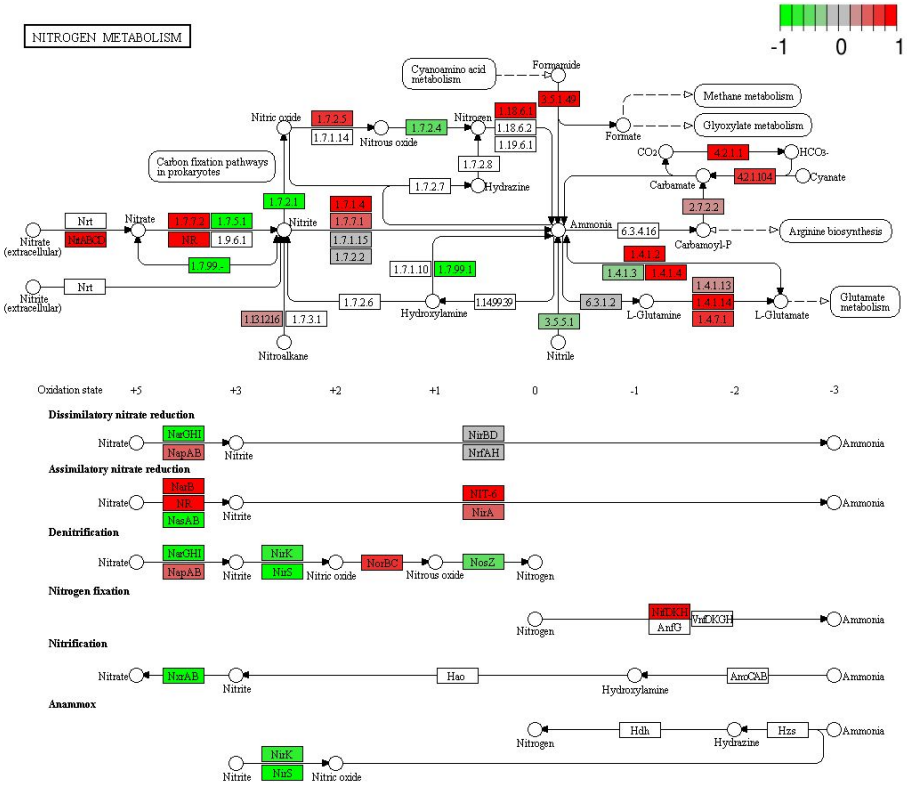
Annotation of genomes



By Dr. Richard Allen White III

Lecture 5 - Sep 24th, 2019

Zoom! 404-899-586

Annotation

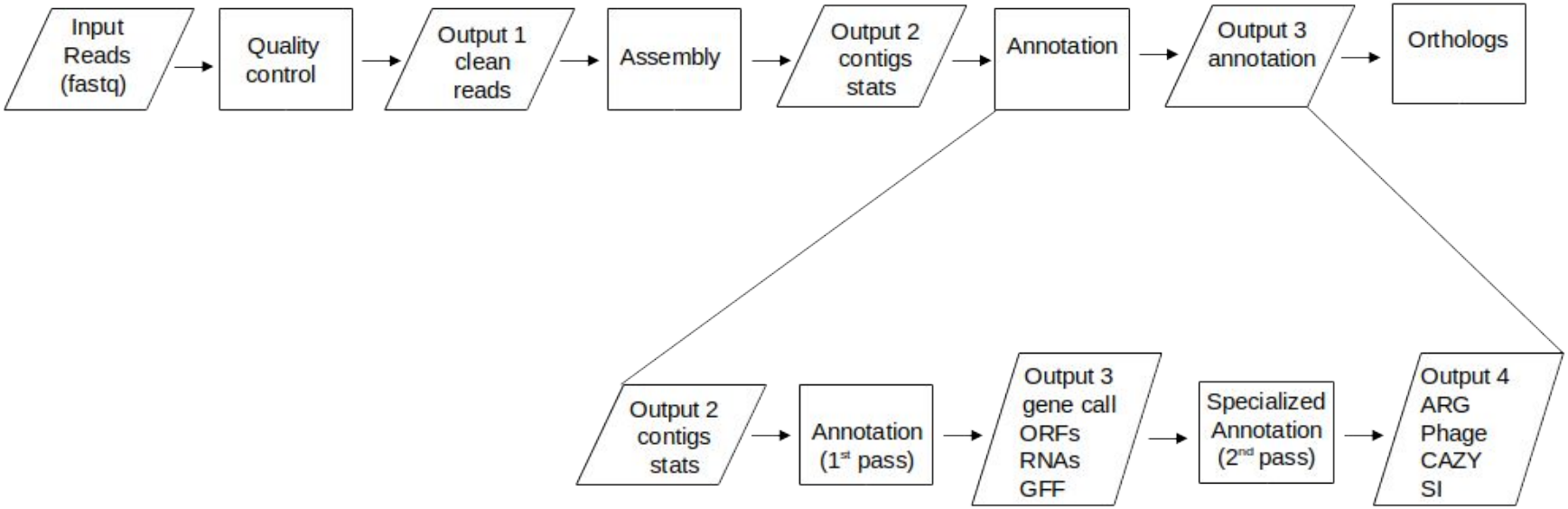
Concepts:

- Open reading frame (ORF) calling
- Reference databases
- Automatic annotation

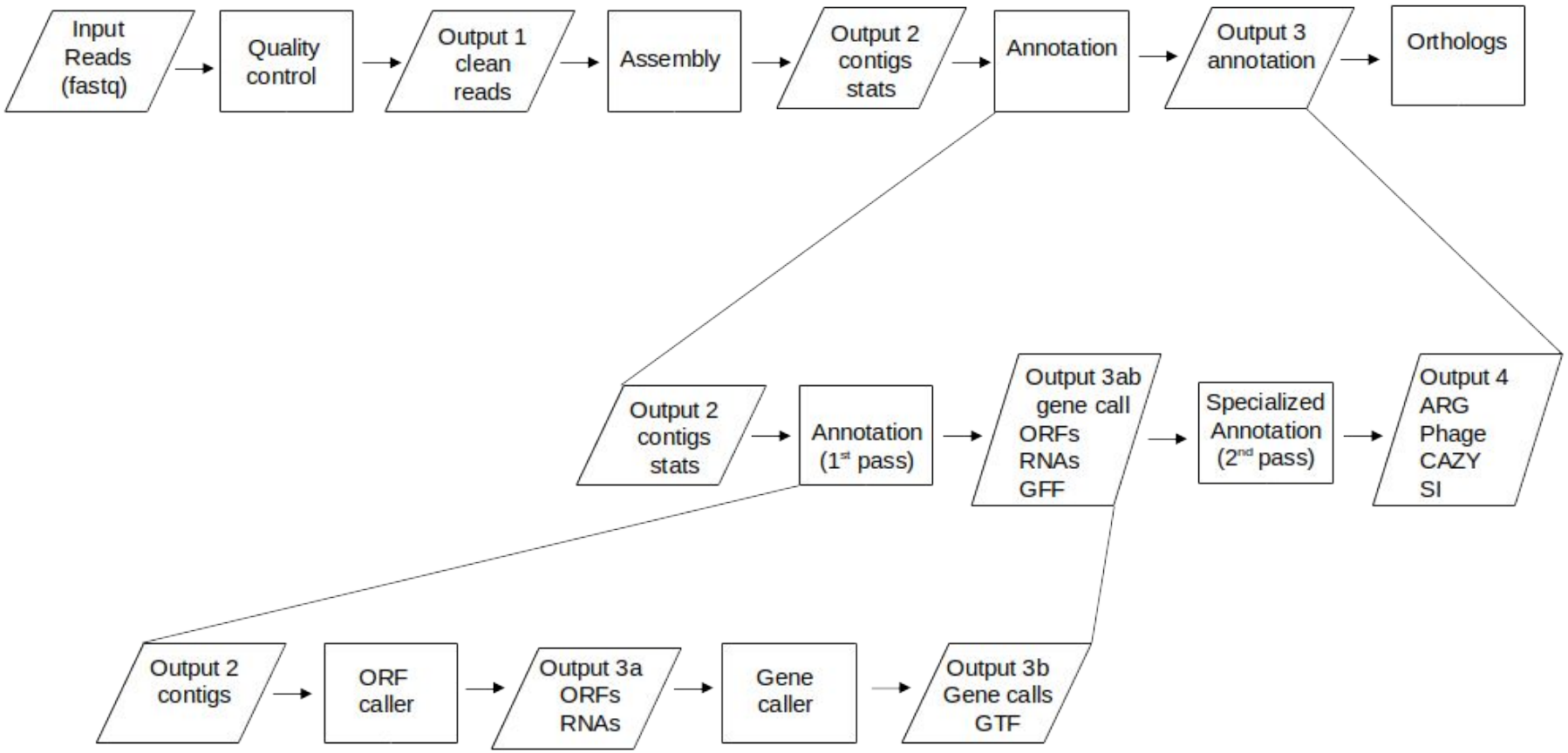
Learning Objectives:

- Ways and programs to call ORFs
- Your only as good as your database
- Finding the best annotation

