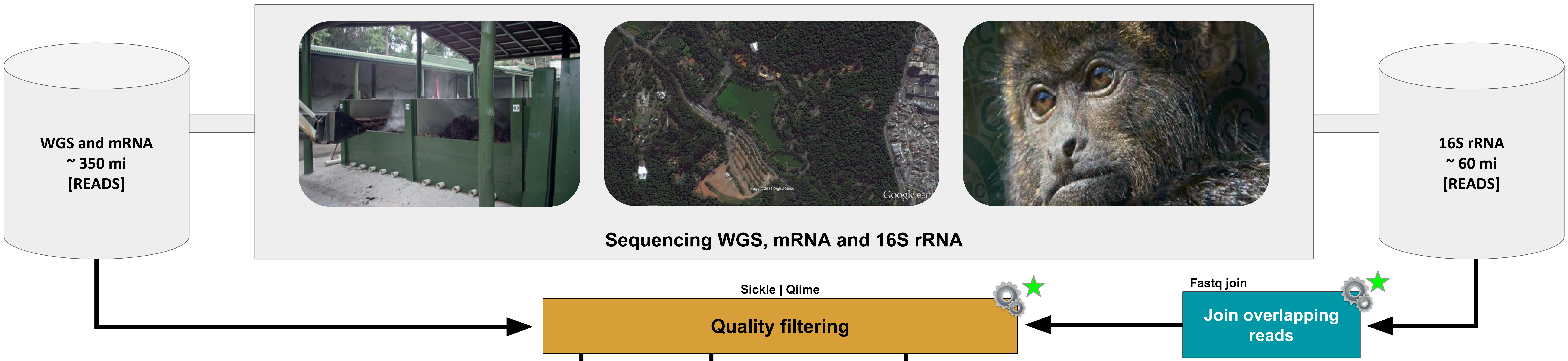
Computational methods for metagenomic data processing in the Metazoo Project

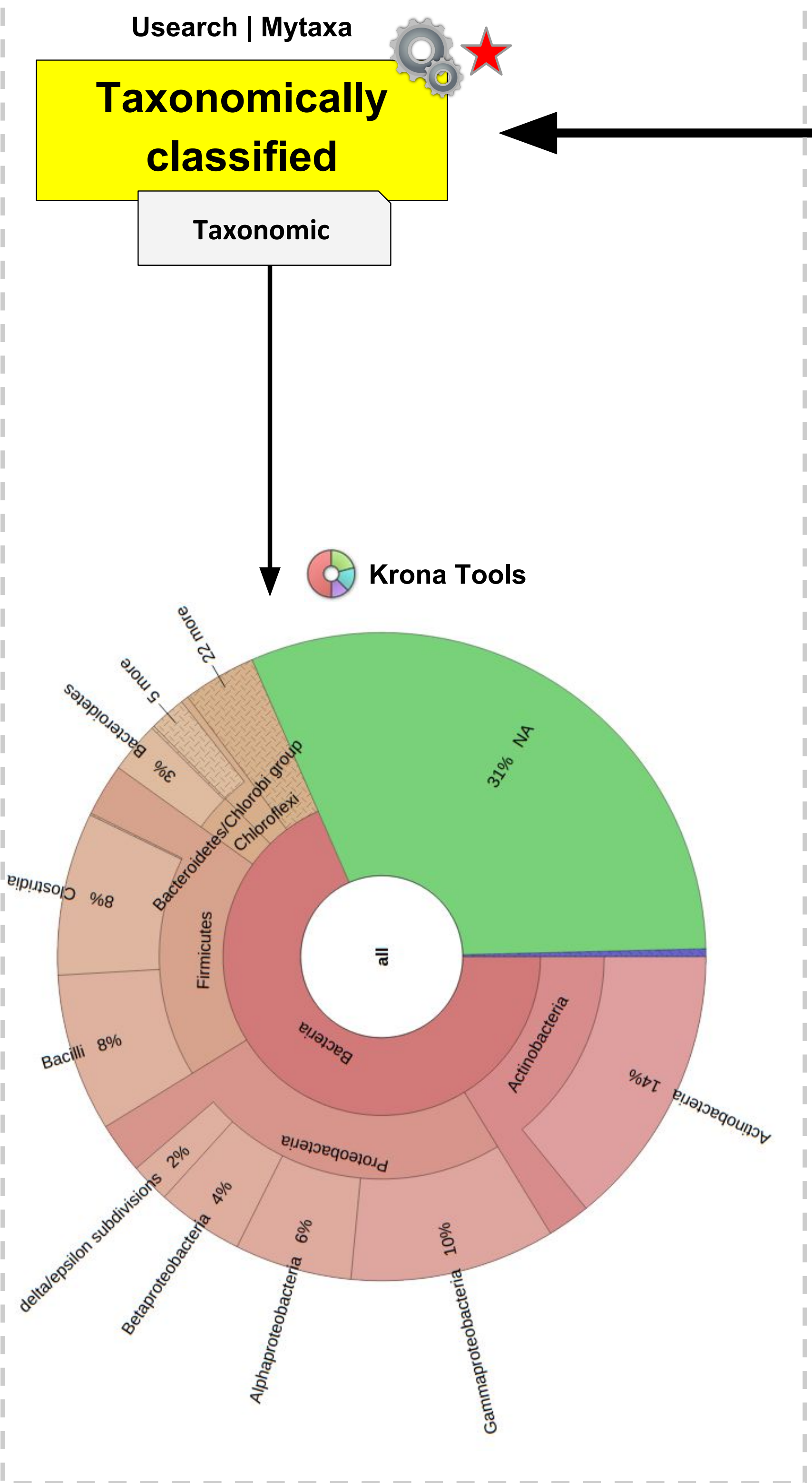
GMM Silva[1,2], FP Lima[1,2,3], AM Thomas[1,2], LN Lemos[1,2], DE Amgarten[1,2], D Barbosa[1,2], C Morais[1], LP Antunes[1], AM da Silva[1,2], JC Setubal[1,2]

