



# MONTAGEM, ANOTAÇÃO E FILOGENÔMICA MITOCONDRIAL DE FORMIGAS DA SUBFAMÍLIA PSEUDOMYRMECINAE USANDO DADOS PÚBLICOS

**Aluno:** Gabriel Alves Vieira

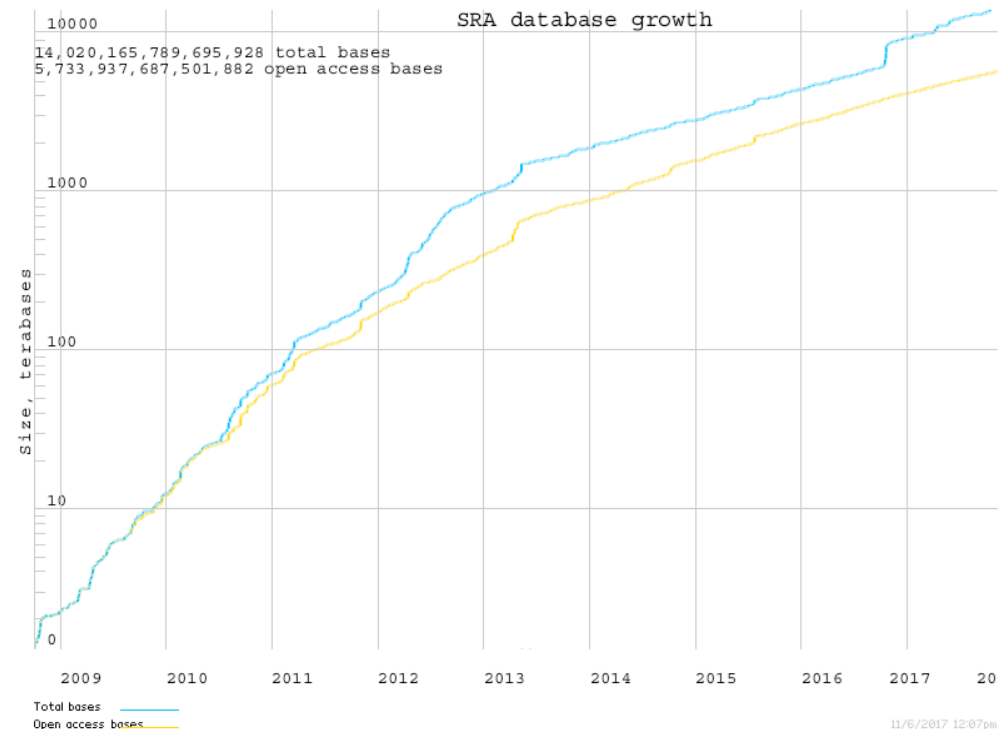
**Orientador:** Francisco Prosdocimi



AAGTCAAGCTGCTCTGTGGGCTGTGATCTGCCTCAAACCCACAGCCTGGGTAGCAGG  
AGGACCTTGATGCTCCTGGCACAGATGAGGAGAATCTCTTTTCTCCTGCTTGAAG  
GACAGACATGACTTTGGATTTCCTCCAGGAGGAGTTGGCAACCACTTCCAAAAGGCT  
GAAACCATCCCTGCTCCTCCATGAGATGATCCAGCAGATCTTCAATCTCTTCAGCACA  
AAGGACTCATCTGCTGCTTGGGATGAGACCCCTCCTAGACAAATCTTACACTGAATC  
TACCAGCAGCTGAATGACCTGGAAGCCTGTGTGATACAGGGGGTGGGGTGACAGAG  
ACTCCCTGATGAAGGAGGACTCCATTCTGGCTGTGAGGAAATACTTCCAAAGAATC  
ACTCTCTATCTGAAAGAGAAGAAATACAGCCCTTGTGCTGGGAGGTTGTGACAGCA  
GAAATCATGAGATCTTTTCTTTGTCAACAACTTGCAAGAAAGTTTAAGAAGTAAG  
GAATGA, TGTGATCTGCCTCAAACCCACAGCCTGGGTAGCAGGAGGACCTTGATGC  
TCCTGGCACAGATGAGGAGAATCTCTTTTCTCCTGCTTGAAGGACAGACATGACT  
TTGGATTTCCCTCCAGGAGGTTGGCAACCACTTCCAAAAGGCTGAAACCATCCCTG  
TCCTCCATGAGATGATCCAGCAGATCTTCAATCTCTTCAGCACAAGGACTCATCTG  
CTGCTTGGGATGAGACCTCCTAGACAAATCTACACTGAACTTACACGACGCTGA