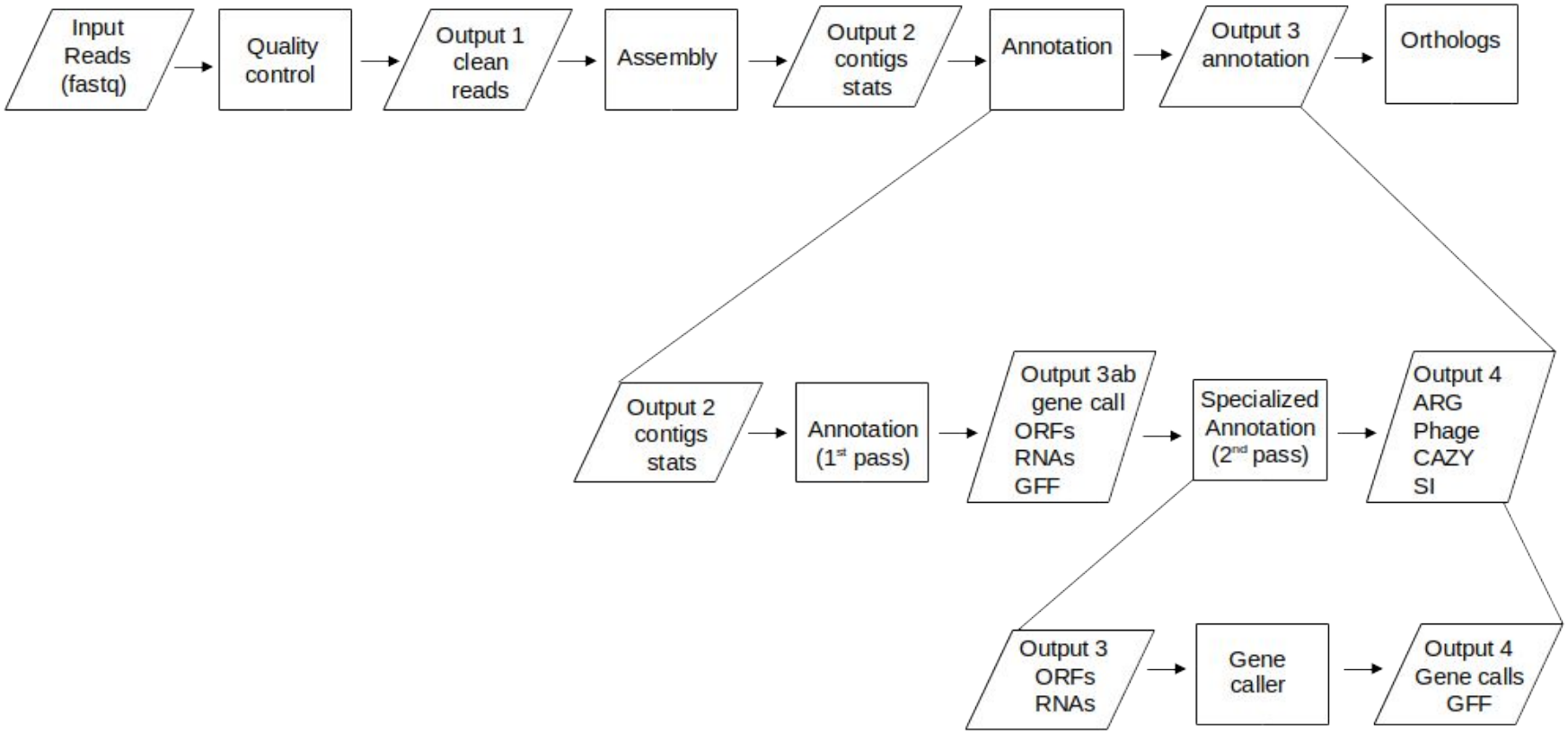
Annotation - flowgraph



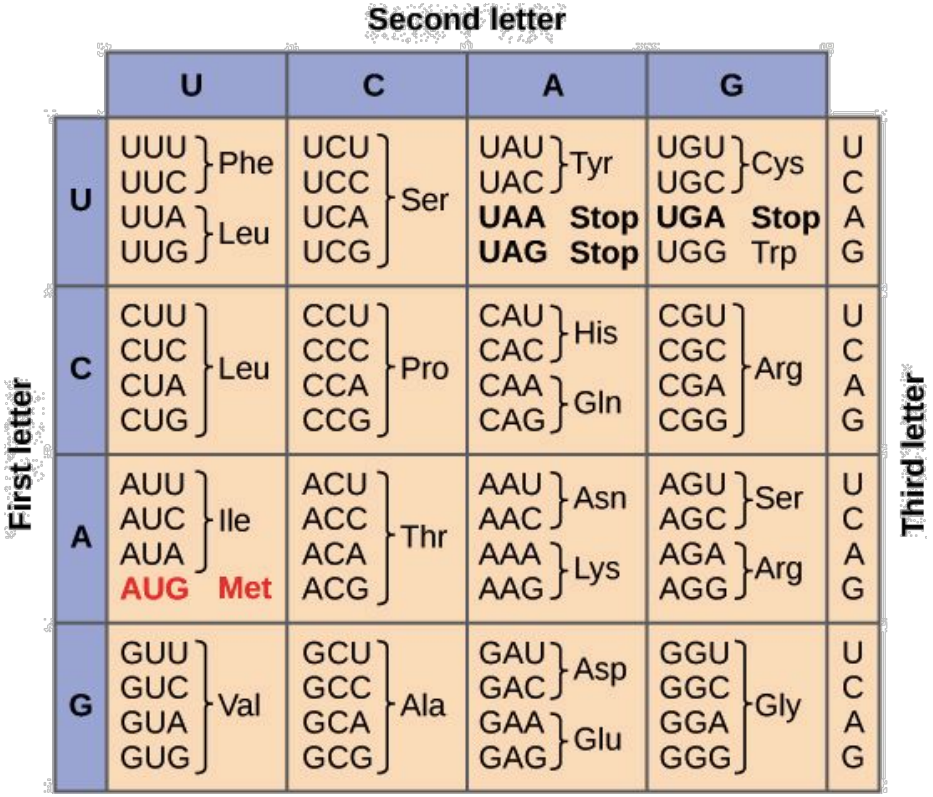
Annotation - flowgraph



Annotation - flowgraph

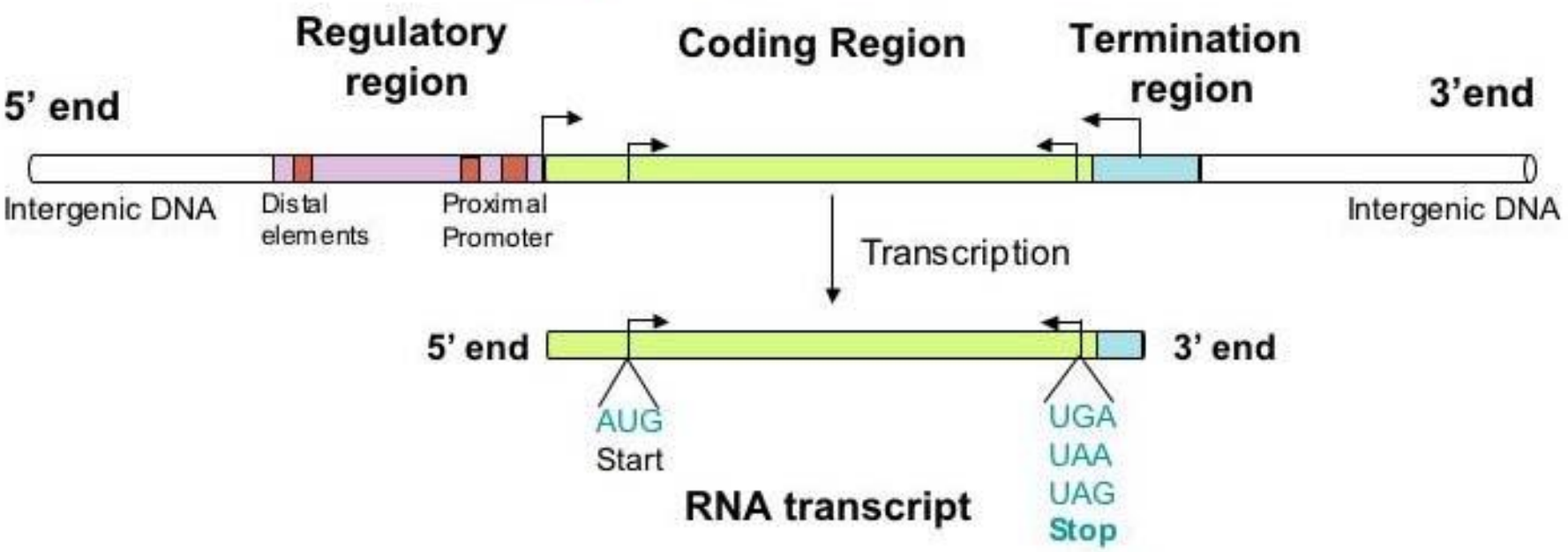


Annotation step 1 -ORF calling



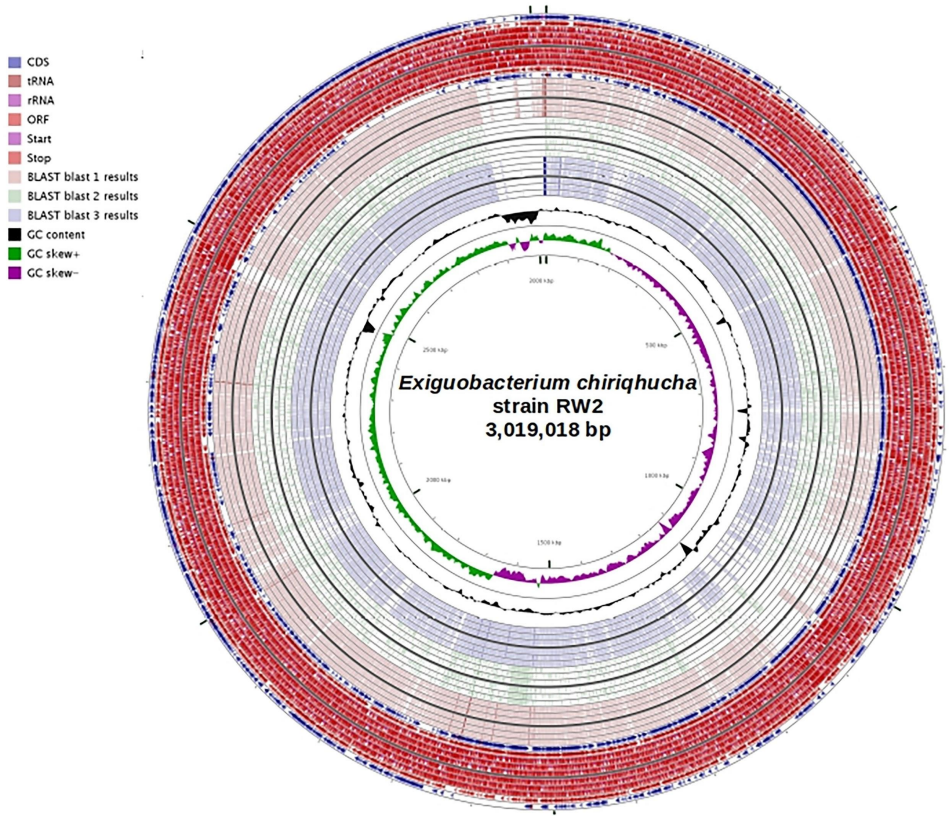
- Conversion of nucleotide contigs to gene calls
  - Find genes (start and stop codons), 1 start with 3 stops
  - Convert into amino acid sequences
  - Make map of location and size of genes (GFF or GTF)

Annotation step 1 -ORF calling



- Prokaryotic gene structure
  - High gene density (85% coding in *E. coli* K12)
  - Six-base consensus sequence is AGGAGG for translation is?
  - Is ORF a gene?
  - 29.5% have an imputed function, 2.1% have a phenotype and 19.5% have no function assignment (*E. coli* K12)
  - How much is coding in the human genome? (10%, 90%, 1%)

Annotation step 1 -ORF calling



- Finding ORFs
  - More ORFs than genes
    - RW2
      - >4,500 ORFs
      - 3,075 genes/3,022 proteins
    - K12
      - >6000 ORFs
      - 4,566 genes/4,242 proteins
  - If random, one stop per every 64 divided by 3 = 21 codons on average
  - Average protein = ~300 codons
  - Problems
    - Hard to search and find long ORFs
    - Short genes
    - Overlapping on opposite strands

White 3rd et al., 2019.
https://www.frontiersin.org/articles/10.3389/fmicb.2018.03189/full

Annotation step 1 -ORF calling

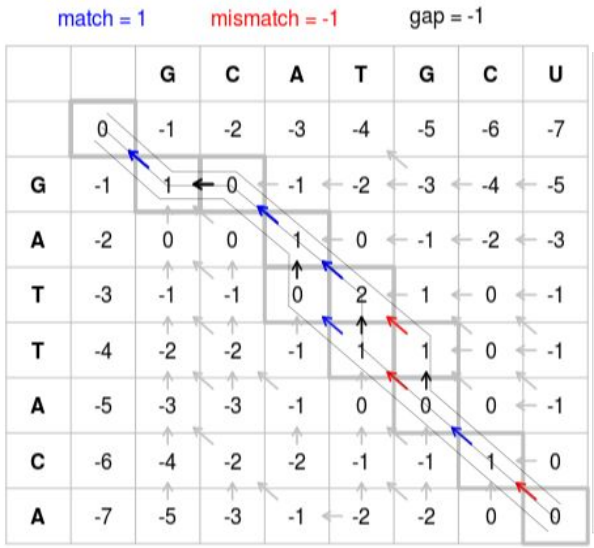
- Methods for ORF calling
  - 6-frame translation
    - Sixpack (http://emboss.sourceforge.net/apps/cvs/emboss/apps/sixpack.html)
  - Interpolated Markov models (IMMs)
    - Glimmer (http://ccb.jhu.edu/software/glimmer/index.shtml)
  - GC bias information + Shine-Dalgarno RBS motiff
    - Prodigal (https://github.com/hyattpd/Prodigal)
  - Heuristic in homogeneous Markov models with hidden Markov model (HMM)
    - Genemark (http://opal.biology.gatech.edu/GeneMark/genemarks.cgi)
  - Sequencing error models and codon usages in a hidden Markov model (HMM)
    - Fraggenescan (https://sourceforge.net/projects/fraggenescan/)
  - Aho–Corasick algorithm to find regions uninterrupted by stop codons
    - orfM (https://github.com/wwood/OrfM)

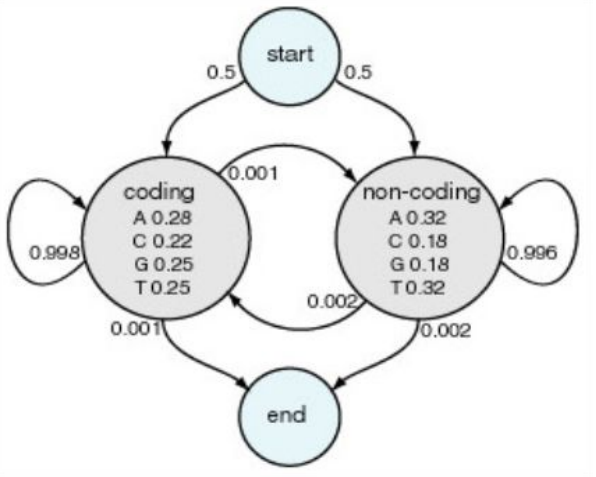
Annotation step 2 -Gene calling

- Gene calling or Gene prediction
  - Figure out what the proteins are coding for based on reference
  - Alignment based (BLAST - local alignments)
  - Deep searching (hmmer - HMMs)

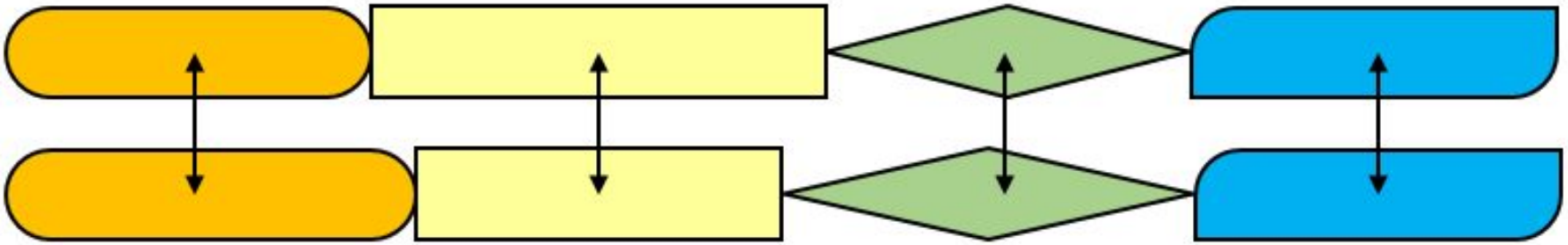
$H =$

	–	A	C	A	C	A	C	T	A
–	0	0	0	0	0	0	0	0	0
A	0	2	1	2	1	2	1	0	2
G	0	1	1	1	1	1	1	0	1
C	0	0	3	2	3	2	3	2	1
A	0	2	2	5	4	5	4	3	4
C	0	1	4	4	7	6	7	6	5
A	0	2	3	6	6	9	8	7	8
C	0	1	4	5	8	8	11	10	9
A	0	2	3	6	7	10	10	10	12

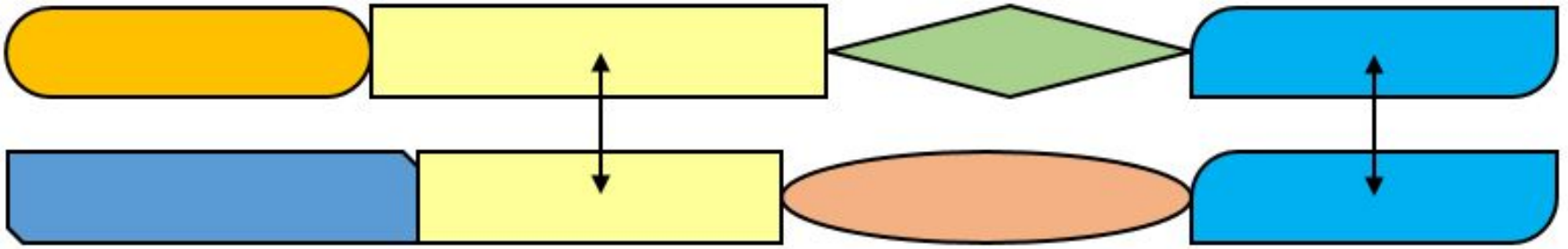




Annotation step 2 -Gene calling

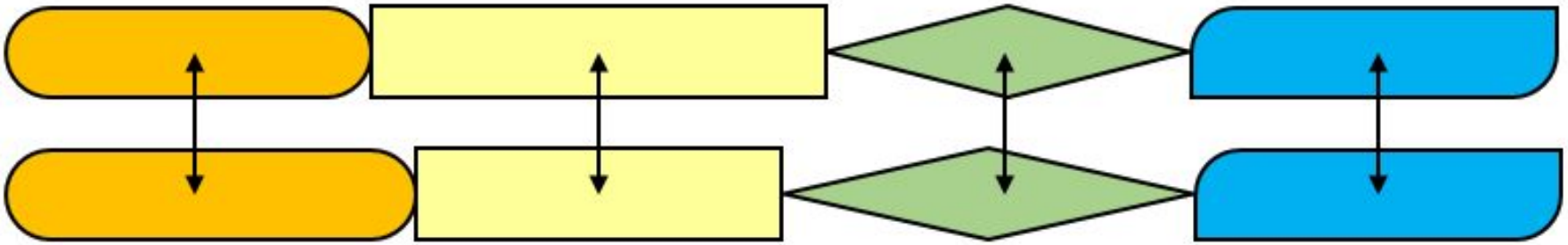


Type? Algorithm?