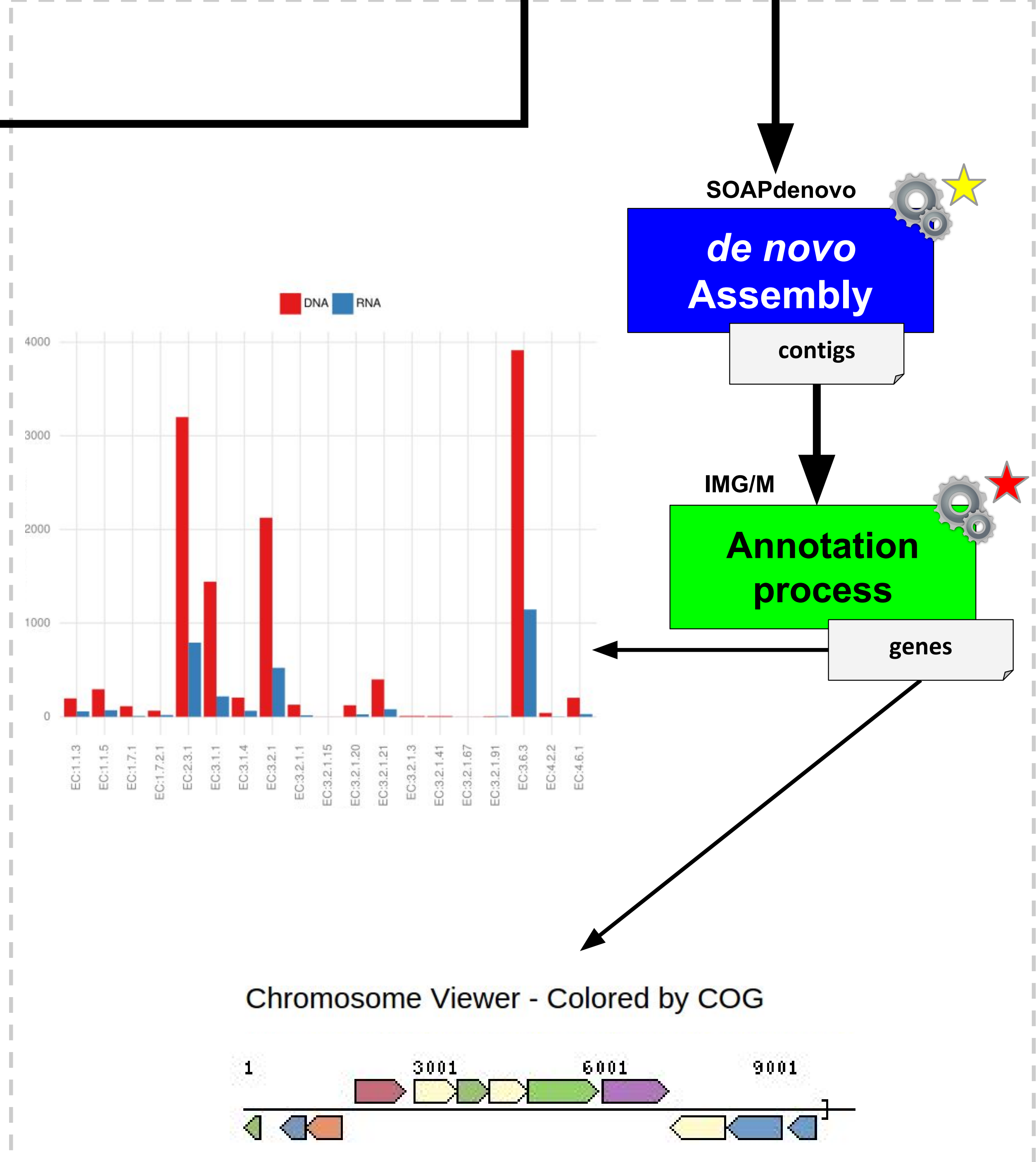
website: http://lbi.iq.usp.br/metazoo/ | e-mail: gianlucamajor@gmail.com, setubal@iq.usp.br
1 - Departament of BioChemistry – Chemistry Institute – USP, SP, Brazil
2 - Inter-institutional Grad Program on Bioinformatics – USP, SP, Brazil
3 - Federal Institute of Alagoas, AL, Brazil

