No budget mitogenomics: Montagem, anotação, genômica comparativa e evolutiva de mitocôndrias de insetos usando dados públicos

Aluno: Gabriel Alves Vieira

Orientador: Francisco Prosdocimi

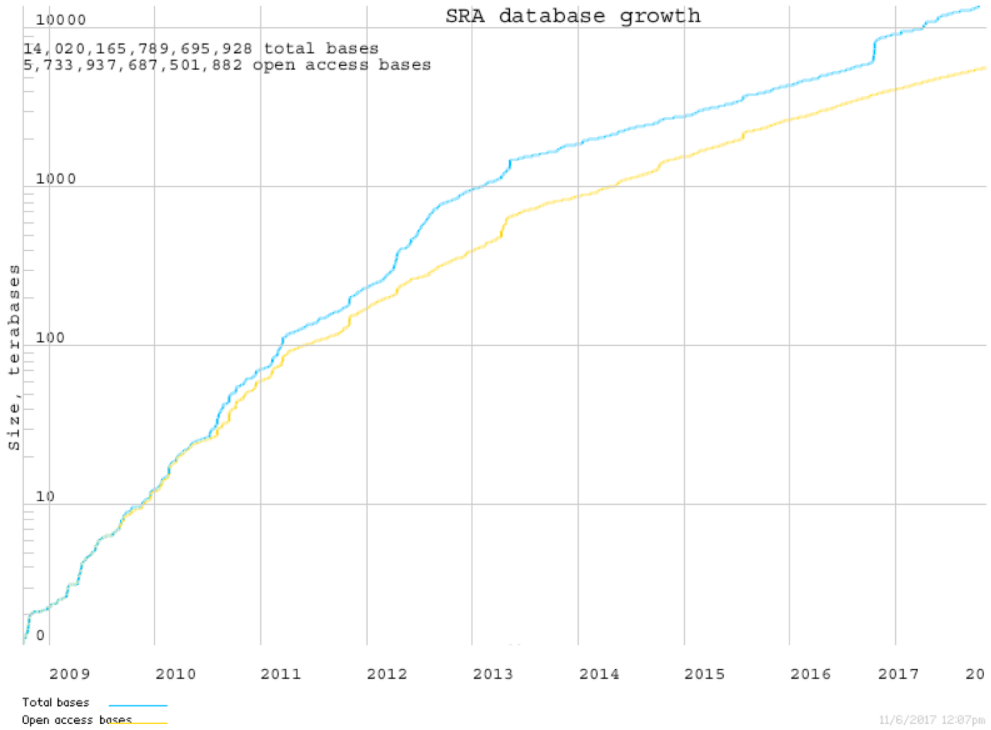
AAGTCAAGCTGCTCTGTGGGCTGTGATCTGCCTCAAACCCACAGCCTGGGTAGCAGG
AGGACCTTGATGCTCCTGGCACAGATGAGGAGAATCTCTCTTTTCTCCTGCTTGAAG
GACAGACATGACTTTGGATTTCCCCAGGAGGAGTTTGGCAACCAGTTCCAAAAGGCT
GAAACCATCCCTGTCCTCCATGAGATGATCCAGCAGATCTTCAATCTCTTCAGCACA
AAGGACTCATCTGCTGCTTGGGATGAGACCCTCCTAGACAAATTCTACACTGAACTC
TACCAGCAGCTGAATGACCTGGAAGCCTGTGTGATACAGGGGGTGGGGGTGACAGAG
ACTCCCCTGATGAAGGAGGACTCCATTCTGGCTGTGAGGAAATACTTCCAAAGAATC
ACTCTCTATCTGAAAGAGAAGAAATACAGCCCTTGTGCCTGGGAGGTTGTCAGAGCA
GAAATCATGAGATCTTTTTCTTTGTCAACAAACTTGCAAGAAAGTTTAAGAAGTAAG
GAATGA, TGTGATCTGCCTCAAACCCACAGCCTGGGTAGCAGGAGGACCTTGATGC
TCCTGGCACAGATGAGGAGAATCTCTCTTTTCTCCTGCTTGAAGGACAGACATGACT
TTGGATTTCCCCAGGAGGAGTTTGGCAACCAGTTCCAAAAGGCTGAAACCATCCCTG
TCCTCCATGAGATGATCCAGCAGATCTTCAATCTCTTCAGCACAAAGGACTCATCTG
CTGCTTGGGATGAGACCCTCCTAGACAAATTCTACACTGAACTCTACCAGCAGCTGA

Carga horária

- Bioética e Biossegurança – 30h
- Fundamentos em Macromoléculas – 45h
- Fundamentos em Biologia Molecular – 45h
- Fundamentos em Metabolismo – 45h
- Iniciação à Ciência: 2º Grau I – 90 h
- No-Budget Science – 15h
- Big Data - 30h
- Ciência e arte – 75h
- Estratégias de Divulgação Científica – 45h
- Atividades didáticas I -45h
- 20/25 ALVs – Seminários Plenos – 30h
- 23/25 Journals - Seminários Gerais – 30h
- Tópicos Avançados – 15h
- **TOTAL: 540h**

International Nucleotide Sequence Database Colaboration

- Crescimento dos bancos de dados públicos de sequenciamento



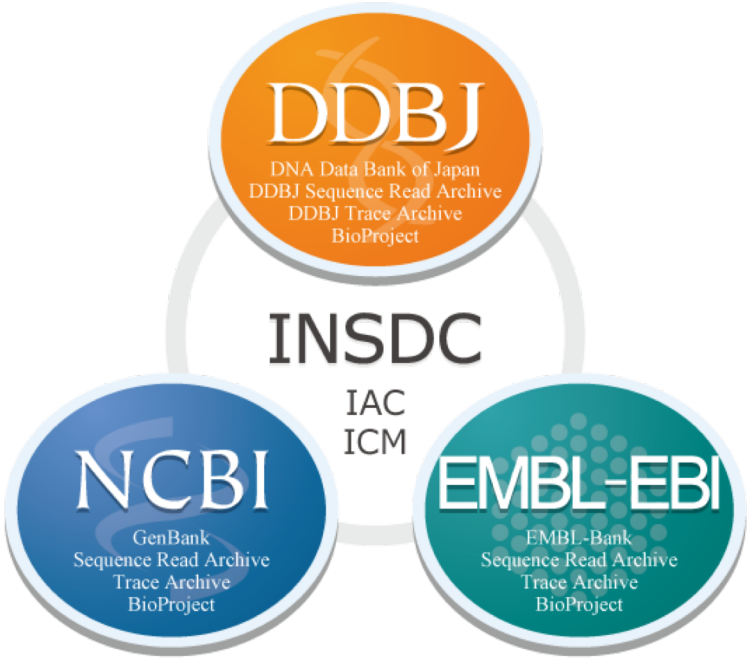
- Sequenciamento de Nova Geração (NGS)
  - ↑ produção de dados
  - ↓ custo por base

- Revistas: Incentivam disponibilização dos dados
  - Reprodutibilidade

Next Generation Sequencing

Sequence Read Archive (SRA)

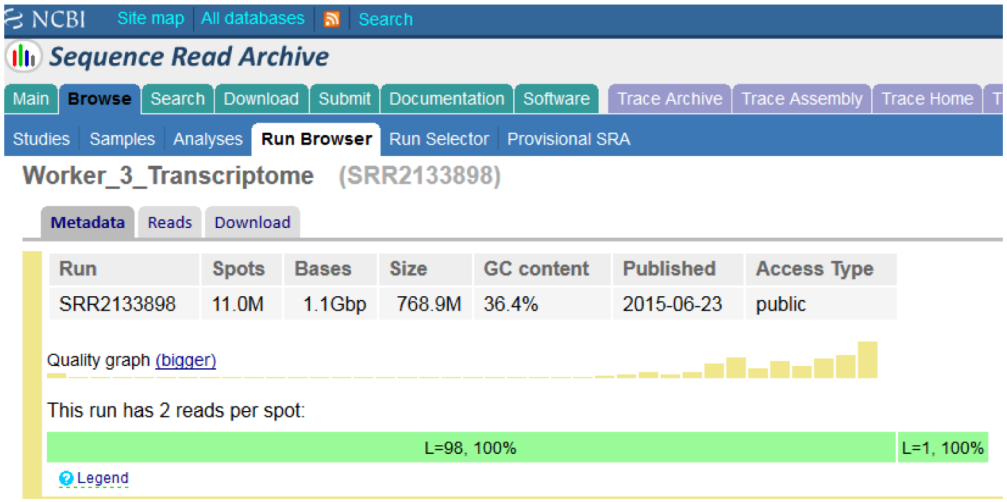
- Parte de colaboração internacional
  - 3 bancos de dados:
    - SRA: NCBI Sequence Read Archive
    - ERA: EBI Sequence Read Archive
    - DRA: DDBJ Sequence Read Archive
  - Sincronizados