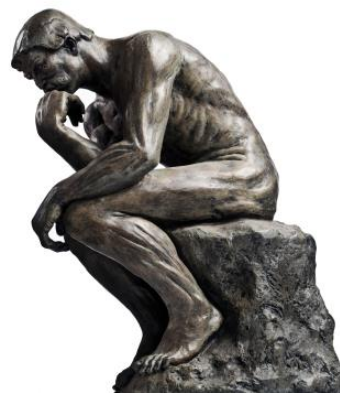
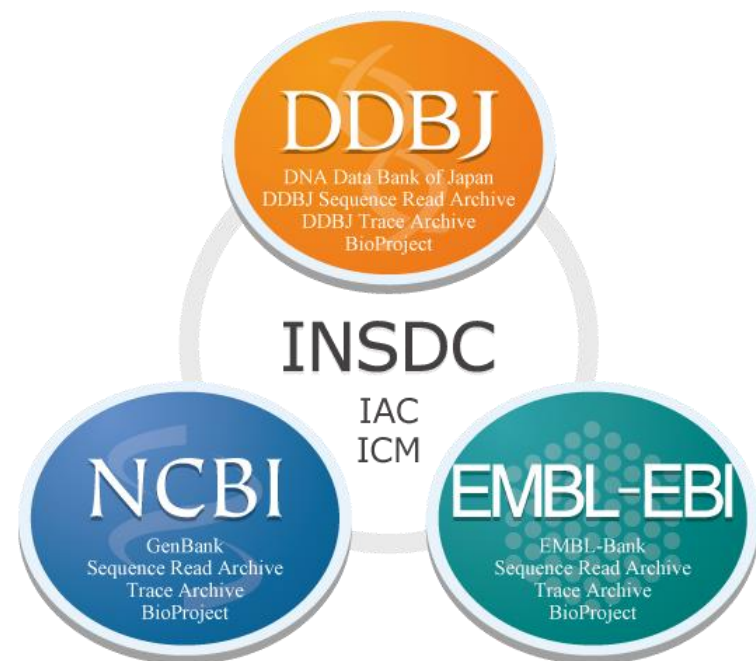
# International Nucleotide Sequence Database Collaboration

- Sequenciamento de Nova Geração (NGS) : 2005
  - ↑ produção de dados
  - ↓ custo por base
- Crescimento dos bancos de dados públicos de sequenciamento
- Revistas: Incentivam disponibilização dos dados
  - Reprodutibilidade



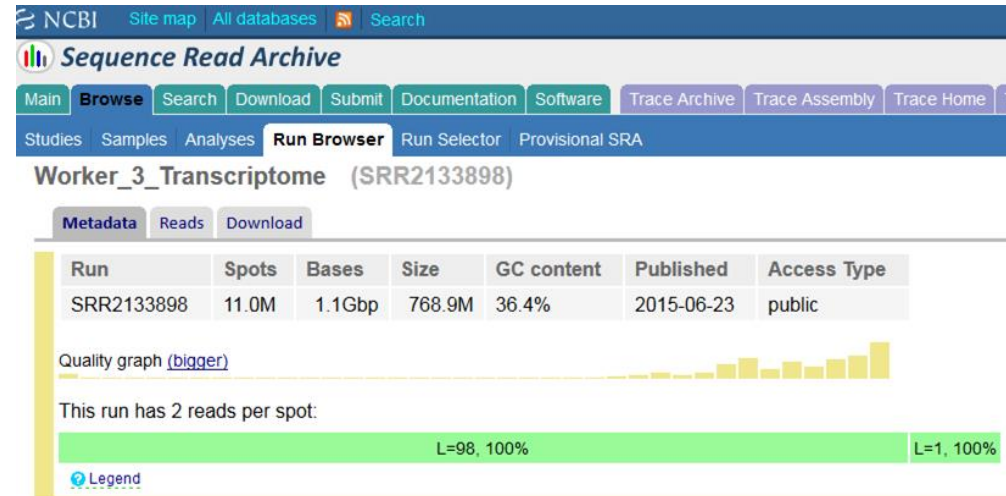
# Sequence Read Archive (SRA)

- Parte de colaboração internacional
  - 3 bancos de dados:
    - SRA: NCBI Sequence Read Archive
    - ERA: EBI Sequence Read Archive
    - DRA: DDBJ Sequence Read Archive
  - Sincronizados semanalmente
- Diferentes tipos de dados:
  - Um pouco de reflexão e imaginação
  - Gerar novos conhecimentos



# Dados públicos

- Dados usados para responder uma pergunta
  - Sem relação com a mitocôndria
- Sequenciamento
  - Sequências mitocondriais presentes
  - 1 em cada 200 / 1 em cada 1000 sequências é mitocondrial
  - Espécies com dados públicos e sem mitogenoma descrito
- Esses dados podem ser usados para obter o mitogenomas
  - Além de outras análises...



