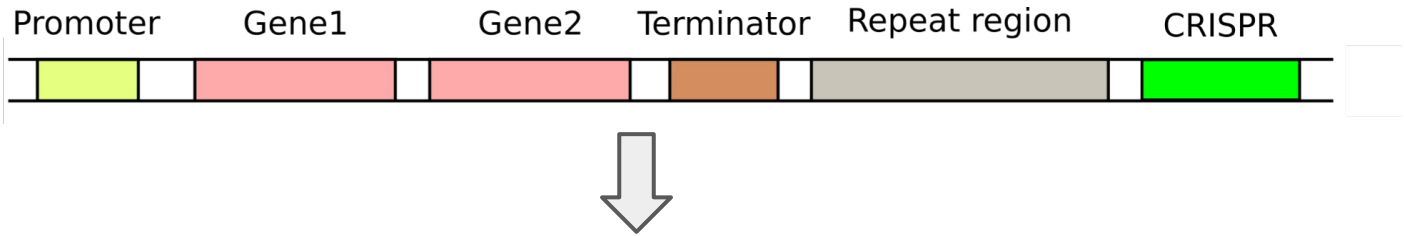
Genome annotation

Endrews Delbaje
29.03.2022

Genome annotation

The standardized identification and registry of functional elements in a genome sequence.



Labels, coordinates, functions…

It requires:

- Identification of all potential coding regions (CDS);
- Start and stop coordinates of the genes/structure in the genome;
- Identification of functions by homology (or if the function is unknown).

Identification of coding regions

Finding ORFs (Open Reading Frames) - Localization of start-stop codons

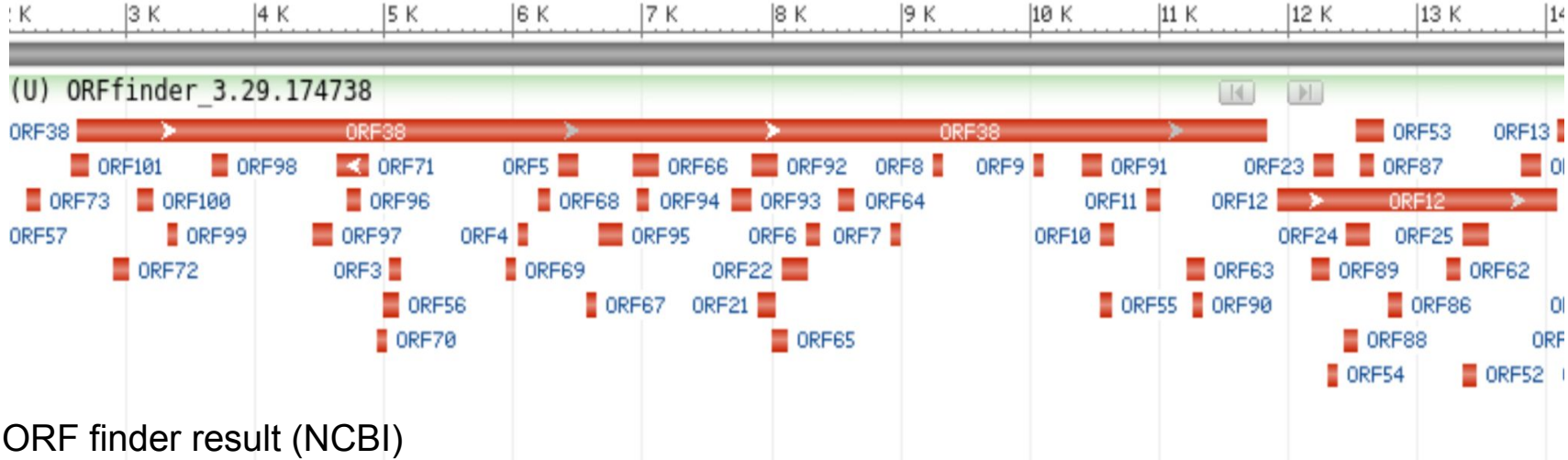
Example:

```
   ATGAGGTGACACCGCAAGCCTTATATTAGCTAA

 3 ATG AGG TGA CAC CGC AAG CCT TAT ATT AGC TAA
 2 A TGA GGT GAC ACC GCA AGC CTT ATA TTA GCT AA
 1 AT GAG GTG ACA CCG CAA GCC TTA TAT TAG CTA A

-1 TA CTC CAC TGT GGC GTT CGG AAT ATA ATC GAT T
-2 T ACT CCA CTG TGG CGT TCG GAA TAT AAT CGA TT
-3 TAC TCC ACT GTG GCG TTC GGA ATA TAA TCG ATT
```

