

## Session 5.1 – Annotation

**Pedro J. Sola Campoy**

**BU-ISCIII**

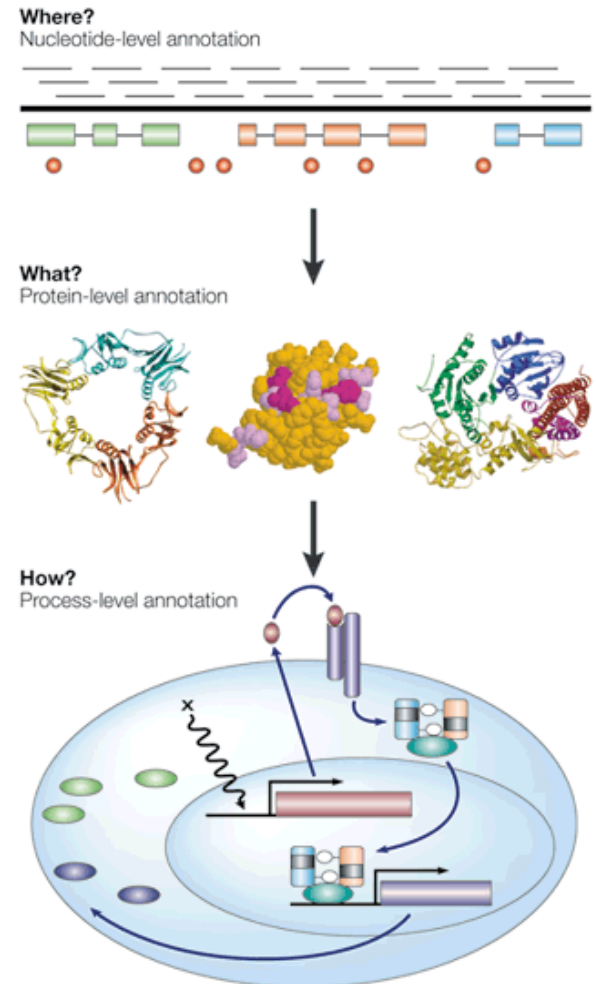
**Unidades Comunes Científico Técnicas – SGSAFI-ISCIII**

05-09 Noviembre 2018, 1ª Edición  
Programa Formación Continua, ISCIII

# Annotation

Genome annotation is the process of **attaching biological (and positional) information to sequences**. It consists of three main steps:

- identifying portions of the genome that **do not code for proteins**
- identifying elements on the genome, a process called **gene prediction**
- attaching **biological information** to these elements



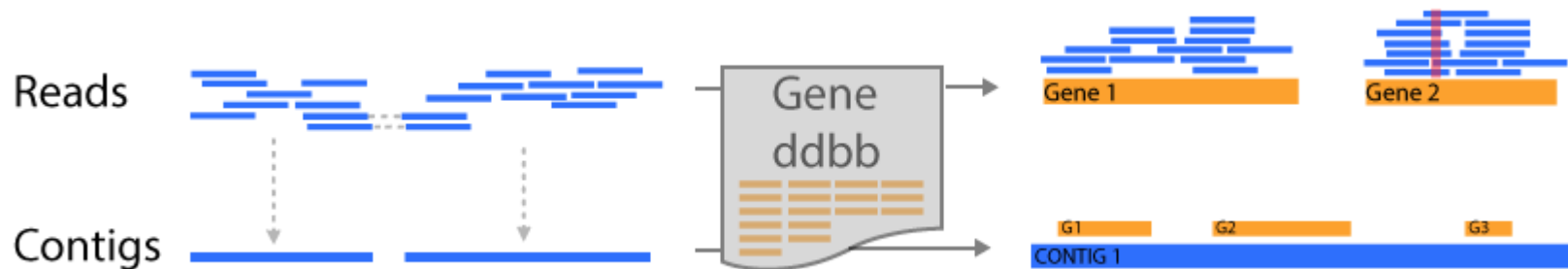
<https://galaxyproject.github.io/training-material/topics/genome-annotation/tutorials/genome-annotation/tutorial.html>

# Main categories

- **Structural annotation** – Finding genes and other biologically relevant sites with **specific locations but unknown function**
  - ORFs
  - Coding sequences(cds)
  - Promoters and regulatory regions
- **Functional annotation** – Elements are used in **database searches** to attach biologically relevant information to whole sequence and individual objects
- Do they depend on each other?