Introduction

The Metazoo project aims to study the **microbial communities** in three different environments of São Paulo's Zoo Park: a **composting process**, the **Sao Francisco Lake**, and **feces of resident howler monkeys**, by using a metagenomic approach. Different computational methodologies are being used to analyze the sequence datasets that were generated from sequencing total DNA (WGS - whole genome shotgun), amplicon 16S rRNA and mRNA (RNA-seq) with Roche-454 and/or Illumina. Besides analyzing the Metazoo project's datasets using computational methods, we are actively creating new computational tools, such as a metagenome browser and a conceptual framework and respective database to ease the integration of different types of data and metadata generated by metagenomic projects.

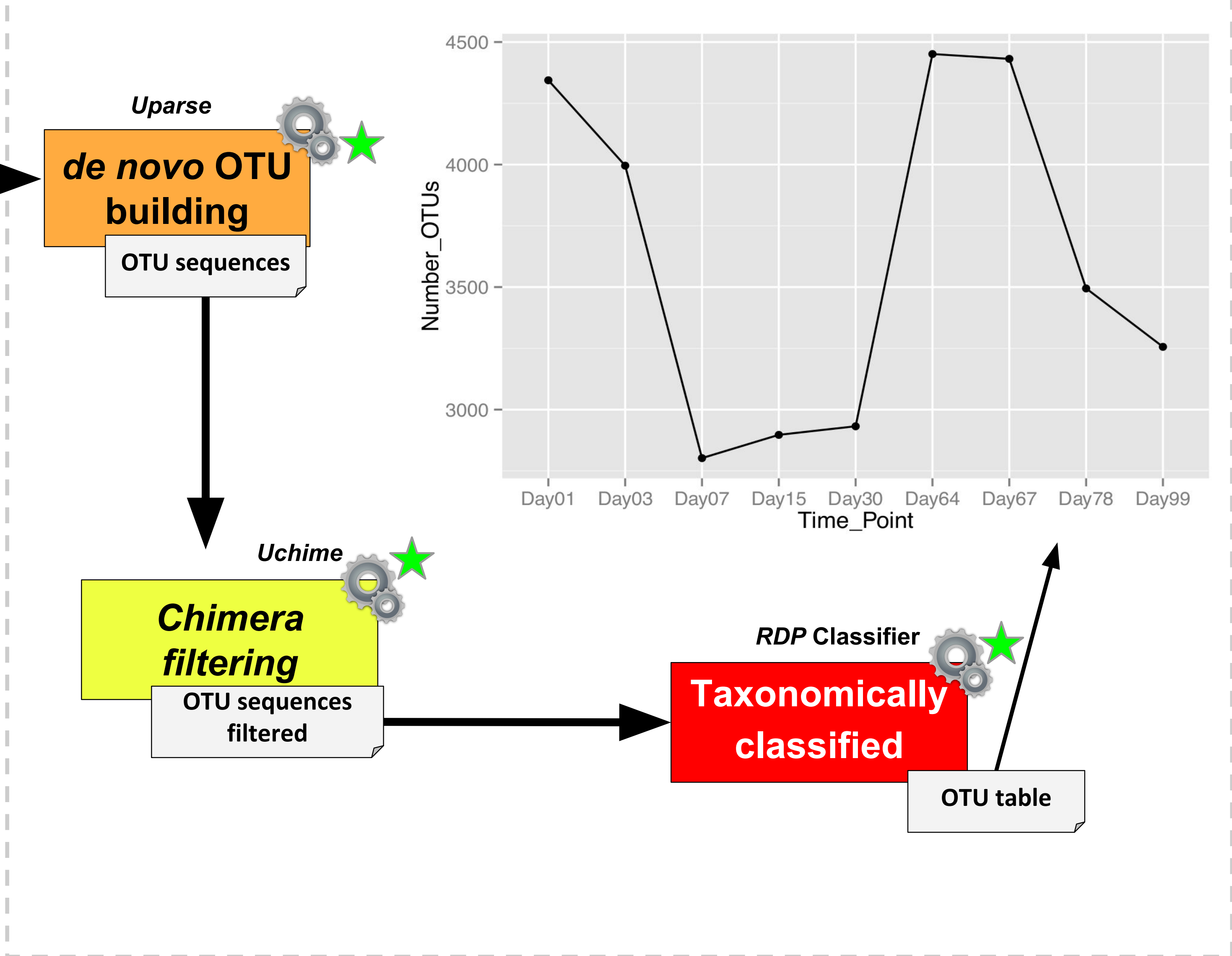




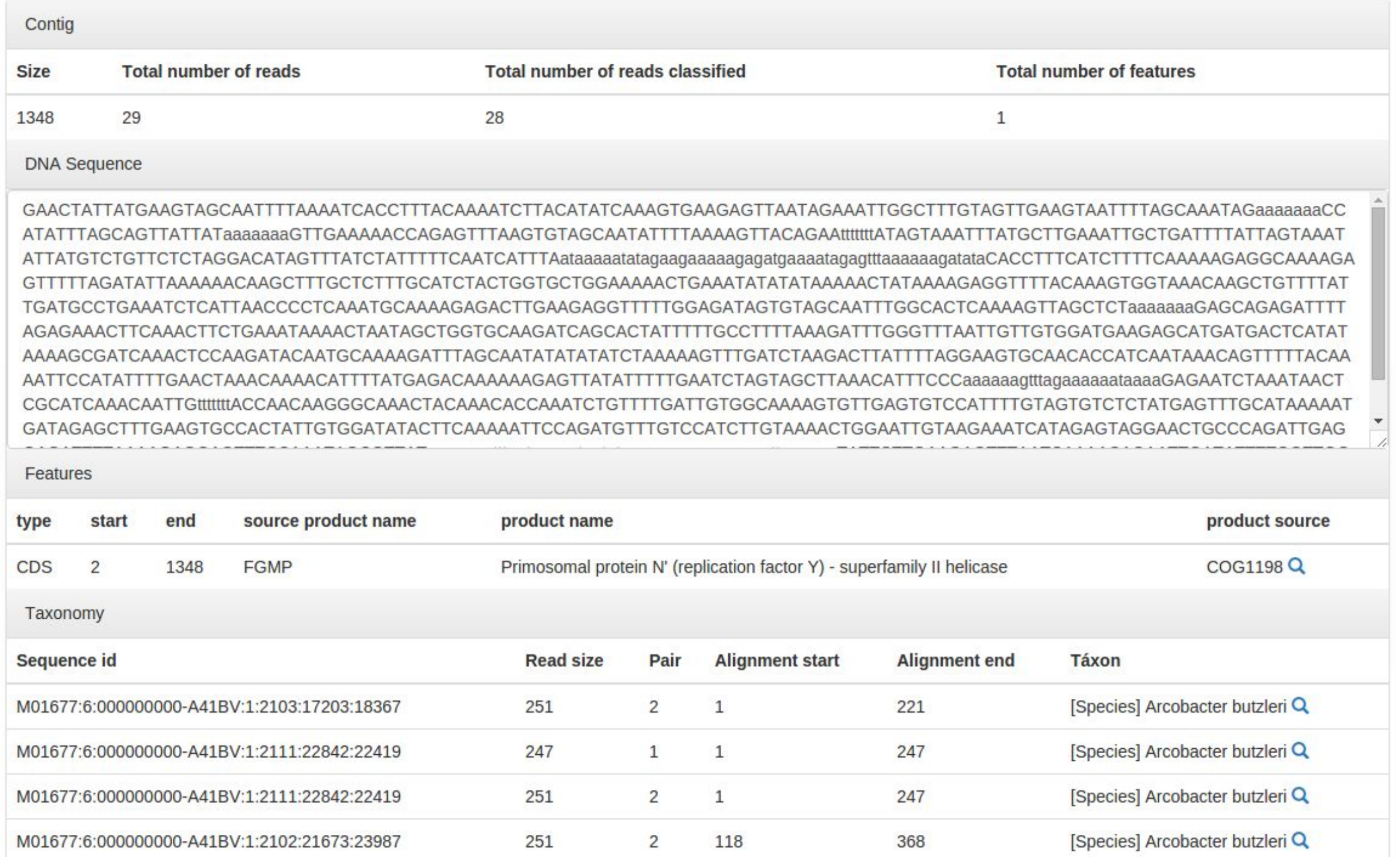


Chromosome Viewer - Colored by COG

★ ★ ★
low | medium | high
CPU cost



CARAVELA: a new approach to metagenomic browsing



Contig

Size	Total number of reads	Total number of reads classified	Total number of features
1348	29	28	1

DNA Sequence

```
GAACTATTATGAAGTAGCAATTTTAAAATCACCTTTACAAAATCTTACATATCAAAGTGAAGAGTTAATAGAAATTGGCTTTGTAGTTGAAGTAATTTTAGCAAATAGaaaaaaaCC
ATATTTAGCAGTTATTATaaaaaaaGTTGAAAAACCAGAGTTTAAGTGTAGCAATATTTTAAAAGTTACAGAAttttttATAGTAAATTTATGCTTGAAATTGCTGATTTTATTAGTAAAT
ATTATGTCTGTTCTCTAGGACATAGTTTATCTATTTTTCAATCATTTAataaaaatatagaagaaaaagagatgaaaatagagtttaaaaaagatataCACCTTTCATCTTTTCAAAAAGAGGCAAAAGA