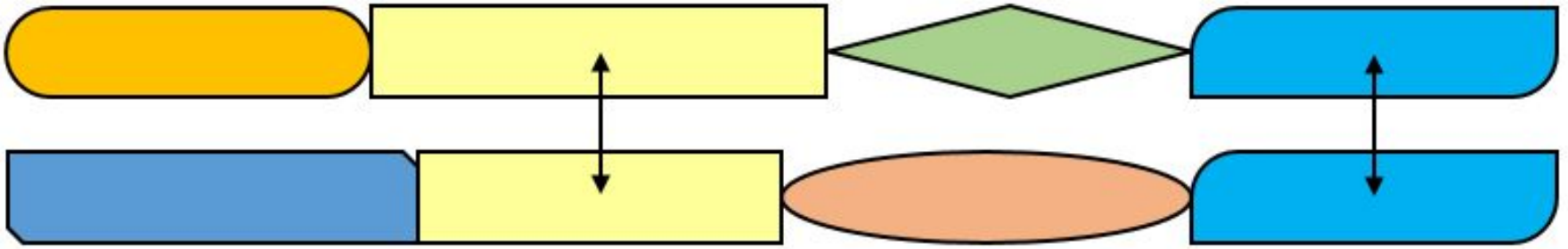


Type? Algorithm?

Annotation step 2 -Gene calling

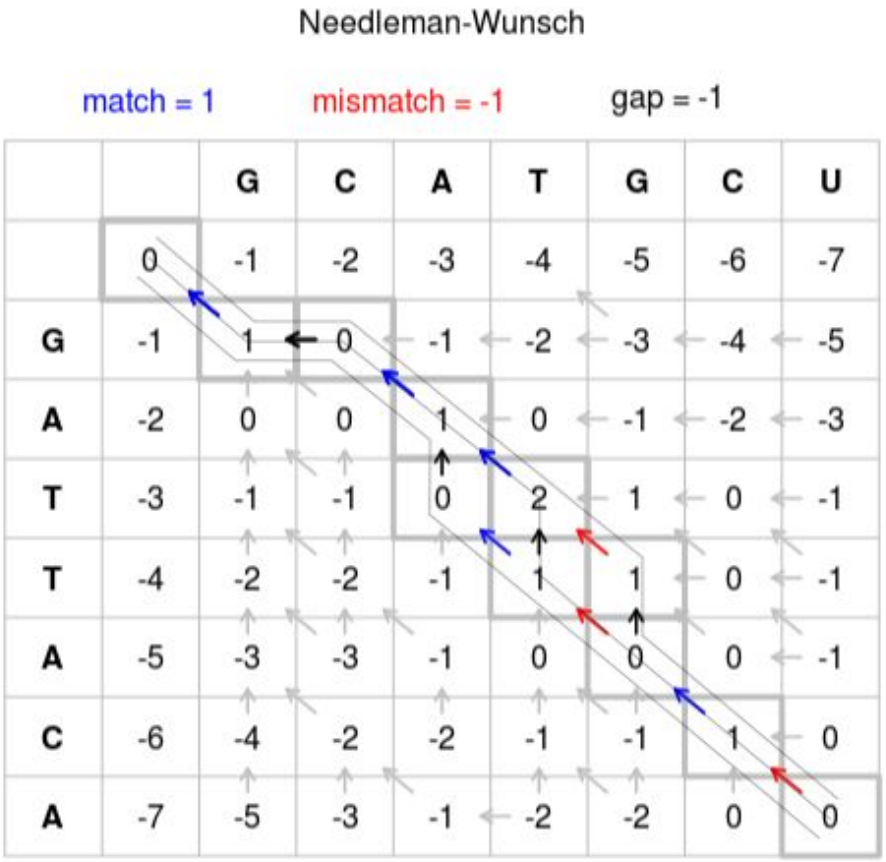


Global, Algorithm: Needleman-Wunsch



Local, Algorithm: Smith-waterman

Annotation step 2 -Gene calling

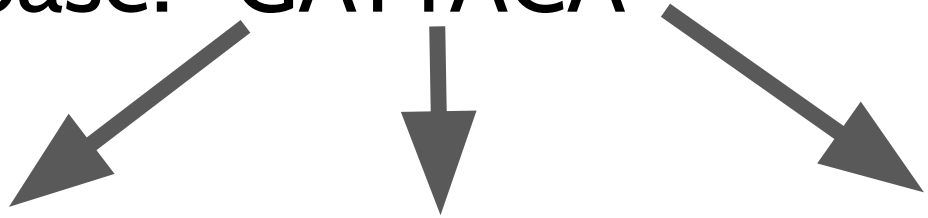


		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

Global, Algorithm: Needleman-Wunsch

Annotation step 2 -Gene calling

Query: GCATGCU
Database: GATTACA



GCATG-CU
G-ATTACA

GCA-TGCU
G-ATTACA

GCAT-GCU
G-ATTACA

Alignments

Global, Algorithm: Needleman-Wunsch

Annotation step 2 -Gene calling

Alignment: +GCATGCU
+GATTACA
Score: 0 // Handle matches handle, doesn't win any score

Alignment: +GCATGCU
+GATTACA
Score: 0 // 1 gap, score -1

Alignment: +GCATGCU
+GATTACA
Score: 0 // 2 gaps, score -2

Alignment: +GCATGCU
+GATTACA
Score: 0 // 3 gaps, score -3

Alignment: +GCATGCU
+GATTACA
Score: 0 // 4 gaps, score -4

Global,
Algorithm: Needleman-Wunsch scoring

Annotation step 2 -Gene calling

Initialize the scoring matrix

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0								
G	0								
T	0								
T	0								
G	0								
A	0								
C	0								
T	0								
A	0								

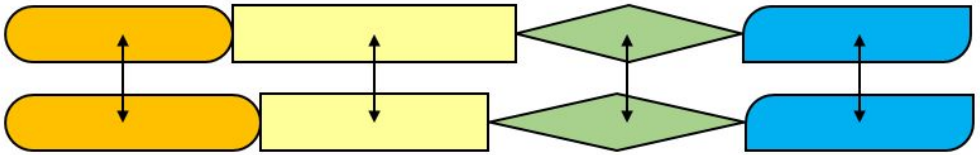
Substitution matrix: $S(a_i, b_j) = \begin{cases} +3, & a_i = b_j \\ -3, & a_i \neq b_j \end{cases}$

Gap penalty: $W_k = kW_1$

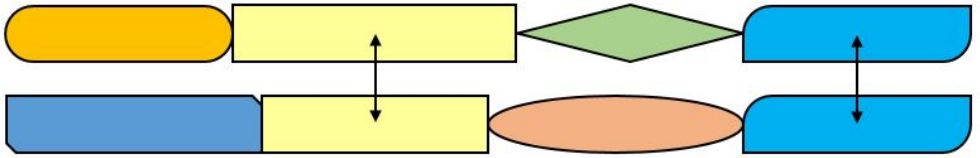
$W_1 = 2$

Local, Algorithm: Smith-Waterman

Annotation step 2 -Gene calling



Global Alignment

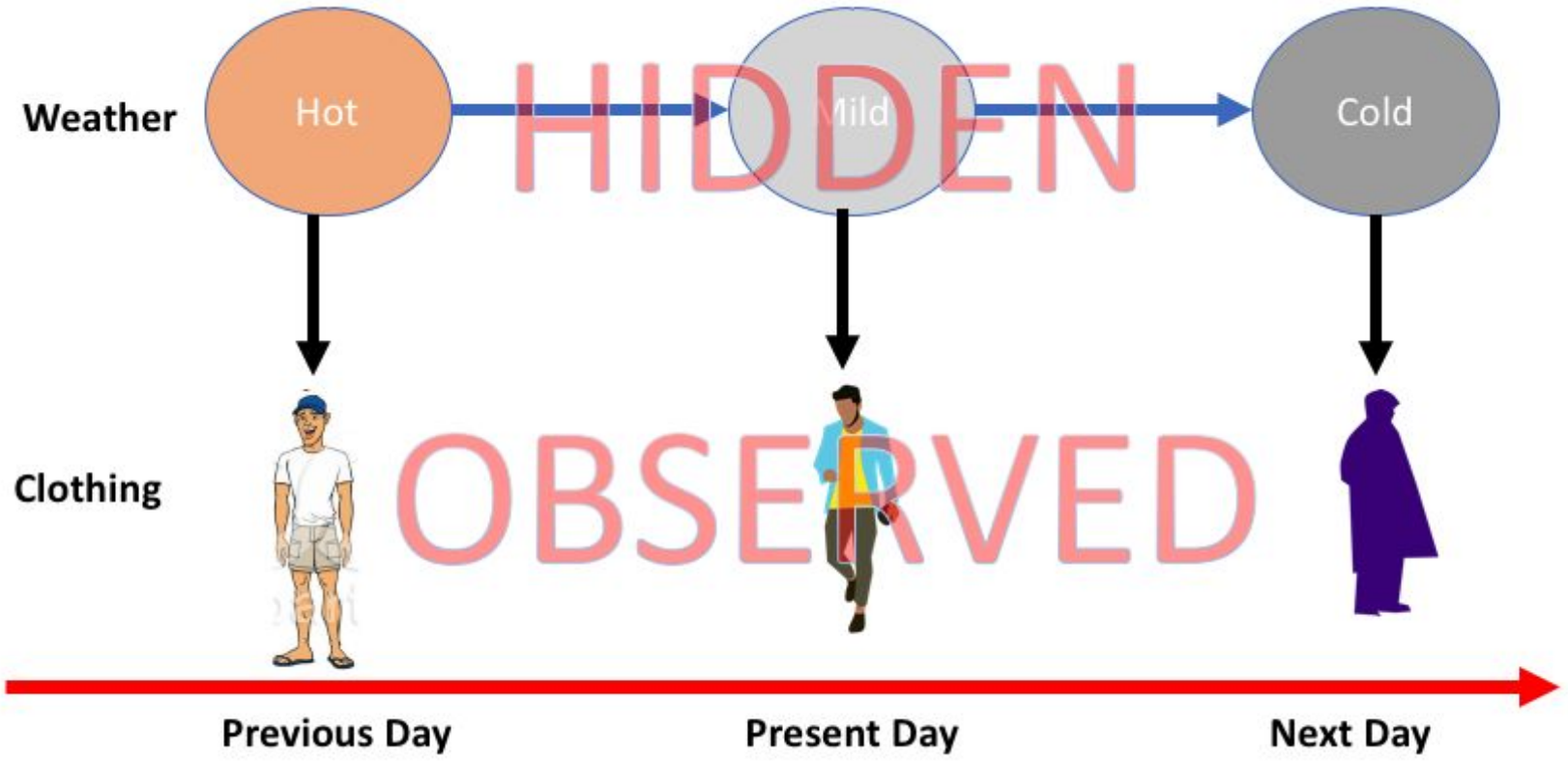


Local Alignment

	Smith-Waterman algorithm	Needleman-Wunsch algorithm
Initialization	First row and first column are set to 0	First row and first column are subject to gap penalty
Scoring	Negative score is set to 0	Score can be negative
Traceback	Begin with the highest score, end when 0 is encountered	Begin with the cell at the lower right of the matrix, end at top left cell

Local vs. Global

Annotation step 2 -Gene calling (HMM)

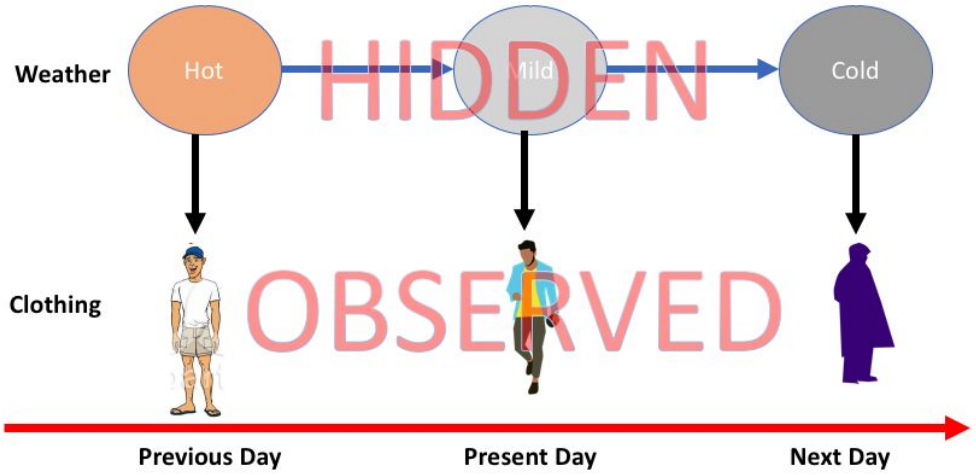


Annotation step 2 -Gene calling (HMM)

1) Transition data — the probability of transitioning to a new state conditioned on a present state.

