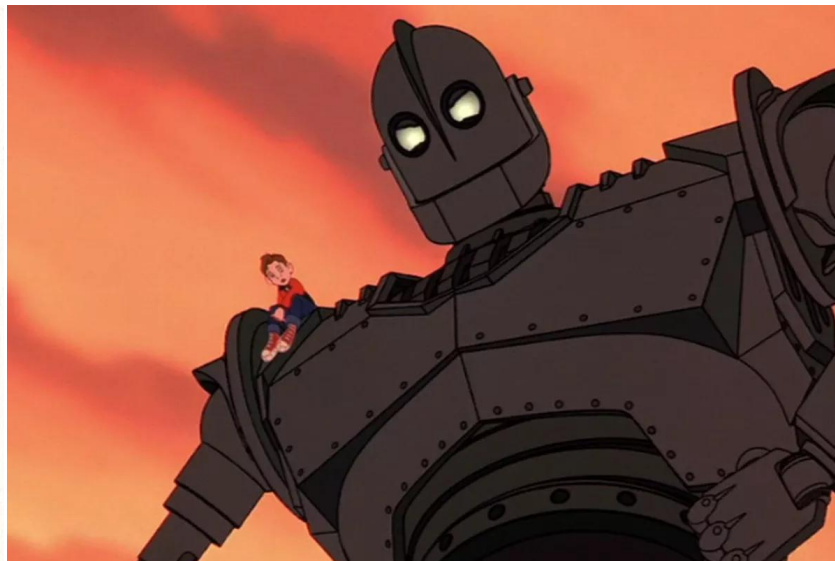
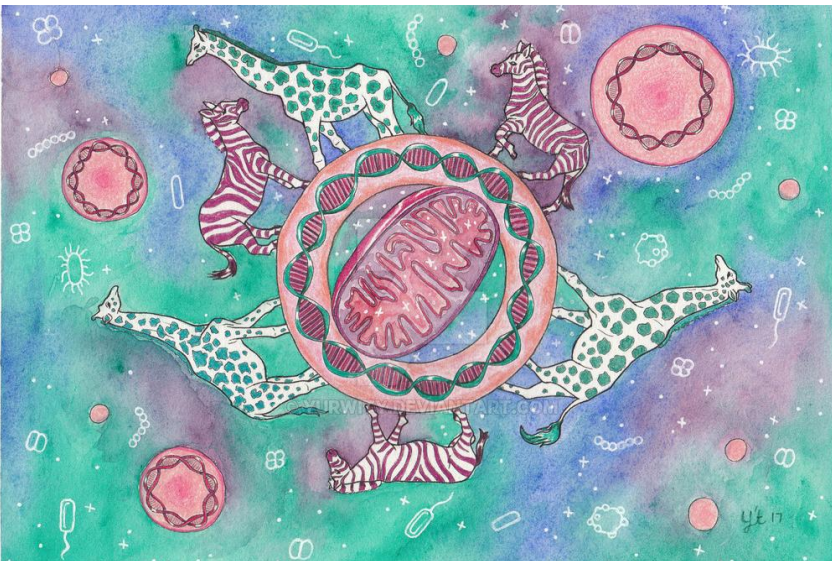


Sobre os ombros de gigantes: Ferramentas gratuitas e trabalhos no-budget em bioinformática

Gabriel Alves Vieira - UFRJ



```
// Node stopping error is caught below in the select.
if err := t.rpcContext.Stopper.RunTask(
    stream.Context(), "storage.RaftTransport: processing batch",
    func(ctx context.Context) {
        t.rpcContext.Stopper.RunWorker(ctx, func(ctx context.Context) {
            errCh <- func() error {
                var stats *raftTransportStats
                stream := &lockedRaftMessageResponseStream{MultiRaft_RaftM
                for {
                    batch, err := stream.Recv()
                    if err != nil {
                        return err
                    }
                    if len(batch.Requests) == 0 {
                        continue
                    }

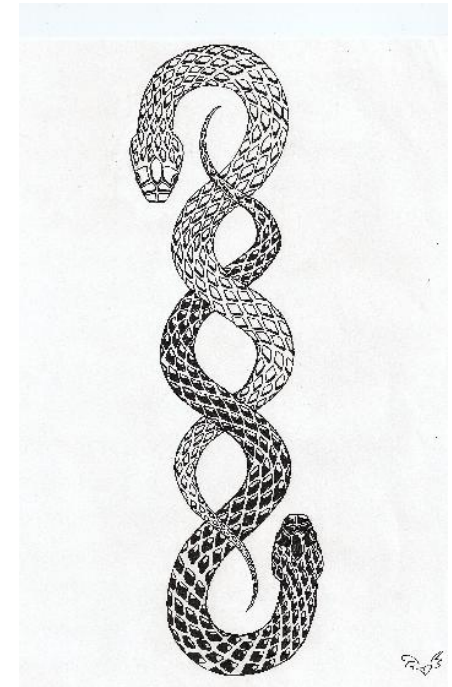
                    if stats == nil {
                        stats = t.getStats(batch.Requests[0].FromReplica.NodeID)
                    }

                    for i := range batch.Requests {
                        req := &batch.Requests[i]
                        atomic.AddInt64(&stats.serverRecv, 1)
                        stream.Send(req)
                        stats.handleRaftRequest(req, pErr)
                    }
                }
            }()
        })
    },
    1)

```

Sumário

1. Bioinfo e eu
2. Montagem, anotação e mitogenômica no-budget
3. Programação (Python)
4. Git e Github
5. Sobre os ombros de gigantes
 - Ver mais longe
 - Vertigem



Se eu vi mais
longe, foi por
estar sobre
ombros de
gigantes.

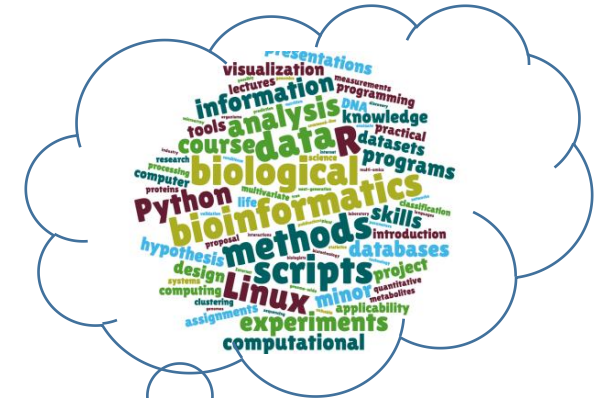
Isaac Newton

 PENSADOR



Desventuras em série...

- Eu:
 - Na vida: Sovina
 - Graduação em Biologia: Péssimo em bancada e campo
- Ouro Preto: Simpósio de Biotecnologia
 - Bioinformática
 - Driblar meus pontos negativos
 - Possibilidade de fazer pesquisa sem financiamento



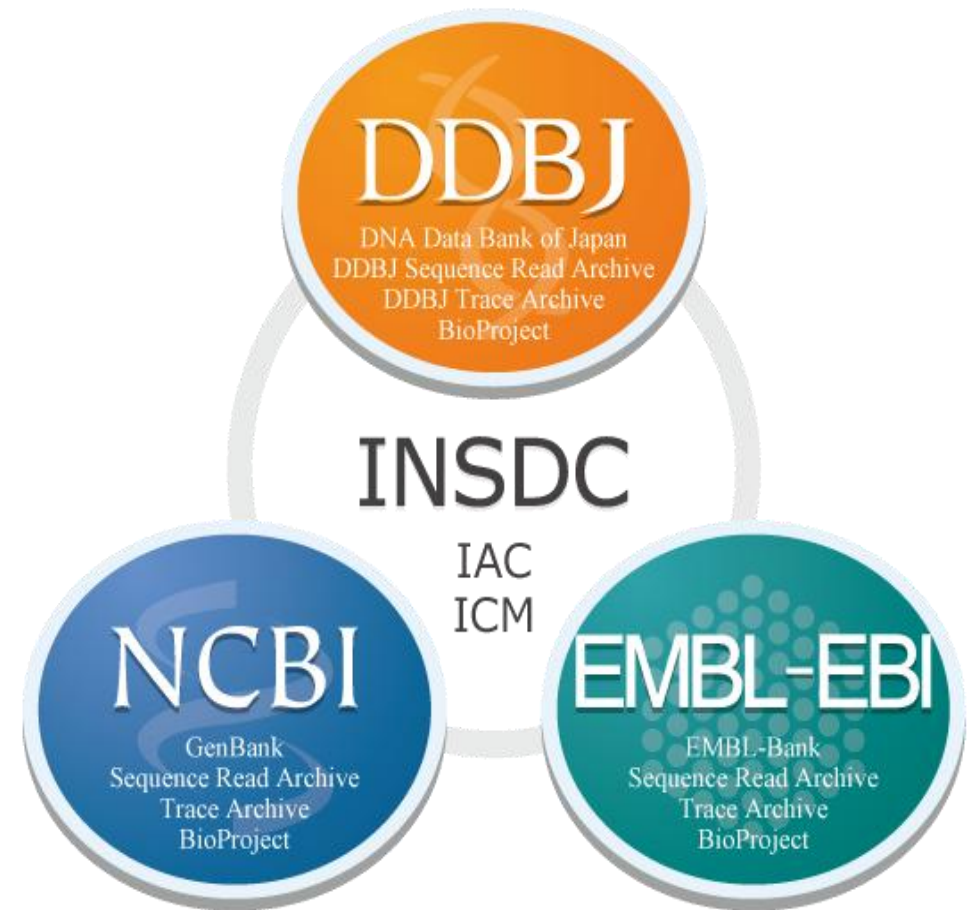
Bioinformática (Genômica)