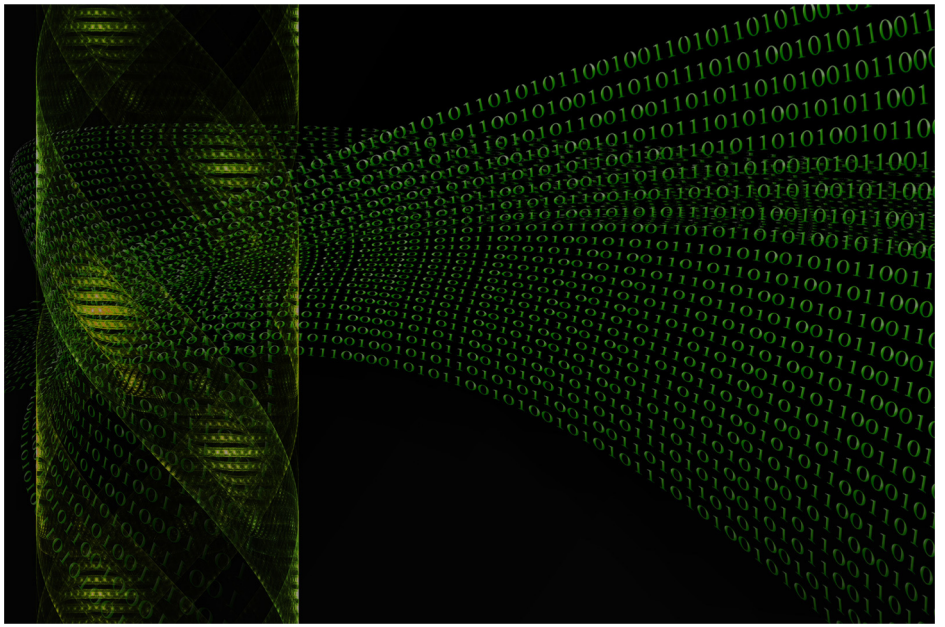


- Diferentes tipos de dados:
  - DNA-seq; Bisulfite-seq; RNA-seq...
  - Um pouco de reflexão e imaginação
  - Gerar novos conhecimentos

Dados públicos

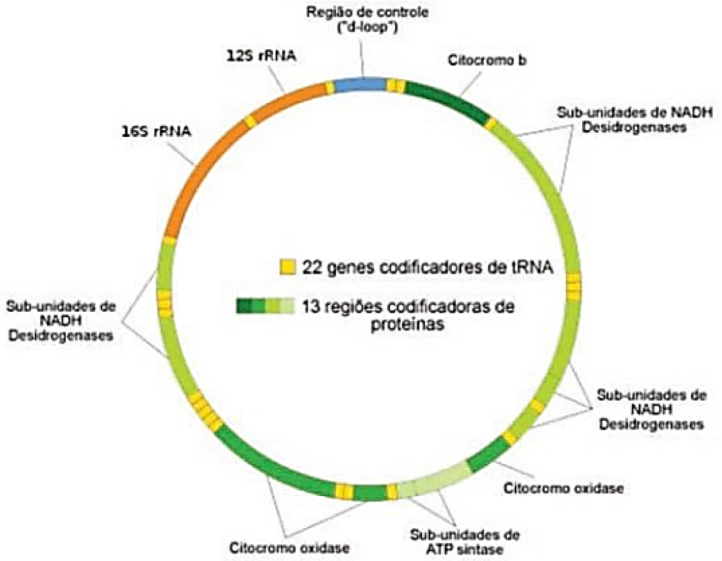
- Dados usados para responder uma pergunta
  - Sem relação com a mitocôndria
- Sequenciamento
  - Não separa a mitocôndria
  - ~0,1%-0,5% dos dados é mitocondrial
  - Espécies com dados públicos e sem mitogenoma descrito
- Esses dados podem ser usados para obter o mitogenomas
  - Além de outras análises...





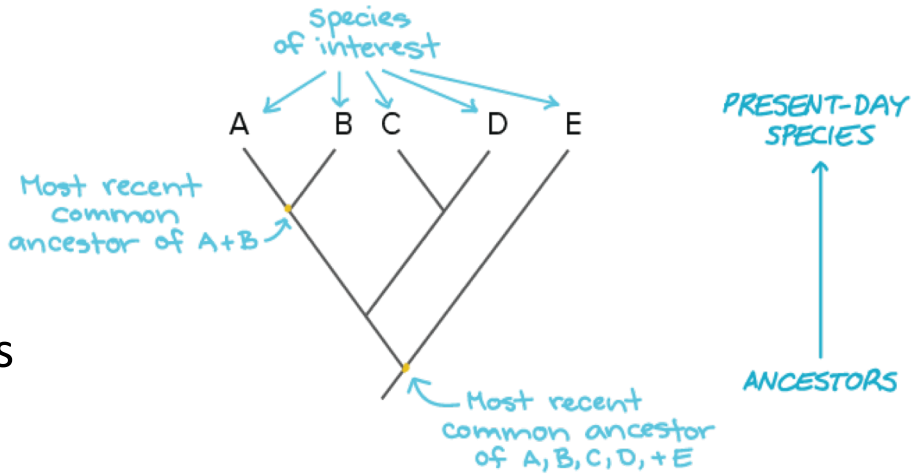
Mas por que o mitogenoma?

- Menor genoma da célula (≅ 16kb)
  - Relativamente “simples” de montar
  - Excelente treino para iniciantes
  - Obtido de diversos tipos de datasets
    - WGS, RNA-Seq, Targeted Sequencing (UCE)



- Utilizado como marcador molecular
  - Sem recombinação
  - Herança materna

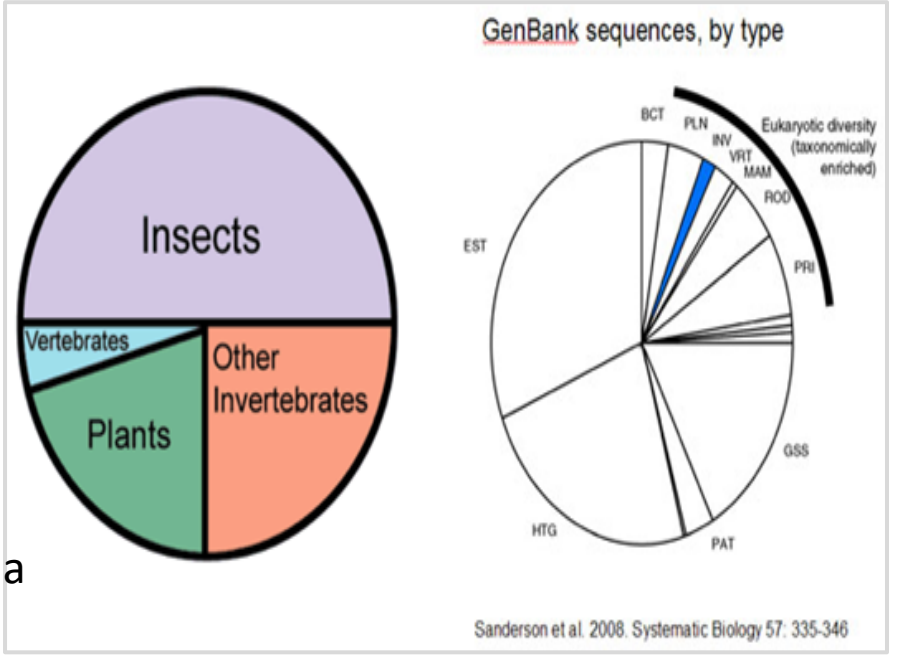
Filogenia e Filogeografia



- Uso de genes mitocondriais para estudar relações evolutivas
  - Corroborar ou questionar filogenias anteriores

Formigas (Hymenoptera: Formicidae)