Identification of coding regions

Finding ORFs (Open Reading Frames) - Localization of start-stop codons

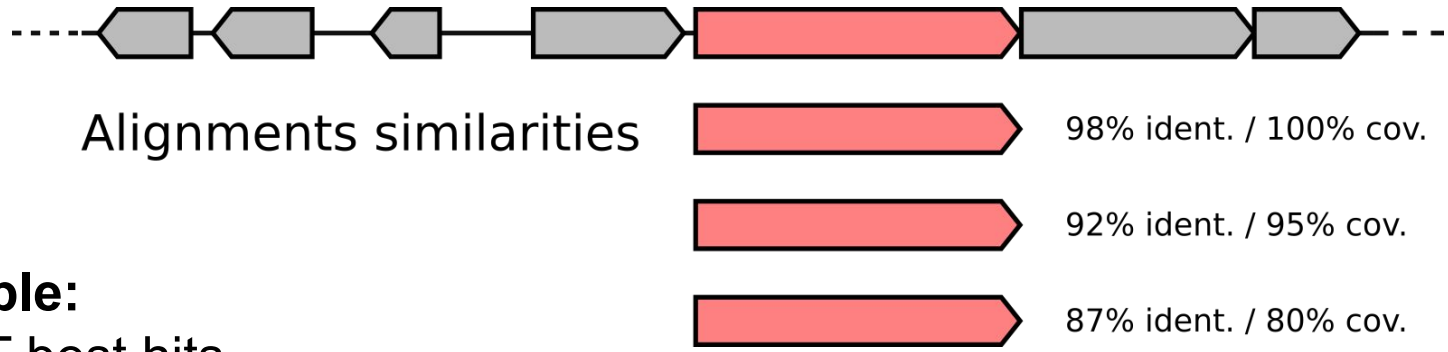
For the same DNA sequence there can be many possibilities of ORFs:



ORF finder result (NCBI)

Function assignment - Database search

Functional assignment by homology using the best database hits:



Example:
BLAST best hits

Public available databases

NCBI (GenBank): Varied sequence community oriented database;
ENA: Varied sequence community oriented database;
KEGG: Gene database curated and organized for pathways;
Pfam: Protein database organized by protein families;
UniProt: Partially curated protein database;
…

Nowadays the process is automatized - Genome annotation by programs/platforms:

PROKKA
Genome Annotation Pipeline (PGAP)
EggNOG
…

Genome annotation formats

GFF (GFF3) (general feature format)

One line per feature and 9 columns

Example:

seqname	source	feature	start	end	score	strand	phase	attribute
scaffold1	prokka	CDS	12000	12980	.	+	1	Amoa
scaffold1	prokka	tRNA	13000	13082	.	-	2	tRNA-Leu

…

Other formats:

GBK

Tables

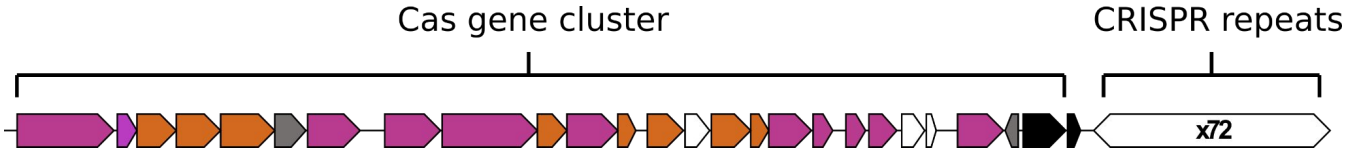
…

Specialized annotation

Normally for complex regions or meta-features

Biosynthetic gene clusters (e.g. AntiSMASH program):

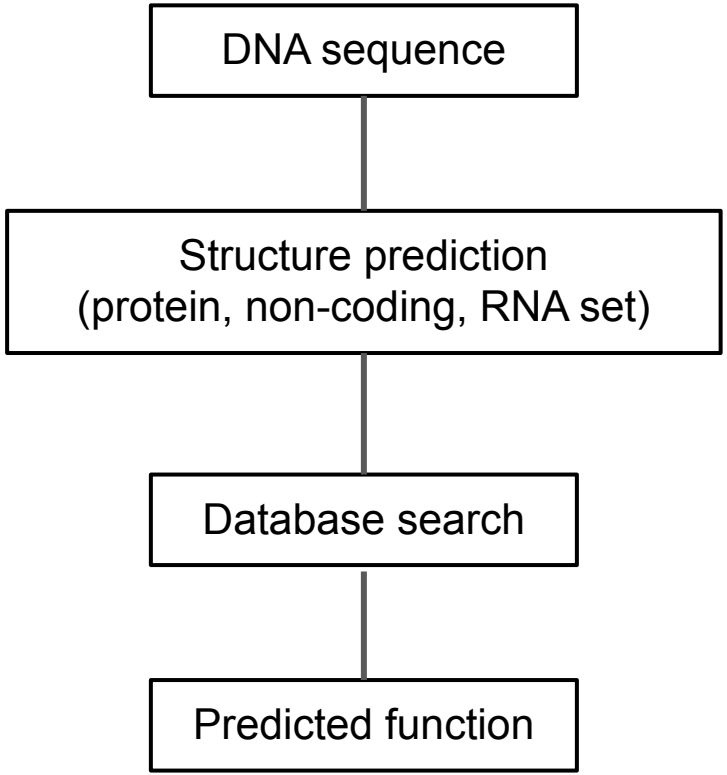
CRISPR/Cas (e.g. CRISPRone program):



Viral sequences, transposons, microRNA, etc.

Prokka: rapid prokaryotic genome annotation

Workflow:



Prodigal: ORF finding and translation;
Aragorn: tRNA;
Barrnap: rRNA.

Search with **BLAST+** and **HMMR3** in the databases:
ISfinder: transposases;
NCBI Bactarial antimicrobial;
UniProtKB: curated protein database.

Genome annotation with Prokka: Output files

```
strain193.faa
strain193.ffn
strain193.fna
strain193.fsa
strain193.gbf
strain193.gff
strain193.log
strain193.sqn
strain193.tbl
strain193.tsv
strain193.txt
strain193.val
```

Genome annotation with Prokka: Results summary

```
contigs: 1
bases: 4495168
CDS: 3873
gene: 3927
rRNA: 8
repeat_region: 9
tRNA: 45
tmRNA: 1
```

PROKKA: Annotation table (.tbl)

```
5776    5234    CDS
                EC_number 7.1.1.6
                db_xref   COG:COG0723
                gene    petC_1
                inference ab initio prediction:Prodigal:002006
                inference similar to AA sequence:UniProtKB:P0C8N8
                locus_tag GNOHDOCP_00006
                product   Cytochrome b6-f complex iron-sulfur subunit
```