- Custo:
 - Obtenção de dados (sequenciamento)
 - Infraestrutura - Servidores robustos
- Publicação dos dados:
 - Exigência - várias revistas
 - Reprodutibilidade
- Abundância de dados públicos + Ferramentas gratuitas = Potencial No-budget
- Sequências montadas e anotadas - Genbank
- Dados brutos de sequenciamento - SRA (Sequence Read Archive)



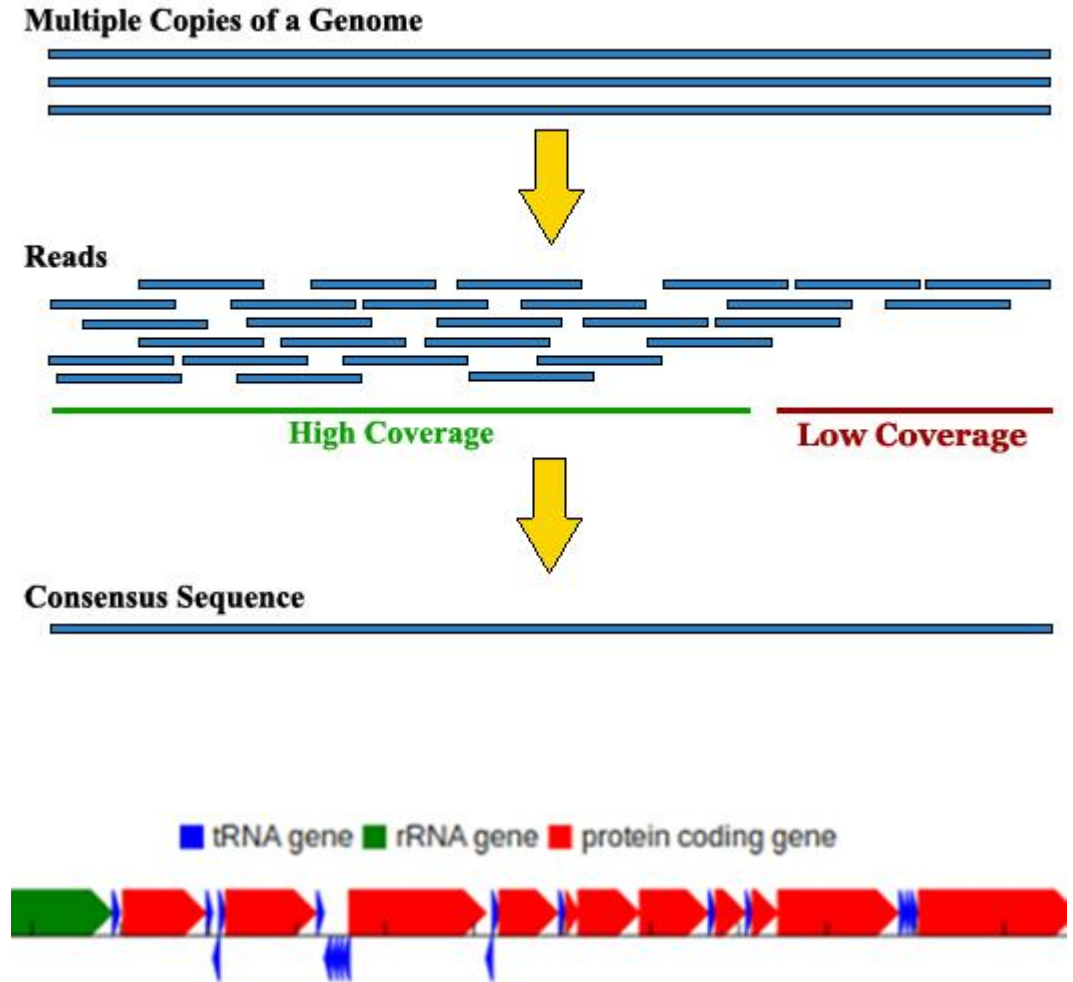
Sequence Read Archive (SRA)

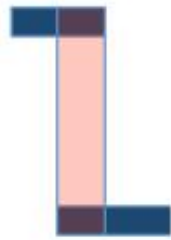
- Parte de colaboração internacional
 - 3 bancos de dados:
 - SRA: NCBI Sequence Read Archive
 - ERA: EBI Sequence Read Archive
 - DRA: DDBJ Sequence Read Archive
 - Sincronizados entre si
- Diferentes tipos de dados:
 - Avaliação de polimorfismos e SNPs
 - Impacto de procedimentos sobre dados (e.g. Trimming)
 - Teste de novos programas de bioinformática
 - **Montagem e anotação de mitogenomas**

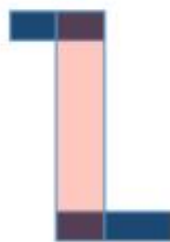


MONTAGEM E ANOTAÇÃO

- Sequenciamento:
 - Obtenção de sequências curtas de DNA (*reads*)
- Montagem
 - Sobreposição das *reads*
 - Montagem de um quebra-cabeça
 - De milhares a bilhões de peças
 - Ferramentas computacionais (montadores)
- Anotação
 - Identificar regiões no genoma (Ex: genes)

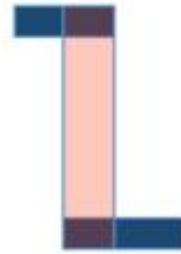






Sequência contínua ou contig





**Fragmentos
alinhados**

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

Contig consenso

```
ACGCGATTCAGGTTACCACGCGTAGCGCATTACACAGATTAG
```

Montagem

- Resultado da montagem:
 - Arquivo fasta (.fasta ou .fa)
- Arquivo de texto
- Apresenta apenas:
 - 1ª linha: Cabeçalho
 - Sequência
- Vantagem: Muito leve
- Desvantagem: Pouca informação

```
>Paraponera clavata mitochondrion, complete genome
taactaatatattattttatttaattatcttcccccttaattattttataatatattagtt
attaataaaaaatttaataatcttattatattatattatccttaaaataattatttttaatt
cttctatgtaataatatttaaatagctatattttttttattttatttttttaactttta
aatttaataaaaatacaatattgatttttttaataaagttttatagataaatttatatttt
ataataataaattaatagttattttttatctattaaactaattttataaatatataatttaa
ttaaaaaattaattaaaatttaattaatattaaatatttttttaatttaataaatatat
tgattattattatttaattgattatctaaagcactttatatataaattatttaatatataat
aaataaatttaattatttaatatatttaattgattaataataaataattatttttaatttaaaga
tttaattatatatatatatataaatatatattcaaaaattataattatttaatatattaaa
aaataattatatatatatatataatgtaattaaaaaaaaatagttaaaaatctaaatt
ttcaatttttaattctaaataaatataaattttcaatattatattttatctttttttatta
ttaaaaaatttaaaaaattatctttaatactcagaggcgcttccccctctcttccccaaa
cattaaatataatcatttttaaaatttaattctatttaaatcataaaatcacttactattatt
tagactaacataataaatttaattcattttatttaattaaatatacttattttttatttatata
aataaaaaaagtcagctaaataaataaagcttttaggttcataccctaatacatagataac|
taatctcttattttttcaatgaaatgcctgataaaaaggattattttgatgaaataaatc
atacaaattttaatttttgttttcatttaagatttatacttaatagaatcaaactatttct
taaaatttcaaaaattttcatgctttttacattaaagtatatataaattttttattat
attaaaaatatatatatatattttttttattataagaattttatctttatctttaaatga
ttttttacaatatgatttttttagaaattttaaaattatttatattttattgtataaa
```