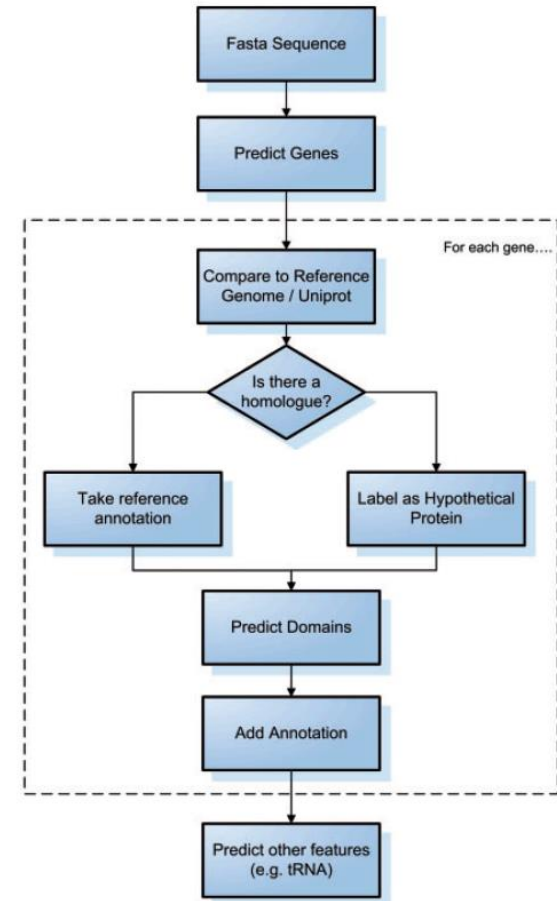
# Mapping vs Assembly

- **Functional annotation based on mapping (srst2)**
  - Pro: more resolute / high quality ddbb
  - Con: Unable to locate genes
- **Functional annotation based on assembly (prokka)**
  - Pro: genes are located / related
  - Depend on assembly (close to repetitive regions)



# Automatic annotation

- Exponential submission of bacterial genomes
- Databases
  - Encyclopedia of DNA elements (ENCODE)
  - Entrez Gene
  - Ensembl
  - GENCODE
  - Gene Ontology Consortium
  - GeneRIF
  - RefSeq
  - Uniprot
  - Vertebrate and Genome Annotation Project (Vega)
  - Pfam
  - etc



# Automatic annotation: limitations

- If sequence homologues are found, may not be functional homologues
- If **no homology found**- limited information can be inferred
- Incorrect annotation can be **propagated** when similarity is over part on sequence not used in annotation
- Problem with multidomain proteins (HMM)
- Inconsistent annotation (**Different names, same protein**)
- Same **gene name, different product** name
- Spelling mistakes
- Looking for **new genes**, not present in DDBB
- Expression experiments / Manual annotation needed

