



Quiz

- Qué es un k-mer (2pts)?
- Cuál es la diferencia entre usar valores de K bajos y valores altos? (4pts).
- Cuál es la ventaja de hacer un ensamblaje híbrido sobre uno sólo con secuencias cortas (4 pts)?
- Extra: Defina el N50 y el L50 (2pts).



UNIVERSIDAD DE
COSTA RICA

Anotación de genomas bacterianos

Curso Genómica de procariontes

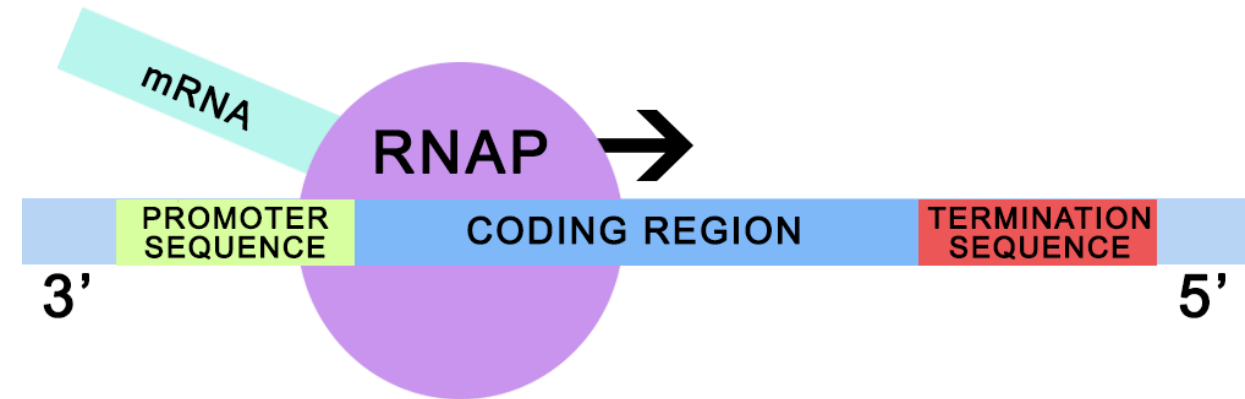
Bradd Mendoza Guido

Anotación

- Identificación de las propiedades funcionales de uno o más genes.
- La mayoría de los softwares de anotación se basan en bases de dato existentes (clase 2).
- Se realiza una comparación entre su secuencia y las presentes en la base de datos con funciones conocidas.



Identificación de secuencias codificantes



Adding biological info to sequences

ribosome binding site

delta toxin
PubMed: 15353161

transfer RNA
Leu-(UUR)

tandem repeat
CCGT x 3

homopolymer
10 x T

ACCGGCC **TAT** GCGAGCATATGCAGGAAGCGGCAGGAATAAGGA
 AAAGCAGCCTCCTGACTTTCTCGCTTGGTGGTTTGAGTGGACCTC
 CCAGGCCAGTGCCGGGCCCTCATAGGAGAGGAAGTCGGGAGGTG
 GCCAGGCGGCAGGAAGGCGCACCCCCCAGCAATCCGCGCGCCGGG
 ACAGAAATGCCCTGCAGGAACCTCTCTAGAAGACCTTCTCCTCCTG
 CAAATAAACCTCACCCATGAATGCTCACGCAAGTTTAATTACAGA
 CCTGAAACA **AGATGCCATTGTCCCCGGCCTCCTGCTGCTGCT**
 CT **CCGTCCGTCCGT** GGGCCACGGCCACCGC **TTTTTTTTTT** GCC

Blast

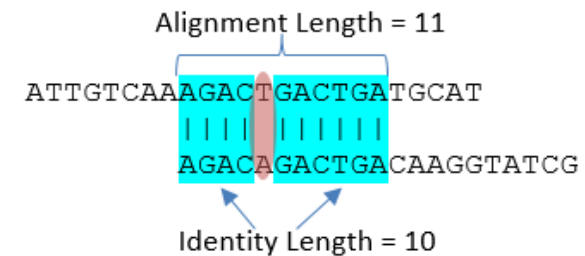
1. The **query** sequence is broken into "**words**" that will act as seeds in alignments



