie MR.** 2017. The mitochondrial genome of *Octostroma stenognatha* and its phylogenetic implications. *Insectes Sociaux* **64**:149–154 DOI [10.1007/s00040-016-0525-8](https://doi.org/10.1007/s00040-016-0525-8).
- Tang M, Tan M, Meng G, Yang S, Su X, Liu S, Song W, Li Y, Wu Q, Zhang A, Zhou X.** 2014. Multiplex sequencing of pooled mitochondrial genomes—a crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research* **42**:e166–e166 DOI [10.1093/nar/gku917](https://doi.org/10.1093/nar/gku917).
- Thompson JD, Gibson TJ, Higgins DG.** 2003. Multiple Sequence Alignment Using ClustalW and ClustalX. *Current Protocols in Bioinformatics* **1**:2.3.1–2.3.22 DOI [10.1002/0471250953.bi0203s00](https://doi.org/10.1002/0471250953.bi0203s00).
- Tian Y, Smith D.** 2016. Recovering complete mitochondrial genome sequences from RNA-Seq: a case study of Polytomella non-photosynthetic green algae. *Molecular Phylogenetics and Evolution* **98**:57–62.
- Tilak M-K, Justy F, Debiais-Thibaud M, Botero-Castro F, Delsuc F, Douzery EJP.** 2014. A cost-effective straightforward protocol for shotgun Illumina libraries designed to assemble complete mitogenomes from non-model species. *Conservation Genetics Resources* **7**:37–40 DOI [10.1007/s12686-014-0338-x](https://doi.org/10.1007/s12686-014-0338-x).
- Timmermans MJTN, Viberg C, Martin G, Hopkins K, Vogler AP.** 2016. Rapid assembly of taxonomically validated mitochondrial genomes from historical insect collections. *Biological Journal of the Linnean Society* **117**:83–95 DOI [10.1111/bij.12552](https://doi.org/10.1111/bij.12552).
- Ulianó-Silva M, Americo JA, Costa I, Schomaker-Bastos A, de Freitas Rebello M, Prosdocimi F.** 2016. The complete mitochondrial genome of the golden mussel *Limnoperna fortunei* and comparative mitogenomics of Mytilidae. *Gene* **577**:202–208 DOI [10.1016/j.gene.2015.11.043](https://doi.org/10.1016/j.gene.2015.11.043).
- Van Dijk EL, Auger H, Jaszczyzyn Y, Thermes C.** 2014. Ten years of next-generation sequencing technology. *Trends in Genetics* **30**:418–426 DOI [10.1016/j.tig.2014.07.001](https://doi.org/10.1016/j.tig.2014.07.001).
- Vanecek T, Vorel F, Sip M.** 2004. Mitochondrial DNA D-loop hypervariable regions: czech population data. *International Journal of Legal Medicine* **118**:14–18 DOI [10.1007/s00414-003-0407-2](https://doi.org/10.1007/s00414-003-0407-2).
- Ward P.** 1999. Systematics biogeography and host plant associations of the *Pseudomyrmex viduus* group (Hymenoptera: Formicidae), *Triplaris*—and *Tachigali*-inhabiting ants. *Zoological Journal of the Linnean Society* **126**:451–540 DOI [10.1006/zjls.1998.0158](https://doi.org/10.1006/zjls.1998.0158).

- Ward PS.** 1989. Systematic studies on pseudomyrmecine ants: revision of the *Pseudomyrmex oculatus* and *P. subtilissimus* species groups. *with taxonomic comments on other species. Quaestiones Entomologicae* **25**(4):393–468.
- Ward PS.** 1991. Phylogenetic analysis of pseudomyrmecine ants associated with domatia-bearing plants. In: *Ant-plant interactions*. Oxford: Oxford University Press 335–352.
- Ward PS.** 1993. Systematic studies on Pseudomyrmex acacia-ants (Hymenoptera: Formicidae: Pseudomyrmecinae). *Journal of Hymenoptera Research* **2**:117–168.
- Ward PS.** 2011. Integrating molecular phylogenetic results into ant taxonomy (Hymenoptera: Formicidae). *Myrmecological News* **15**:21–29.
- Ward PS.** 2017. A review of the *Pseudomyrmex ferrugineus* and *Pseudomyrmex goeldii* species groups: acacia-ants and relatives (Hymenoptera: Formicidae). *Zootaxa* **4227**:524–542 DOI [10.11646/zootaxa.4227.4.3](https://doi.org/10.11646/zootaxa.4227.4.3).
- Ward PS, Brady SG, Fisher BL, Schultz TR.** 2015. The evolution of myrmicine ants: phylogeny and biogeography of a hyperdiverse ant clade (Hymenoptera: Formicidae): phylogeny and evolution of myrmicine ants. *Systematic Entomology* **40**:61–81 DOI [10.1111/syen.12090](https://doi.org/10.1111/syen.12090).
- Ward PS, Branstetter MG.** 2017. The acacia ants revisited: convergent evolution and biogeographic context in an iconic ant/plant mutualism. *Proceedings of the Royal Society B: Biological Sciences* **284**:20162569 DOI [10.1098/rspb.2016.2569](https://doi.org/10.1098/rspb.2016.2569).
- Ward P, Downie D.** 2004. The ant subfamily Pseudomyrmecinae (Hymenoptera: Formicidae): phylogeny and evolution of big-eyed arboreal ants. *Systematic Entomology* **30**:310–335.
- Wolstenholme DR.** 1992. Animal mitochondrial DNA: structure and evolution. *International Review of Cytology* **141**:173–216 DOI [10.1016/S0074-7696\(08\)62066-5](https://doi.org/10.1016/S0074-7696(08)62066-5).
- Yang S, Li X, Cai L-G, Qian Z-Q.** 2015. Characterization of the complete mitochondrial genome of *Formica selsyi* (Insecta: Hymenoptera: Formicidae: Formicinae). *Mitochondrial DNA* **27**(5):3378–3380 DOI [10.3109/19401736.2015.1018229](https://doi.org/10.3109/19401736.2015.1018229).
- Zwick A, Regier JC, Zwickl DJ.** 2012. Resolving discrepancy between nucleotides and amino acids in deep-level arthropod phylogenomics: differentiating serine codons in 21-Amino-acid models. *PLOS ONE* **7**(11):e47450 DOI [10.1371/journal.pone.0047450](https://doi.org/10.1371/journal.pone.0047450).