

Bioinformatic tools for Genome annotation

Course: Microbiomes (2021)

Sandra Godinho Silva
Instituto Superior Técnico
21/04/2021

About me: Sandra Godinho Silva



PhD Student @ Institute for Bioengineering and Biosciences

Instituto Superior Técnico
Lisbon, Portugal

 [Orcid](#)  [@sandragodinhosilva](#)

Table of contents

- **Part 1:** Bioinformatics: what are the tools available?
- **Part2:** Bioinformatic tools for Genome Annotation
- **Hands-on 1:** Annotate genomes with the COG database
 - **1.1:** Rast-WebMGA workflow
 - **1.2:** Join COG annotation with a R script
 - **1.2.1:** R Setup
 - **1.2.2:** Get started with R and RStudio

Part 1

Bioinformatics: what are the tools available?

Bioinformatics

Bioinformatics is the science that conjugates
biology + statistics + computer sciences
to study biological issues through the analysis of data.

Bioinformatics is a constantly changing and updating field.

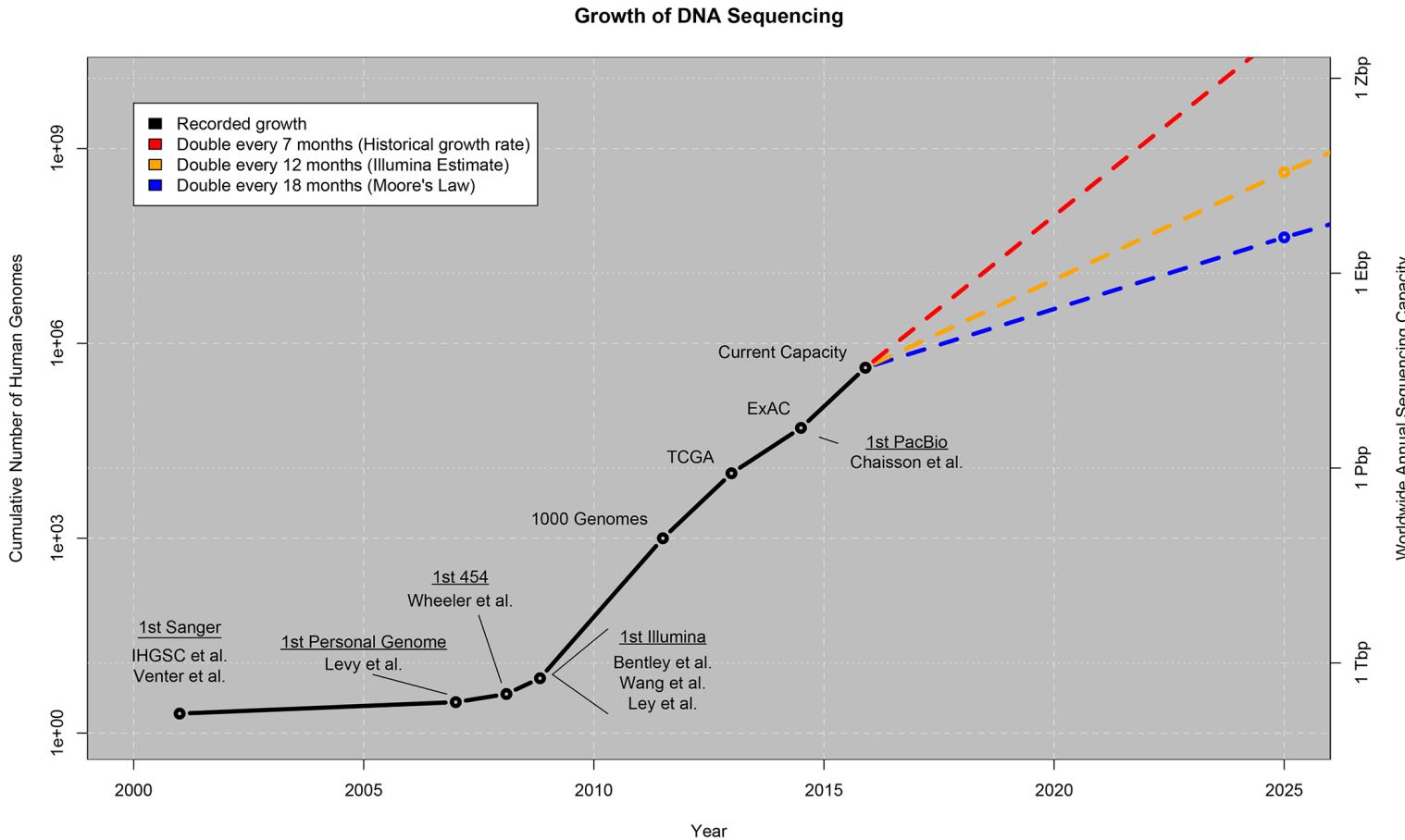
Fast developments in:

- 
- software
 - computing hardware
 - high throughput technologies

are creating a significant bottleneck:

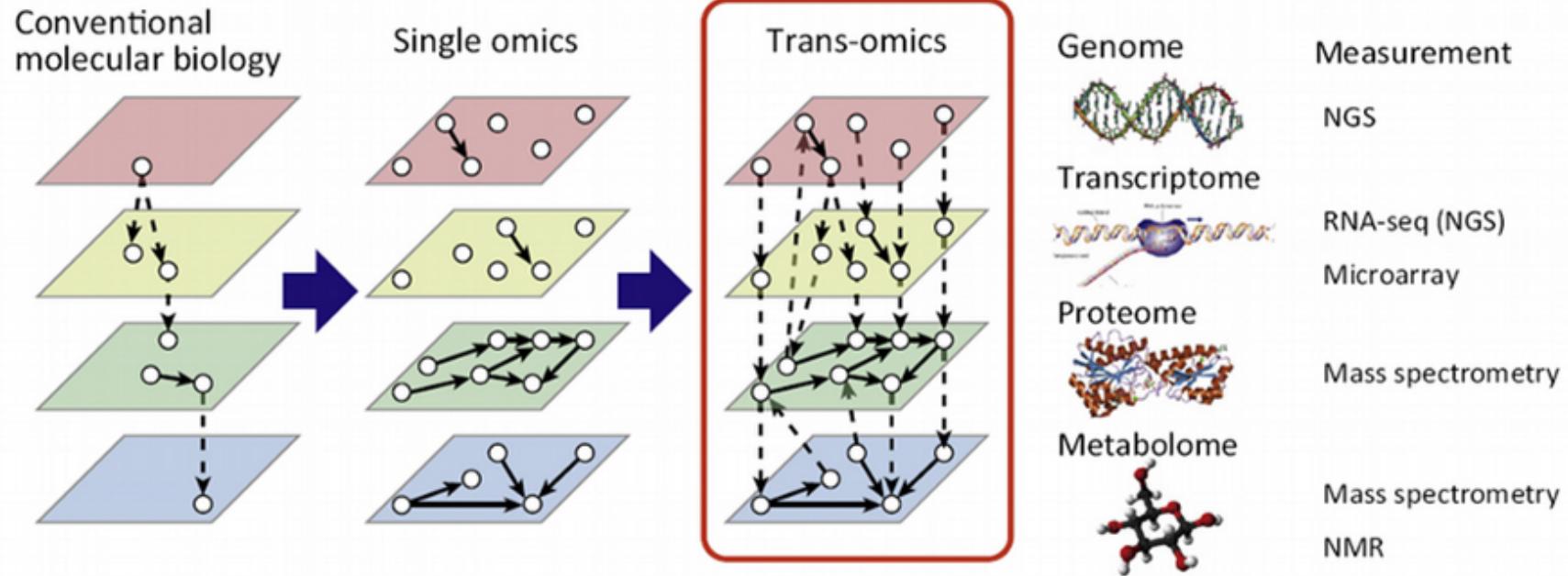
How to thoroughly analyze this massive amount of data?

How to thoroughly analyze this massive amount of data?



Stephens et al. PLoS Biology, 2015

Integrative OMICS



Yugi et al., Trends Biotechnol. 2016 Apr; 34(4):276–290

Choose the most appropriate route for you

Nowadays, bioinformaticians and researchers working in life scientists can choose from an overwhelming collection of exciting technologies and programming languages.

Several options:

- Web-based platforms
- Scripting/programming

Web-based platforms

Allow **accessible**, **reproducible** and **transparent computational research**.

KBase

Galaxy



The U.S. Department of Energy Systems Biology Knowledgebase **KBase** is an open-source software and data platform designed to meet the grand challenge of systems biology — predicting and designing biological function from the biomolecular (small scale) to the ecological (large scale).

Kbase: <https://www.kbase.us/>

Web-based platforms

Allow **accessible**, **reproducible** and **transparent computational research**.

KBase

Galaxy



Galaxy provides a system that enables researchers without informatics expertise to perform computational analyses through the web. A user interacts with Galaxy through the web by uploading and analyzing the data. Galaxy interacts with underlying computational infrastructure (servers that run the analyses and disks that store the data) without exposing it to the user. [Galaxy project - Europe: https://usegalaxy.eu/](https://usegalaxy.eu/)

Scripting & programming

[Bash](#)[Python \(1/2\)](#)[Python \(2/2\)](#)[R \(1/3\)](#)[R \(2/3\)](#)[R \(3/4\)](#)[R \(4/4\)](#)

- Default login shell for most Linux distributions;
- Simple bash scripts are really useful for data manipulation.

```
ubuntu@ubuntu-VirtualBox:~/code$ bash if_with_OR.sh
Enter any number
40
You lost the game
ubuntu@ubuntu-VirtualBox:~/code$ bash if_with_OR.sh
Enter any number
15
You won the game
ubuntu@ubuntu-VirtualBox:~/code$
```

Scripting & programming

Bash

Python (1/2)

Python (2/2)

R (1/3)

R (2/3)

R (3/4)

R (4/4)



Python

- Clear and powerful object-oriented programming language;
- Portable – runs just about anywhere;
- Clear syntax – relatively easy to learn. Source:
<https://www.python.org/>

Scripting & programming

Bash

Python (1/2)

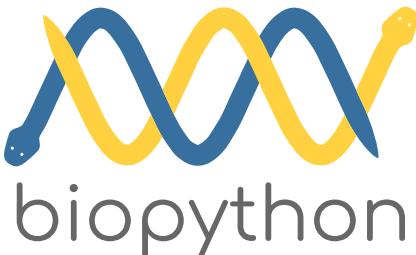
Python (2/2)

R (1/3)

R (2/3)

R (3/4)

R (4/4)



Biopython

- Collection of [Python](#) modules that provide functions to deal with DNA, RNA & protein sequence operations;
- It has sibling projects like BioPerl, BioJava and BioRuby.
Source: <https://biopython.org/>

Scripting & programming

Bash

Python (1/2)

Python (2/2)

R (1/3)

R (2/3)

R (3/4)

R (4/4)



R

- R is a powerful, popular open-source scripting language
- More than 20 years old - fairly mature - and growing in popularity.

Scripting & programming

Bash

Python (1/2)

Python (2/2)

R (1/3)

R (2/3)

R (3/4)

R (4/4)

Why is R so popular?

- **Statistical Language:** R is widely used in biology, genetics as well as in statistics.
- **Vast array of packages:** With over 10,000 packages in the CRAN repository, the number is constantly growing. These packages appeal to all the areas of industry.
- **Quality Plotting and Graphing:** The popular libraries like ggplot2 and plotly advocate for aesthetic and visually appealing graphs that set R apart from other programming languages.