2) Emission data — the probability of transitioning to an observed state conditioned on a hidden state.

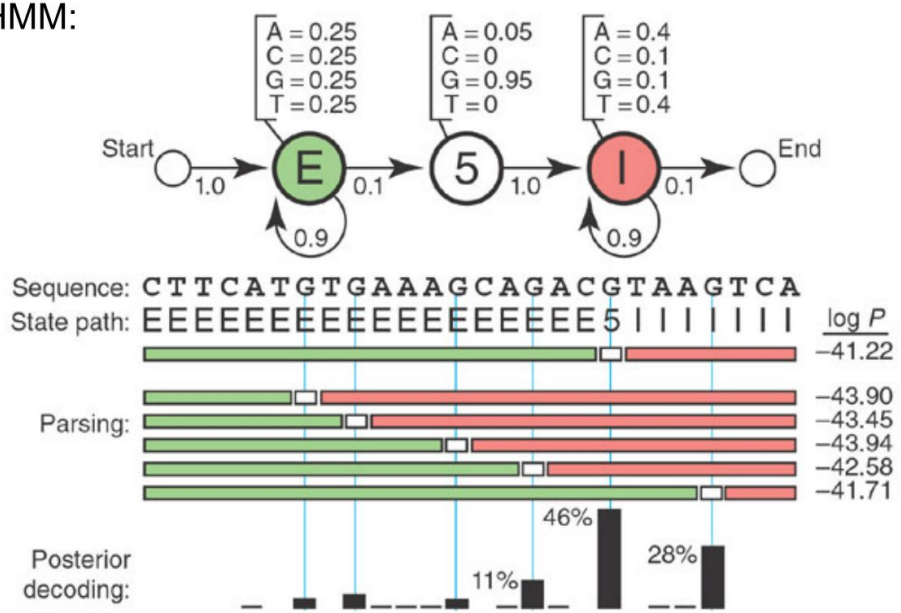
3) Initial state information — the initial probability of transitioning to a hidden state. This can also be looked at as the prior probability.



Annotation step 2 -Gene calling (HMM)

§ A ***Markov model*** (or ***chain***) is a probability model comprised of one or more ***states*** that generate a sequence of symbols (e.g., DNA: A, C, G, T)

Example HMM:



§ In a first order or simple Markov chain, the next state chosen depends only on the current state (not on any past states)

§ In a ***Hidden Markov Model (HMM)***, we cannot directly determine the current state from the generated (or emitted) symbol

Figure from SR Eddy *Nature Biotechnology* 22, 1315 - 1316 (2004)

Annotation step 2 -Gene calling (HMM)

§ A protein sequence profile HMM, the states correspond to ***matches***, ***insertions***, or ***deletions*** at different positions in the profile

Profile HMM:

Multiple Sequence Alignment:

1 2 3

C A F
C G W
C D Y
C V F
C K Y

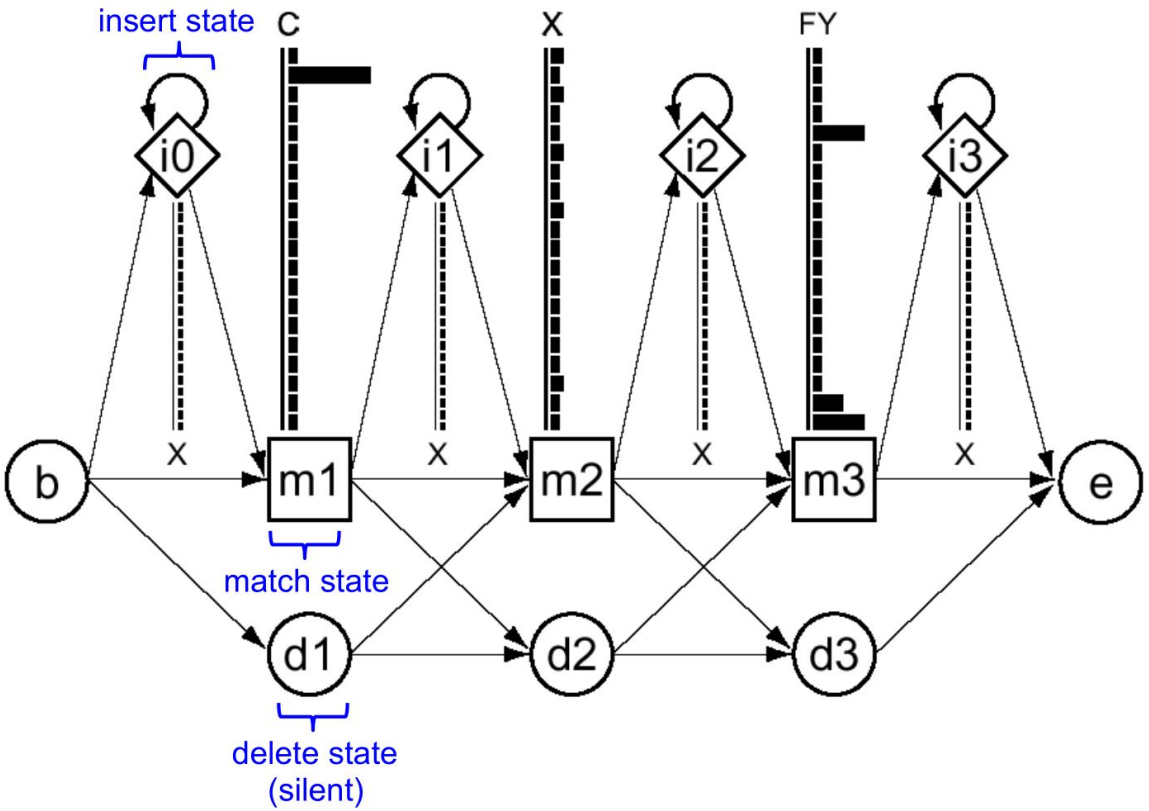
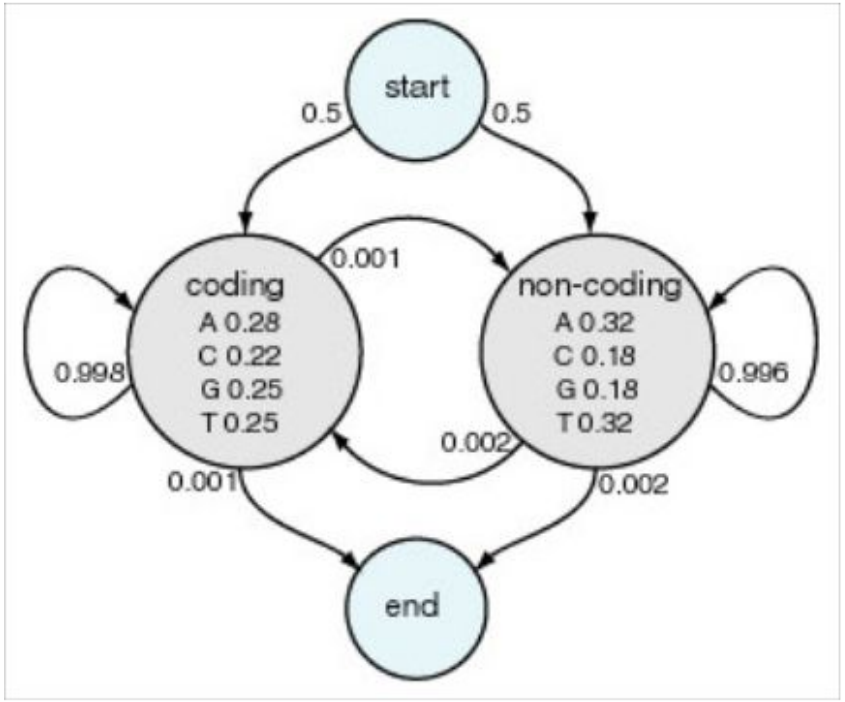


Figure from SR Eddy (1998) *Bioinformatics* 14:755-763
downloaded from: http://www.quretec.com/u/vilo/edu/2005-06/Text_Algorithms/L10_Probabilistic/HMM_intro_2.png

Annotation step 2 -Gene calling (HMM)



$$\Phi = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.5 & 0.998 & 0.002 & 0 \\ 0.5 & 0.001 & 0.996 & 0 \\ 0 & 0.001 & 0.002 & 0 \end{bmatrix}$$

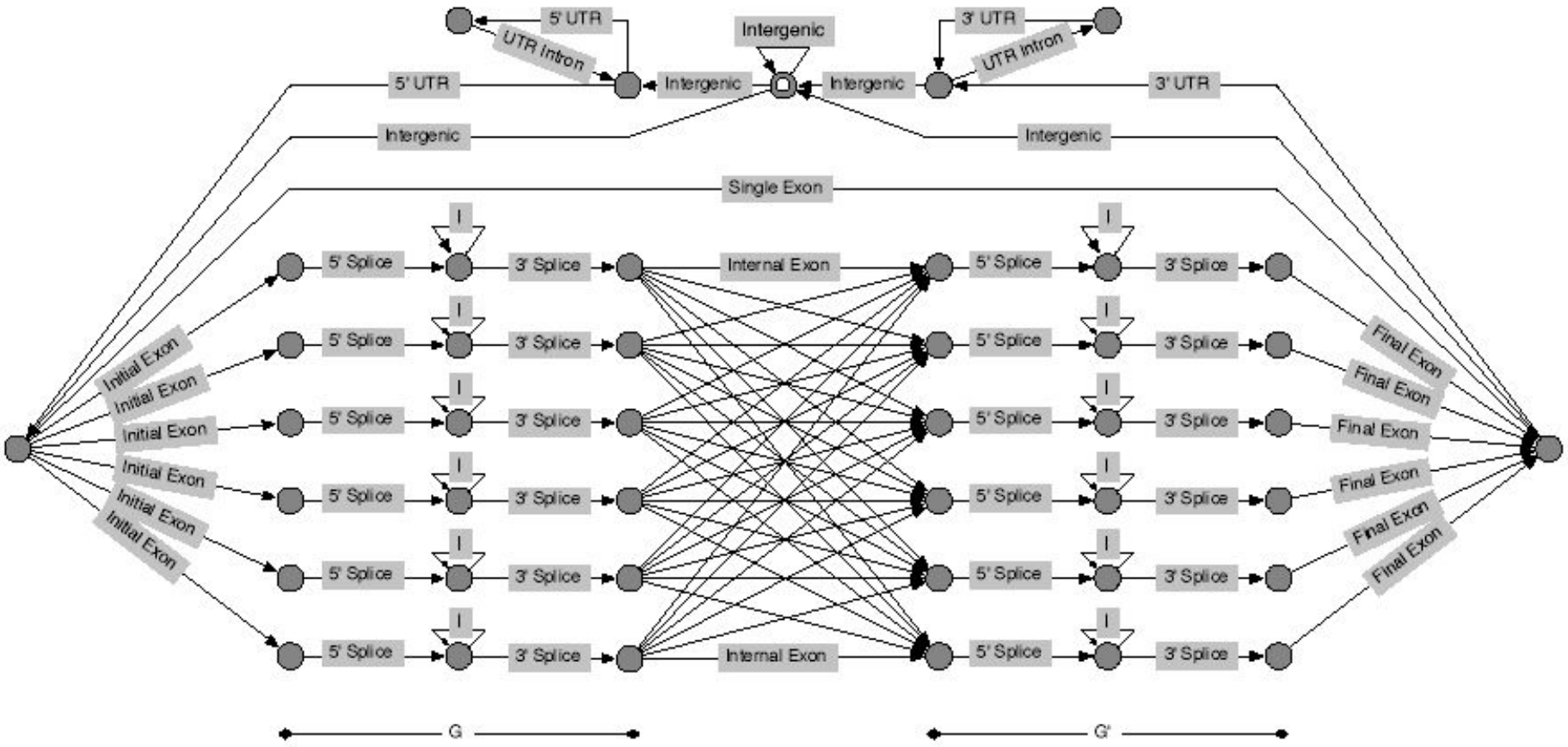
$$H = \begin{bmatrix} 0.28 & 0.32 \\ 0.22 & 0.18 \\ 0.25 & 0.18 \\ 0.25 & 0.32 \end{bmatrix}$$

$x_m(i)$ = probability of being in state m at position i;

$H(m,y_i)$ = probability of emitting character y_i in state m;

Φ_{mk} = probability of transition from state k to m.