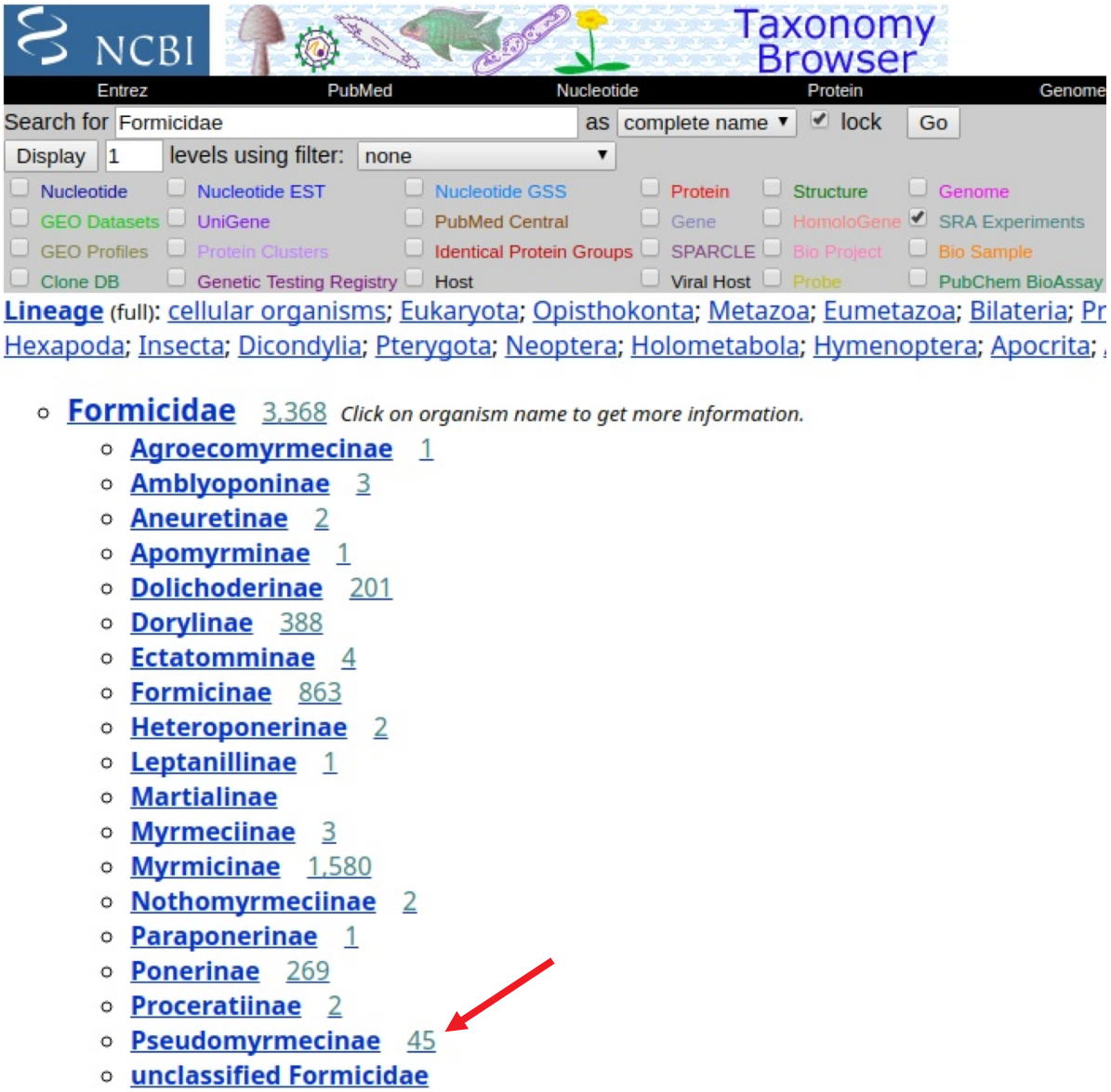
- Engenheiras ecossistêmicas
  - Alteram características do solo
  - Serviços ecossistêmicos
- Importância econômica
  - Pragas
  - Agentes de controle biológico
- Grande biodiversidade
  - ≈ 13.000 spp.
  - GenBank
    - 15 mitogenomas completos descritos



Desconhecimento da biodiversidade

Formigas (Hymenoptera: Formicidae)

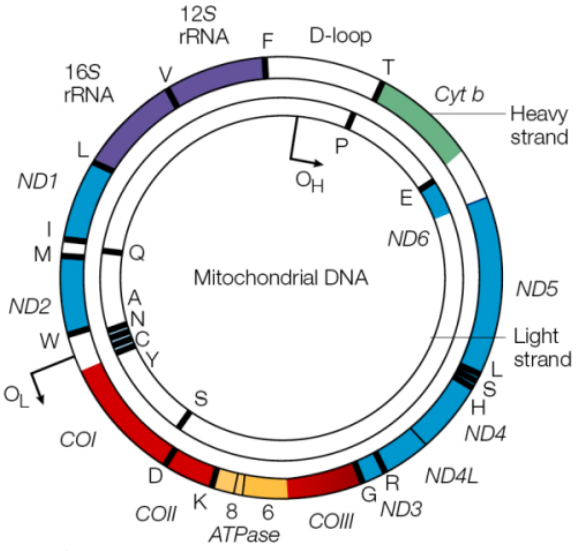
- Grande quantidade de informação para o clado:
  - ≈ 3.91 Terabytes
- Todas as subfamílias
  - Exceto Martialinae

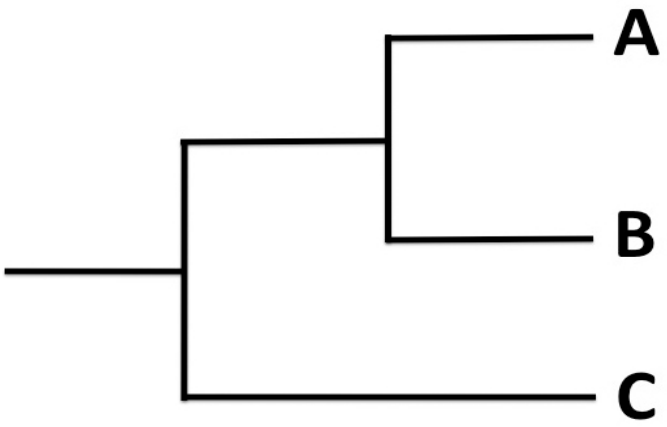


Objetivos

- Montar, anotar e disponibilizar mitogenomas da família Formicidae usando dados públicos
  - Qual a história evolutiva que a mitocôndria nos conta?
  - Auxiliar a resolver as relações filogenéticas, especialmente as de organismos não modelo

- Estabelecer metodologia para montagem de mitocôndrias em computadores de pequeno porte



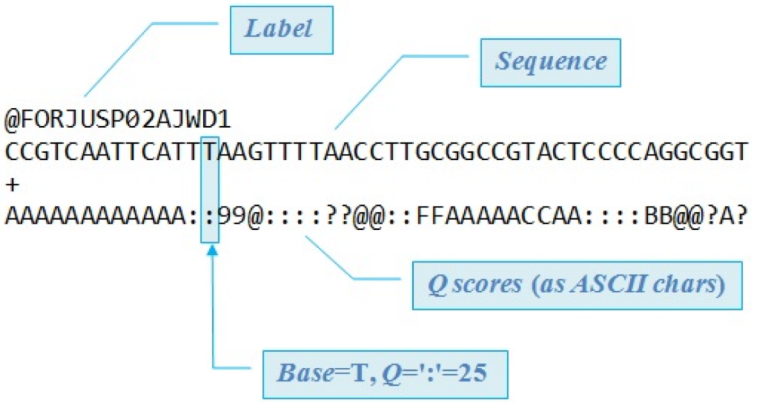


Obtenção dos dados

- Arquivos *.sra*
  - Utilizado pelos 3 bancos
  - Disponíveis para download direto
  - Precisa ser convertido (SRA Toolkit) para *.fastq*

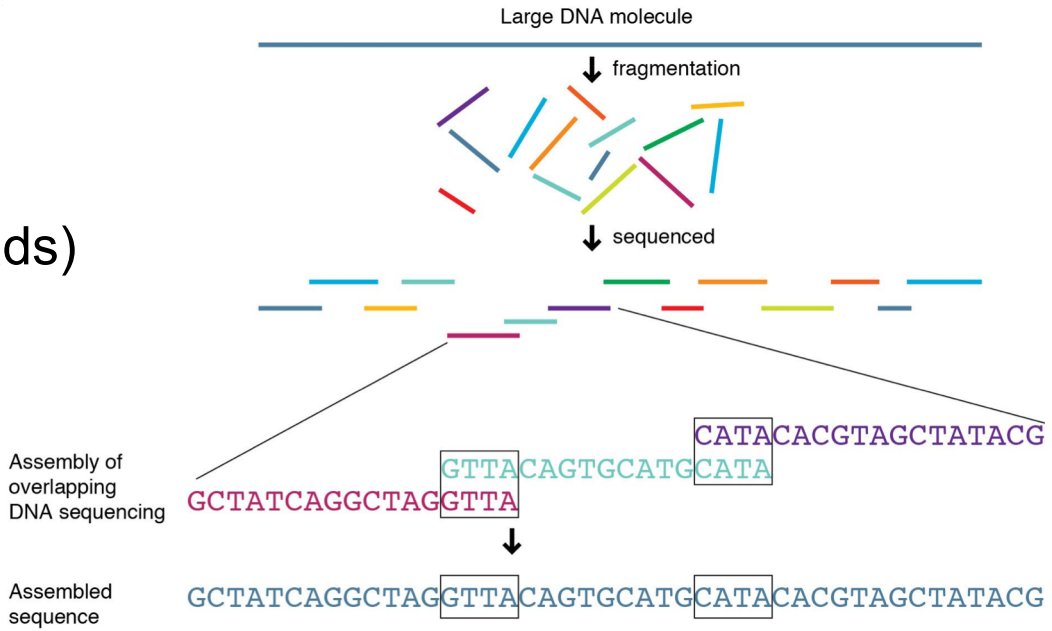
Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)	FASTQ files (Galaxy)	Submitted files (FTP)	Submitted files (Galaxy)	NCBI SRA file (FTP)	NCBI SRA file (Galaxy)	CRAM Index files (FTP)	CRAM Index files (Galaxy)
PRJNA360290	SAMN06208930	SRS1901018	SRX2468701	SRR5150611	55425	Paraponera clavata	Illumina HiSeq 2500	PAIRED	File 1 File 2	File 1 File 2			File 1	File 1		

- Arquivos *.fastq*
  - Cabeçalho, sequência e qualidade
  - Aceitos pela maioria dos programas