Scripting & programming

Bash

Python (1/2)

Python (2/2)

R (1/3)

R (2/3)

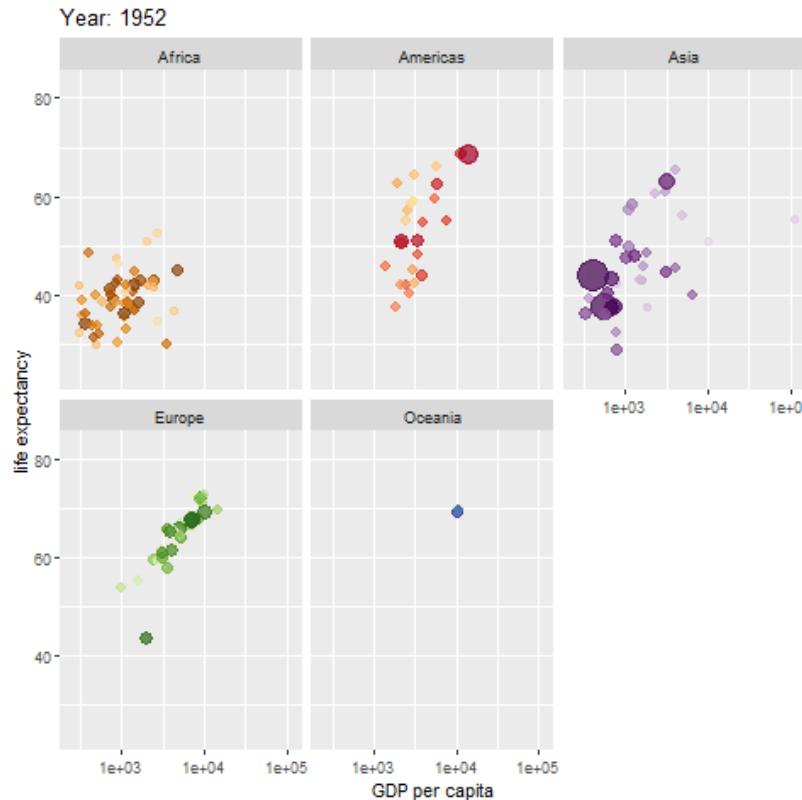
R (3/4)

R (4/4)

Dynamic graph:

ggplot2 +
ggridges

Source: [link](#)



Scripting & programming

Bash

Python (1/2)

Python (2/2)

R (1/3)

R (2/3)

R (3/4)

R (4/4)



R Studio

- Integrated Development Environment (IDE).
- Provides a graphical interface to R, making it more user-friendly, and providing dozens of useful features.

Source: <https://www.rstudio.com/>

And much more!





Part 2

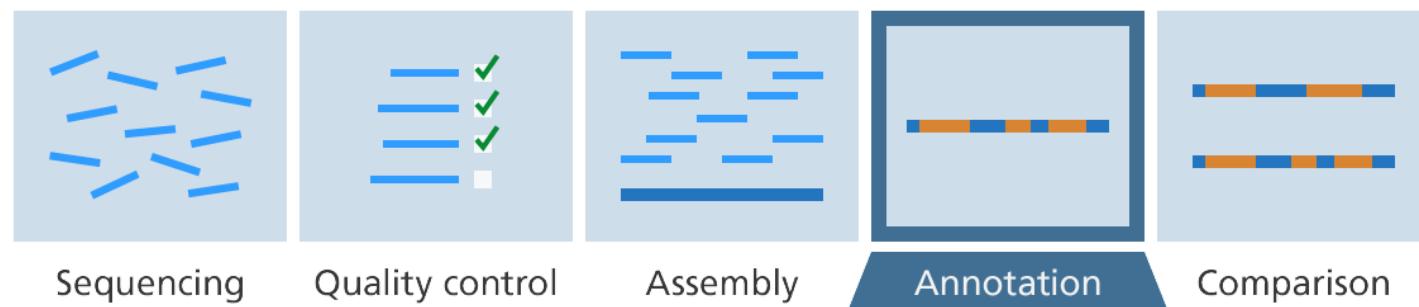
Bioinformatic tools for Genome Annotation

Genome Annotation

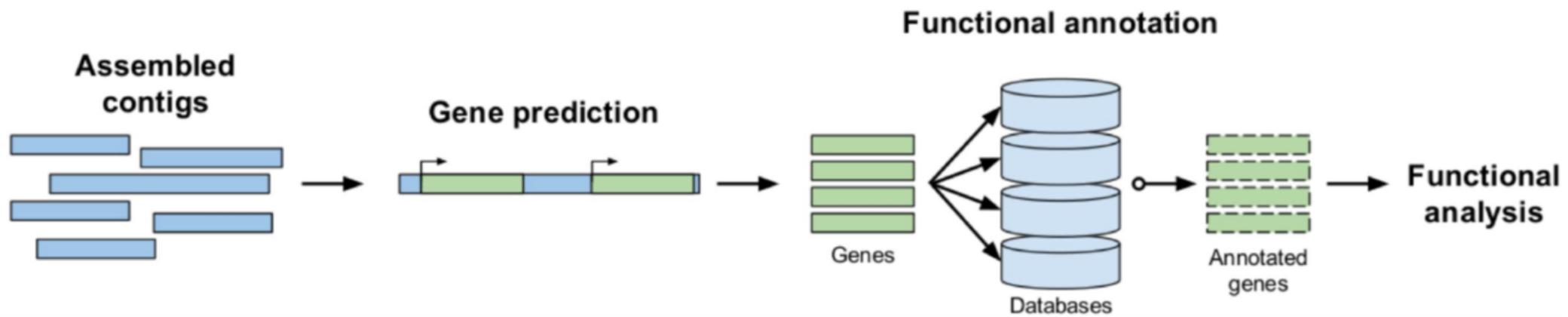
The process of attaching biological information to sequences.