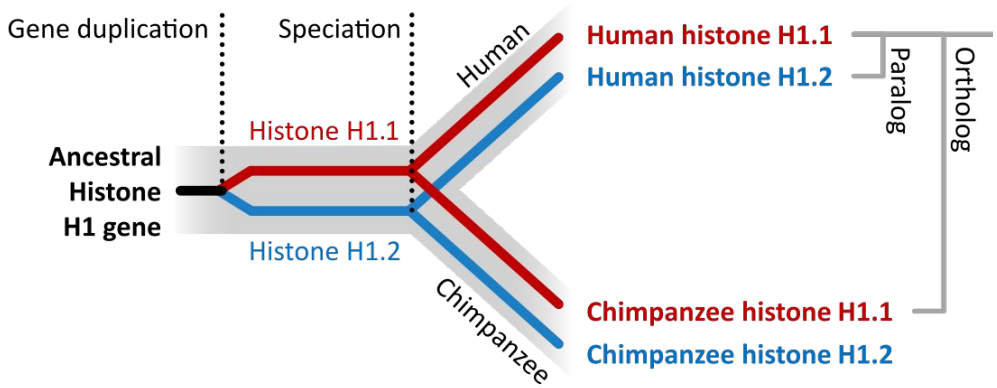
Enzyme Commission number (E.C.)

Numerical classification based on the chemical reactions a enzyme catalyze:

Class	Reaction catalyzed	Typical reaction	Enzyme example(s) with trivial name
EC 1 ***Oxidoreductases***	Oxidation/reduction reactions; transfer of H and O atoms or electrons from one substance to another	$AH + B \rightarrow A + BH$ (**reduced**) $A + O \rightarrow AO$ (**oxidized**)	Dehydrogenase, oxidase
EC 2 ***Transferases***	Transfer of a functional group from one substance to another. The group may be methyl-, acyl-, amino- or phosphate group	$AB + C \rightarrow A + BC$	Transaminase, kinase
EC 3 ***Hydrolases***	Formation of two products from a substrate by hydrolysis	$AB + H_2O \rightarrow AOH + BH$	Lipase, amylase, peptidase, phosphatase
EC 4 ***Lyases***	Non-hydrolytic addition or removal of groups from substrates. C-C, C-N, C-O or C-S bonds may be cleaved	$RCOCOOH \rightarrow RCOH + CO_2$ or [X-A+B-Y] → [A=B + X-Y]	Decarboxylase
EC 5 ***Isomerases***	Intramolecule rearrangement, i.e. isomerization changes within a single molecule	$ABC \rightarrow BCA$	Isomerase, mutase
EC 6 ***Ligases***	Join together two molecules by synthesis of new C-O, C-S, C-N or C-C bonds with simultaneous breakdown of ATP	$X + Y + ATP \rightarrow XY + ADP + P_i$	Synthetase
EC 7 ***Translocases***	Catalyse the movement of ions or molecules across membranes or their separation within membranes		Transporter

Database of Clusters of Orthologous Genes (COGs)

Orthologous genes - Recapitulating:



The orthologous group in the database is labeled using a code. **Example:**
COG0105 - Nucleoside diphosphate kinase

Each COGs includes proteins that are inferred to be orthologs (direct evolutionary counterparts)

Database of Clusters of Orthologous Genes (COGs)

We can divide the genes in categories using the COGs codes:

A	RNA processing and modification
B	Chromatin Structure and dynamics
C	Energy production and conversion
D	Cell cycle control and mitosis
E	Amino Acid metabolis and transport
F	Nucleotide metabolism and transport
G	Carbohydrate metabolism and transport
H	Coenzyme metabolism
I	Lipid metabolism
J	Tranlsation
K	Transcription
L	Replication and repair
M	Cell wall/membrane/envelop biogenesis
N	Cell motility
O	Post-translational modification, protein turnover, chaperone functions
P	Inorganic ion transport and metabolism
Q	Secondary Structure
T	Signal Transduction
U	Intracellular trafficing and secretion
Y	Nuclear structure
Z	Cytoskeleton
R	General Functional Prediction only
S	Function Unknown