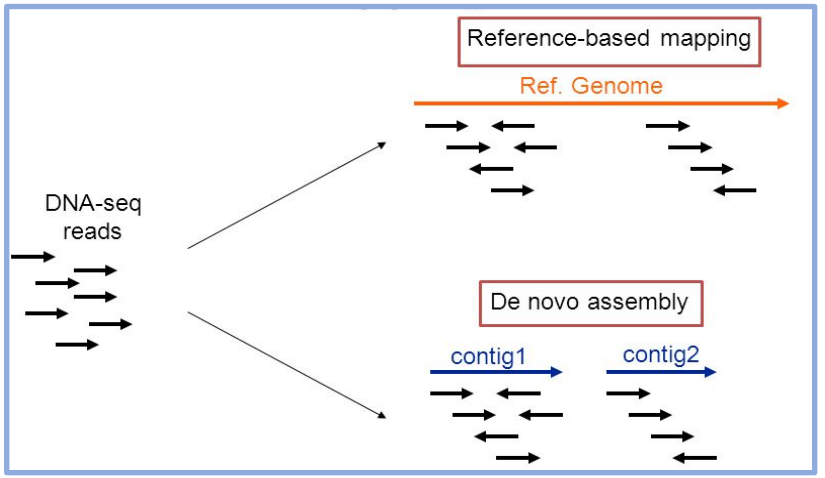


MONTAGEM

- Quebra de múltiplas cópias de DNA
- Obtenção de pequenas sequências (reads)
- Quebra-cabeça – Ferramentas computacionais
  - NOVOPlasty, MIRA, MITObim

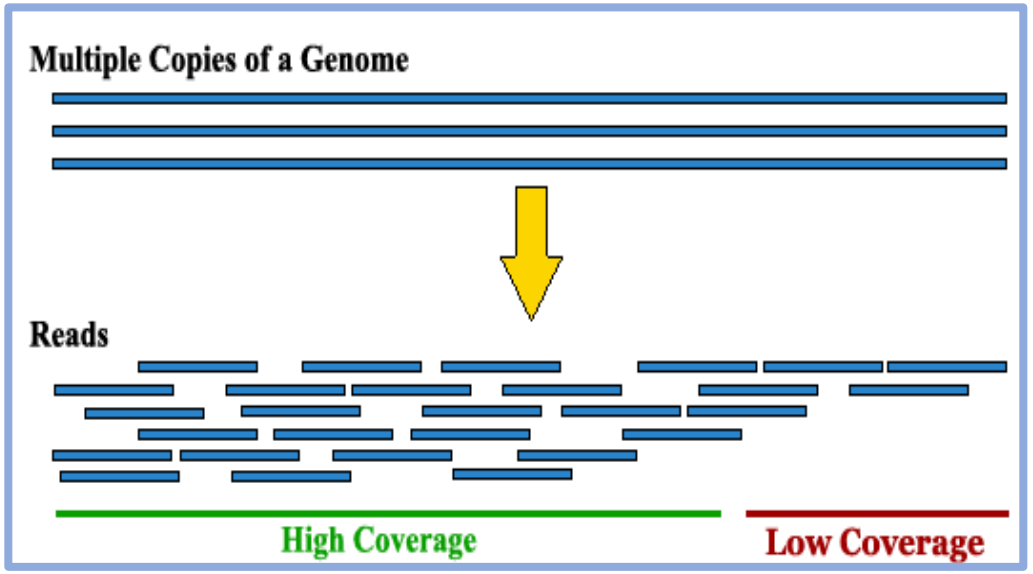


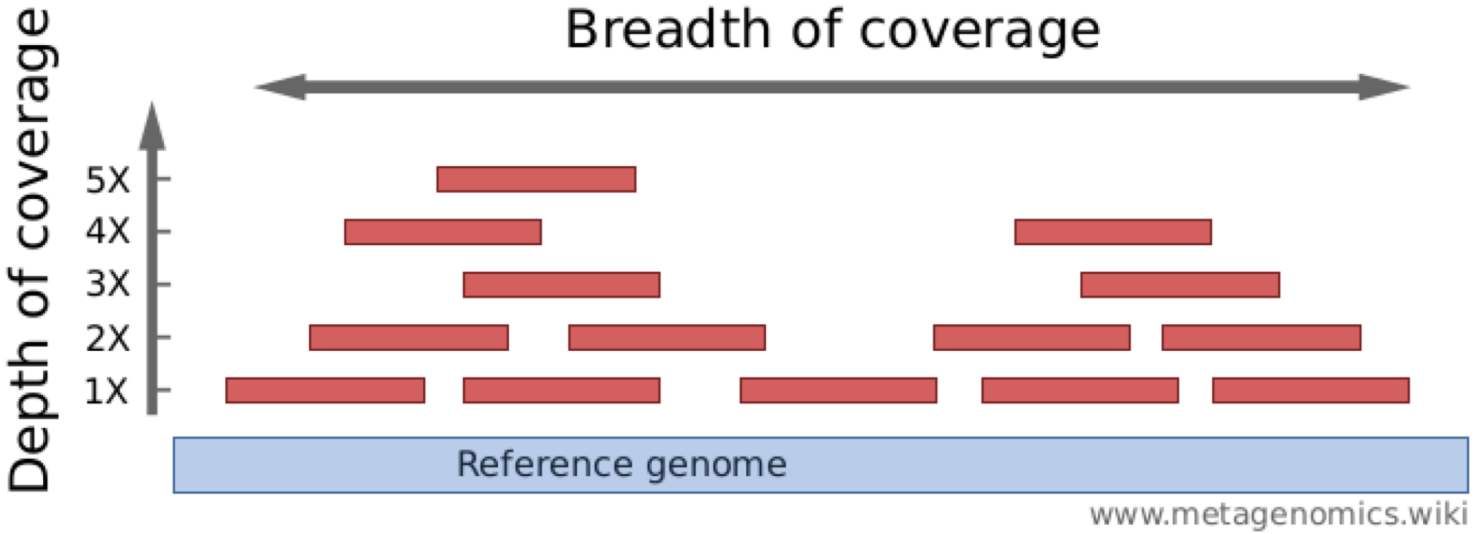
- Dois tipos de montagem
  - *De novo*
    - Sobreposição
  - Referência
    - Mapeamento com sequência
    - + fácil; menos gaps
    - Qualidade da referência



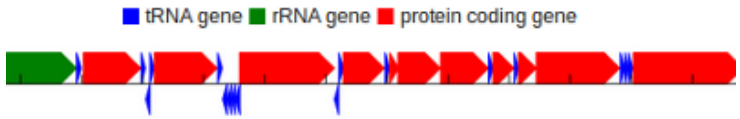
MONTAGEM

- Checagem (TABLET)
  - Cobertura
    - Quantas sequências corroboram a montagem
  - Circularização

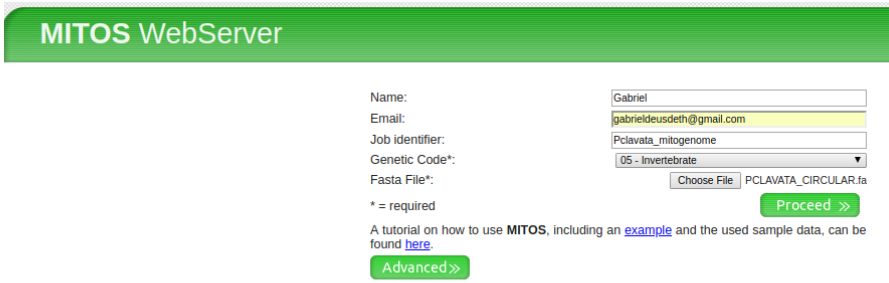




ANOTAÇÃO



- ANOTAÇÃO
  - Identificar regiões no genoma (Ex: genes)
    - Automática: MITOS Web Server
    - Manual: Artemis



- Submissão ao GenBank
  - BankIt (Plataforma Online)
  - Third Party Annotation (TPA)

NCBI BankIt

- Sequência disponível online no formato genbank (.gb)