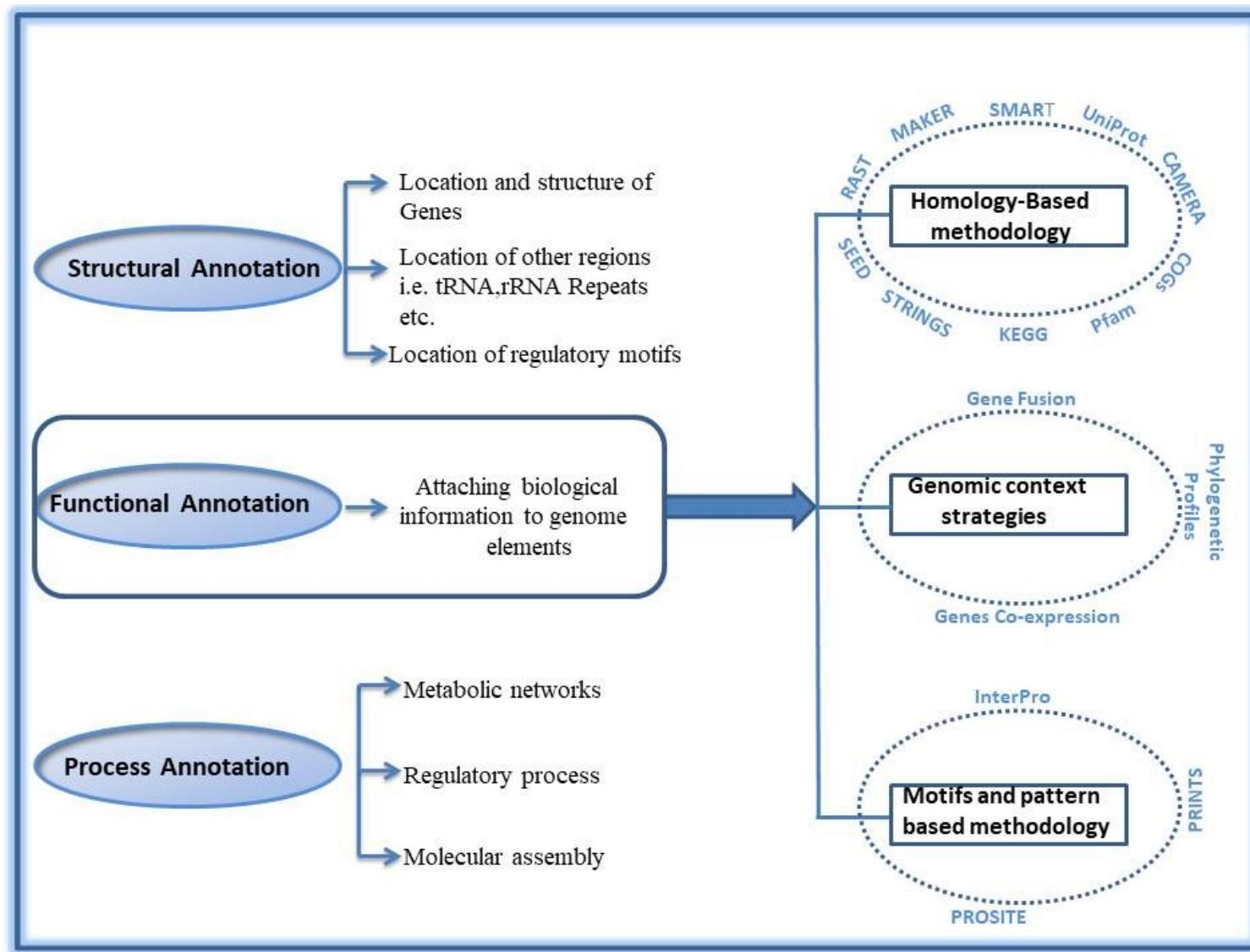
Consists of two main steps:

- structural annotation: identifying genomic elements;
- functional annotation: attaching biological information to these elements.



Source: [link](#)





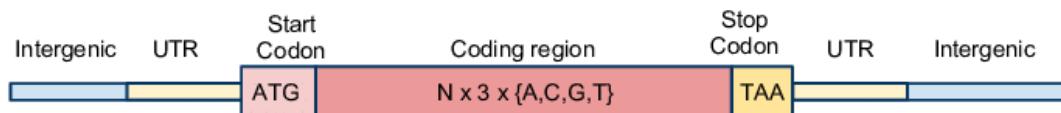
Source: <https://www.ssbs.edu.in/genome-annotation.htm>

Structural Annotation

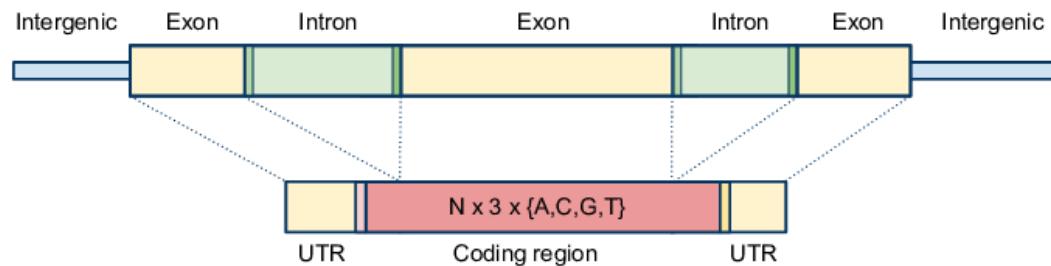
The process of identifying genomic elements such as:

- open reading frames (ORFs) and their localization;
- gene structure;
- coding regions;
- regulatory motifs.