Montagem

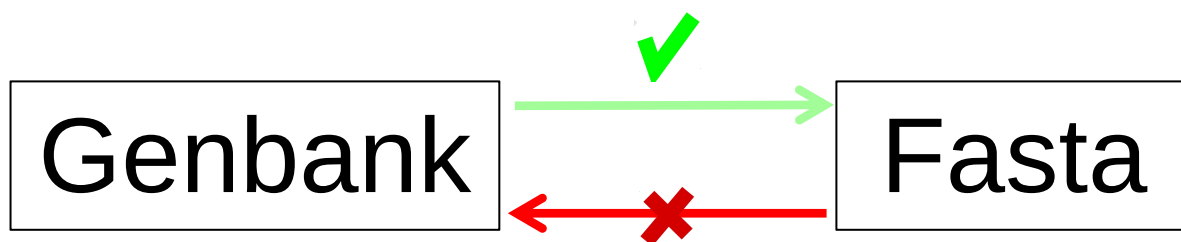
TCTACCGCAAACAACACCGTCGAGGTGACTTACGAGGATACTCTCCCAAGGTTGTACCGACTAACGCAGAGTGAAGTGGTCGTGAGGCGCGGAGAGAGAGGG
GGGGGTGAGGAGTTGGTATGCTTTCCCTACCTGTGCGTCTACCATTGGCAGTGTAGTCCTCTGAGATGCCTCATTGGCGGTACTCTGACCTCGCTTTTCTCTCTC
CCGTTTCGCTGAGCCGTGACTTCTGCTCTGGTTGTGGCGGTAGGCTAGGGGAGGGAGGGGGGAGGGGGGACGGCGGTGTTTCGTGCCTCCCCCTCTTGCTTGGA
GATGCAGGCATGTTTTGCATGTGCTGAGGACAACGATGTACACTGGGCTGCAACCTGGAAGTGTGGGGCAGAGAGAGAGAGGCTCGAGGAATGGAGCATGGAGG
CTCATGGAGAGGTCTTGTTGGTGCGTTTGATGTGGTGGCATCACACACCTCACTCGAAACTACTCGCCAGCGAGCCTCTCCTTTCCCGCATGCGTGTTGGCTCTCT
TCCTCTTCTCTCTTCCGGTCTCGGCTCTTCTTCAGGGCAGCGCCAACCAGCCGCAAAAACAAAGCGAGGGCACAGAGGAGGAGTACTCACGACACGAGTAATGCC
GAAGCAGGTCAATTCATGCAGCAAACATGCCCCGCGAGGAGAAACGCCATGGCGCTGTCCGCCGGAAGTGCAGGTGTGCCTATGGCGAGCGCCCCCTCCCCCTCA
CTGAGCGCGTGCGTGCCGGCACGCCTGCGCACCATCGCCACGCGTCCTTTTTTGGTTTGAGGGCATTGGGCTCTTTCGTCCTCTCAACCTTCACGACATTTGCGC
CTCCGTCTCCTGCCTCGCACACTCCCTCACCTCCTCCCTCCCTCTTCTCTCTTCCCTTCCCTTCCTTCGACGCCGGCGCCACGCGCACACAGGCACAGGGACAGAC
ACACCTATGGACAGCGTGCTCGTGTCGAAACATGTGGCAGACTCTGCGGCAGCGATATCGTCTTCAGCCACATCCCTCGCAGCCTTCCTGGAGTCCAACCTGCAC
GGTGTGGAAGTCTCCGACGGCGGCGCTACACCGCAACTCGAGCCTCGTCGTCGCCGAAACCAACAGCCTCTGCATCTTCGAAAACCTCGTCGCAGAGGAATCACT
CGCTCGTGGTCTTTCCCTCGACAAGACGCAGCGGCTGCTCCTCTTCTACGTGAACCGCCCCTGCGTCGTGGTGGGCGCAACCAGAACCTCTTCCAGGAGGTGGC
GCTGCGGGCGGGCGGCTGCCGACGGCGTTTCCGTGCGTCGCCGTGCCCTCTGGCGGTGGCGCCGTCTTTCATGACGAGGGGAACCTCTGCCTCTGCTTCATCACGC
ACCGCACGCGCTATGCACCGGAGAAGACAATCCAGCTGATCCGCCTTGGGCTCTGTGTGAATTATGCGATCGACCTGCACGGCTGACCACGACAAGGCGGCAC
GACCTCTTCTTGACGGAAAAAAGATCACCGGATCTGCAATGCGGGTGCAGCGCGAAATTGCGTACCACCACTGCACGCTGCTCGTCGATACCCCACTGGCGTC
GCTCGGTGCTACCTTACCCCCGAAGGGGAGTACGTGGCGTTCAAGACGTGTCGGTGGGCTCGGTGCGAAGCCCTGTACCACACTCGCGGAGTCGGTCCACA
TTGCGAGTGGGCAGGGCGCCATGGCCTCACTCAAGAGGAACATGGCGGAGTTTTTTCTAACTGAGGGTGATCGAGTGCTGGAAGCAGCGGCACCGTGGGAGCT
CGACGTGAGGGAGCTGCGGCAGTCCTTCGCCACCGCACGTCAGTCTGCGCGGACACGCCTCTCTTTCCCTTGATGTTGTCGGAGCCGTGCGGGCGGACATGTC
TTTCATCGAGGGCGAGGGCAGGGCGGGCGGCCAGTGGTGACCTCGCCACCCTTGGCGAGGCGGTTCAAAAGGCTGCGTCGAAGGACTGGGCCTACGCGATGCC
AGCCTTACATCCACGGTGCTCCTCAGCAGCGGTGAGCTCCAGCGGCGCCTACAAGCGCTTTCGTTTGGCCGGATGTAGTGCGGCTATCCTCTCTCGCGGAGGA
GCAGTTGCTGGCGGCACTGCAGCAATGTGTCTTTCGGGACTTGGTCGGCTGCGGGGAAGTCGTGGCAGAGACCGAAGTGGGGTTGCACCTGCTTACCACAGTC