Pseudomyrmecinae

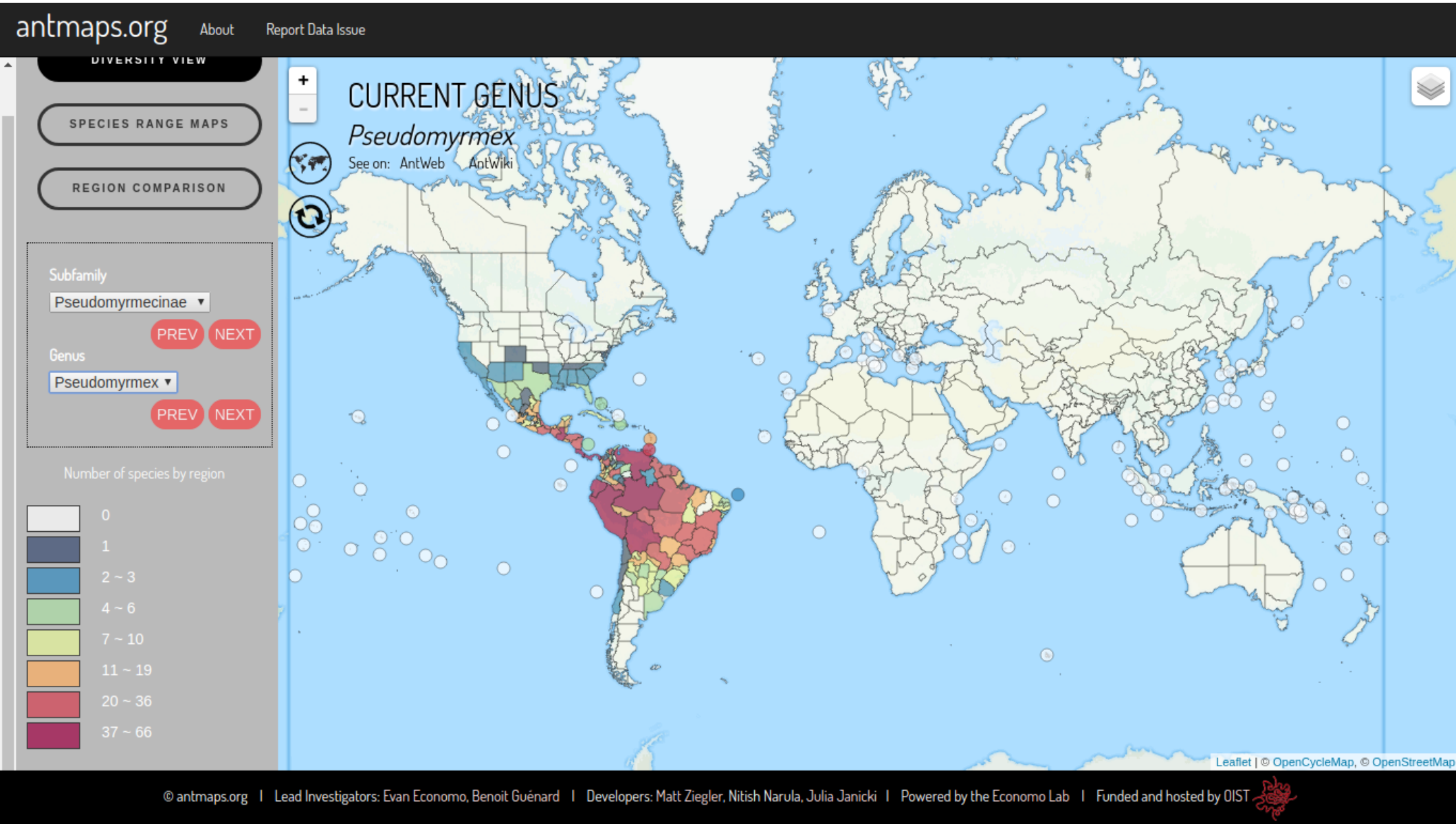
- 231 espécies – Novo Mundo

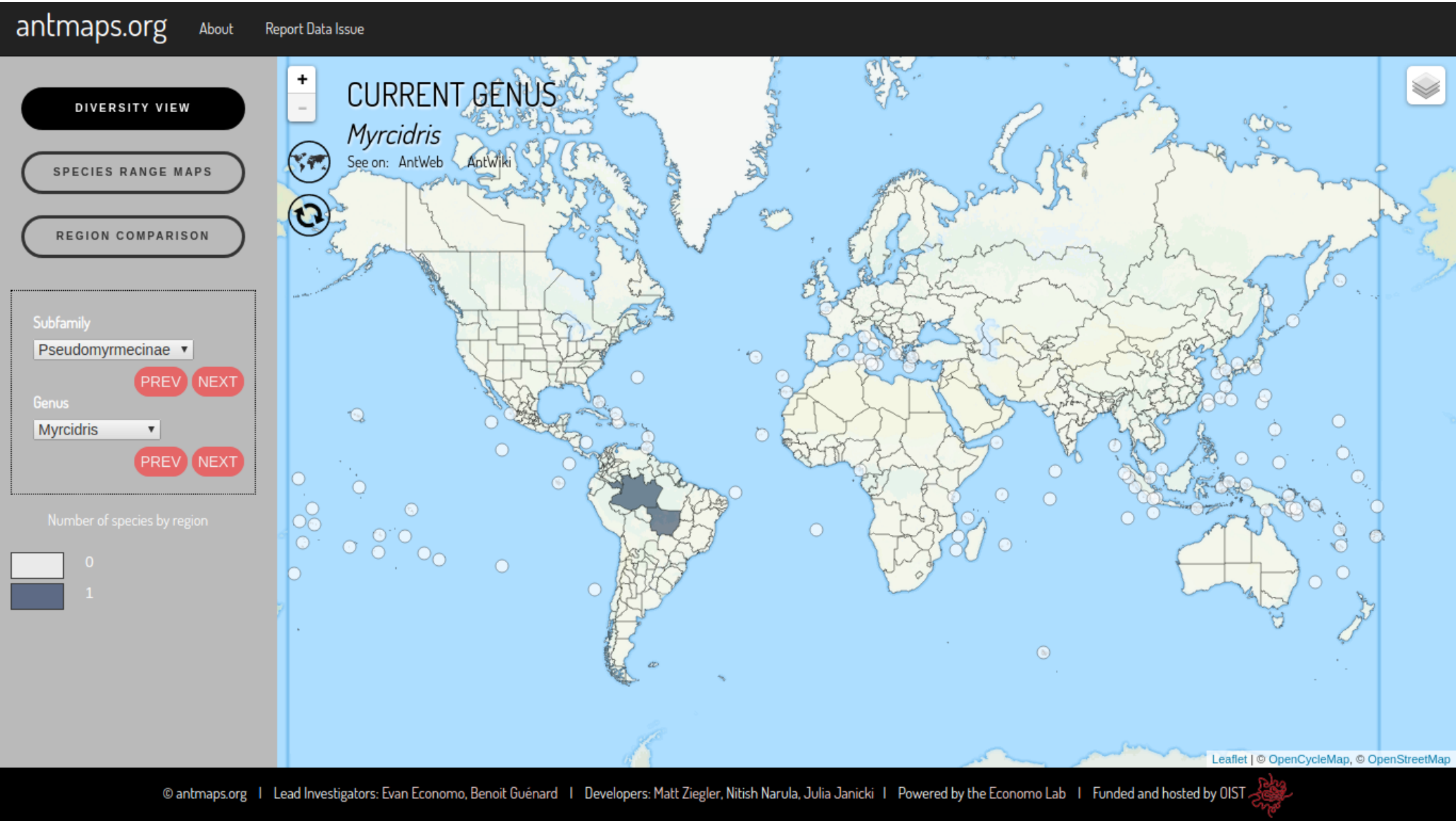
- 3 gêneros:
  - *Pseudomyrmex* (~137 spp.)
    - Mais diverso e estudado
    - Dividido em 10 grupos (morfologia)
  - *Tetraponera* (93 spp.)
    - Únicas não encontradas nas américas
    - Paleotrópicas
  - *Myrcidris* (1 espécie)