A) Prokaryotic Gene



B) Eukaryotic Gene



Prokaryotic and Eukaryotic gene structure. Source: [link](#)

Structural Annotation: Methods

- **Similarity**

Similarity between sequences

```
>Synth12 on TGACv1_scaffold_221301_3B dna:scaffold:1:198209:1
Length=198209
Score = 2805 bits (3110), Expect = 0.0
Identities = 1559/1560 (99%), Gaps = 1/1560 (0%)
StrandPlus/Minus

Query 1 TATTCAGAACATTAGTGTACTAAATAATTATTAGTATTGGTCAAAT 60
Sbjct 142977 TTATTCAGAACATTAGTGTACTAAATAATTAGTATTGGTCAAAT 142977

Query 61 ATGGTCAAACCTGTTGCAAACTATGTTAAACTGTGCCAAATATGGTCAAAT 120
Sbjct 142917 ATGGTCAAACCTGTTGCAAACTATGTTAAACTGTGCCAAATATGGTCAAAT 142858

Query 121 CAAACATGTTAAACACTATTCAAGAACATTAGTGTACTAAATAATTATTAT 180
Sbjct 142857 CAAACATGTTAAACACTATTCAAGAACATTAGTGTACTAAATAATTAT 142798

Query 181 TTAGAATAATAGTTTAAACCTAAACAGTAAACGTTGACTTCTAGTCAGGCTAAACT 240
Sbjct 142797 TTAGAATAATAGTTTAAACCTAAACAGTAAACGTTGACTTCTAGTCAGGCTAAACT 142738

Query 241 CCTGGGGTTAATAGGTTAACCTTAGCTTAACTATTGTCAGGGAAACACAAAGTGA 300
Sbjct 142737 CCTGGGGTTAATAGGTTAACCTTAGCTTAACTATTGTCAGGGAAACACAAAGTGA 142678

Query 301 GACTTGAAAATGGGGAAATGAACCCAGAAGTTAAGCGCTCAGCGCTGGAGTAGTG 360
Sbjct 142677 GACTTGAAAATGGGGAAATGAACCCAGAAGTTAAGCGCTCAGCGCTGGAGTAGTG 142618

Query 361 GAGGATGGGTGACCGCCGAGAAGTTAGATGATTGGAAATGATGAGGGGGATTAAGGAT 420
Sbjct 142617 GAGGATGGGTGACCGCCGAGAAGTTAGATGATTGGAAATGATGAGGGGGATTAAGGAT 142558

Query 421 TAGAGGTTAAATTAACAGTGAGTGGTGTATTAGAGattaaatgtaaaaatattcg 480
Sbjct 142557 TAGAGGTTAAATTAACAGTGAGTGGTGTATTAGAGattaaatgtaaaaatattcg 142498

Query 481 aaatggaaaattggaaaaaattaaaaaaaattcaaaaaataaaaaatttgcggaaaaatttg 540
Sbjct 142497 AAATGGAAAATTGGAAAAAATTAAAAAAATTCAAAAAATAATTTCGGAAAATTTG 142438

Query 541 TTACCAACCGGACTAAAGTGGAGCTCCAGACAGCCGCCGCTGGAGGGCCCTTAGTCC 600
Sbjct 142437 TTACCAACCGGACTAAAGTGGAGCTCCAGACAGCCGCCGCTGGAGGGCCCTTAGTCC 142378
```

- *ab-initio* prediction

Genes are predicted based on gene content and signal detection (e.g. start/stop codon; Ribosome Biding Site (RBS), etc.).

		Second letter					
		U	C	A	G		
First letter	U	UUU } Phe UUC UUA } Leu UUG	UCU } Ser UCC UCA UCG	UAU } Tyr UAC UAA Stop UAG Stop	UGU } Cys UGC UGA Stop UGG Trp	U	C
	C	CUU } CUC CUA CUG } Leu	CCU } CCC CCA CCG } Pro	CAU } His CAC CAA } Gln CAG	CGU } CGC CGA CGG } Arg	U	C
	A	AUU } AUC AUA AUG } Ile Met	ACU } ACC ACA ACG } Thr	AAC } Asn AGC } Ser AAA } Lys AAG } Arg	AGU } AGC AGA AGG } Ser	U	A
	G	GUU } GUC GUA GUG } Val	GCU } GCC GCA GCG } Ala	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC GGA GGG } Gly	U	G

Examples of *ab-initio* tools:

- Glimmer
- GenemarkHMM
- **PRODIGAL**
PROkaryotic DYnamic programming Gene-finding ALgorithm

- Predicts protein-coding genes
- Handles gaps and partial genes
- Identifies translation initiation sites
- Handles finished genomes, draft genomes and metagenomes.
- Runs quickly
- Runs unsupervised: is an **unsupervised machine learning algorithm**.
Automatically learns the properties of the genome from the sequence itself, including RBS motif usage, start codon usage, and coding statistics.
- Source: <https://github.com/hyattpd/Prodigal>

Functional Annotation

The process of **attaching biological information to genomic elements** by describing the biochemical and biological function of proteins.

Possible analysis:

- similarity searches;
- gene cluster prediction for secondary metabolites;
- identification of transmembrane domains in protein sequences;
- finding gene ontology terms;
- pathway information.

Functional roles are assigned to coding sequences (CDSs).

Functional Assignment Databases

NCBI Nucleotide and Proteins databases

NCBI hosts constantly updated databases of proteins and DNA from several sources that include most of the newly sequenced organisms.

Queries to this database can be performed through the BLAST web server, also hosted by the NCBI.

BLAST can be considered the basic level of annotation for finding similarities.