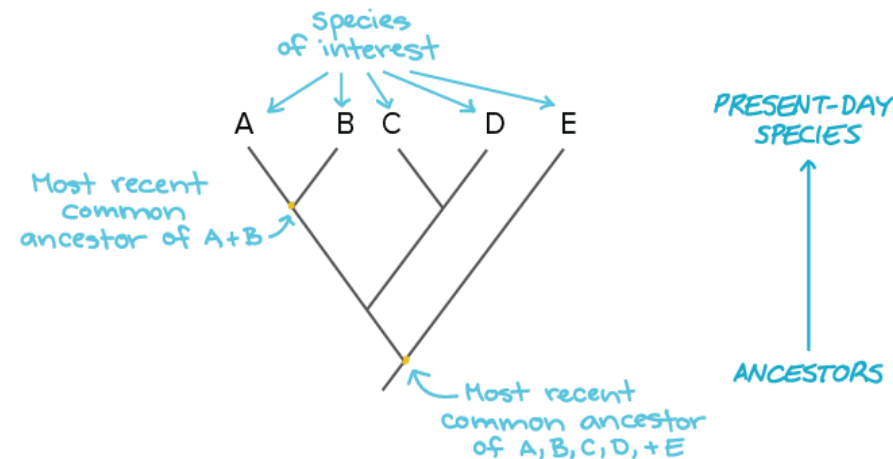
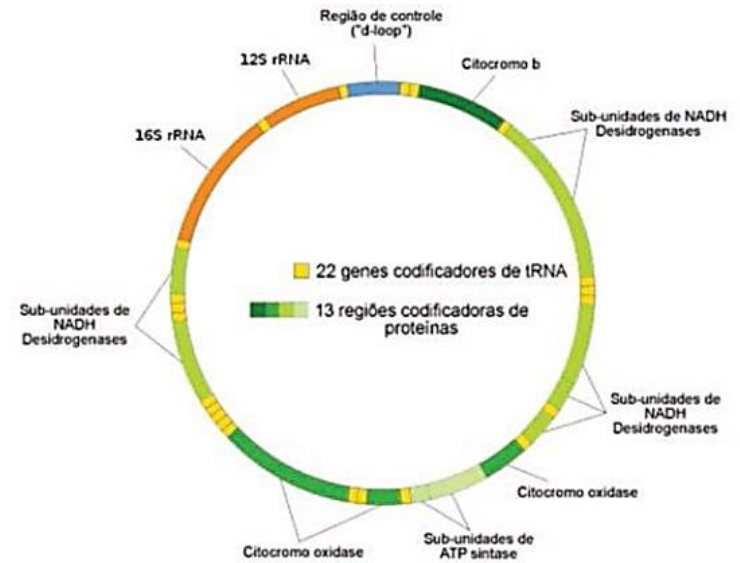


# Mas por que o mitogenoma?

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

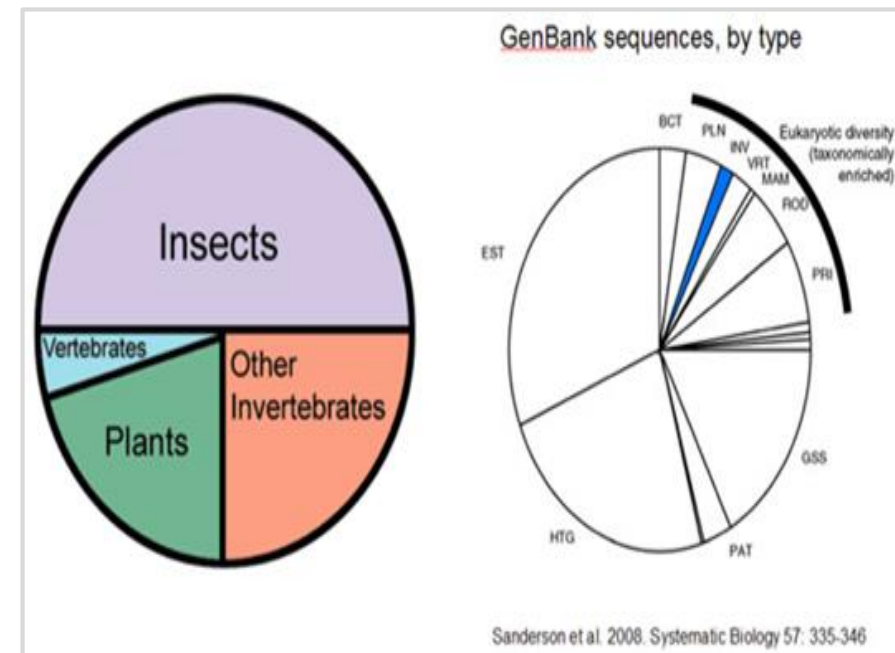
- Utilizado como marcador molecular
    - Sem recombinação
    - Herança materna
- Filogenia e  
Filogeografia