2. BLAST searches for matches (or synonyms) in **target** entries in the database



3. If a **target** entry has two or more matches to "**words**" from the query, the alignment is extended in both directions looking for additional similarity



$$\text{Alignment Identity \%} = \frac{\text{Identity Length}}{\text{Alignment Length}} = \frac{10}{11}$$

$$\text{Query Identity \%} = \frac{\text{Identity Length}}{\text{Query Length}} = \frac{10}{25}$$

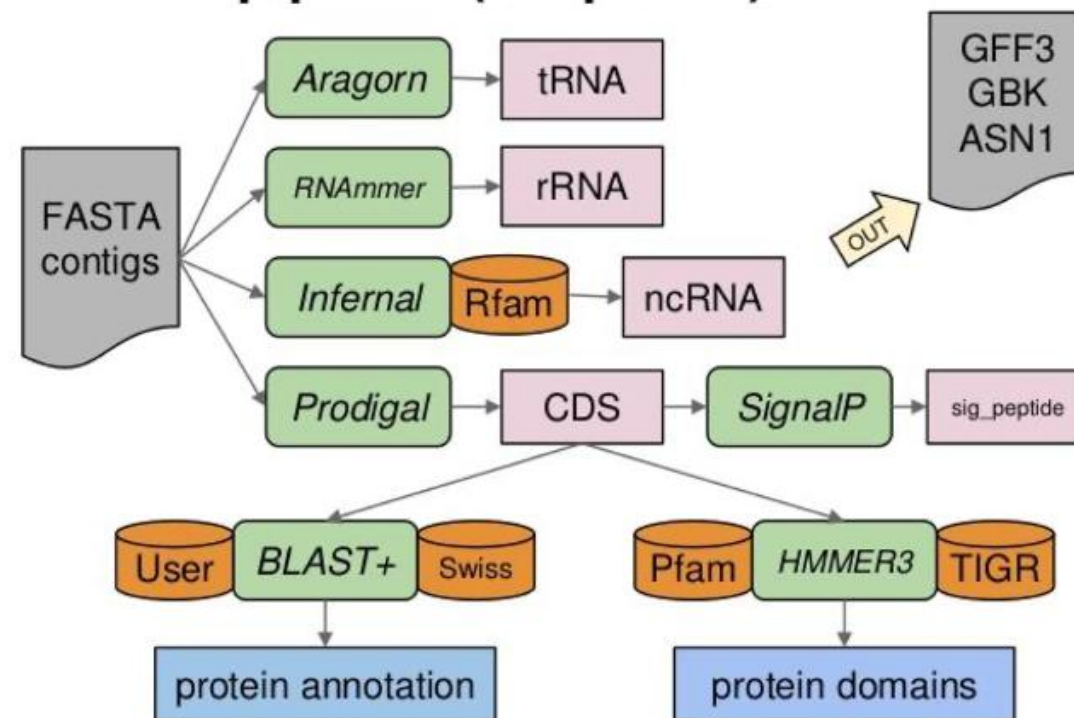
$$\text{Query Coverage \%} = \frac{\text{Alignment Length}}{\text{Query Length}} = \frac{11}{25}$$

$$\text{Subject Identity \%} = \frac{\text{Identity Length}}{\text{Subject Length}} = \frac{10}{21}$$

$$\text{Subject Coverage \%} = \frac{\text{Alignment Length}}{\text{Subject Length}} = \frac{11}{21}$$