Anotação

TCTACCGCAAACAACACCGT**CGAGGTGACTTACGAGGA** **Regiões reguladoras** **CTAACGCAGAG**TGAAGTGGTCGTGAGGCGCGGAGAGAGAGGG
GGGGGTGAGGAGTTGGTATGCTTTCCTACCTGTGCGTCTACCATTGGCAGTGTAGTCTCTGAGATGCCTCATTGGCGGTACTCTGACCTCGCTTTTCCTCTCTC
CCGTTTCGCTGAGCCGTGACTTCTGCTCTGTTGTGGCGGTAGGCTAGGGGAGGGAGGGGGAGGGGGGACGGCGGTGTTCTGTCCTCCCC**CTTTGCTTGTGGA**
GA **CDS** **GAGGACAACGATGTACACTGGGCTGCAACCTGGAAGTGTGGGGCAGAGAGAGAGGGCTCGAGGAATGGAGCATGGAGG**
CTCATGGAGAGGTCTTGTGGTGCCTTTGATGTGGTGGCATCACACACCTCACTCGAAAATACTCGCCAGCGAGCCTCTCCTTCCCGCATGCGTGTGGCTCTCT
TCCTCTTCTCTTCCGGTCTCGGCTCTTCTTCAGGGCAGCGCCAACCAGCCGCAAAAACAAAGCGAGGCAC**AGAGGAGGAGTACTCACGACACGAGTAATGCC**
GAAGCAGGTCAATTCATGCAGCAAACATGCCCGGCAGGAGAAACGCCATGGCGCTGTCCGCCGGAAGTGCAGGTGTGCCTATGGCGAGCGCCCCCTCCCCCTCA
CTGAGCGCGTGCGTGCCGGCACGCCTGCGCACCATCGCCACGCGTCCTTTTTTGGTTTGCAGGCGATTGGGCTCTTTCGTCCTCTCAACCTTCACGACATTTGCGC
CTCCGTCTCCTGCCTCGCACACTCCCTCACCT**CTCCCTCCC** **rRNA** **TCGACGCCGGCGCCACGCGCACACAGGCACAGGGACAGAC**
ACACCTATGGACAGCGTGCTCGTGTGAAACATGTGGCAGACTCTGCGGCAGCGATATCGTCTTCAGCCACATCCCTCGCAGCCTTCCTGGAGTCCAACCTGCAC
GGTGTGGAAGTCTCCGACGGCGGCGCTACACCGCAACTCGAGCCTCGTCGTCGCCGAAACCAACAGCCTCTGCATCTTCGAAAACCTCGTCGCAGAGGAATCACT
CGCTCGTGGTCTTTCCTCGACAAGACGCAGCGGCTGCTCCTCTTCTACGTGAACCGCCCCTGCGTCGTGGTGGGCGCAACCAGAACCTCTTCCAGGAGGTGGC
GCTGCGGCGGGCGGCTGCCGACGGCGTTTCCGTGCTCGCCGTGCCTCTGGCGGTGGCGCCGTCTTTCATGACGAGGGGAACCTCTGCCTC**TGCTTCATCACGC**
tRNA **GAAGACAATCCAGCTGATCCGCCTTGGGCTCTGTGTGAATTATGCGATCGACCCTGCACGGCTGACCACGACAAGGCGGCAC**
GACCTCTTCTTGACGGAAAAAAGATCACCGGATCTGCAATGCGGGTGCAGCGCGAAATTGCGTACCACCACTGCACGCTGCTCGTCGATACCCCACTGGCGTC
GCTCGGTGCTACCTTACCCCGAAGGGGAGTACGTGGCGTTCAAGACGTCGTCGGTGGGCTCGGTGCGAAG**CCC** **Padrões específicos** **TCCACA**
TTGCGAGTGGGCAGGGCGCCATGGCCTCACTCAAGAGGAACATGGCGGAGTTTTTTCTAACTGAGGGTGATCGAGTGCTGGAAGCAGCGGCACCGTGGGAGCT
CGACGTGAGGGAGCTGCGGCAGTCTTCGCCACCGCACGTCACTGCTGCGCGGACACGCCTCTTTTTCCCTTGATGTTGTCGGAGCCGTCGCGGCGGACATGTC
TTTCATCGAGGGCGAGGGCAGGCGGGCGGCCAGTGGTGACCTCGCCACCTTGGCGAGGCGGTTACAAGGCTGCGTCGAAGGACTGGGCCTACGCGATGCC
AGCCTTCACATCCACGGTGCTCCTCAGCAGCGGTGAGCTCAGCGGCGCCTACAAG **Elementos repetitivos** **GTGCGGCTATCCTCTCTCGCGGAGGA**
GCAGTTGCTGGCGGCACTGCAGCAATGTGTCTTTGCGGACTTGGTCGGCTGCGGGGAAGTCGTGGCAGAGACCGAAGTGGGGTTGCACCTGCTTACCACAGTC

Formato genbank

- Apresenta:
 - Metadados (header)
 - Anotação (FeatureTable)
 - Sequencia
- Muito mais informação que o fasta
 - + pesado também
 - Conversão: via de mão única



Header

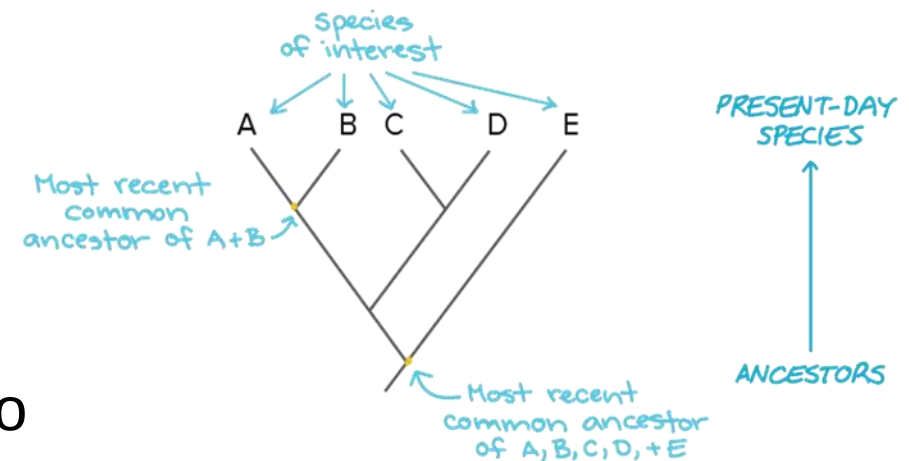
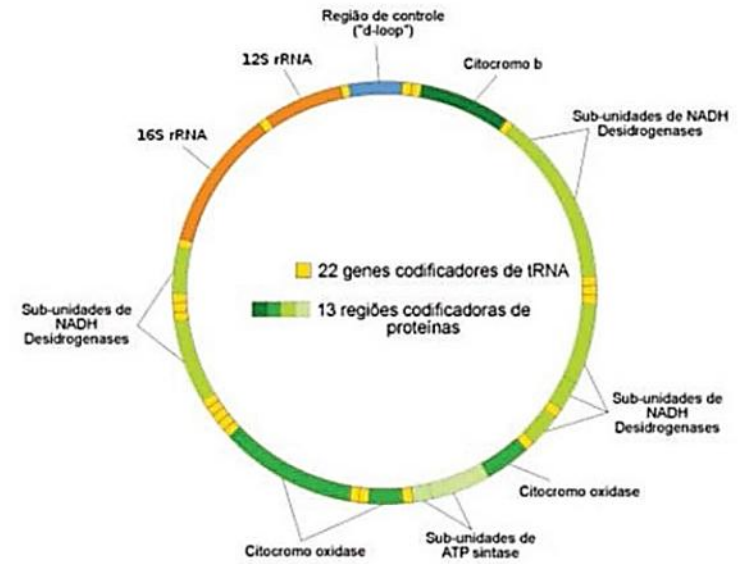
Feature Table

Sequence

```
LOCUS      KU985485                658 bp    DNA     linear     INV 25-OCT-2016
DEFINITION Pseudomyrmex gracilis voucher BCCISEC0010109 cytochrome oxidase
            subunit 1 (COI) gene, partial cds; mitochondrial.
ACCESSION  KU985485
VERSION    KU985485.1
KEYWORDS   BARCODE.
SOURCE     mitochondrion Pseudomyrmex gracilis
  ORGANISM Pseudomyrmex gracilis
            Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;
            Pterygota; Neoptera; Holometabola; Hymenoptera; Apocrita; Aculeata;
            Vespoidea; Formicidae; Pseudomyrmecinae; Pseudomyrmex.
REFERENCE  1 (bases 1 to 658)
AUTHORS    Dominguez,D.F., Bustamante,M., Albuja,R.A., Castro,A., Lattke,J.E.
            and Donoso,D.A.
TITLE      C digos de barras (COI barcodes) para hormigas (Hymenoptera:
            Formicidae) de los bosques secos del sur del Ecuador
JOURNAL     Unpublished
REFERENCE  2 (bases 1 to 658)
AUTHORS    Dominguez,D.F., Bustamante,M., Albuja,R.A., Castro,A., Lattke,J.E.
            and Donoso,D.A.
TITLE      Direct Submission
JOURNAL     Submitted (29-MAR-2016) Ciencias Agropecuarias, Universidad de
            Cuenca, Av. 12 de Abril s/n, Cuenca, Azuay, Ecuador
FEATURES   Location/Qualifiers
     source          1..658
                     /organism="Pseudomyrmex gracilis"
                     /organelle="mitochondrion"
                     /mol_type="genomic DNA"
                     /specimen_voucher="BCCISEC0010109"
                     /db_xref="BOLD:DRYLO015-15.COI-5P"
                     /db_xref="taxon:219809"
                     /country="Ecuador: Loja, Macara, Reserva Laipuna"
                     /lat_lon="4.21 S 79.88 W"
                     /collection_date="01-Oct-2014"
                     /collected_by="M. Velez, C. Gomez, M. Tuza, JE Lattke. G.
                     Piedra"
                     /identified_by="John E. Lattke"
                     /PCR_primers="fwd_seq: attcaaccaatcataaagatattgg, rev_seq:
                     taaacttctggatgtccaaaaatca"
     gene            <1..>658
                     /gene="COI"
     CDS             <1..>658
                     /gene="COI"
                     /codon_start=2
                     /transl_table=5
                     /product="cytochrome oxidase subunit 1"
                     /protein_id="AOX21864.1"
                     /translation="ILYFMFAMWAGMIGSSMSMIIRIELGSCGSIINNDQLYNSIVTG
                     HAFIMIFFMVPFMIGGFNGLVPLMIGSPDMAYPRMNMNSFWLPPSIMLLTSSFI
                     NSGAGTGWTVYPPLSSSIHFHSGASVDLAIFSLHIAGISSINGAINFISTIINMTHKNF
                     SMDKTPLMVWSILITAVLLLSLPVLGAIITMLLTDRNLNTSFFDPAGGGDPILYQHL
                     F"
ORIGIN
1 aattctatac tttatattg ctatatgagc aggtataatc ggatcatcaa taagaataat
61 tattcgaatt gatttaggat catcggaatc tattattaat aacgaccaac tatacaactc
121 tatcgtaaca ggacatgcat ttattataat ttcttttata gttataccat ttataatcgg
181 aggatttggat aactttctag ttccactaat aattggatca ccagatatag cttaccctcg
241 tataaataac ataagatttt gactcttacc cccatcaatt atacttctca ctttaagaag
301 atttattaat tcaggagctg gaactggctg aacagtatat cctctcttat cttcaagaat
361 ttttcattagg ggagcctcag tagatttagc aattttctct ctacatatg caggaatttc
421 atcaatcata ggagctatta acttcattct tacaattatt aataataacc ataaaaattt
481 ttcaatagac aaaactcctt taatagtctg atctatttta attcacgacg ttctactact
541 tctctccctt cctgttctag cggagccaat tacaatacta ttaacagatc gaaatcttaa
601 tacttcattt ttgaccagag caggtggagg ggaccaaat ctttacaac acttattc
//
```