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



# Formigas (Hymenoptera: Formicidae)

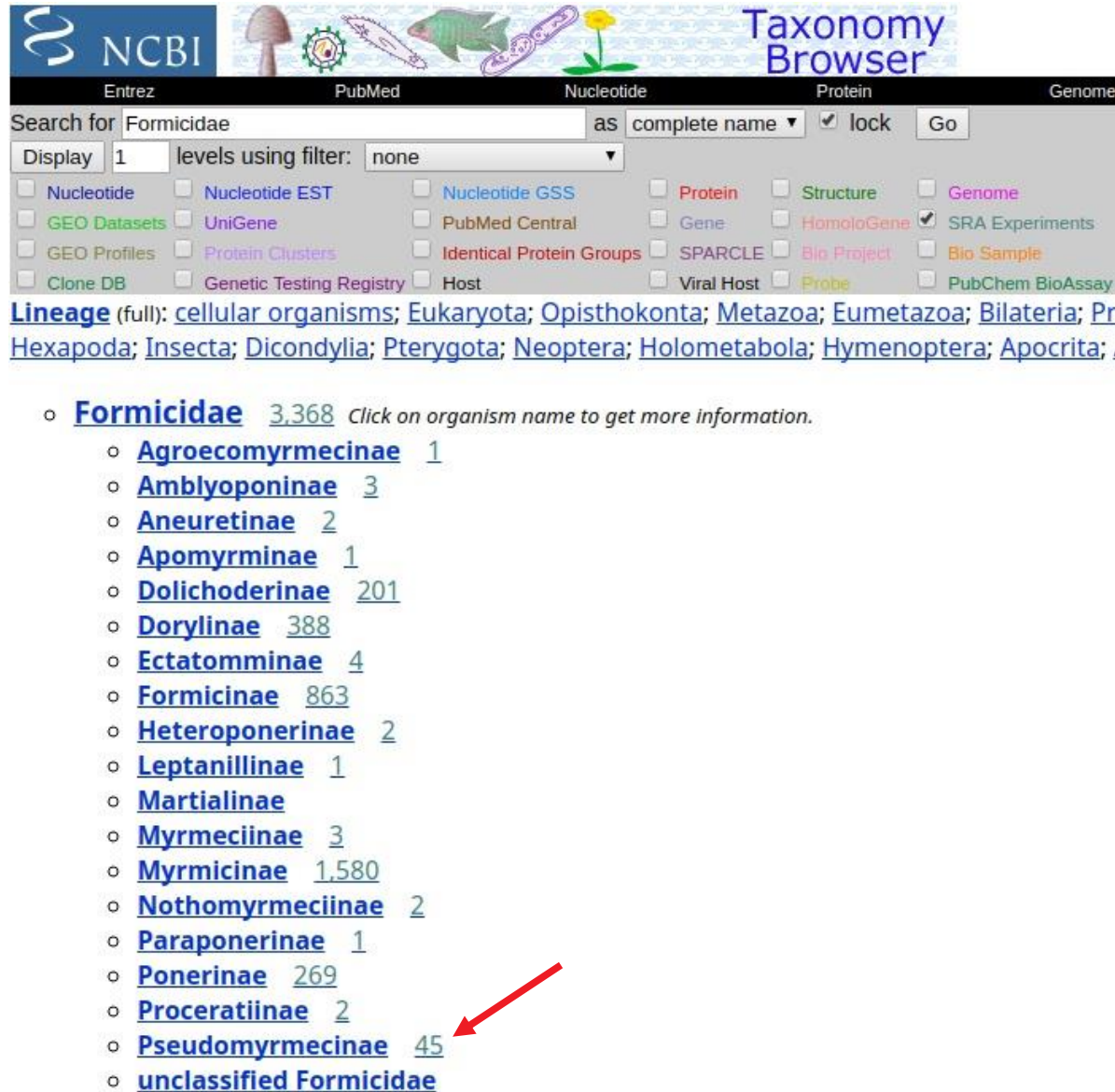
- Engenheiras ecossistêmicas
  - Alteram características do solo
  - Serviços ecossistêmicos
- Importância econômica
  - Pragas
  - Agentes de controle biológico
- Grande biodiversidade
  - ≈ 13.000 spp.
  - GenBank
    - 15 mitogenomas completos descritos



Desconhecimento da biodiversidade molecular

# Formigas (Hymenoptera: Formicidae)

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



The screenshot shows the NCBI Taxonomy Browser interface. The search bar contains "Formicidae" and the results are displayed as a list of subfamilies with their respective counts. A red arrow points to the "Pseudomyrmecinae" subfamily, which has a count of 45.

NCBI Taxonomy Browser

Search for: Formicidae as complete name lock Go

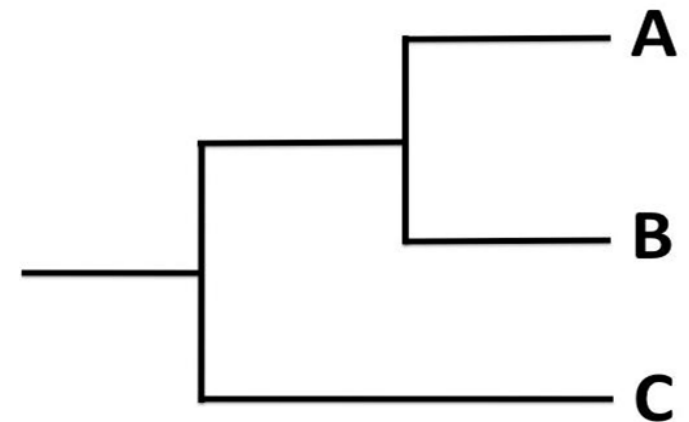
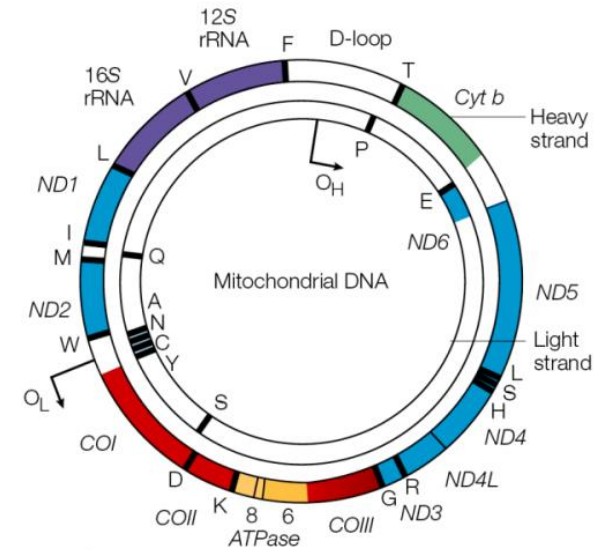
Display: 1 levels using filter: none