BLAST (Basic Local Alignment Search Tool)

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

The screenshot shows the NCBI BLAST homepage. At the top, there's a banner for "BLAST+ 2.6.0 released" with a link to "More BLAST news...". Below the banner, there are sections for "Web BLAST" featuring "Nucleotide BLAST" (nucleotide → nucleotide), "blastx" (translated nucleotide → protein), and "tblastn" (protein → translated nucleotide); and "Protein BLAST" (protein → protein). At the bottom, there's a search bar for "BLAST Genomes" with a placeholder "Enter organism common name, scientific name, or tax id" and a "Search" button.

Databases

Specialist databases

Some databases offer extra information and search criteria for specific fields of study, while usually including the sequences already in public databases such as NCBI.

Databases

Specialist databases

PFAM

COG

Kegg

CAZymes



Pfam is a large collection of protein families, represented by multiple sequence alignments and hidden Markov models (HMMs).

Source: <http://pfam.xfam.org/>

Databases

Specialist databases

PFAM

COG

Kegg

CAZymes

COG

Clusters of Orthologous Groups (COGs)

Database of proteins generated by comparing the protein sequences of complete genomes.

Each COG consists of individual orthologous proteins or orthologous sets of paralogs from at least 3 lineages.

Orthologs typically have the same function, allowing transfer of functional information from one member to an entire COG.

Source: <https://www.ncbi.nlm.nih.gov/research/cog-project/>

Databases

Specialist databases

PFAM

COG

Kegg

CAZymes



Kyoto Encyclopedia of Genes and Genomes

Database resource for understanding high-level functions and utilities of biological systems, such as the cell, the organism and the ecosystem, from molecular-level information.

Source: <https://www.genome.jp/kegg/>

Databases

Specialist databases

PFAM

COG

Kegg

CAZymes



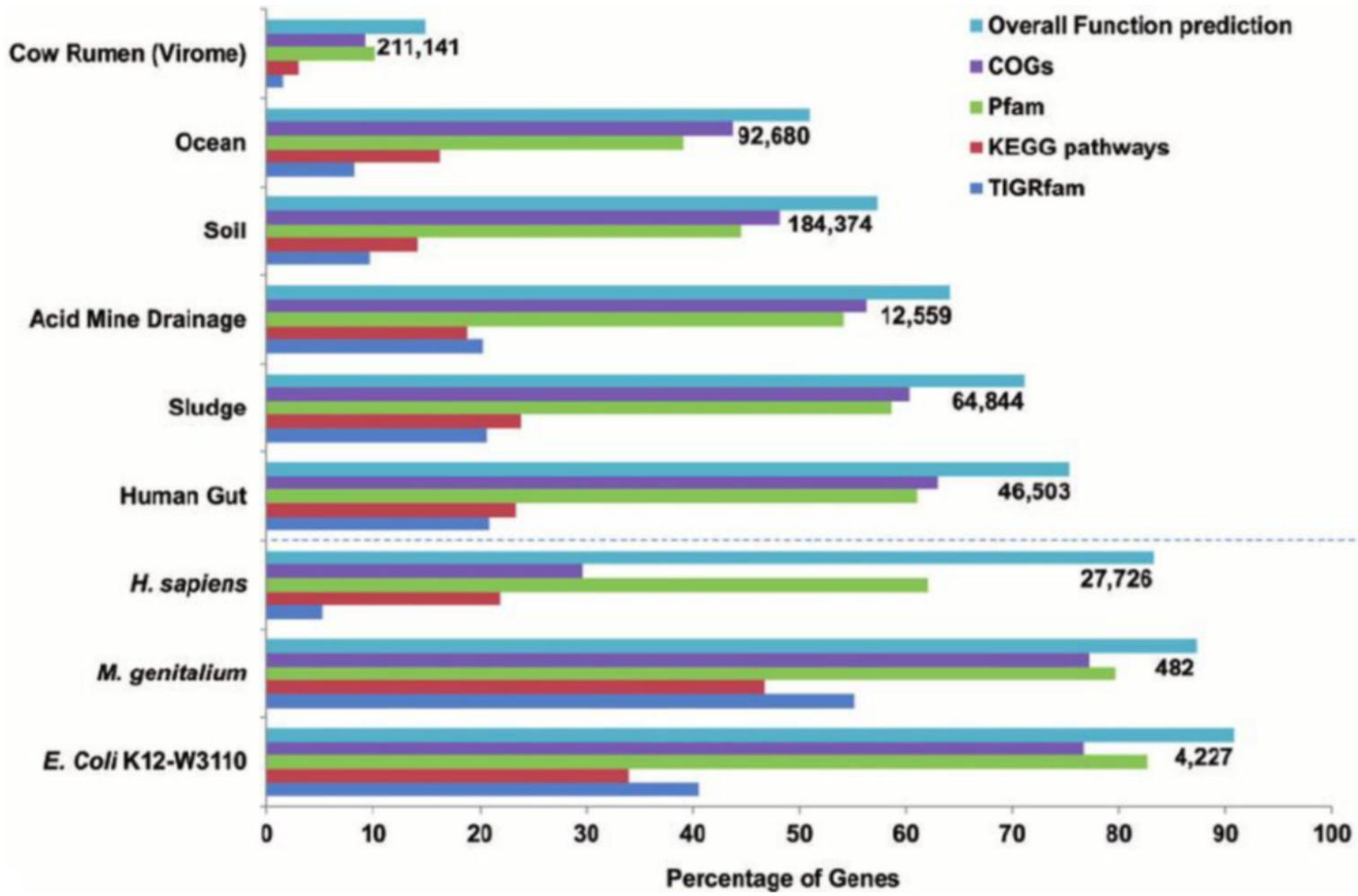
Carbohydrate-Active enZymes

Enzymes involved in the synthesis, metabolism, and transport of carbohydrates.