Mestrado

- Montagem de genomas mitocondriais
 - Menor genoma da célula (≈ 16 kbp)
 - Excelente treino
 - Diversos estudos:
 - Filogenética
 - Filogeografia
 - Genética populacional
 - Conservação
- Dados públicos
 - Várias espécies sem mitogenoma descrito



Formigas (Hymenoptera: Formicidae)

- Mais de 13000 *spp.*
- Grande quantidade de informação para o clado:
 - \approx 3.91 Terabytes
- 15 mitogenomas completos
- Todas as subfamílias
 - Exceto Martialinae
- Pseudomyrmecinae
 - Sem mitogenomas
 - Mutualismo com plantas
 - Abundância de dados

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.

- o [Agroecomyrmecinae](#) 1
- o [Amblyoponinae](#) 3
- o [Aneuretinae](#) 2
- o [Apomyrminae](#) 1
- o [Dolichoderinae](#) 201
- o [Dorylinae](#) 388
- o [Ectatomminae](#) 4
- o [Formicinae](#) 863
- o [Heteroponerinae](#) 2
- o [Leptanillinae](#) 1
- o [Martialinae](#)
- o [Myrmeciinae](#) 3
- o [Myrmicinae](#) 1,580
- o [Nothomyrmeciinae](#) 2
- o [Paraponerinae](#) 1
- o [Ponerinae](#) 269
- o [Proceratiinae](#) 2
- o [Pseudomyrmecinae](#) 45
- o [unclassified Formicidae](#)

Mestrado

- 14 mitogenomas montados, anotados e submetidos ao GenBank
- Análises adicionais:
 - Ordem gênica (Sintenia)
 - Genômica comparativa
 - Filogenômica

- Sem gastos com sequenciamento
- Apenas softwares gratuitos foram utilizados:
 - **Montagem:** NOVOPlasty, MIRA, MITObim
 - **Anotação:** MITOS Web Server, Artemis
 - **Análises:** MEGA7, BRIG, Phylomito



Submitted 7 November 2018
Accepted 10 December 2018
Published 24 January 2019

Corresponding authors
Gabriel A. Vieira,
gabriel.vieira@bioqmed.ufrj.br,
fprosdoci@gmail.com
Francisco Prosdoci,
prosdoci@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 Prosdoci

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

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

Gabriel A. Vieira and Francisco Prosdoci

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

NO BUDGET,
SCIENCE



NEM TODA CIÊNCIA
PRECISA DE BANCADA

MITOFree

- Automação do pipeline usado no mestrado
 - Gerar mitogenomas a partir de dados do SRA
 - Modificações - Diminuição do consumo de RAM
 - Em desenvolvimento:
 - Download de datasets SRA ✓
 - Montagem de mitogenomas ✓
 - Anotação ✗
 - Análise filogenômica (Phylomito) ✗
- Montagem:
 - Manual - 40 datasets:
 - 2 semanas; Trabalho intenso
 - Script - 58 datasets
 - Um fds; Na praia...
- Aprendam a programar!!!

##General usage:

```
$> python3 /path/to/mitofree.py dataset_list.txt
```

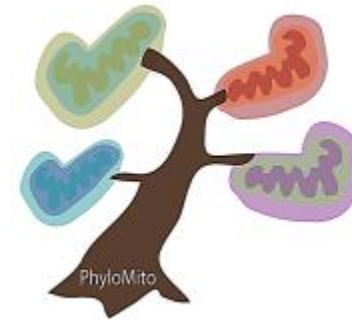
##Example of dataset_list.txt

##Each line corresponds to a different assembly

##Three tab-separated columns:

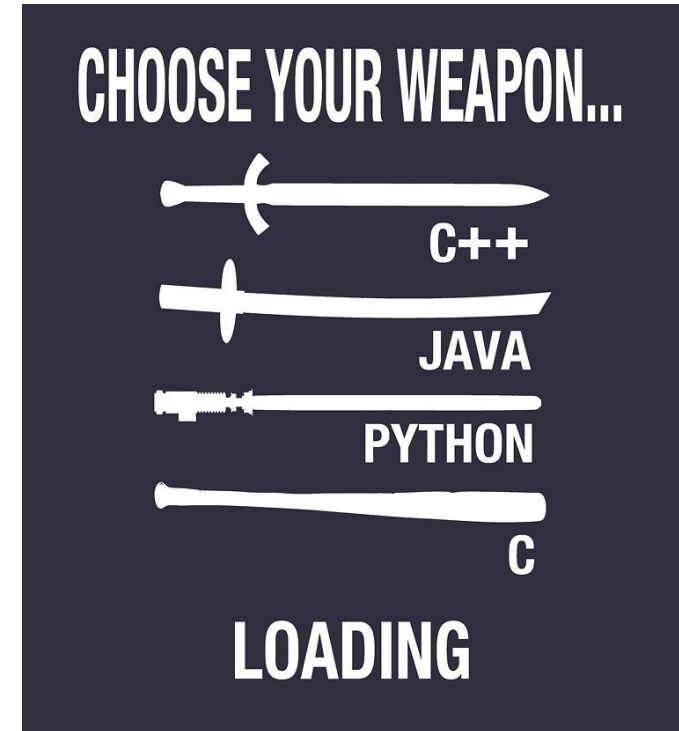
##SRA_ACCESSION SPECIES_NAME SEED_GENBANK ACCESSION

ERR1306022	Species1	MK297287
ERR7295163	Species2	MK297241
ERR1306034	Species3	MK291745
SRR4409513	Species4	MK291678



Onde aprender a programar?

- Muitos cursos e materiais pela internet afora
 - Qualquer linguagem
- Udemy, Datacamp, Youtube, Khan Academy...
- Coursera (<https://www.coursera.org/>):
 - Fazer o curso como ouvinte - Grátis
 - Programação, história, filosofia...
 - App - Baixar vídeos e assistir offline
- **Equilíbrio entre teoria e prática**



Introdução à ciência da programação usando python (Fabio Kon - USP) - PT

The screenshot shows the Coursera interface for a Portuguese course. At the top, the Coursera logo is on the left, followed by a blue 'Explorar' button with a dropdown arrow, and a search bar containing the text 'O que você deseja aprender?'. Below this is a teal banner with the breadcrumb 'Navegar > Ciência da Computação > Desenvolvimento de Software'. The main title 'Introdução à Ciência da Computação com Python Parte 1' is displayed in large white font. Below the title, there are five yellow stars, the rating '4.9', and the text '2,331 classificações • 667 avaliações'. A white button labeled 'Ir para o curso' is on the left, and the text 'Já inscrito' and 'Auxílio financeiro disponível' is on the right. At the bottom left, it says '68.573 já inscritos!'.