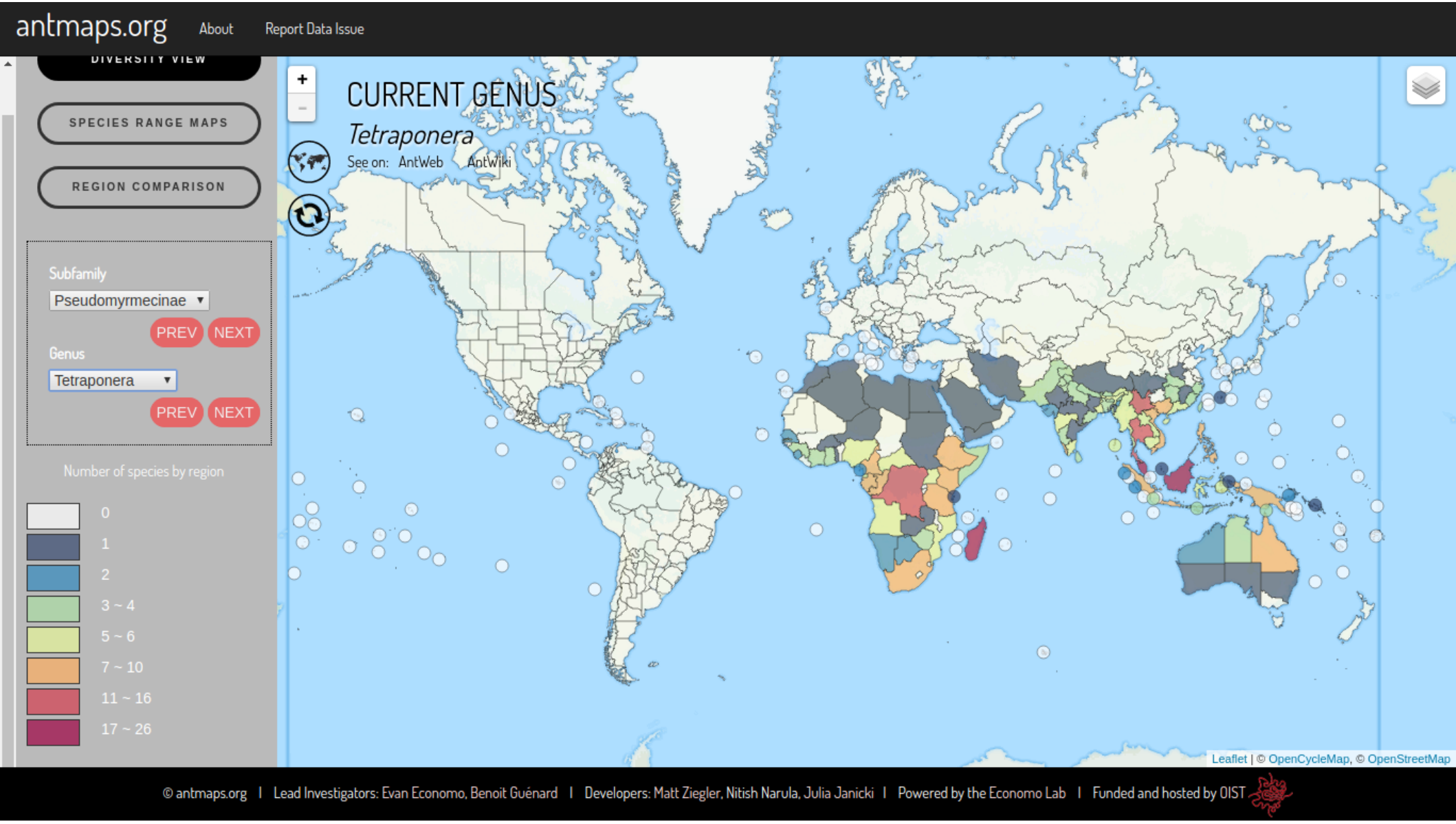
Generalista: *Pseudomyrmex gracilis*

- Dois grupos principais:
  - Generalistas (geralmente arbóreas)
  - Mutualistas obrigatórios (plant-ant)

Mutalista: *Tetraponera aethiops*

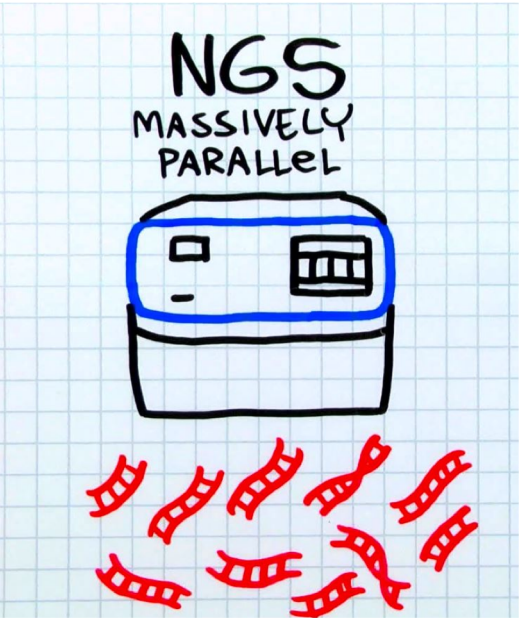

antmaps.org
About
Report Data Issue
DIVERSITY VIEW
SPECIES RANGE MAPS
REGION COMPARISON
Subfamily
Pseudomyrmecinae
PREV
NEXT
Genus
Pseudomyrmex
PREV
NEXT
Number of species by region
0
1
2 ~ 3
4 ~ 6
7 ~ 10
11 ~ 19
20 ~ 36
37 ~ 66
CURRENT GENUS
Pseudomyrmex
See on: AntWeb AntWiki
Leaflet | © OpenCycleMap, © OpenStreetMap
© antmaps.org | Lead Investigators: Evan Economo, Benoit Guénard | Developers: Matt Ziegler, Nitish Narula, Julia Janicki | Powered by the Economo Lab | Funded and hosted by OIST


antmaps.org
About
Report Data Issue
DIVERSITY VIEW
SPECIES RANGE MAPS
REGION COMPARISON
Subfamily
Pseudomyrmecinae
PREV
NEXT
Genus
Myrcidris
PREV
NEXT
Number of species by region
0
1
CURRENT GENUS
Myrcidris
See on: AntWeb AntWiki
Leaflet | © OpenCycleMap, © OpenStreetMap
© antmaps.org | Lead Investigators: Evan Economo, Benoit Guénard | Developers: Matt Ziegler, Nitish Narula, Julia Janicki | Powered by the Economo Lab | Funded and hosted by OIST


antmaps.org
About
Report Data Issue
DIVERSITY VIEW
SPECIES RANGE MAPS
REGION COMPARISON
Subfamily
Pseudomyrmecinae
PREV
NEXT
Genus
Tetraponera
PREV
NEXT
Number of species by region
0
1
2
3 ~ 4
5 ~ 6
7 ~ 10
11 ~ 16
17 ~ 26
CURRENT GENUS
Tetraponera
See on: AntWeb AntWiki
Leaflet | © OpenCycleMap, © OpenStreetMap
© antmaps.org | Lead Investigators: Evan Economo, Benoit Guénard | Developers: Matt Ziegler, Nitish Narula, Julia Janicki | Powered by the Economo Lab | Funded and hosted by OIST

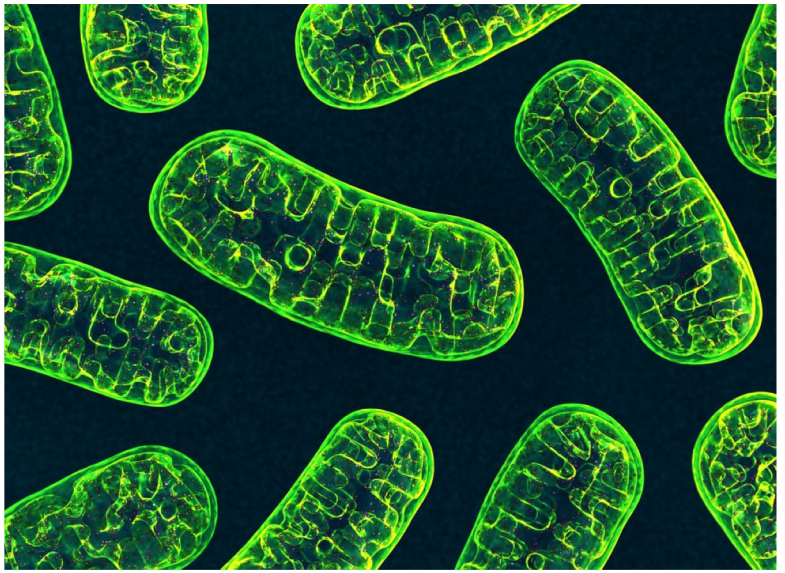
Pseudomyrmecinae

- Sem mitogenomas descritos



- Muitos dados disponíveis (WGS e UCE)

- 14 genomas mitocondriais completos:
  - NOVOPlasty, MIRA, MITObim
  - 12 *Pseudomyrmex* spp.
  - 2 *Tetraponera* spp.
- Análises:
  - Sintenia
  - BRIG (comparação por BLAST)
  - Filogenômica