CAZymes are organized in families in the continuously updated database [CAZy](#).

Includes: glycoside hydrolases (GHs), glycosyltransferases (GTs), polysaccharide lyases (PLs), carbohydrate esterases (CEs) and carbohydrate binding modules (CBMs).

Source: <http://www.cazy.org/>



Prakash, Tulika & Taylor, Todd. (2012). Functional assignment of metagenomic data: Challenges and applications. *Briefings in bioinformatics*. 10.1093/bib/bbs033. ([link](#))

Functional assignment workflows

- Prokaryotic Genomes Automatic Annotation Pipeline (via NCBI)
- InterPro
- Prokka
Command-line tool. Also integrated into K-Base.
- RAST
A web server for annotating bacterial and archaeal genomes that provides annotation results in under a day

InterPro

Initial quality control

Identification of rRNA reads

Binning of 16S rRNA gene reads

QIIME-based taxonomic assignment

Non-RNA reads: ORF prediction

Functional annotation with
InterPro (IPR) Scan

Integrates several other databases
Updated every 8 weeks

The approach:

Prokka uses a **variety of databases** to **assign function to predicted CDS features**.

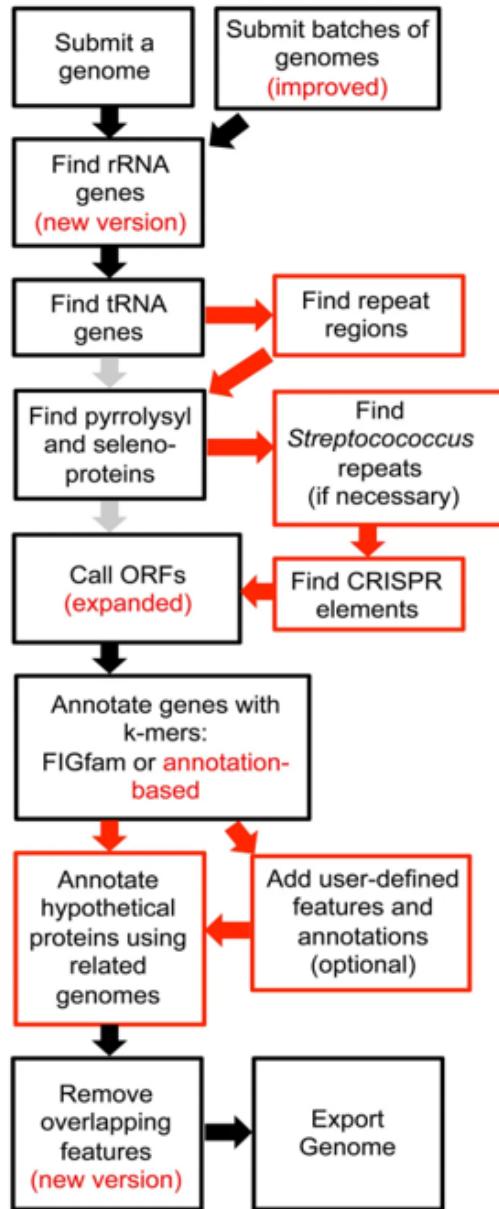
It takes a hierarchical approach to make it fast.

Proteins coding genes are annotated in two stages:

1) Prodigal identifies the coordinates of candidate genes, but does not describe the putative gene product.

2.1) A small, core set of well characterized proteins are first searched using **BLAST+**. This combination of small database and fast search typically completes about 70% of the workload.

2.2) A series of slower but more sensitive HMM databases are searched using **HMMER3**.



Rast

The new RAST-tk annotation pipeline

Very unique “Package” which works with its own database (SEED) and annotation scheme (“subsystems technology”).

Take-home messages: Genome Annotation

What?

Attributes functions to genes predicted from genomes or metagenomes

Why?

To answer the "What do they do?" question

How?

Through the pipeline ORF/Gene calling – Database search – functional analytics

Take-home messages: Genome Annotation

Ideally: Annotation of **all available/acquirable genes** to provide functional overview and hints on possible novel genes

The annotation accuracy is only as good as the available supporting data!

As new data becomes available, gene predictions and functional assignments will change.
Ex: Characterization of hypothetical proteins

Introduction to File Formats

FASTA

GFF3

GENBANK

DNA and protein sequences can be written in FASTA format. First line:> followed by the description. In the second line the sequence starts.

```
>gi|528476558|ref|NC_018928.2| Homo sapiens chromosome 17
TCCACTCACTGAGACAATAGACCCAAGCACATCAGCTCTGAGGCCTC
AGCACATCAGCTCTGAGGCCTCCACTCACTGAGACAATAGACCCAAG
TCACTGAGACAATAGACCCAAGCACATCAGCTCTGAGGCCTCCACTC
AACAGCTCTGAGGGCTCCACTCACTGAGACAACAGACCCAAGAACAA
>gi|528476670|ref|NC_018912.2| Homo sapiens chromosome 1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
CCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAA
ACCCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTA
TAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCCCCAA
```

Introduction to File Formats

[FASTA](#)[GFF3](#)[GENBANK](#)

General feature format (gene-finding format, generic feature format, GFF) is a file format used for describing genes and other features of DNA, RNA and protein sequences.