Ir para o curso

Já inscrito
Auxílio financeiro disponível

68.573 já inscritos!

<https://www.coursera.org/learn/ciencia-computacao-python-conceitos?>

Python For Everybody (Charles Severance - University of Michigan) - EN

The screenshot shows the Coursera interface for an English course. At the top, the Coursera logo is on the left, followed by a blue 'Explorar' button with a dropdown arrow, and a search bar containing the text 'O que você deseja aprender?'. Below this is a dark blue banner with the breadcrumb 'Navegar > Ciência da Computação > Desenvolvimento de Software'. The main title 'Programa de cursos integrados Python para todos' is displayed in large white font. Below the title, there is a subtitle 'Learn to Program and Analyze Data with Python. Develop programs to gather, clean, analyze, and visualize data.' A yellow button labeled 'Cadastre-se gratuitamente' is on the left, with 'Inicia em Jul 25' below it. On the right, the text 'Teste gratuitamente: Inscreva-se para iniciar seu teste gratuito de acesso completo por 7 dias.' is shown, with 'Auxílio financeiro disponível' below it. At the bottom left, it says '229.655 já inscritos!'.

Cadastre-se gratuitamente
Inicia em Jul 25

Teste gratuitamente:
Inscreva-se para iniciar
seu teste gratuito de
acesso completo por 7
dias.

Auxílio financeiro disponível

229.655 já inscritos!

<https://www.coursera.org/specializations/python>

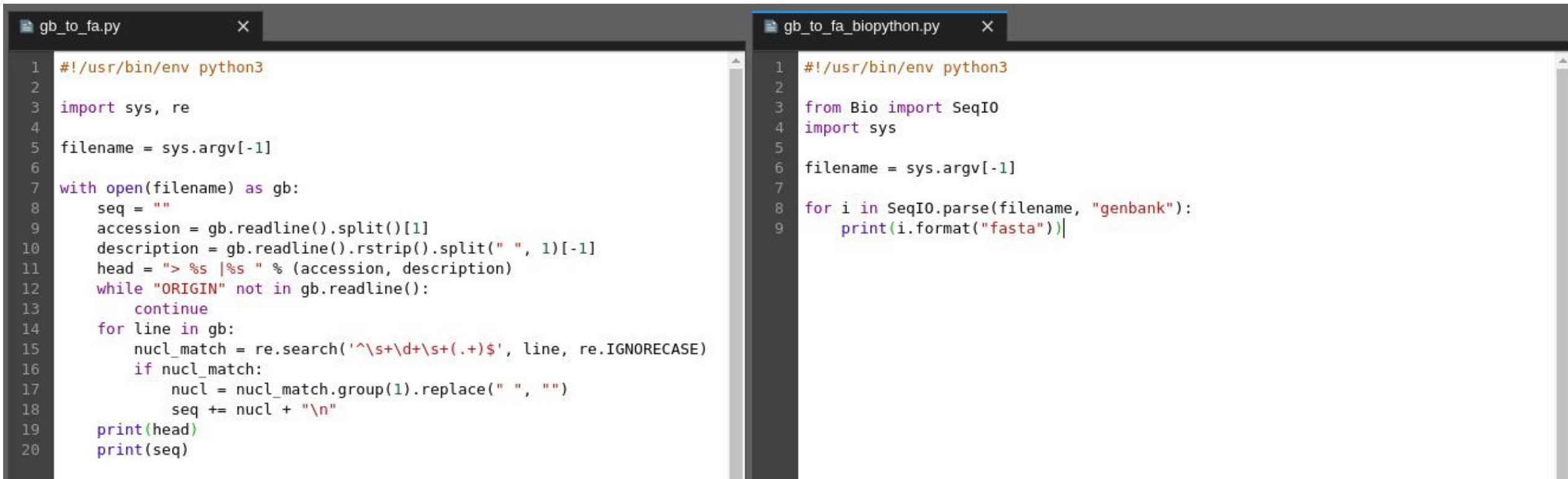
Programação em Bioinformática

- Qual linguagem?
 - O importante é saber programar em alguma
 - Populares entre bioinformatas:
 - Python, Perl, R - Diversos pacotes/módulos para bioinfo.
- No meu caso: Python
 - Muito recomendada para iniciantes
 - Simples
 - General purpose
 - Comunidade muito ativa
 - Biopython - manipulação de arquivos



`pythonic.love()`

Conversão Genbank para Fasta



```
gb_to_fa.py
1  #!/usr/bin/env python3
2
3  import sys, re
4
5  filename = sys.argv[-1]
6
7  with open(filename) as gb:
8      seq = ""
9      accession = gb.readline().split()[1]
10     description = gb.readline().rstrip().split(" ", 1)[-1]
11     head = "> %s |%s " % (accession, description)
12     while "ORIGIN" not in gb.readline():
13         continue
14     for line in gb:
15         nucl_match = re.search('^\\s+\\d+\\s+(\\.+)$', line, re.IGNORECASE)
16         if nucl_match:
17             nucl = nucl_match.group(1).replace(" ", "")
18             seq += nucl + "\\n"
19     print(head)
20     print(seq)
```

```
gb_to_fa_biopython.py
1  #!/usr/bin/env python3
2
3  from Bio import SeqIO
4  import sys
5
6  filename = sys.argv[-1]
7
8  for i in SeqIO.parse(filename, "genbank"):
9      print(i.format("fasta"))
```

Python:

- 20 linhas (poderia ser menos)
- Difícil de entender
- Demorado para escrever

Biopython:

- 9 linhas
- Mais legível
- Rápido e fácil de escrever

Python - Anaconda

- Distribuição Python
 - Python, programas, módulos
 - Data science
- Conda
 - Gerenciador de pacotes/programas
 - Software hospedados por canais
 - Bioconda - Canal de softw. de bionformática
 - tRNAscan-SE
 - Instalação manual: difícil
 - Usando conda: **\$> conda install -c bioconda trnascan-se**



Python - Jupyter

- Jupyter
 - Notebook (**Web** App - Documento de texto)
 - Lab (**Web** IDE)
 - Hub (Plataforma Multiusuário **Web**)
 - Roda direto no navegador
- Instalação: Anaconda ou standalone
- Suporte multilinguagem (Python, R, Ruby...)
- Documento dividido em células
- Mesclar texto **formatado** e código **executável**
 - Output visualizável - Gráficos
- Reproduzir resultados dentro do Notebook
- Ferramenta produtividade - Relatórios
- Potencial educacional - Tutoriais





Simple spectral analysis

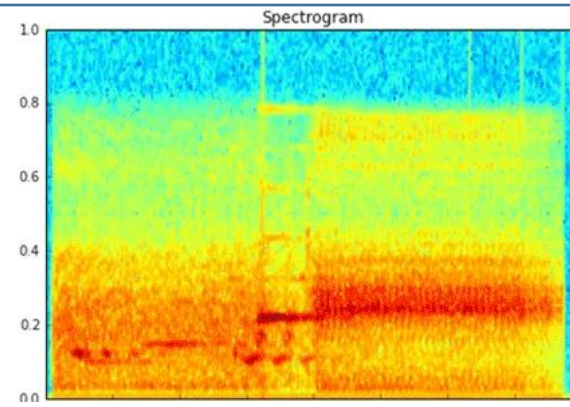
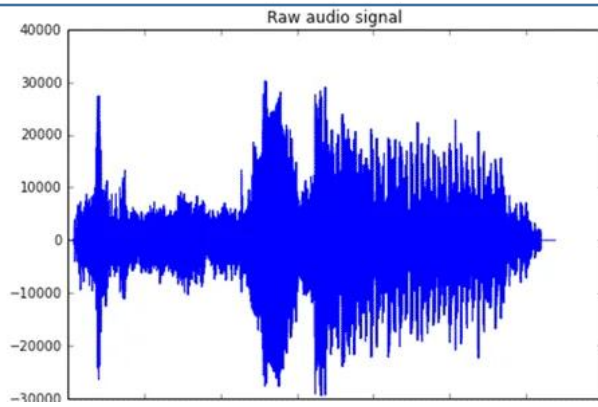
An illustration of the [Discrete Fourier Transform](#)

$$X_k = \sum_{n=0}^{N-1} x_n \exp \frac{-2\pi i}{N} kn \quad k = 0, \dots, N-1$$

```
In [2]: from scipy.io import wavfile
rate, x = wavfile.read('test_mono.wav')
```