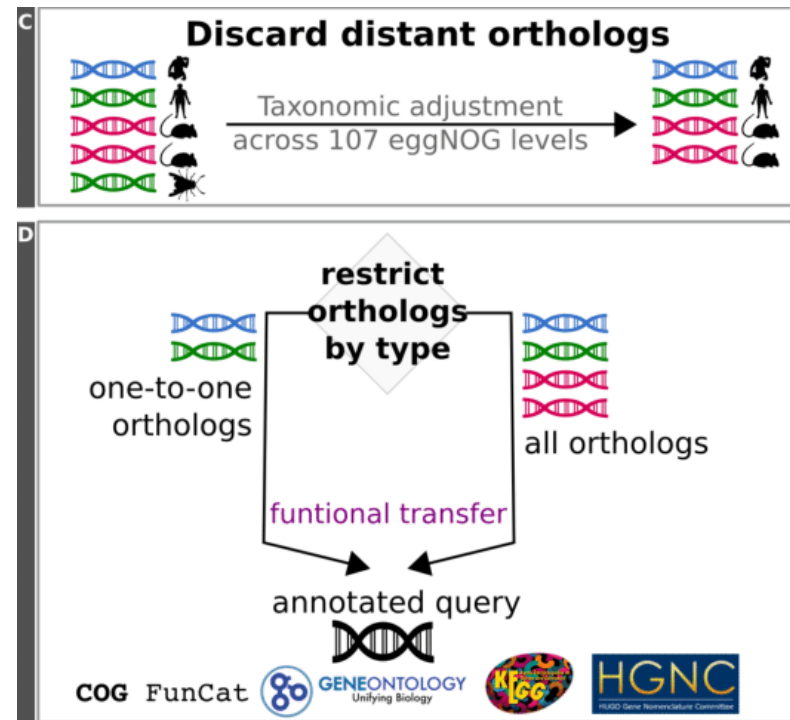
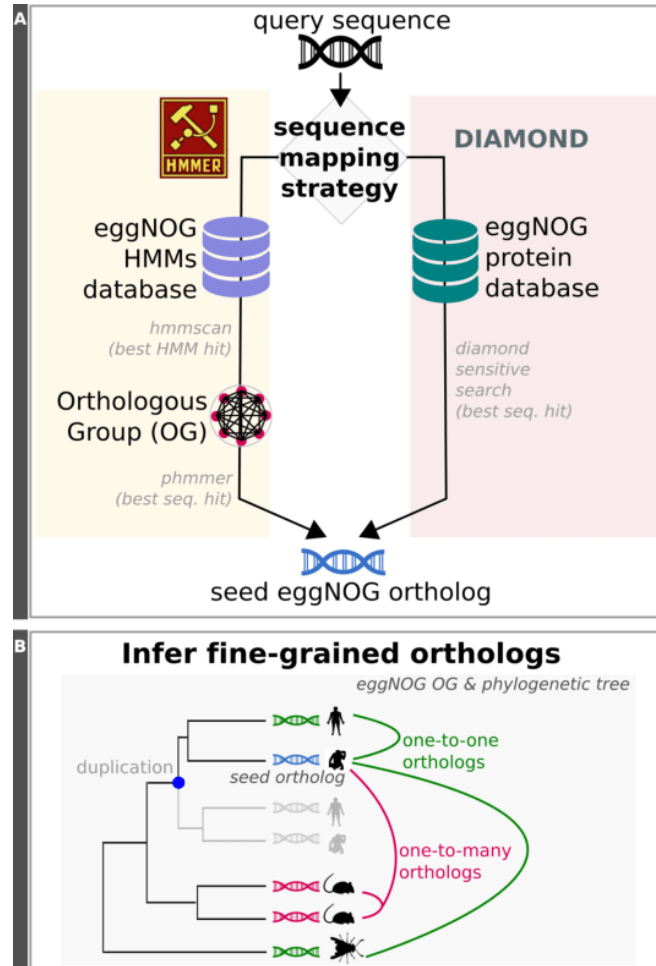
Prokka

Prokka pipeline (simplified)



Más sensitivo*

eggNOG-mapper



Herramientas especializadas

JOURNAL ARTICLE


NCycDB: a curated integrative database for fast and accurate metagenomic profiling of nitrogen cycling genes FREE

Qichao Tu ✉, Lu Lin ✉, Lei Cheng, Ye Deng, Zhili He

Bioinformatics, Volume 35, Issue 6, March 2019, Pages 1040–1048,

<https://doi.org/10.1093/bioinformatics/bty741>

Published: 28 August 2018 **Article history** ▼



dbCAN3 automated carbohydrate-active enzyme & substrate annotation

Home | Annotate | [GitHub](#) | Download | Example result | Help | About us | [AWS mirror site](#)

Cite us: [dbCAN3](#) | [dbCAN2](#) | [dbCAN](#)