GTTTTTAGATATTAAAAAACAAGCTTTGCTCTTTGCATCTACTGGTGCTGGAAAAACTGAAATATATATAAAAACTATAAAAGAGGTTTTACAAAGTGGTAAACAAGCTGTTTTAT
TGATGCCTGAAATCTCATTAACCCCTCAAATGCAAAAGAGACTTGAAGAGGTTTTTGGAGATAGTGTAGCAATTTGGCACTCAAAAGTTAGCTCTaaaaaaaGAGCAGAGATTTT
AGAGAAACTTCAAACTTCTGAAATAAAACTAATAGCTGGTGCAAGATCAGCACTATTTTTGCCTTTTAAAGATTTGGGTTTAATTGTTGTGGATGAAGAGCATGATGACTCATAT
AAAAGCGATCAAACTCCAAGATACAATGCAAAAGATTTAGCAATATATATATCTAAAAAGTTTGATCTAAGACTTATTTTAGGAAGTGCAACACCATCAATAAACAGTTTTTACAA
AATTCCATATTTTGAACTAAACAAAACATTTTATGAGACAAAAAAGAGTTATATTTTTGAATCTAGTAGCTTAAACATTTCCCaaaaaagtttagaaaaaataaaaGAGAATCTAAATAACT
CGCATCAAACAATTGttttttACCAACAAGGGCAAACTACAAACACCAAATCTGTTTTGATTGTGGCAAAAGTGTTGAGTGTCCATTTTGTAGTGTCTCTATGAGTTTGCATAAAAAT
GATAGAGCTTTGAAGTGCCACTATTGTGGATATACTTCAAAAATTCCAGATGTTTGTCCATCTTGTAAAACTGGAATTGTAAGAAATCATAGAGTAGGAACTGCCCAGATTGAG
```