# start gene CM000169.1.g1									
CM000169.1	AUGUSTUS	gene	3168	4292	0.99	+	.	ID=CM000169.1.g1	
CM000169.1	AUGUSTUS	transcript	3168	4292	0.99	+	.	ID=CM000169.1.g1.t1;Parent=CM000169.1.g1	
CM000169.1	AUGUSTUS	start_codon	3168	3170	.	+	0	Parent=CM000169.1.g1.t1	
CM000169.1	AUGUSTUS	intron	3339	3422	1	+	.	Parent=CM000169.1.g1.t1	
CM000169.1	AUGUSTUS	CDS	3168	3338	0.99	+	0	ID=CM000169.1.g1.t1.cds;Parent=CM000169.1.g1.t1	
CM000169.1	AUGUSTUS	CDS	3423	4289	1	+	0	ID=CM000169.1.g1.t1.cds;Parent=CM000169.1.g1.t1	
CM000169.1	AUGUSTUS	stop_codon	4290	4292	.	+	0	Parent=CM000169.1.g1.t1	

Introduction to File Formats

FASTA

GFF3

GENBANK

The genbank sequence format is a rich format for storing sequences and associated annotations.

The screenshot shows the NCBI Nucleotide search results for the sequence *Aspergillus fumi* CM000169.1. The sequence is described as a "whole genome shotgun sequence" from a "linear CON 23-MAR-2015 whole genome shotgun". The sequence viewer displays the first 1,000 bases. On the left, detailed information is provided for the sequence, including Locus (CM000169), Definition (Aspergillus fumi CM000169.1), Accession (CM000169.1), Version (CM000169.1), DBLINK (BioProject, BioSample: SAMN00115746), Keywords (WGS), Source (Aspergillus fumigatus Af293), Organism (Aspergillus fumigatus Af293), and Reference (1 bases 1 to 4918979). Authors listed are Nierman W C, Pain A, Anderson M J, Wortman J R, Kim H S. The right side of the interface includes a "Display Settings" dropdown set to "GenBank", a "Format" dropdown showing options like "Summary", "GenBank (full)" (which is selected), "FASTA", etc., and an "Apply" button. Other sections include "Change region shown", "Customize view" (set to "Customize"), "Basic Features" (set to "Default features"), "Display options" (checkboxes for "Show sequence" (checked) and "Show reverse complement"), and "Analyze this sequence" buttons for "Run BLAST" and "Pick Primers".

Hands-on 1

Genome annotation with the
Clusters of Orthologous Genes (COG) Database

Worflow

Genome annotation with Clusters of Orthologous Genes (COG) Database

K-Base

1) Download bins from K-Base in fasta format

Rast

2) Upload bins in Rast
3) Perform gene annotation
4) Download fasta aminoacid files

WebMGA

5) Upload downloaded bins in WebMGA
6) Perform COG annotation
7) Download COG annotation from WebMGA

R

8) Join COG annotation into single table - R script

K-Base

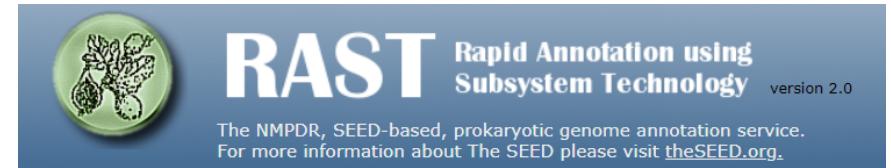
1) Download bins from K-Base in fasta format;

[Link](#)



Rast

- 2) Upload bins (fasta files) in Rast;
- 3) Perform genome annotation;
- 4) Download fasta aminoacid files;



Link

WebMGA

- 5) Upload genomes (fasta aminoacid files) in WebMGA;
- 6) Perform COG annotation;
- 7) Download COG annotation from WebMGA.

Link

The screenshot shows the WebMGA web interface. At the top, there is a navigation bar with links for Home, Server, Results, Scripts, Help, and Contact Us. Below the navigation bar, there is a decorative background of DNA sequence data. On the left side, there is a sidebar with several options: clustering, orf prediction, function annotation (which is currently selected and highlighted in orange), cog, kog, prk, pfam (which is also highlighted in orange), and tigrfam. The main content area has a yellow header titled "function annotation (PFAM)". Below the header, it says: "This program performs function annotation by using HMMER 3.0 program on PFAM database." Under "Inputs:", it says: "Protein FASTA file (required), can be in .gz format". Under "Outputs:", it says: "output.zip will be produced with a README file describing the output files and format". There is a form for uploading a sequence file, an optional email field, parameters (-E 0.001, show description), and a "Submit" button. Below the form, there is a link "Show an example". At the bottom, there is a section for "Program/Database References" with two items: "Profile hidden Markov models", S. R. Eddy Bioinformatics (1998) 14(9):755-763, and "The Pfam protein families database", R. D. Finn, et al. Nucleic Acids Research (2010) 38: D211-D222. At the very bottom, there is a copyright notice: "© 2010-2021 Weizhong Li's Group" and a citation: "Please kindly cite: S. Wu, Z. Zhu, L. Fu, B. Niu and W. Li, "WebMGA: a Customizable Web Server for Fast Metagenomic Sequence Analysis", BMC Genomics 2011, 12:444".

Hands-on 1.2

1.2.1 Install R and RStudio

[Link: 1.2.1 R setup](#)

1.2.2 Introduction to R and RStudio

[Link: Getting started with R](#)

1.2.3 Run R script to join COG annotation files

Workflow to join COG annotation in a single file

For each bin:

- 1) Download COG annotation from WebMGA (will be a zipped file)
- 2) Unzip file;
- 3) Look for the .txt file named "cog";
- 4) Rename this file with the bin's name (Ex: cog_SAMPLEX_bin1)

For all bins:

- 5) Move all cog annotation files to the same folder;
- 6) Download R script and move it to this folder;
- 7) Open all cog annotation files and delete the character "#" from the first line;
- 8) Run R script.

Final output: COG table with annotations for all bins

Download script: [Script_Merge_COGs.R](#)

Note R extension!

That is all for today!

Any questions?



Thank you!

Here's where you can find me...

 sandragodinhosilva.netlify.com
 @SandraGodSilva

Slides created via the R package **xaringan**