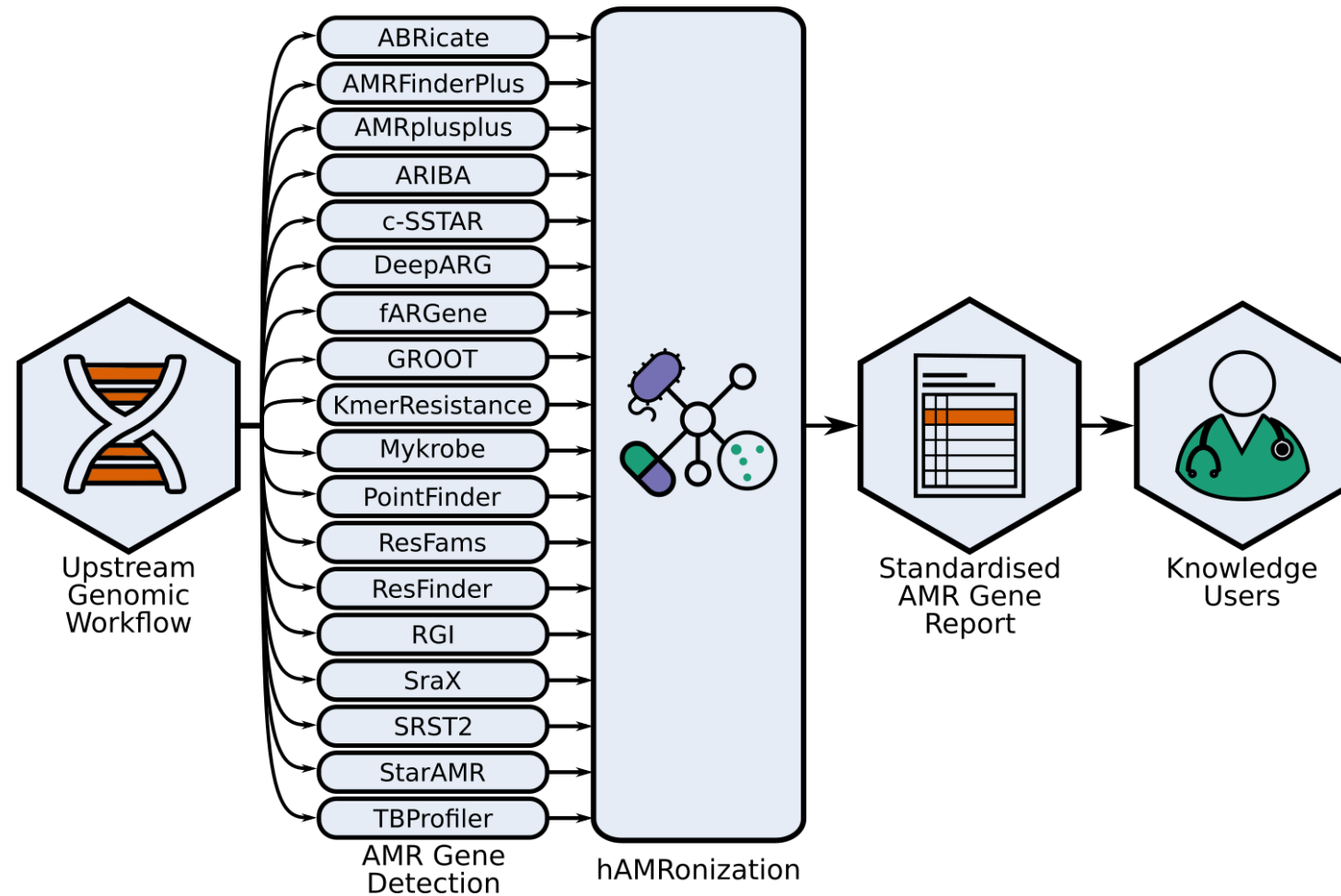
What is dbCAN3

dbCAN3 server is a web server for automated Carbohydrate-active enzyme **AN**notation, funded by the NSF (DBI-1933521) and NIH (R01GM140370). Similar resources on the web include [CAZy](#), [CAT](#) (obsolete), and [CUPP](#). dbCAN3 server is an updated version of [dbCAN](#) (obsolete) and [dbCAN2](#) (obsolete), and has the following [new features](#) (thanks to dbCAN users all over the world for suggestions):