Features

type	start	end	source product name	product name	product source
CDS	2	1348	FGMP	Primosomal protein N' (replication factor Y) - superfamily II helicase	COG1198

Taxonomy

Sequence id	Read size	Pair	Alignment start	Alignment end	Táxon
M01677:6:000000000-A41BV:1:2103:17203:18367	251	2	1	221	[Species] Arcobacter butzleri
M01677:6:000000000-A41BV:1:2111:22842:22419	247	1	1	247	[Species] Arcobacter butzleri
M01677:6:000000000-A41BV:1:2111:22842:22419	251	2	1	247	[Species] Arcobacter butzleri
M01677:6:000000000-A41BV:1:2102:21673:23987	251	2	118	368	[Species] Arcobacter butzleri

The CARAVELA tool is able to display a list of contigs and for each **contig**, it displays **annotated genes**, reads participating in its composition and **taxa associated with each read** (when such association exists).

Such a capability should enable manual/automated curation of assembly (identification of mis-assemblies) as well as taxonomic assignments (detection of possible mis-assignments)

References

IMG/M: Markowitz, Victor M., et al. "IMG/M: a data management and analysis system for metagenomes." *Nucleic acids research* 36.suppl 1 (2008): D534-D538.

Sickle: Najoshi G. A windowed adaptive trimming tool for FASTQ files using quality. https://github.com/najoshi/sickle.

Qiime: Caporaso, J. Gregory, et al. "QIIME allows analysis of high-throughput community sequencing data." *Nature methods* 7.5 (2010): 335-336.

Usearch: Edgar, Robert C. "Search and clustering orders of magnitude faster than BLAST." *Bioinformatics* 26.19 (2010): 2460-2461.

Mytaxa: Luo, Chengwei, Luis M. Rodriguez-R, and Konstantinos T. Konstantinidis. "MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences." *Nucleic acids research* (2014): gku169.

Krona tools: Ondov, Brian D., Nicholas H. Bergman, and Adam M. Phillippy. "Interactive metagenomic visualization in a Web browser." *BMC bioinformatics* 12.1 (2011): 385.

SoapDeNovo: Xie, Yinlong, et al. "SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads." *Bioinformatics* 30.12 (2014): 1660-1666.

Fastq-join: Aronesty, Erik. "Comparison of sequencing utility programs." *The Open Bioinformatics Journal* 7 (2013): 1-8.

Uparse: Edgar, Robert C. "UPARSE: highly accurate OTU sequences from microbial amplicon reads." *Nature methods* 10.10 (2013): 996-998.

Uchime: Edgar, Robert C., et al. "UCHIME improves sensitivity and speed of chimera detection." *Bioinformatics* 27.16 (2011): 2194-2200.

RDP Classifier: Wang, Qiong, et al. "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy." *Applied and environmental microbiology* 73.16 (2007): 5261-5267.

Support:

FAPESP
CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico
CAPES
ZOOLÓGICO DE SÃO PAULO
UNIFESP Universidade Federal de São Paulo 1933
Universidade de São Paulo Instituto de Química
BIOINFORMÁTICA USP