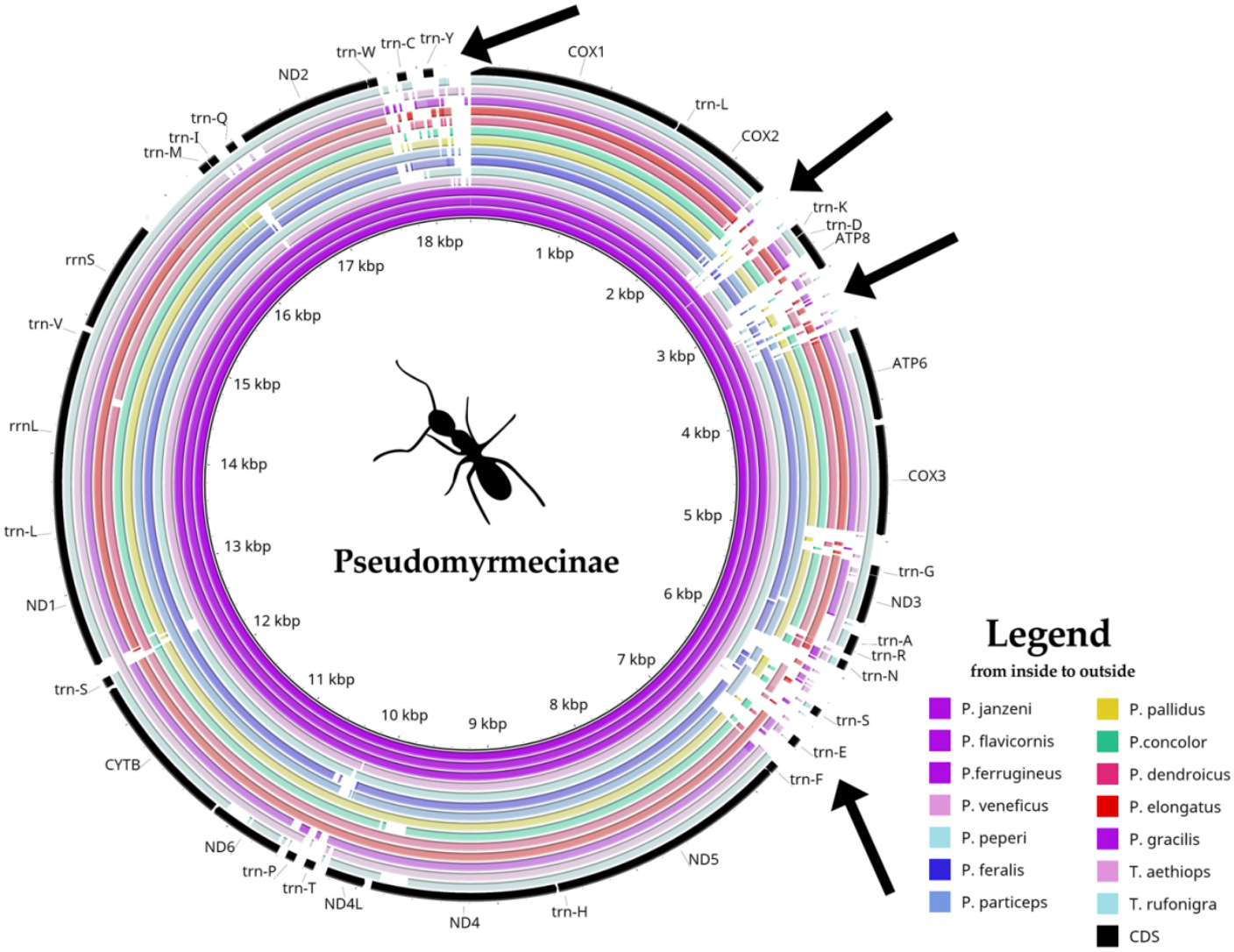


Sintenia

Synteny of all complete Formicidae mitogenomes available on Genbank. The gene arrangements inside the horizontal rectangle are present in most species analyzed and believed to be an ancestral feature for their clades, while the ones outside correspond to derivate unique gene orders, encountered in a single species. Vertical rectangles and arrows indicate regions where syteny changes occurred and the asterisk (*) and arrow in the trn-Y of W. auropunctata indicates that it is the only feature in Formicinae mitochondria that changed its coding strand and transcription direction.

BRIG (Blast Ring Image Generator)

- Tamanho de sequência (bp) variável em Pseudomyrmecinae:
  - 15704 (*Pseudomyrmex gracilis*) até 18835 (*Pseudomyrmex feralis*)
- Hipótese: Inserções de nucleotídeos
- Comparação por BLAST:
  - 4 possíveis regiões de inserção

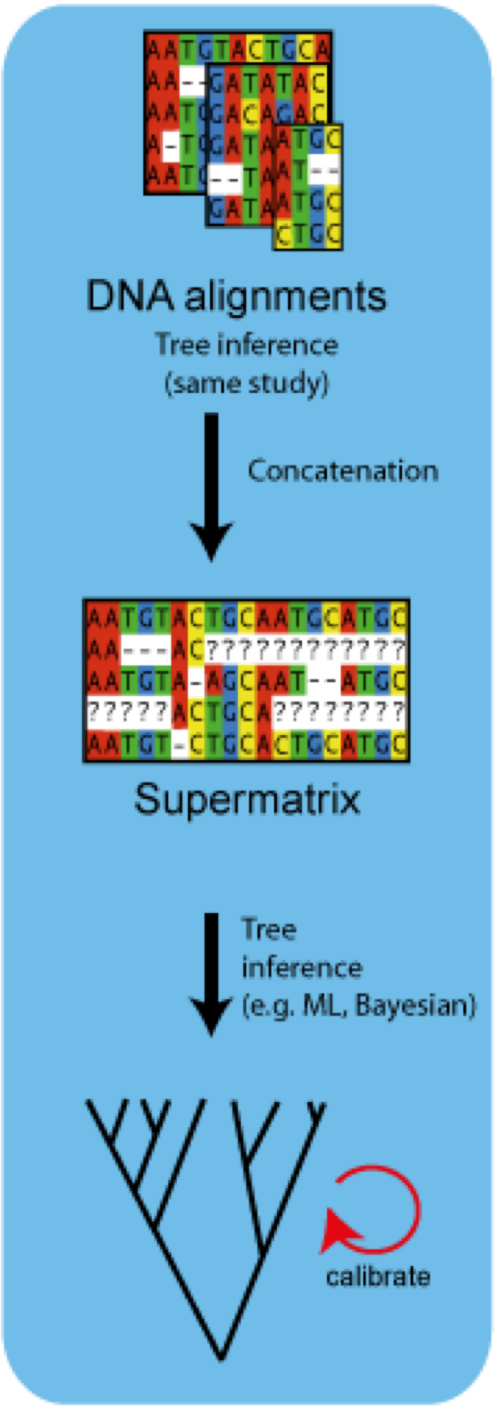


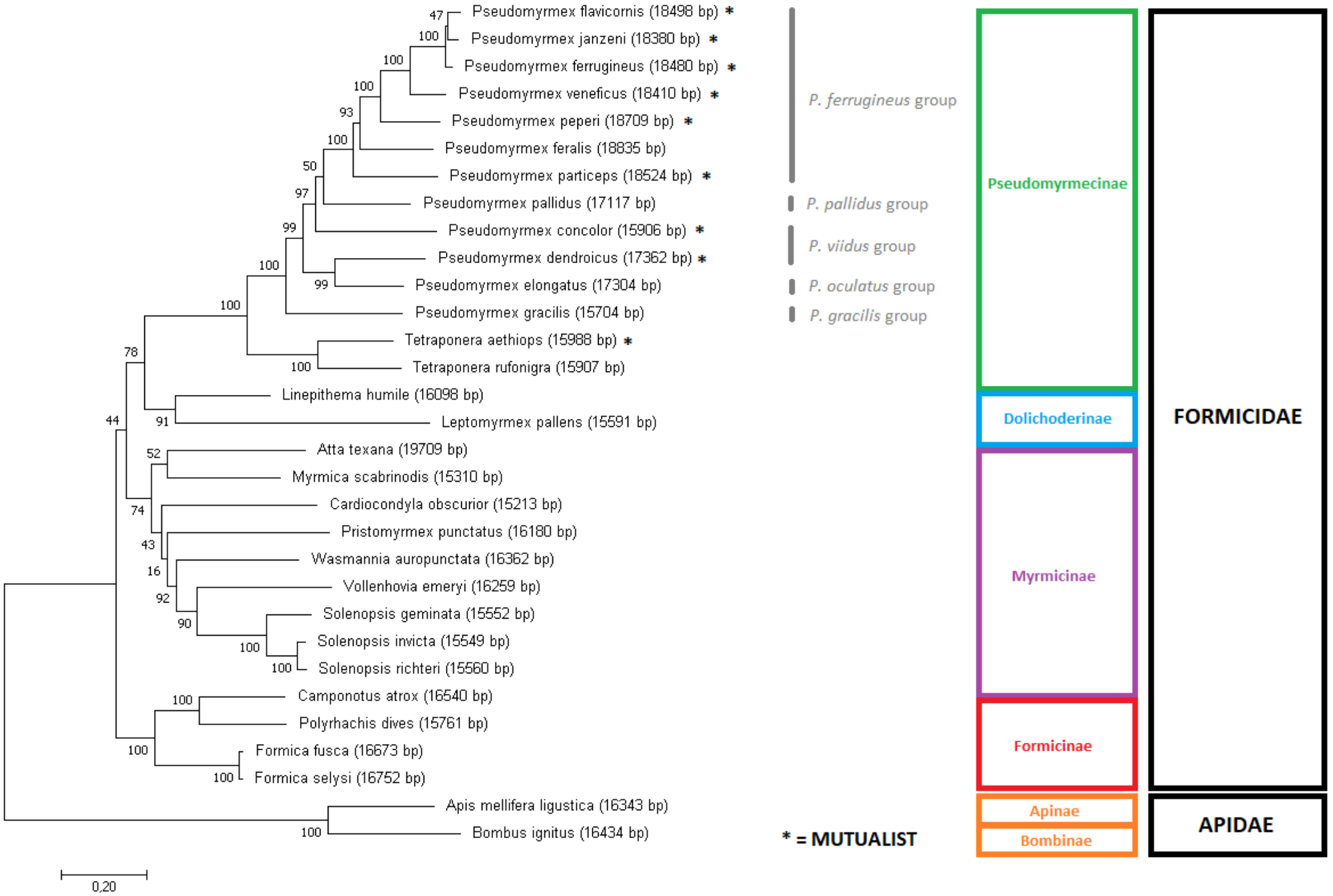
Filogenômica

- 29 mitogenomas de formiga
  - 14 novos
  - 15 já descritos
  - +2 outgroups (abelhas)

- Supermatriz:
  - 13 genes mitocondriais
  - Alinhados e concatenados

- Inferência filogenômica
  - Máxima verossimilhança
  - Reamostragem: Bootstrap (1000 repetições)