Annotation step 2 -Gene calling (HMM)

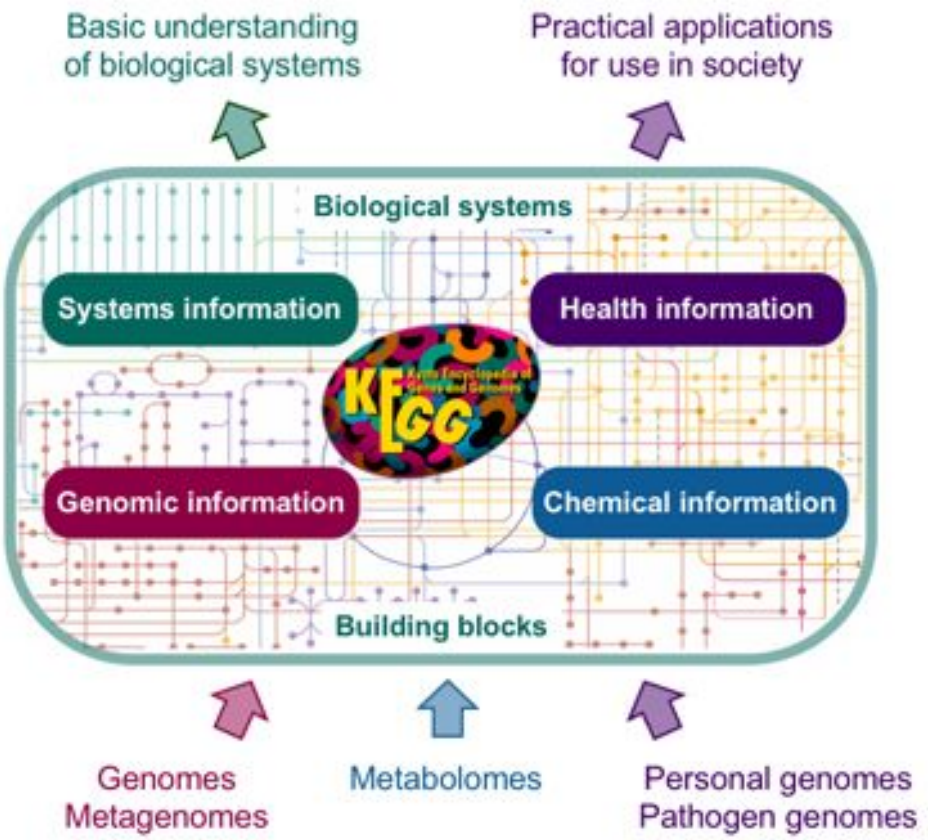


Kulp, D., PhD Thesis, UCSC 2003

Only as good as your database!!

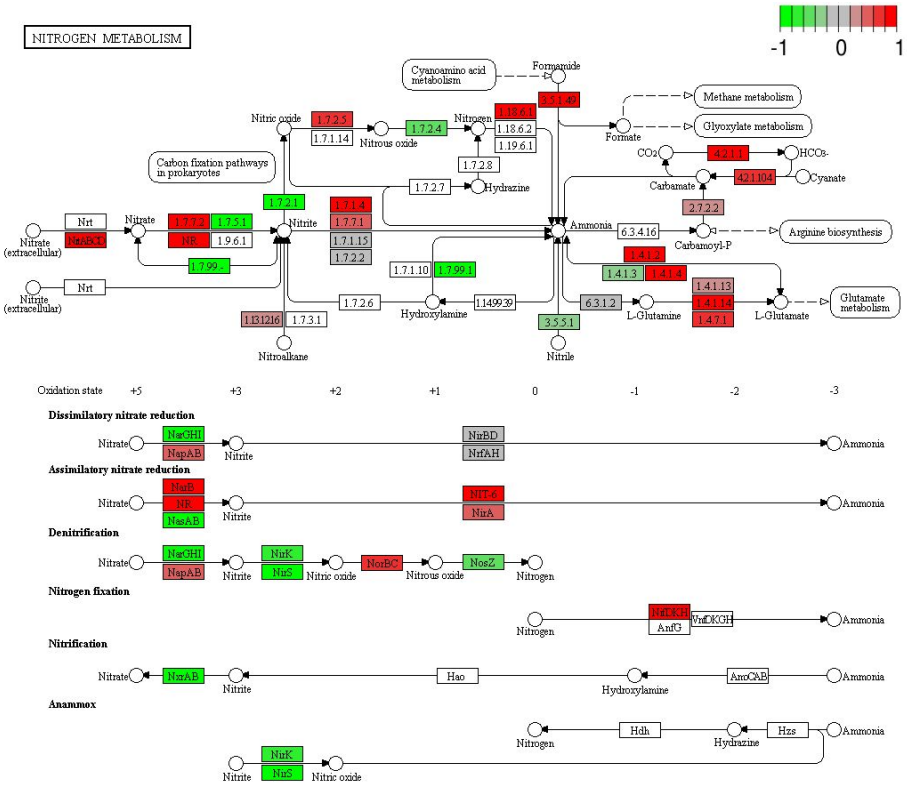
- Only as good as your database reference
  - Functional gene database
    - SEED
    - KEGG (KO)
    - COG
    - NOG, now eggNOG
    - Uniprot
    - CaZy
    - MetaCyc
  - Taxonomic identification database
    - GenBank
    - RefSeq
    - Uniprot
    - GTDB

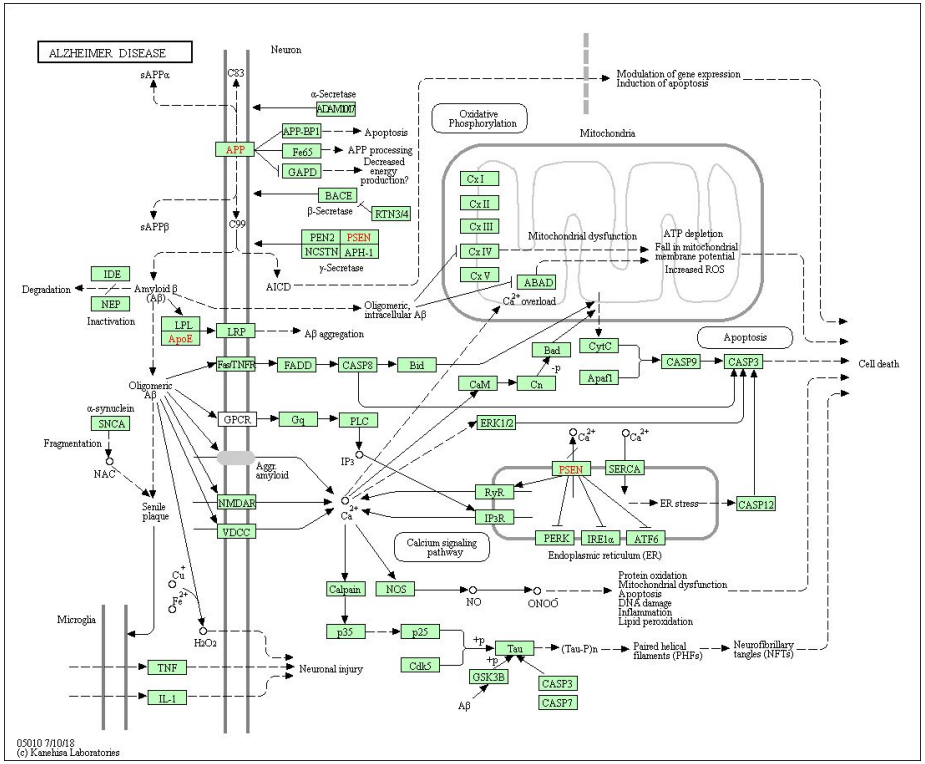
Only as good as your database (KEGG)



https://www.genome.jp/kegg/kegg1a.html

Only as good as your database (KEGG)

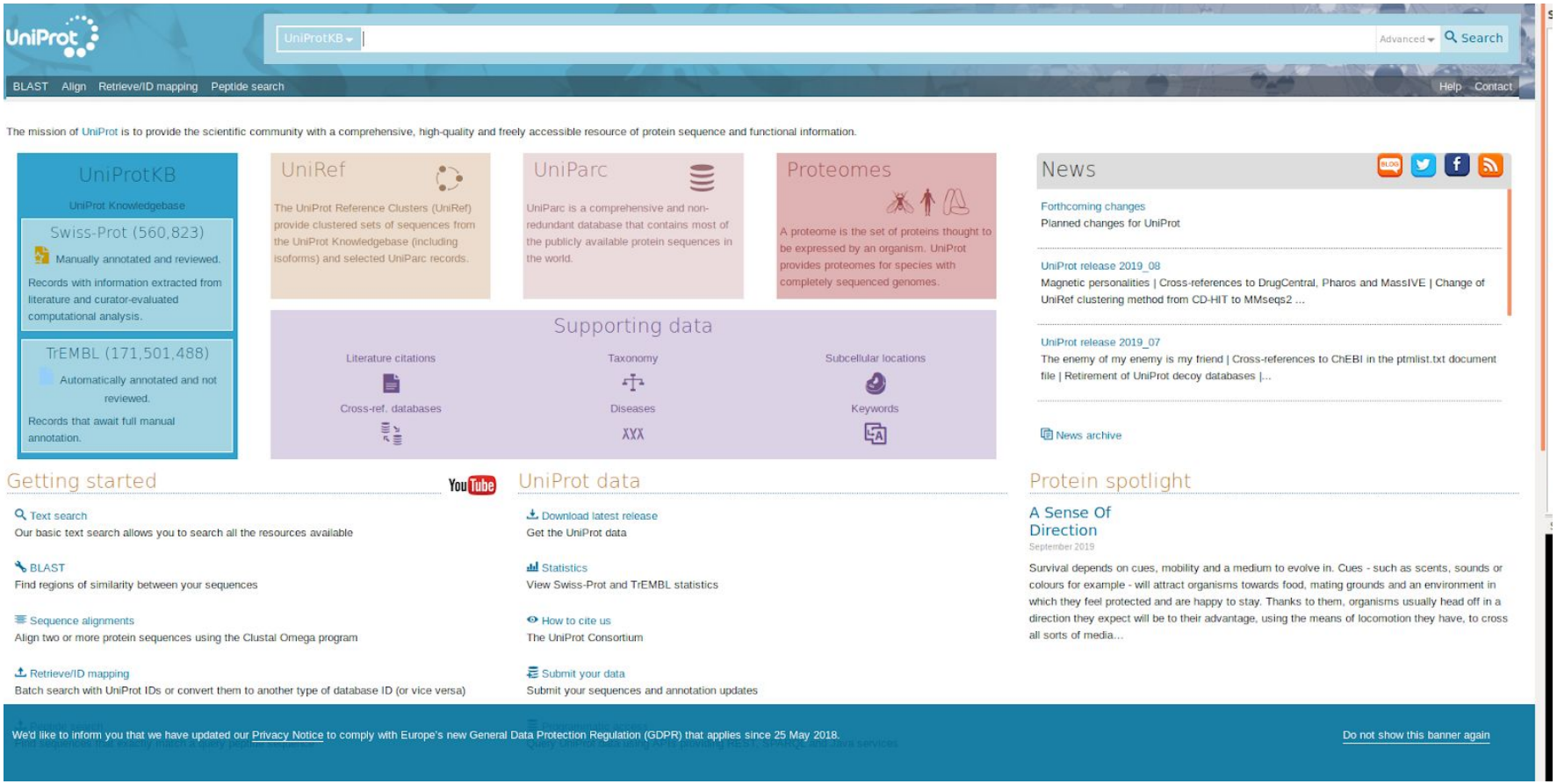




https://www.genome.jp/kegg/kegg1a.html

$5k a year!

Only as good as your database (Uniprot)



https://www.uniprot.org/

Only as good as your database (Uniprot)

UniProtKB

UniProt Knowledgebase

Swiss-Prot (560,823)

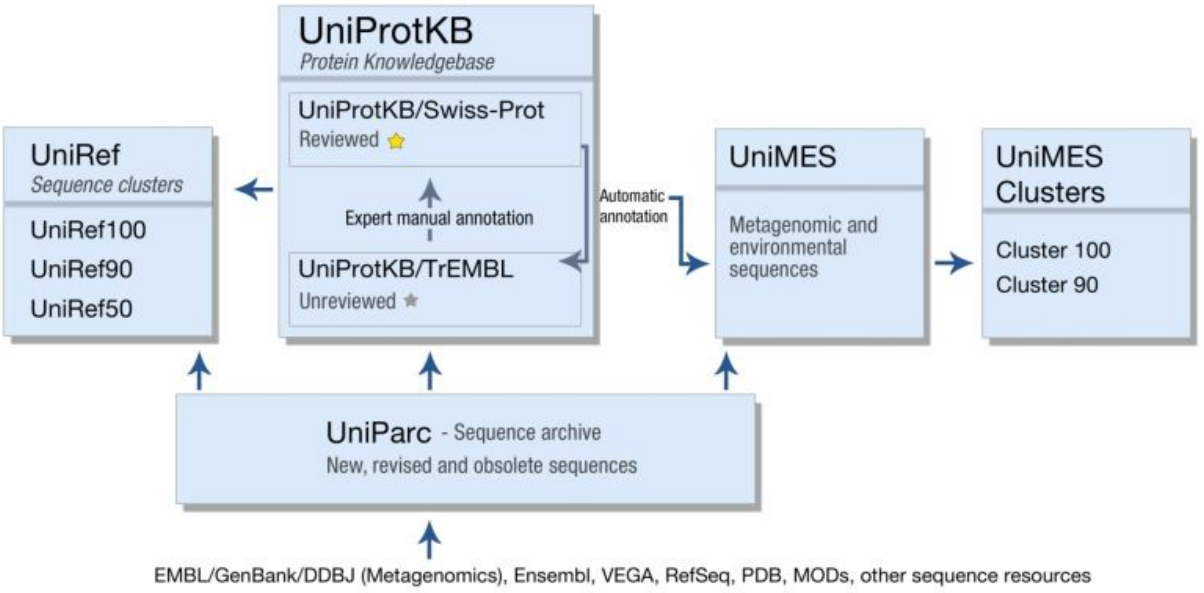
Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (171,501,488)

Automatically annotated and not reviewed.