And we can easily view it's spectral structure using matplotlib's builtin specgram routine:

```
In [5]: fig, (ax1, ax2) = plt.subplots(1,2,figsize(16,5))
ax1.plot(x); ax1.set_title('Raw audio signal')
ax2.specgram(x); ax2.set_title('Spectrogram');
```



Célula de texto

Célula de código

Output

Reproducible Research is more than Publishing Research Artefacts: A Systematic Analysis of Jupyter Notebooks from Research Articles

Max Schröder^{1,2}, Frank Krüger¹, and Sascha Spors¹

¹ *Institute of Communications Engineering, University of Rostock*

² *University Library, University of Rostock*

E-Mail: {max.schroeder, frank.krueger, sascha.spors}@uni-rostock.de

36 papers do Pubmed

- Publicaram os notebooks associados ao seu código/análises em repositórios públicos
- Maior parte hospedada no GitHub

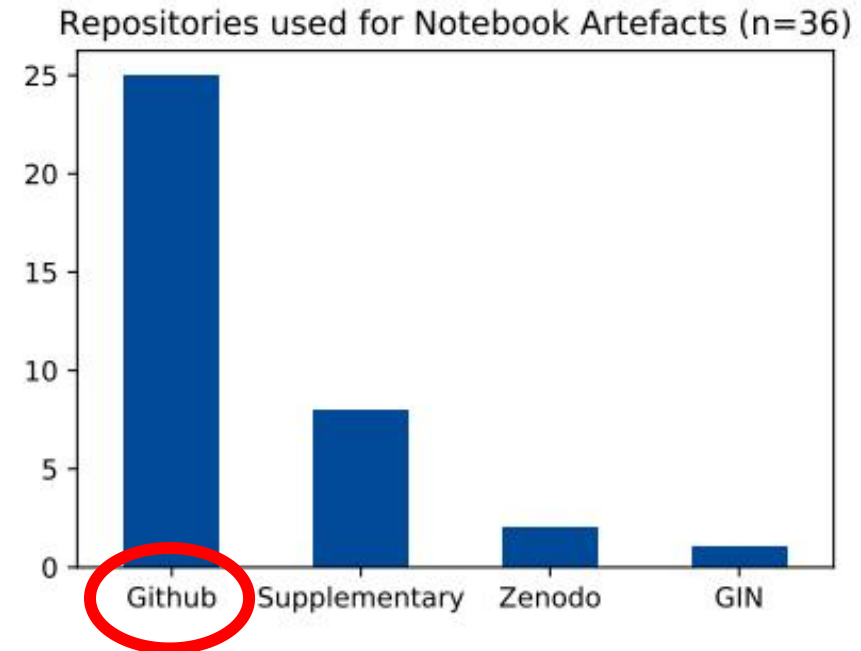


Figure 3: Which repositories are used to publish source code artefacts i.e., Jupyter notebooks?

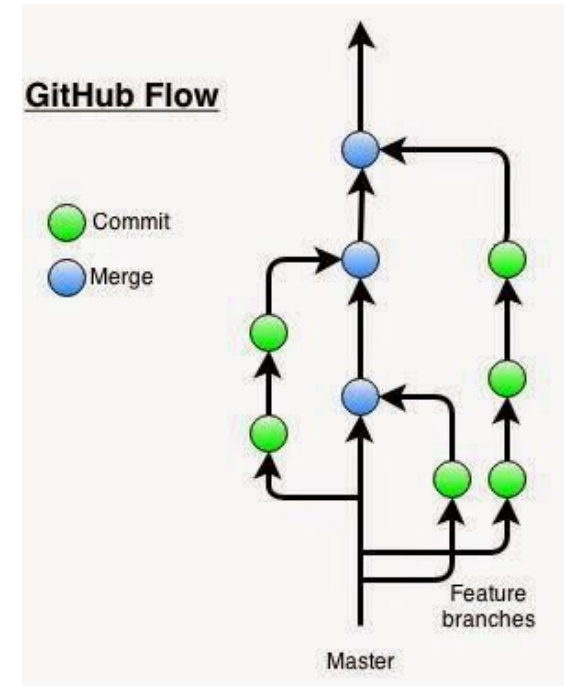
Git e GitHub

- Git:
 - Ferramenta de controle de versão
 - Programa rodado localmente no seu PC
 - Pastas - repositórios
 - Durante desenvolvimento - snapshots (commits)
 - Commits podem ficar no seu PC, ou...
- GitHub:
 - Serviço de hospedagem de repositórios git
 - Plataforma de desenvolvimento de software



Git e GitHub

- Desenvolvimento paralelo de features (ramos):
 - Todo repositório possui um ramo principal (master)
 - Programa estável - nova feature : novo ramo
 - Master - commit em destaque na página do github
 - Após implementação - Fundir o ramo secundário no master
 - Múltiplas pessoas - vários ramos - agilidade
- Repositórios públicos - Projetos open-source
 - Encontrar/colaborar com novos programas/algoritmos
- GitHub pages - Site p/ seu projeto - ↑ Visibilidade
- Educacional - Tutoriais, cursos...



Git e GitHub

GitHub (PT)

The screenshot shows the GitHub repository page for 'Abduzidos / Aprenda-Git'. The repository has 65 commits, 2 branches, 0 releases, and 9 contributors. The main branch is 'master'. The repository is described as 'Uma introdução rápida e prática a aos conceitos básicos do Git'. The commit history shows a merge pull request #24 from Peedrohj/pedro, and a list of files including Alunos, Arts, Scripts, assets, README.md, and via-classroom.md.

File	Commit Message	Time Ago
Alunos	Merge pull request #24 from Peedrohj/pedro	10 months ago
Arts	Rename art	10 months ago
Scripts	Merge pull request #24 from Peedrohj/pedro	10 months ago
assets	Add files via upload	11 months ago
README.md	Merge branch 'master' into master	11 months ago
via-classroom.md	Create via-classroom.md	11 months ago

<https://github.com/Abduzidos/Aprenda-Git>

Software Carpentry

The screenshot shows the Software Carpentry website. The main heading is 'Version Control with Git'. The text describes the challenges of working on plans at the same time and the importance of version control. It lists several benefits of version control, including the ability to go back in time, know who made changes, and collaborate with others. The prerequisites section states that Git is used from the Unix Shell, and some previous experience with the shell is expected, but not mandatory.

Version Control with Git

Wolfman and Dracula have been hired by Universal Missions (a space services spinoff from Euphoric State University) to investigate if it is possible to send their next planetary lander to Mars. They want to be able to work on the plans at the same time, but they have run into problems doing this in the past. If they take turns, each one will spend a lot of time waiting for the other to finish, but if they work on their own copies and email changes back and forth things will be lost, overwritten, or duplicated.

A colleague suggests using [version control](#) to manage their work. Version control is better than mailing files back and forth:

- Nothing that is committed to version control is ever lost, unless you work really, really hard at it. Since all old versions of files are saved, it's always possible to go back in time to see exactly who wrote what on a particular day, or what version of a program was used to generate a particular set of results.
- As we have this record of who made what changes when, we know who to ask if we have questions later on, and, if needed, revert to a previous version, much like the "undo" feature in an editor.
- When several people collaborate in the same project, it's possible to accidentally overlook or overwrite someone's changes. The version control system automatically notifies users whenever there's a conflict between one person's work and another's.

Teams are not the only ones to benefit from version control: lone researchers can benefit immensely. Keeping a record of what was changed, when, and why is extremely useful for all researchers if they ever need to come back to the project later on (e.g., a year later, when memory has faded).

Version control is the lab notebook of the digital world: it's what professionals use to keep track of what they've done and to collaborate with other people. Every large software development project relies on it, and most programmers use it for their small jobs as well. And it isn't just for software: books, papers, small data sets, and anything that changes over time or needs to be shared can and should be stored in a version control system.