Pseudomyrmex flavicornis (18498 bp) *
Pseudomyrmex janzeni (18380 bp) *
Pseudomyrmex ferrugineus (18480 bp) *
Pseudomyrmex veneficus (18410 bp) *
Pseudomyrmex peperi (18709 bp) *
Pseudomyrmex feralis (18835 bp)
Pseudomyrmex particeps (18524 bp) *
Pseudomyrmex pallidus (17117 bp)
Pseudomyrmex concolor (15906 bp) *
Pseudomyrmex dendroicus (17362 bp) *
Pseudomyrmex elongatus (17304 bp)
Pseudomyrmex gracilis (15704 bp)
Tetraponera aethiops (15988 bp) *
Tetraponera rufonigra (15907 bp)
Linepithema humile (16098 bp)
Leptomyrmex pallens (15591 bp)
Atta texana (19709 bp)
Myrmica scabrinodis (15310 bp)
Cardiocondyla obscurior (15213 bp)
Pristomyrmex punctatus (16180 bp)
Wasmannia auropunctata (16362 bp)
Vollenhovia emeryi (16259 bp)
Solenopsis geminata (15552 bp)
Solenopsis invicta (15549 bp)
Solenopsis richteri (15560 bp)
Camponotus atrox (16540 bp)
Polyrhachis dives (15761 bp)
Formica fusca (16673 bp)
Formica selysi (16752 bp)
Apis mellifera ligustica (16343 bp)
Bombus ignitus (16434 bp)
47
100
100
100
93
100
50
97
99
100
99
100
78
100
44
91
52
74
43
16
92
90
100
100
100
100
100
100
0,20
P. ferrugineus group
P. pallidus group
P. viidus group
P. oculatus group
P. gracilis group
Pseudomyrmecinae
Dolichoderinae
Myrmicinae
Formicinae
Apinae
Bombinae
FORMICIDAE
APIDAE
* = MUTUALIST

Trabalho escrito

- Terminando de organizar a discussão
- Submissão em Outubro/2018
- PLoS ONE?
  - Impacto 2.8 (não temos dinheiro)
  - Current Zoology

No budget Mitogenomics:
a case study on the ant subfamily Pseudomyrmecinae

Gabriel Alves Vieira[1], Francisco Prosdocimi[1]

[1] Laboratório de Genômica e Biodiversidade, 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 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 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 identified possible sites in which nucleotidic insertions happened in some mitogenomes. Using a data mining/bioinformatics approach, the current work increased the number of complete mitochondrial genomes available for ants from 16 to 30, demonstrating the unique potential of public databases for mitogenomics studies.

Keywords: Pseudomyrmecinae;; Mitogenomics; Phylogenomics; evolutionary biology of ants; Bioinformatics; Data mining.

Montagem de mitogenomas usando dados públicos em computadores padrão

- É possível montar mitogenomas sem a necessidade de supercomputadores?
  - **Trabalho metodológico**

- Adaptação do *pipeline* usado para montar as 14 pseudomyrmecines

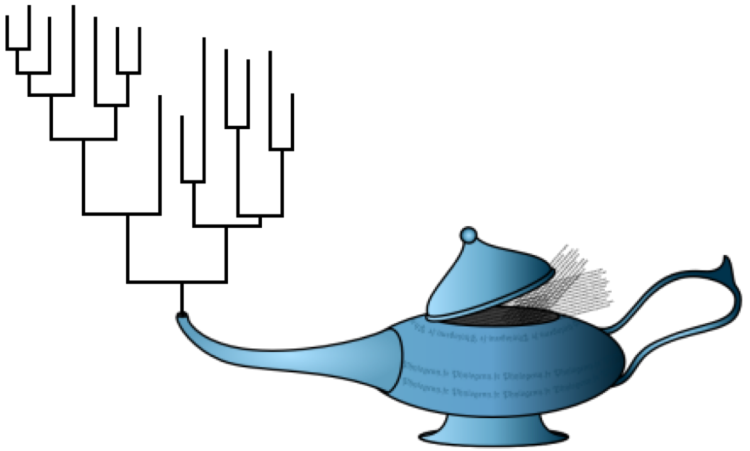
- Otimização do consumo de RAM
  - Não usa MIRA
  - NOVOPlasty e MITObim

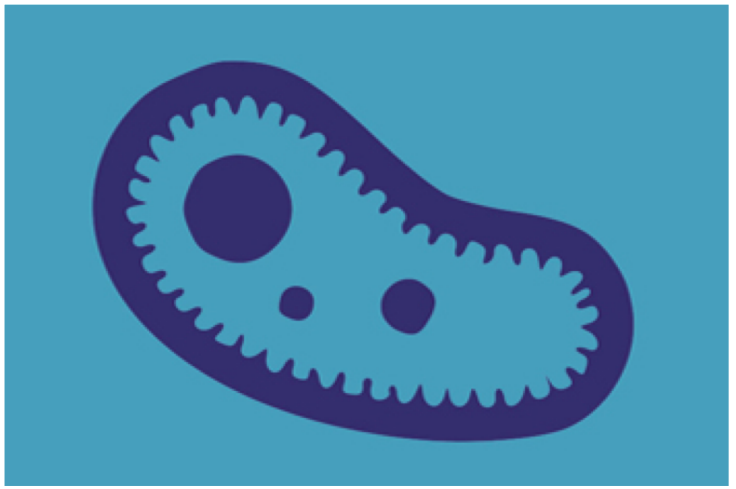
- Prova de conceito:
  - Montar 100 mitocôndrias usando no máximo 8 GB de RAM
  - 40 já montadas
  - 7 alunos de graduação envolvidos

Leveza (~ R$ 6300,00) :
Intel® Core™ i7-3930K x 12
32 GB de RAM
3 TB de HD

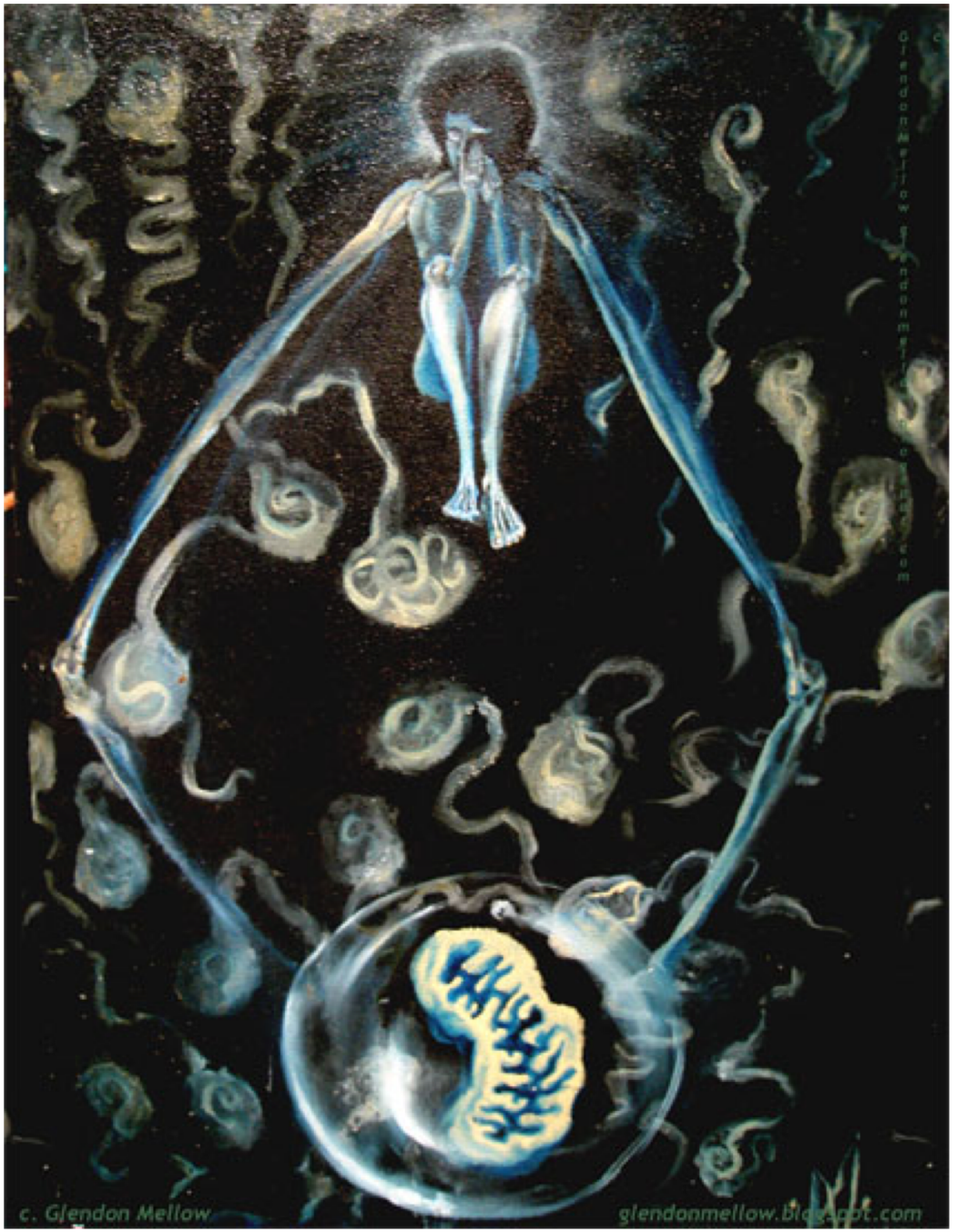
Perspectivas

- Submeter o artigo sobre as Pseudomyrmecinae
- Finalizar o artigo descrevendo a metodologia para montagem otimizada para computadores pessoais
- Criar script que automatize o processo
- Escrever a dissertação
  - Defesa programada para Março (2019)





OBRIGADO!



CAPES