Lineage (full): [cellular organisms](#); [Eukaryota](#); [Opisthokonta](#); [Metazoa](#); [Eumetazoa](#); [Bilateria](#); [Pr](#)  
[Hexapoda](#); [Insecta](#); [Dicondylia](#); [Pterygota](#); [Neoptera](#); [Holometabola](#); [Hymenoptera](#); [Apocrita](#);

- [Formicidae](#) 3,368 Click on organism name to get more information.
  - [Agroecomyrmecinae](#) 1
  - [Amblyoponinae](#) 3
  - [Aneuretinae](#) 2
  - [Apomyrminae](#) 1
  - [Dolichoderinae](#) 201
  - [Dorylinae](#) 388
  - [Ectatomminae](#) 4
  - [Formicinae](#) 863
  - [Heteroponerinae](#) 2
  - [Leptanillinae](#) 1
  - [Martialinae](#)
  - [Myrmeciinae](#) 3
  - [Myrmicinae](#) 1,580
  - [Nothomyrmeciinae](#) 2
  - [Paraponerinae](#) 1
  - [Ponerinae](#) 269
  - [Proceratiinae](#) 2
  - [Pseudomyrmecinae](#) 45
  - [unclassified Formicidae](#)

# 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

Diagram illustrating the structure of a FASTQ file entry:

```
@FORJUSP02AJWD1
CCGTCAATTCAATTAAGTTTAACTTGC GGCCGTACTCCCAGGCGGT
+
AAAAAAAAAAAA::99@:::??@::FFAAAAACCAA:::BB@@?A?
```

Labels:

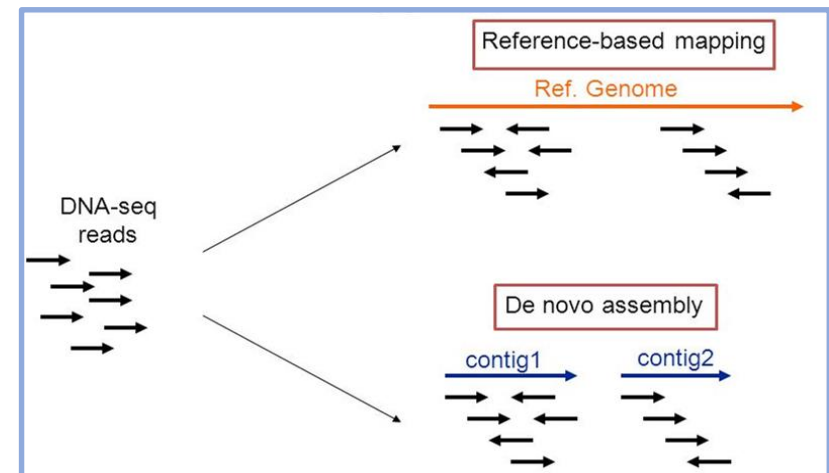
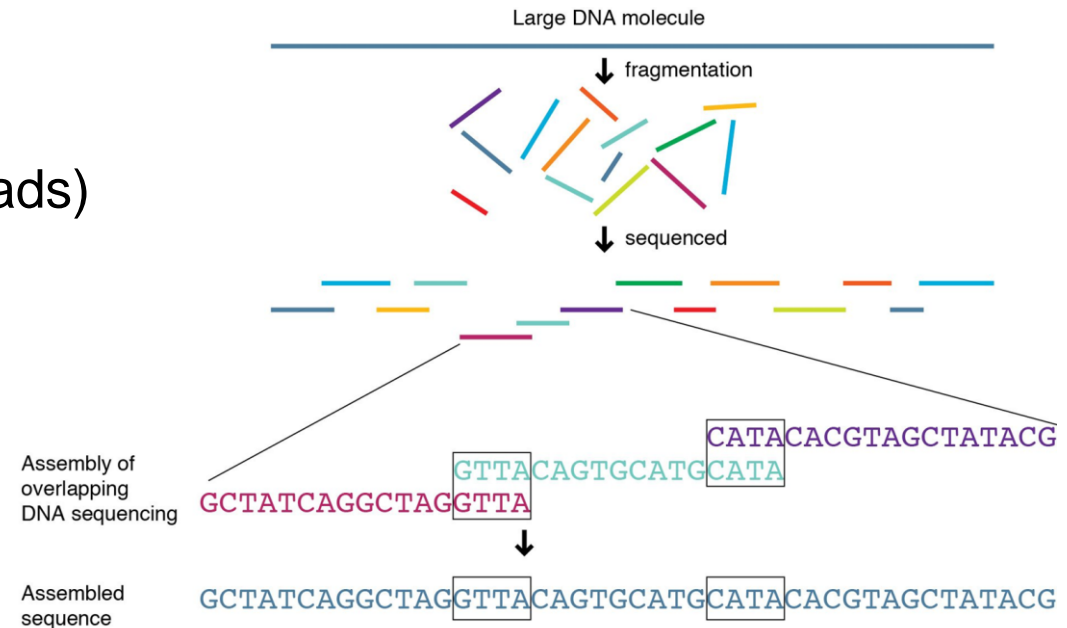
- Label**: @FORJUSP02AJWD1
- Sequence**: CCGTCAATTCAATTAAGTTTAACTTGC GGCCGTACTCCCAGGCGGT
- Q scores (as ASCII chars)**: AAAAAAAAAA::99@:::??@::FFAAAAACCAA:::BB@@?A?
- Base=T, Q=':'=25**: (Indicates the mapping of the character ':' to a quality score of 25)