Prerequisites

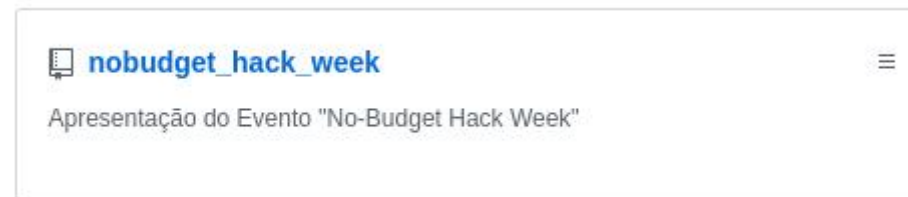
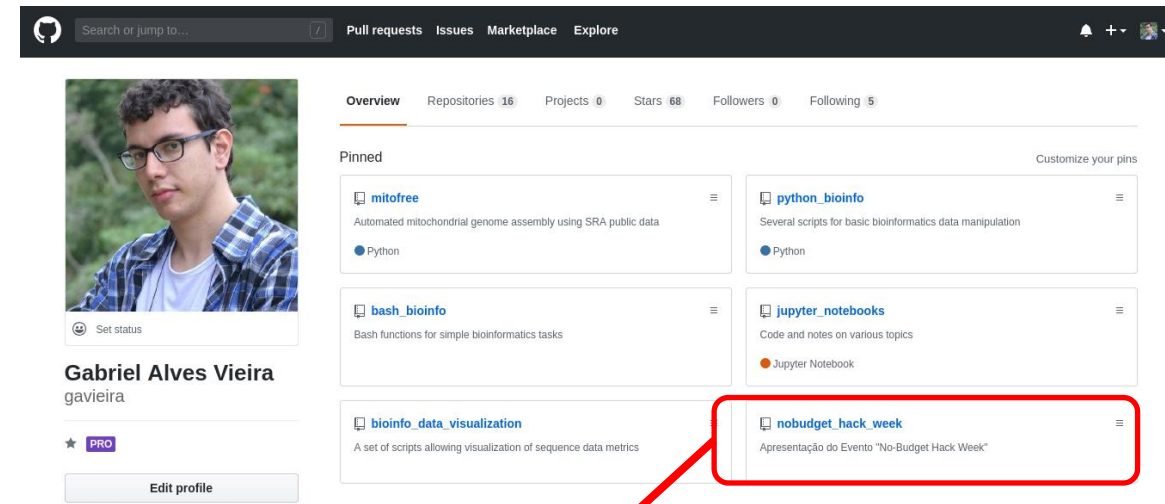
In this lesson we use Git from the Unix Shell. Some previous experience with the shell is expected, *but isn't mandatory*.

<http://swcarpentry.github.io/git-novice/>

GitHub page

Git e GitHub

- MitoFree, scripts, **esta apresentação**
 - Dêem uma mexida, olhem os diferentes commits
 - Evolução da apresentação/programas
- Portfolio
 - Importante para bioinformatas e cientistas de dados
 - Seu trabalho público - habilidades
 - Oportunidades interessantes - colaborações, emprego...
- Fácil de aprender
- Usem o GitHub!!!



Enquanto converte cafeína em código...

- Algum bug bizarro apareceu?
- Não faz idéia de como executar algum passo do algoritmo?
- O Stack Overflow é seu pastor, e nada lhe faltará:
 - Dúvidas sobre as mais diversas linguagens
- Para dúvidas mais específicas de Bioinformática:
 - Biostars, ResearchGate, Reddit
- Be polite, be precise



Fontes de trabalhos no-budget em bioinfo

1. Uso de dados públicos

- Não gasta dinheiro para obter sequência
- Gasto com infraestrutura

2. Software Development

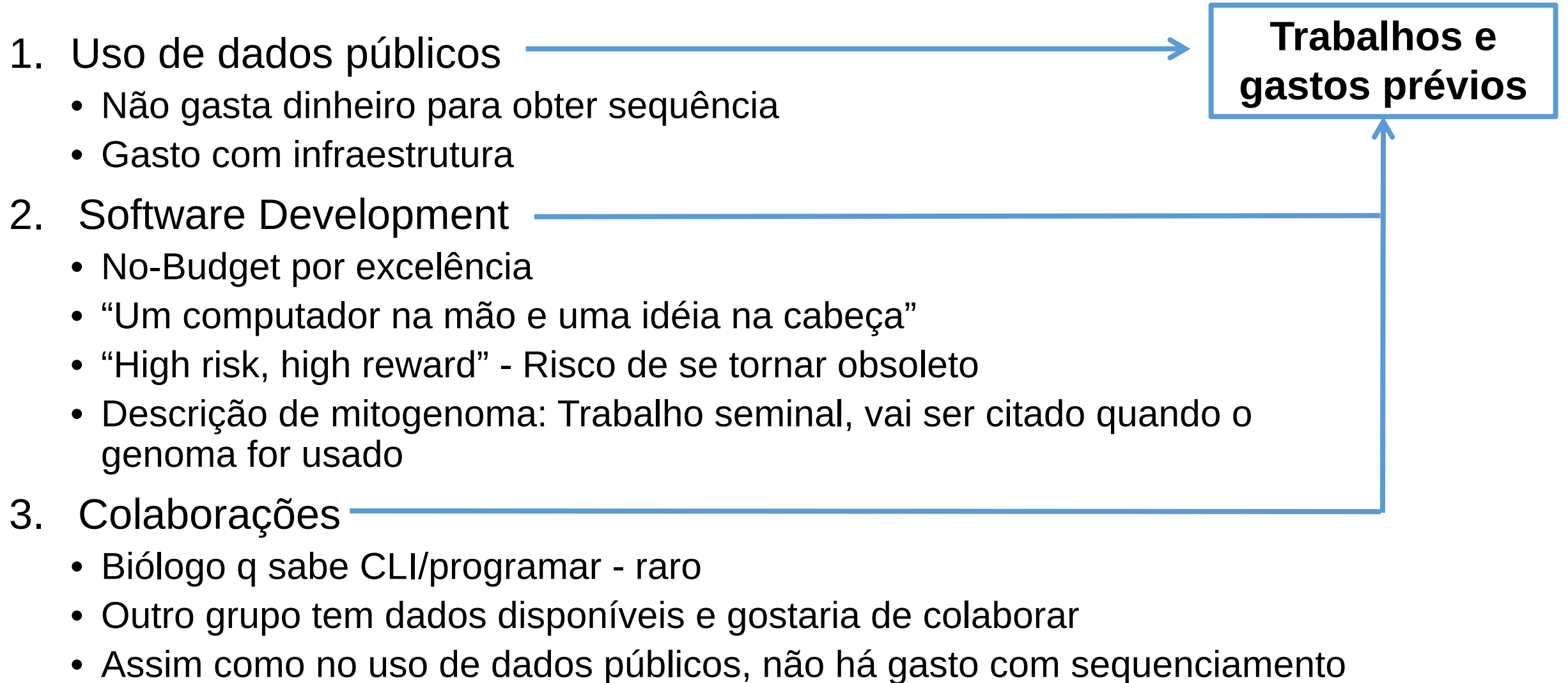
- No-Budget por excelência
- “Um computador na mão e uma idéia na cabeça”
- “High risk, high reward” - Risco de se tornar obsoleto
- Descrição de mitogenoma: Trabalho seminal, vai ser citado quando o genoma for usado

3. Colaborações

- Biólogo q sabe CLI/programar - raro
- Outro grupo tem dados disponíveis e gostaria de colaborar
- Assim como no uso de dados públicos, não há gasto com sequenciamento



Fontes de trabalhos no-budget em bioinfo



Sobre os ombros de gigantes

- Bioinformática:
 - Potencial enorme para trabalhos No-Budget
 - Pautado no trabalho (dados, programas, algoritmos...) de outrem
- Era da informação:
 - Aspectos positivos:
 - Oportunidades para aprender e criar
 - Fruto de trabalho duro de outras pessoas
 - Aspectos negativos:
 - “Pós-verdade” - Palavra do Ano 2016 pelo dicionário Oxford
 - Associada à crise do financiamento científico no Brasil (?)



“Descendo da torre de marfim”

- Abraçar esse legado:
 - Nos (re)aproveitarmos do que é positivo:
 - Aprender novas habilidades, usar o que já existe para gerar novos conhecimentos
 - Combatemos os pontos negativos:
 - Desinformação, desvalorização da ciência
 - Extensão
- Ciência no Brasil sempre foi um ato de resistência



OBRIGADO!

Contato: gabrieldeusdeth@gmail.com



100MITO



Ilustração por Camille Prado