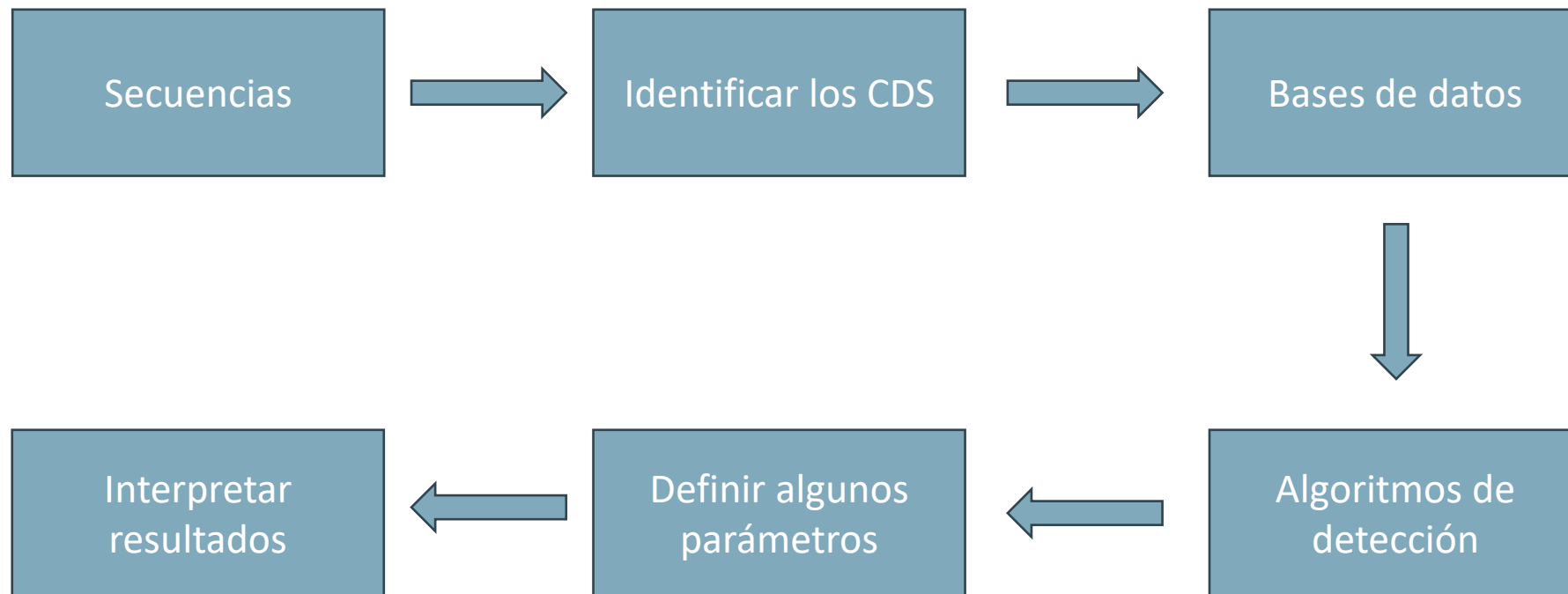
- dbCAN3 server allows users predict glycan substrates for CAZymes by searching against [dbCAN-sub](#), and for CAZyme gene clusters (CGCs) by using two approaches: searching against PULs of [dbCAN-PUL](#) and [dbCAN-sub](#) majority voting
- dbCAN3 server, like dbCAN2, allows submission of nucleotide sequences: prokaryotic genomes (fna file) or metagenome assembled genomes (MAGs); for eukaryotic genomes, please still submit protein seqs (faa file)
- dbCAN3 server, like dbCAN2, integrates three state-of-the-art tools/databases for automated CAZyme annotation:
 1. [HMMER](#) search for CAZyme family annotation vs. [dbCAN CAZyme domain HMM database](#)
 2. [DIAMOND](#) search for BLAST hits in the [CAZy database](#)
 3. [HMMER](#) search for CAZyme subfamily annotation vs. [dbCAN-sub](#) HMM database of CAZyme subfamilies (derived from [eCAMI](#) classification of CAZyDB families)
- dbCAN3 server can identify transcription factors (TFs), transporters (TCs), signal transduction proteins (STPs), and further CAZyme gene clusters (CGCs) using [CGC-Finder](#) if users submit faa+gff files or fna file
- dbCAN3 server combines the results from the three tools and allows visualization of detailed results as tables/graphs

[dbCAN3 server](#) will be updated once a year to use the most updated CAZy database, dbCAN HMMdb and dbCAN-sub HMMdb

Genes de resistencia



Qué necesitamos para un anotación?





Preguntas?