# MONTAGEM

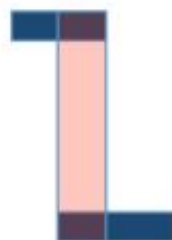
- Obtenção de pequenas sequências (reads)
- Quebra-cabeça – Ferramentas computacionais
  - NOVOPlasty, MIRA, MITObim

- Dois tipos de montagem

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



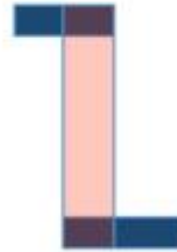




**Sequência contínua ou contig**







**Fragmentos  
alinhados**

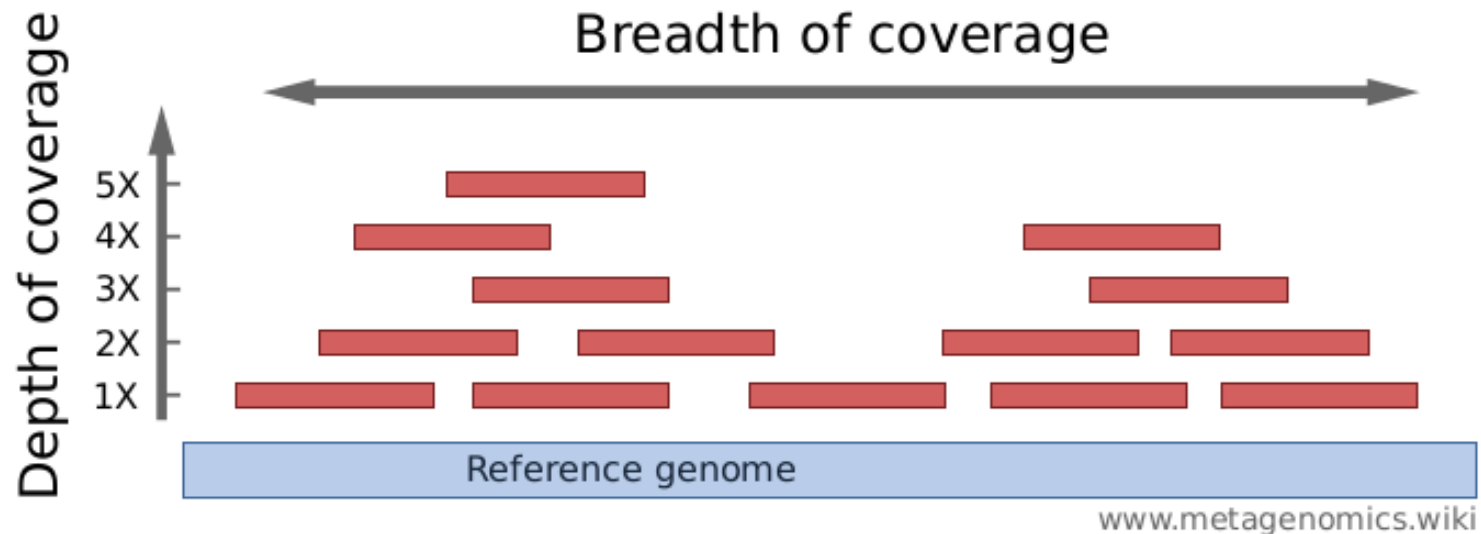
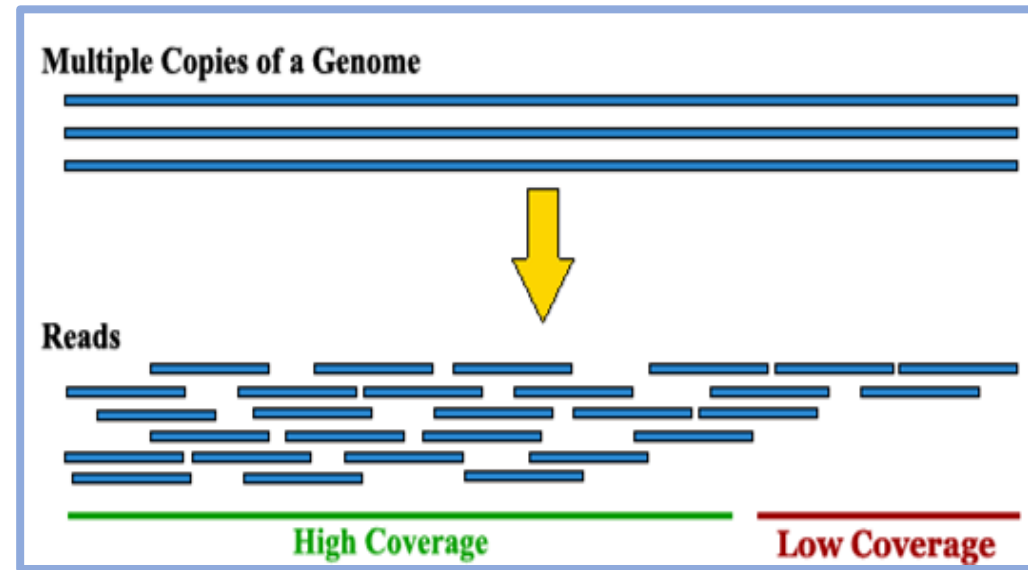
```
ACGCGATTCAGGTTACCACG
GCGATTCAGGTTACCACGCG
GATTCAGGTTACCACGCGTA
TTCAGGTTACCACGCGTAGC
CAGGTTACCACGCGTAGCGC
GGTTACCACGCGTAGCGCAT
TTACCACGCGTAGCGCATT
ACACGCGTAGCGCATTACA
CACGCGTAGCGCATTACACA
CGCGTAGCGCATTACACAGA
CGTAGCGCATTACACAGATT
TAGCGCATTACACAGATTAG
```