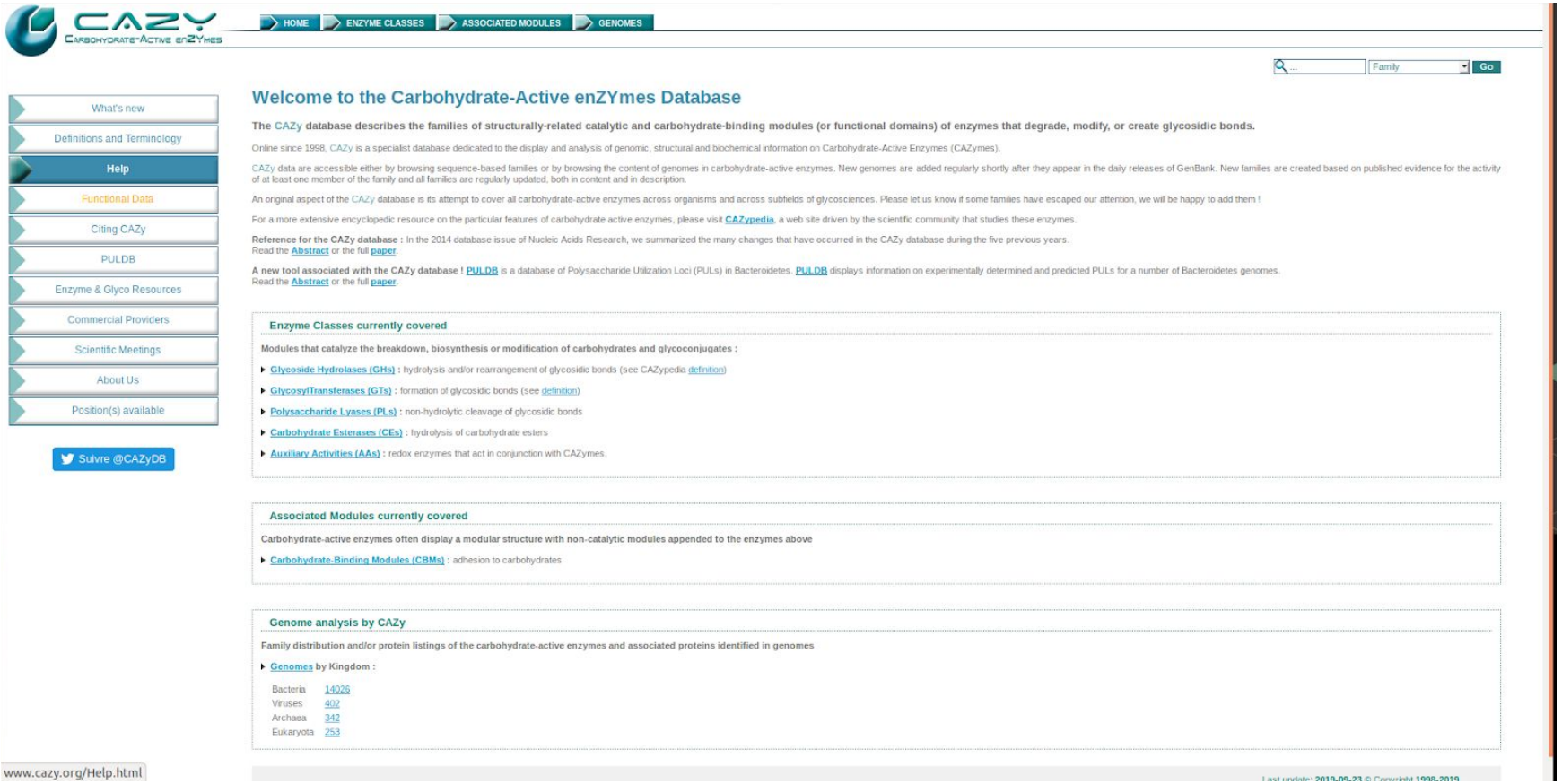
Records that await full manual annotation.



https://www.uniprot.org/

FREE!!!

Only as good as your database (CAZY)



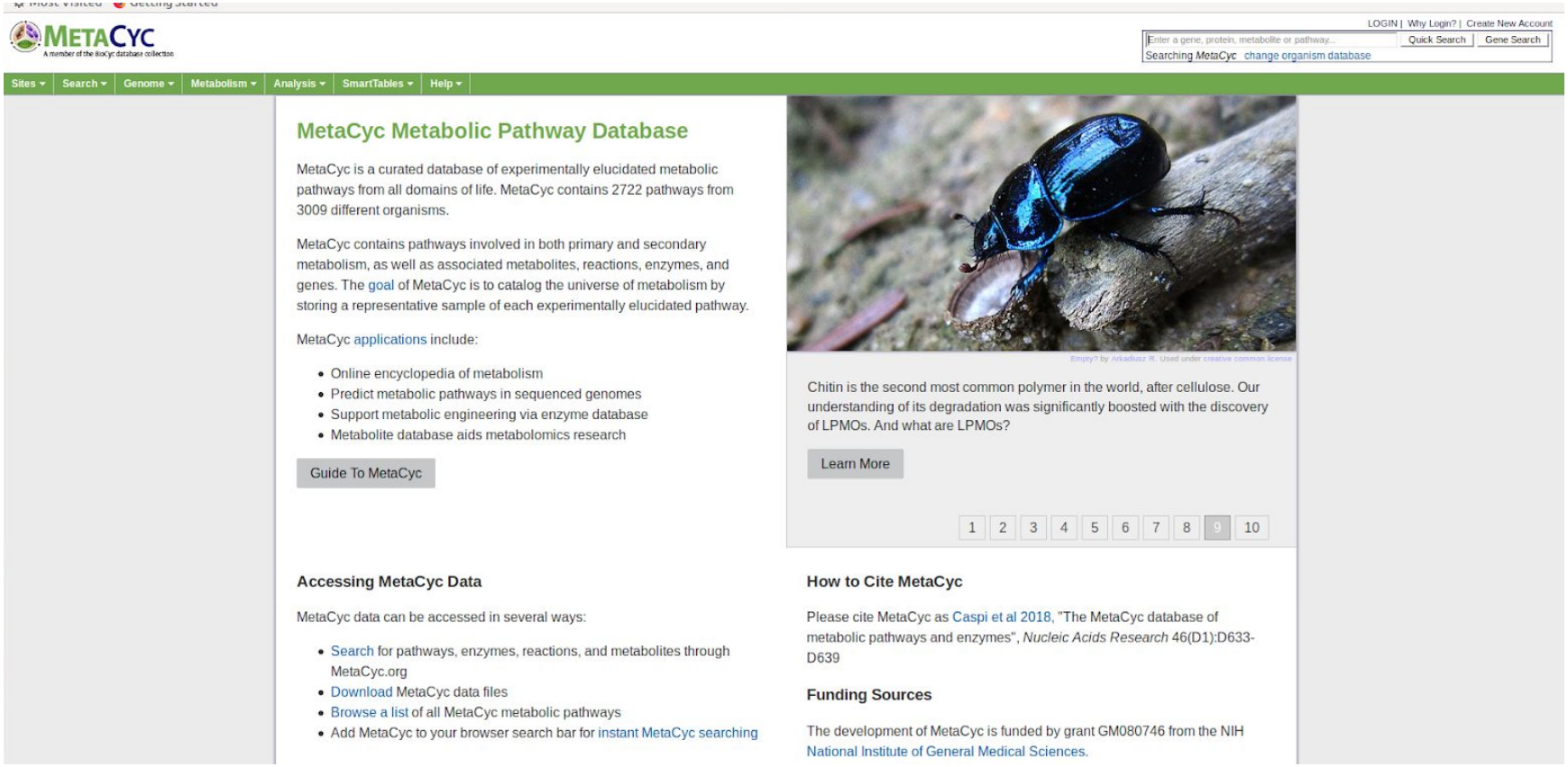
http://www.cazy.org/

FREE!!! BUT?

Only as good as your database (CAZY)

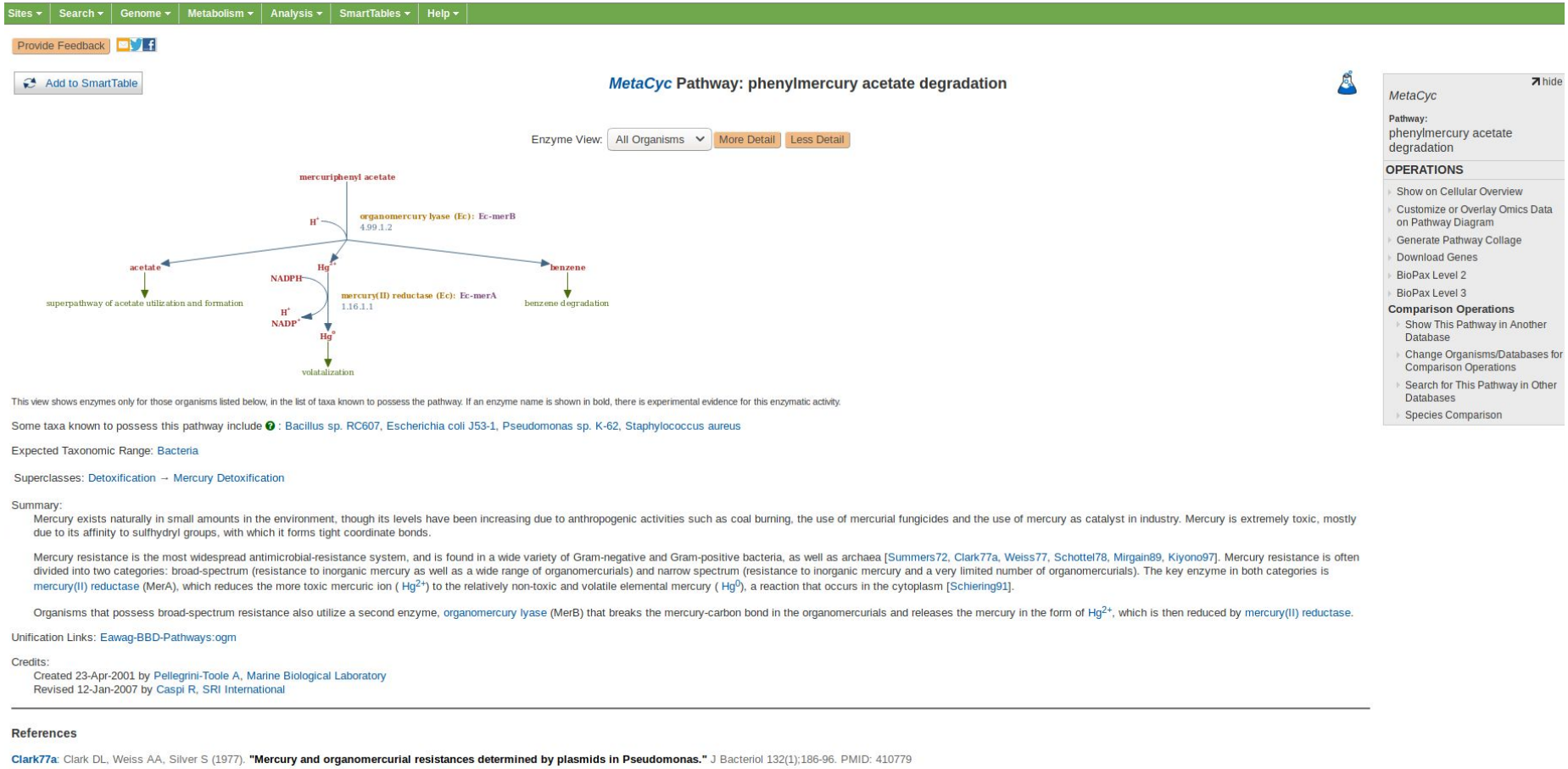
http://bcb.unl.edu/dbCAN2/

Only as good as your database (MetaCyc)



https://metacyc.org/

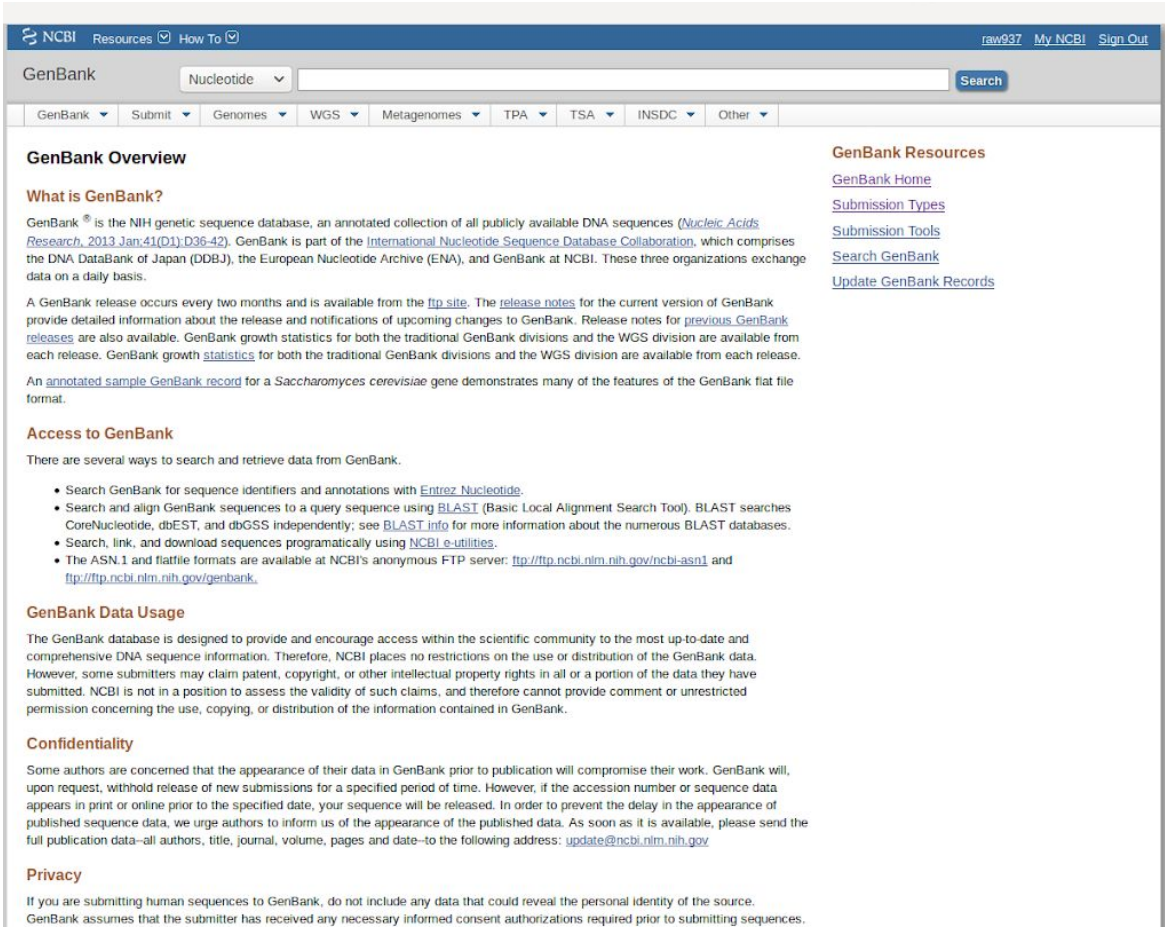
Only as good as your database (MetaCyc)