**Contig consenso**

```
ACGCGATTCAGGTTACCACGCGTAGCGCATTACACAGATTAG
```

# MONTAGEM

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



■ tRNA gene ■ rRNA gene ■ protein coding gene

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



Generalista: *Pseudomyrmex gracilis*

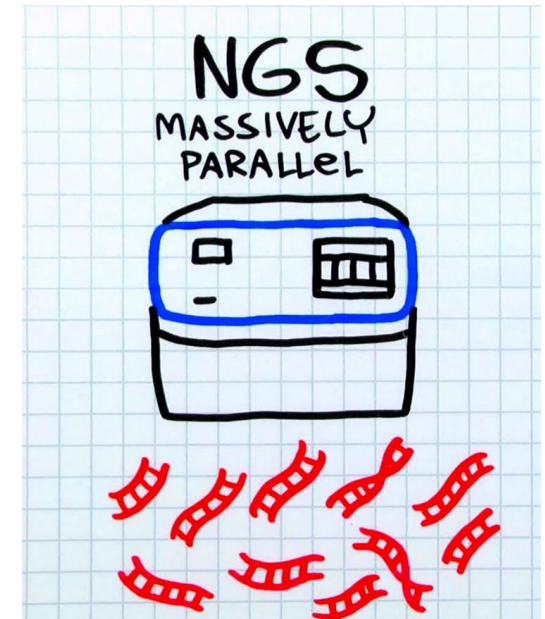


Mutalista: *Tetraponera aethiops*



# Pseudomyrmecinae

- Sem mitogenomas descritos
- Muitos dados disponíveis (WGS e UCE)
  - 45 datasets
- 14 genomas mitocondriais completos:
  - NOVOPlasty, MIRA, MITObim
  - 12 *Pseudomyrmex* spp.
  - 2 *Tetraponera* spp.
- Análises:
  - Sintonia
  - BRIG (comparação por BLAST)
  - **Filogenômica**



# 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)



DNA alignments

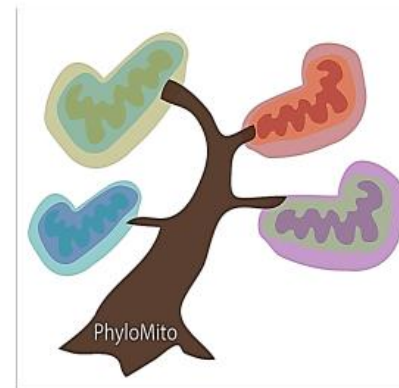
Tree inference  
(same study)