https://metacyc.org/

BioCyc - Pay model like KEGG

Only as good as your database (GenBank)



https://www.ncbi.nlm.nih.gov/genbank/

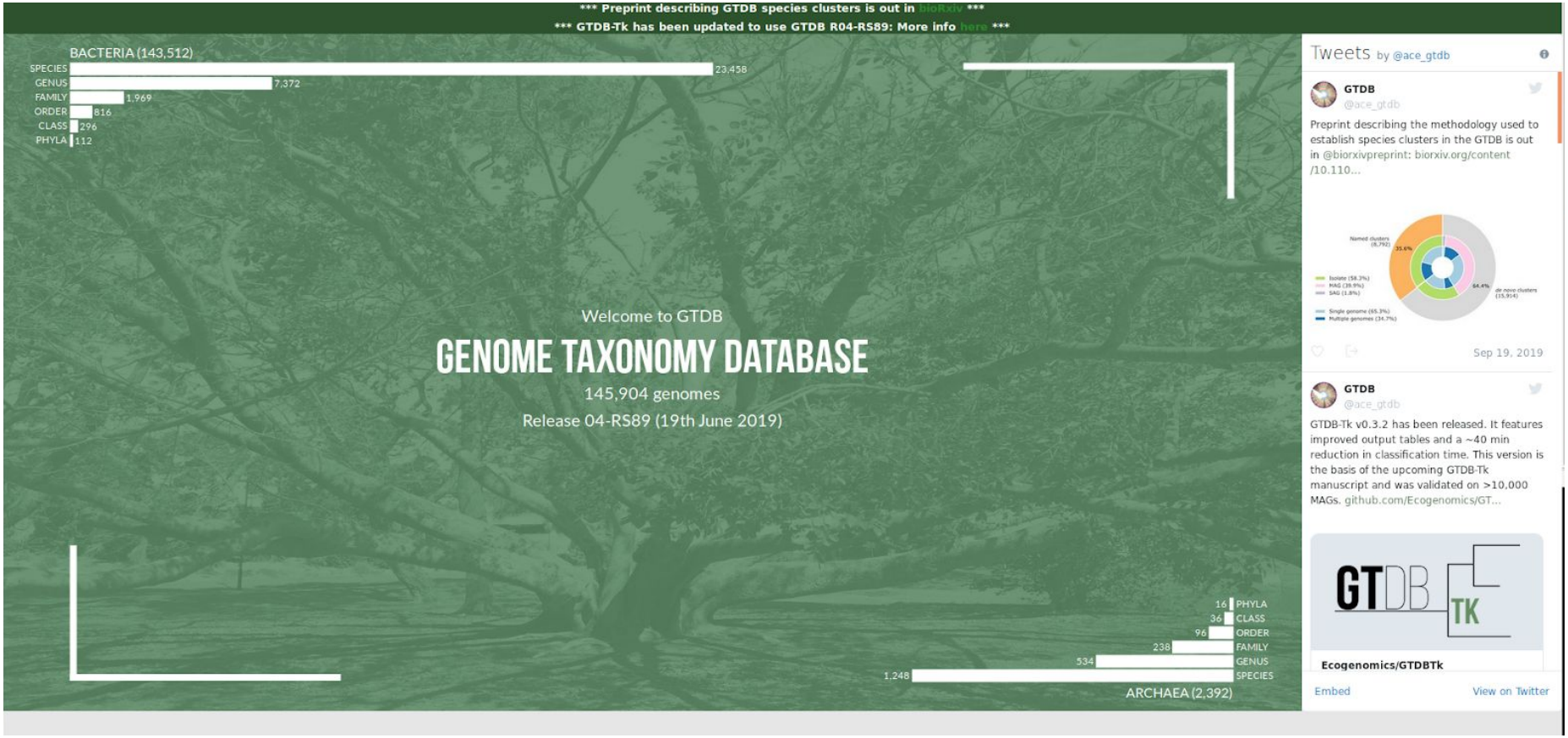
FREE!!! BUT!!

Only as good as your database (GenBank)

https://www.ncbi.nlm.nih.gov/genbank/

Landfill of data - can have errors (phiX)

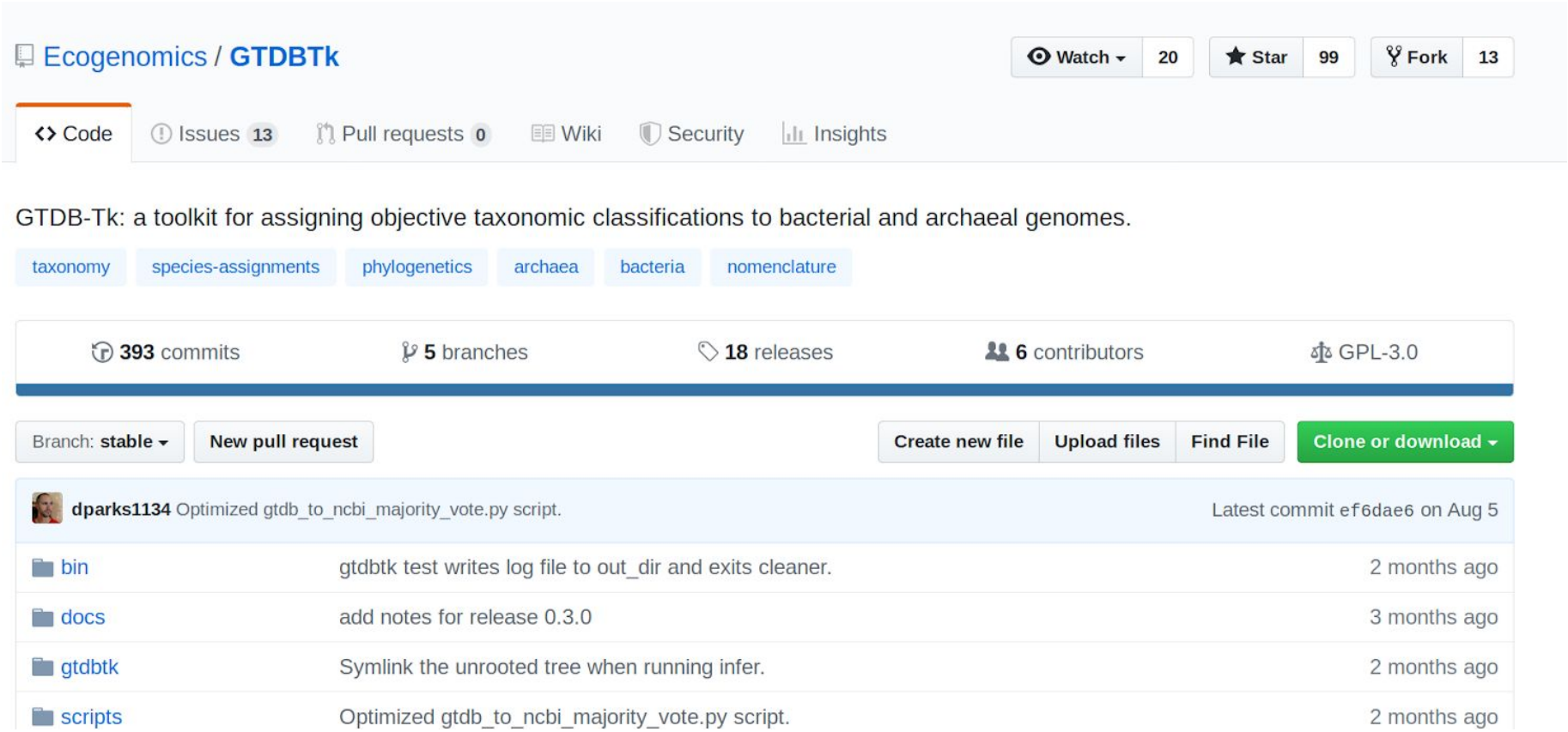
Only as good as your database (GTDB)



https://gtdb.ecogenomic.org/

Use this taxonomy!!

Only as good as your database (GTDB)



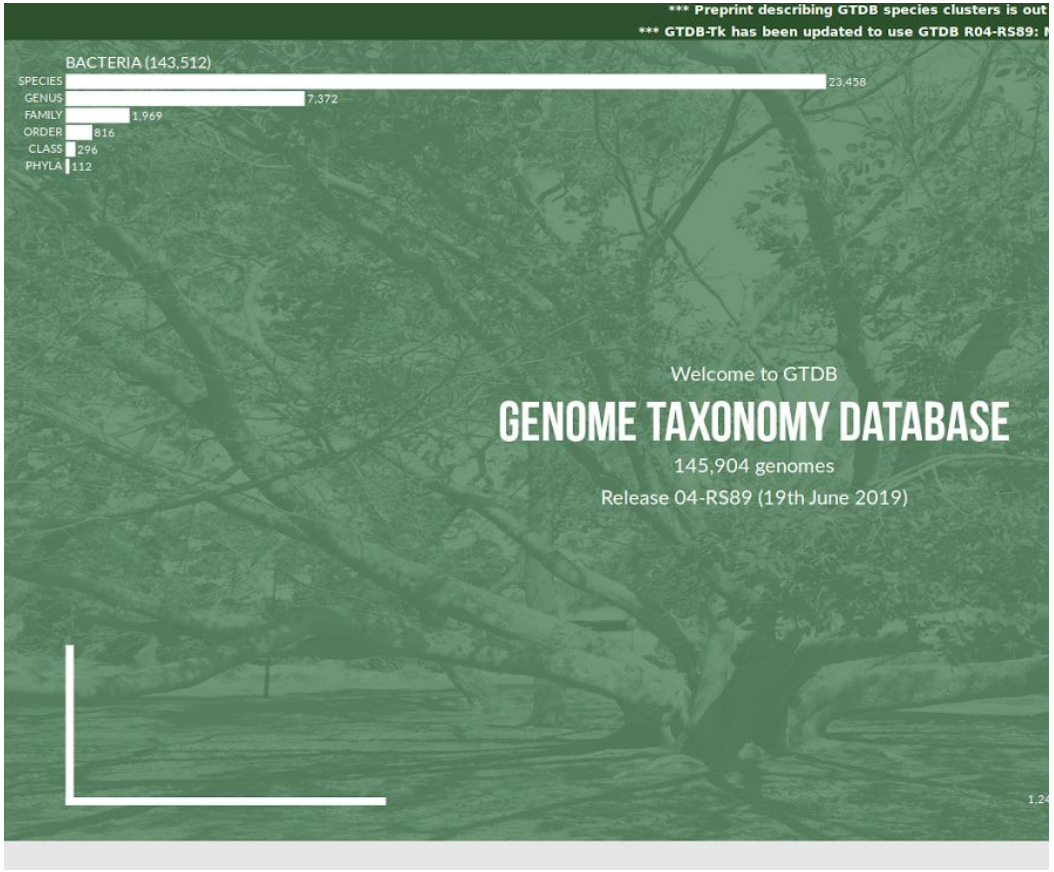
https://github.com/Ecogenomics/GtdbTk

Use this tool!