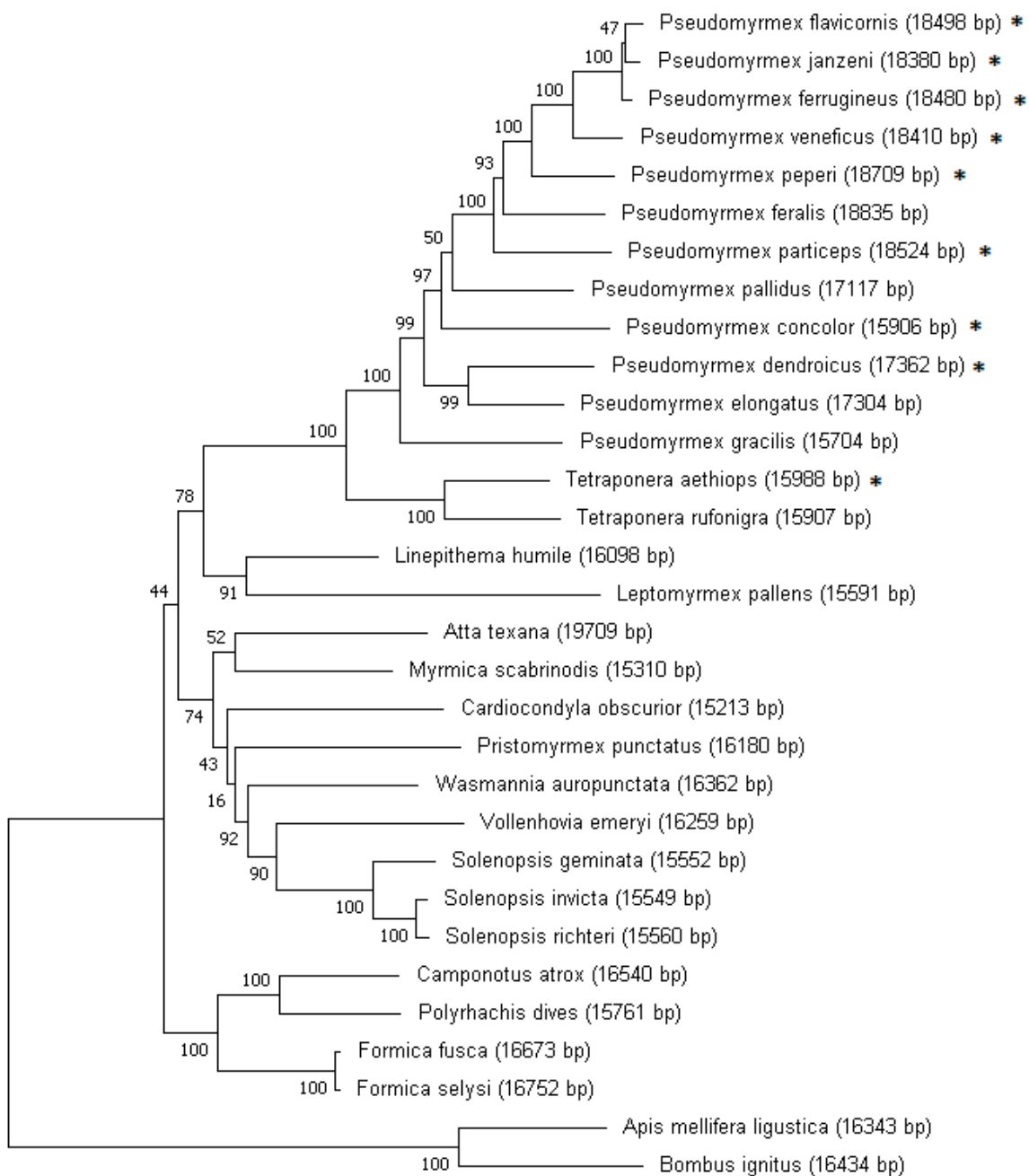
Concatenation



Supermatrix

Tree  
inference  
(e.g. ML, Bayesian)





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
- Current Zoology
  - Impact factor: 2.393

Gabriel Alves Vieira<sup>1</sup>, Francisco Prosdocimi<sup>1</sup>

<sup>1</sup> 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.

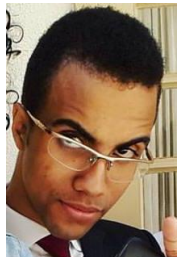
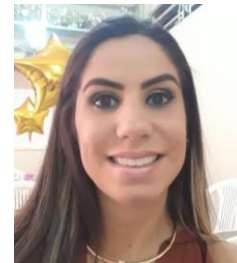


# Montagem de mitogenomas usando dados públicos em computadores padrão

- 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

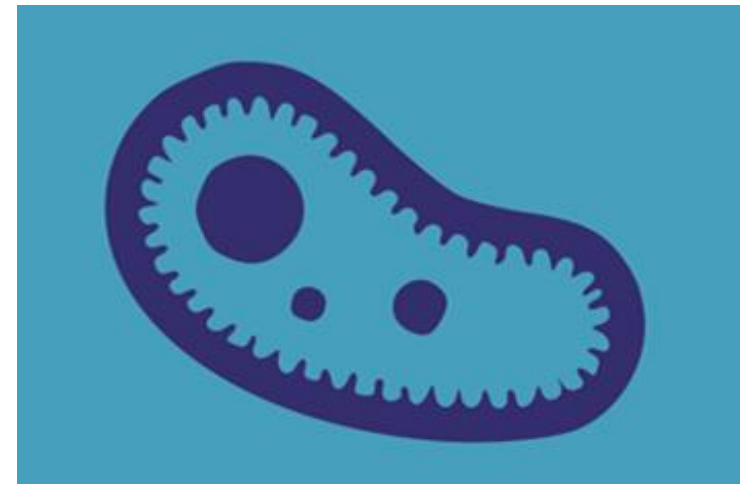
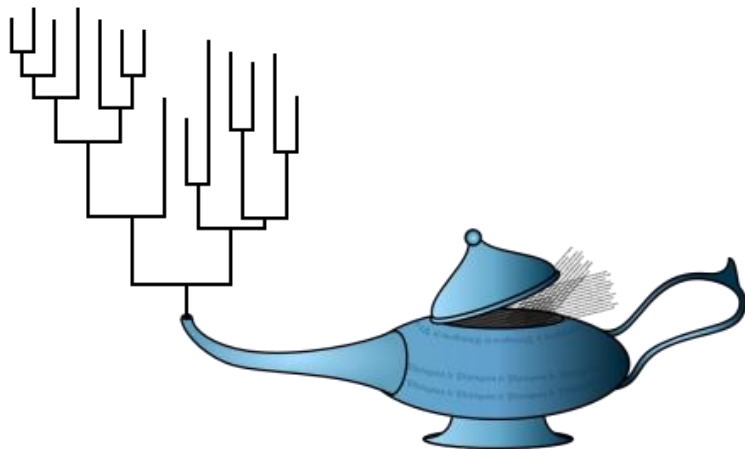


Leveza (~ R\$ 6300,00) :  
Intel® Core™ i7-3930K x 12  
32 GB de RAM  
3 TB de HD



# 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)
- Doutorado:
  - Coordenar alunos de graduação - Abarcar a ampla biodiversidade das formigas
  - Trabalhar na montagem de um genoma completo



# OBRIGADO!





# 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.



# Genômica comparativa

- 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