Only as good as your database (GTDB)

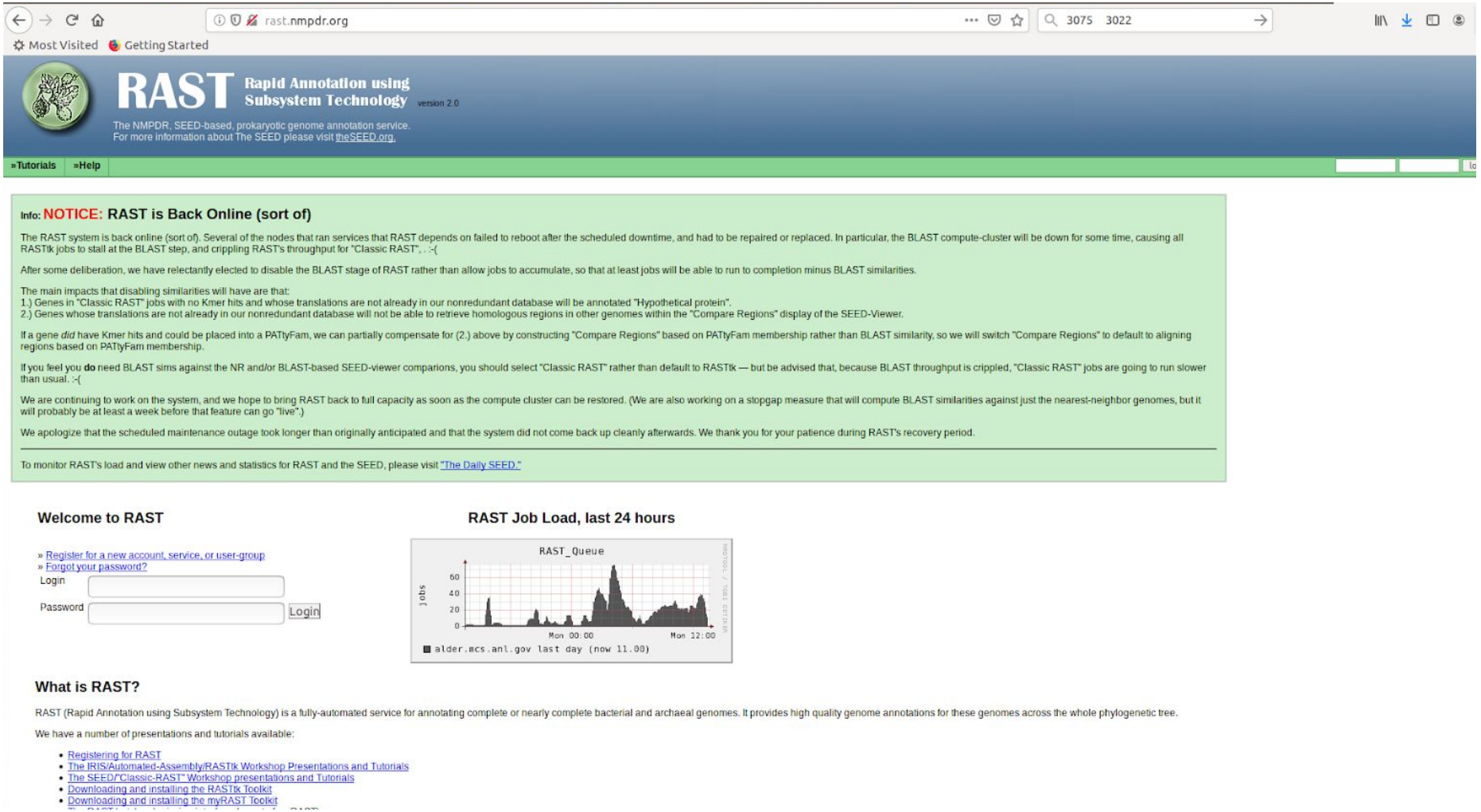
Name	Last modified	Size	Description
Parent Directory		-	
FILE_DESCRIPTIONS	11-Sep-2019 01:12	3.0K	
METHODS	17-Sep-2019 04:16	1.8K	
RELEASE_NOTES	09-Aug-2019 22:55	1.3K	
VERSION	20-Jun-2019 12:12	30	
ar122_metadata_r89.tsv	20-Jun-2019 12:12	2.9M	
ar122_msa_individual_genes_r89.tar.gz	19-Jun-2019 14:45	13M	
ar122_msa_marker_info_r89.tsv	20-Jun-2019 12:12	8.6K	
ar122_msa_mask_r89.txt	19-Jun-2019 14:45	32K	
ar122_msa_r89.faa	19-Jun-2019 14:45	6.1M	
ar122_r89.sp_labels.tree	17-Sep-2019 04:03	68K	
ar122_r89.tree	09-Aug-2019 22:48	56K	
ar122_ssu_r89.fna	20-Jun-2019 12:12	1.2M	
ar122_taxonomy_r89.tsv	20-Jun-2019 12:12	334K	
bac120_metadata_r89.tsv	20-Jun-2019 12:12	197M	
bac120_msa_individual_genes_r89.tar.gz	19-Jun-2019 14:45	343M	
bac120_msa_marker_info_r89.tsv	20-Jun-2019 12:12	8.2K	
bac120_msa_mask_r89.txt	19-Jun-2019 14:45	40K	
bac120_msa_r89.faa	19-Jun-2019 14:45	113M	
bac120_r89.sp_labels.tree	17-Sep-2019 04:03	1.2M	
bac120_r89.tree	19-Jun-2019 14:45	1.0M	
bac120_ssu_r89.fna	20-Jun-2019 12:12	26M	
bac120_taxonomy_r89.tsv	20-Jun-2019 12:12	20M	
gtdb_uba_mags.tar.gz	19-Jun-2019 14:46	2.1G	
gtdb_uba_mags_arc.tar.gz	19-Jun-2019 14:46	93M	
gtdbtk_r89_data.tar.gz	19-Jun-2019 17:53	26G	
metadata_field_desc.tsv	13-Sep-2019 00:18	6.6K	
ncbi_vs_gtdb_r89_archaea.xlxs	30-Aug-2019 07:45	31K	
ncbi_vs_gtdb_r89_bacteria.xlxs	30-Aug-2019 07:45	281K	
sp_clusters_r89.tsv	20-Jun-2019 12:12	6.7M	
ssu_r89.fna	20-Jun-2019 12:12	401M	
synonyms_r89.tsv	30-Aug-2019 11:42	74K	

Apache/2.2.15 (CentOS) Server at data.ace.uq.edu.au Port 443



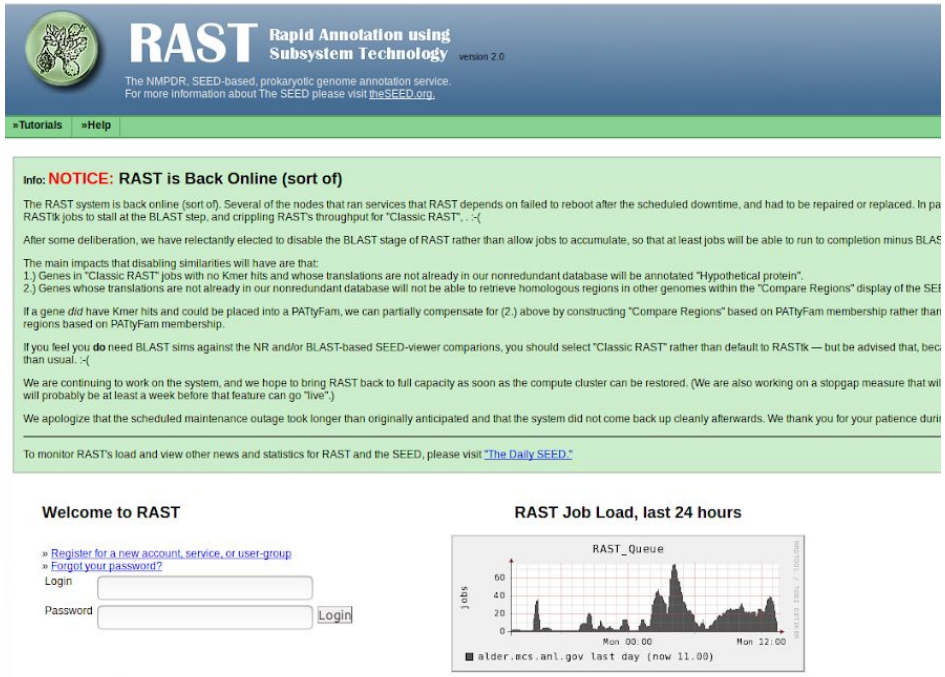
https://data.ace.uq.edu.au/public/gtdb/data/releases/release89/89.0/

Automatic annotator - RAST



http://rast.nmpdr.org/

Automatic annotator - RAST



Methods for RAST

- Glimmer
  - ORFs
- BLAST
  - Gene calling
- Database
  - FigFams
- Framework
  - SEED subsystems

Automatic annotator - PROKKA

README.md

build passing | License GPL v3 | DOI 10.1093/bioinformatics/btu153 | Language Perl 5

Prokka: rapid prokaryotic genome annotation

Introduction

Whole genome annotation is the process of identifying features of interest in a set of genomic DNA sequences, and labelling them with useful information. Prokka is a software tool to annotate bacterial, archaeal and viral genomes quickly and produce standards-compliant output files.

Installation

Brew

If you are using the MacOS Brew or LinuxBrew packaging system:

```
brew install brewsci/bio/prokka
```

https://github.com/tseemann/prokka

Automatic annotator - PROKKA

Prokka: rapid prokaryotic genome annotation

Introduction

Whole genome annotation is the process of identifying features of interest in a set of genomic DNA sequences, and labelling them with useful information. Prokka is a software tool to annotate bacterial, archaeal and viral genomes quickly and produce standards-compliant output files.

Installation

Brew

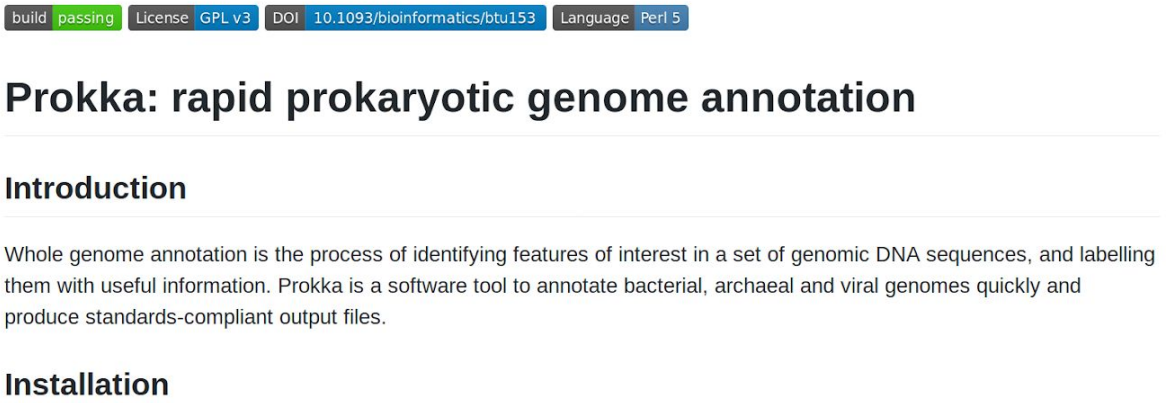
If you are using the MacOS Brew or LinuxBrew packaging system:

```
brew install brewsci/bio/prokka
```

Methods for PROKKA

- Prodigal
  - ORFs
- BLAST+ and hmmer
  - BLAST+
    - 70%
  - Hmmer
    - 30%
- Database
  - UniProtKB/COG
- Framework
  - COG/UniProt

Automatic annotator - PROKKA



Install:
conda install -c conda-forge -c bioconda -c defaults prokka

Command:
prokka Bradyrhizobium_USDA3458.fna --cpu N --outdir USDA3458 --prefix USDA3458 --rfam

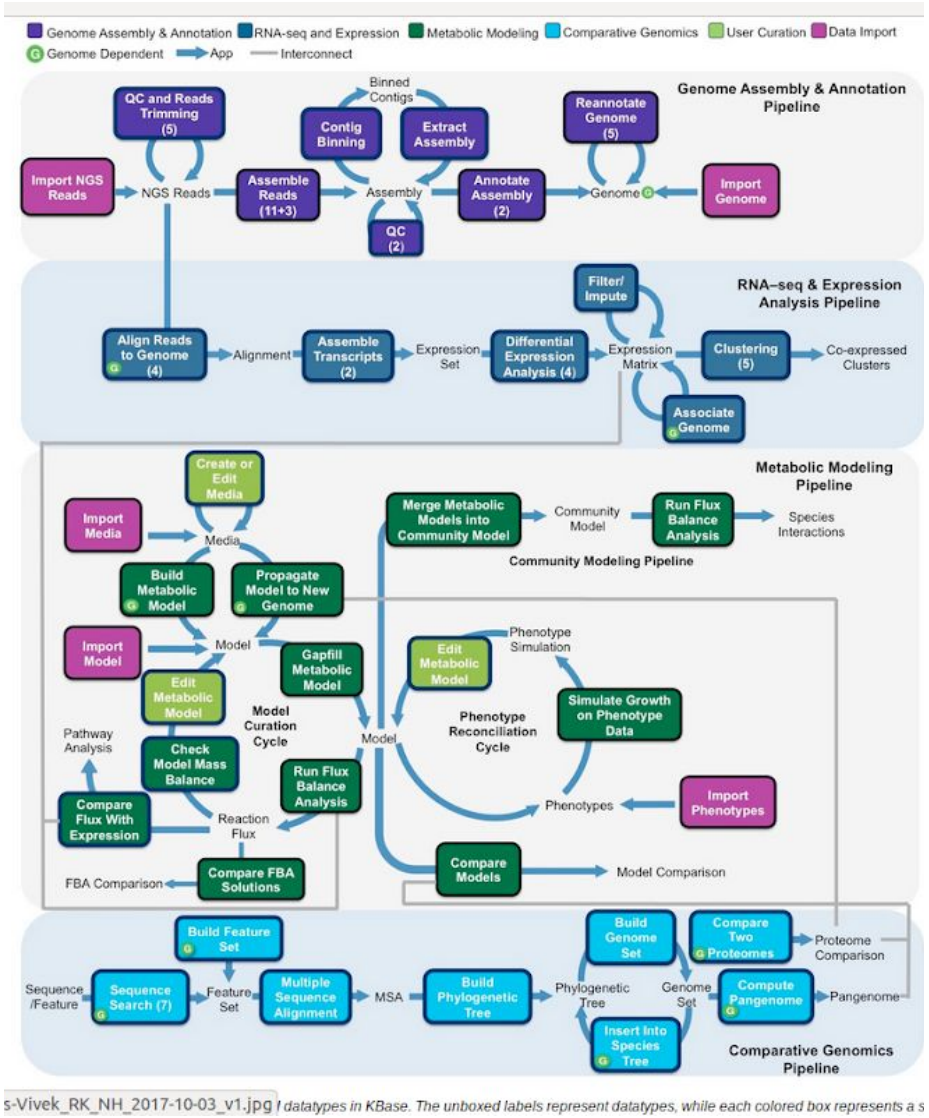
predicted 56 tRNAs, 1 transfer-messenger RNA (tmRNA), 37 noncoding RNAs (misc_RNA), 1 copy of a 5S-16S-23S operon, 0 CRISPRs, and 8,018 coding genes

White III et al., 2019. https://mra.asm.org/content/8/38/e00813-19.full

Automatic annotator - kBASE

https://kbase.us/

Automatic annotator - kBASE



s-Vivek_RK_NH_2017-10-03_v1.jpg *datatypes in KBase. The unboxed labels represent datatypes, while each colored box represents a s*

https://kbase.us/