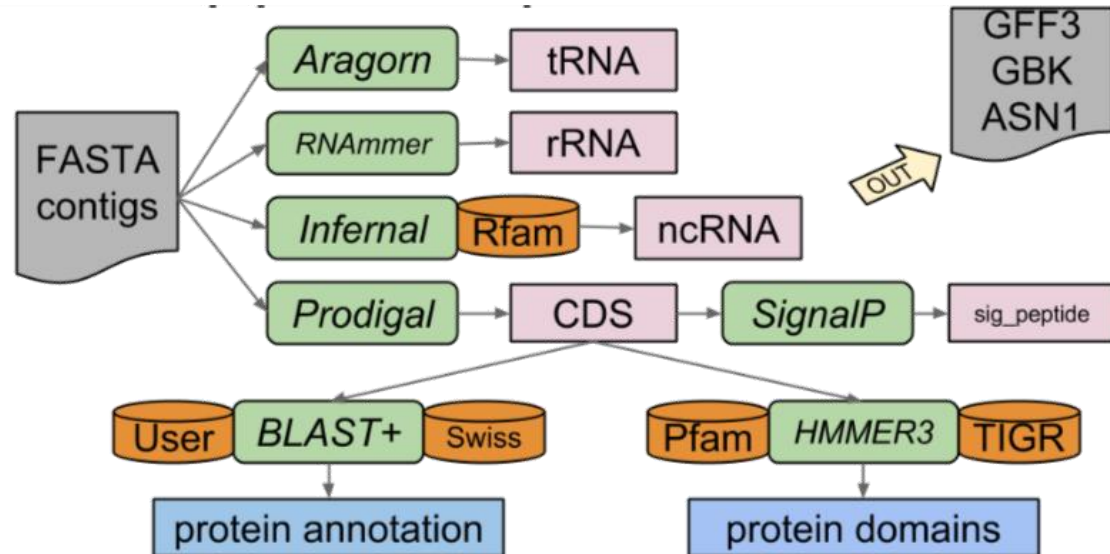
# Automatic annotation: limitations

- RefSeq is one attempt to standardize and improve the quality of genome annotation
  - WP\_ prefix. All identical proteins regardless of species
  - Standard classification

```
beta-lactamase (conceptual)
  class A beta-lactamase (HMM:NF033103)
  metallo-beta-lactamase (HMM:NF012229)
    subclass B1 metallo-beta-lactamase (HMM:NF033088)
      NDM family subclass B1 metallo-beta-lactamase (HMM:NF000259)
        subclass B1 metallo-beta-lactamase NDM-1 (allele)
        subclass B1 metallo-beta-lactamase NDM-2 (allele)
        subclass B1 metallo-beta-lactamase NDM-3 (allele)
      VIM family subclass B1 metallo-beta-lactamase (HMM:NF012100)
      SPM family subclass B1 metallo-beta-lactamase (HMM:NF012150)
    subclass B2 metallo-beta-lactamase (HMM:NF033087)
    subclass B3 metallo-beta-lactamase (HMM:NF033105)
  class C beta-lactamase (HMM:NF033085)
  class D beta-lactamase (conceptual)
    class D beta-lactamase (main branch) (HMM:NF012161)
    class D beta-lactamase (other branch) (HMM:NF000270)
```

# Automatic annotation: Prokka

Tool (reference)	Features predicted
Prodigal ( Hyatt 2010 )	Coding sequence (CDS)
RNAmmmer ( Lagesen et al. , 2007 )	Ribosomal RNA genes (rRNA)
Aragorn ( Laslett and Canback, 2004 )	Transfer RNA genes
SignalP ( Petersen et al. , 2011 )	Signal leader peptides
Infernal ( Kolbe and Eddy, 2011 )	Non-coding RNA
BLAST+ ( Camacho <i>et al.</i> , 2009 )	Specific function or name Personal database



- Optional **user-provided** set of annotated proteins
- All bacterial proteins in **UniProt**
- All proteins from finished bacterial genomes in **RefSeq**
- Hidden Markov model profile databases, **Pfam and TIGRFAMs**
- Hypothetical protein

<https://galaxyproject.github.io/training-material/topics/genome-annotation/tutorials/annotation-with-prokka/slides.html#8>



# Automatic annotation: Prokka output

Suffix	Description of file contents
.fna	FASTA file of original input contigs (nucleotide)
.faa	FASTA file of translated coding genes (protein)
.ffn	FASTA file of all genomic features (nucleotide)
.fsa	Contig sequences for submission (nucleotide)
.tbl	Feature table for submission
.sqn	<b>Sequin</b> editable file for submission
.gbk	Genbank file containing sequences and annotations
.gff	GFF v3 file containing sequences and annotations
.log	Log file of Prokka processing output
.txt	Annotation summary statistics

# Annotation format: gff3

1. Seqid - name
2. Source - program
3. Type - term or SOFA sequence ontology
4. Start
5. End
6. Score
7. Strand - (+/-)
8. Phase - (0/1/2)
9. Attributes
  - Attributes
    - Name
    - Alias
    - Parent
    - Target
    - Gap
    - Derives\_from
    - Note
    - Dbxref
    - Ontology\_term

```
##gff-version 3
ctg123 . operon      1300 15000 . + . ID=operon001;Name=superOperon
ctg123 . mRNA       1300 9000 . + . ID=mrna0001;Parent=operon001;Name=sonichedgehog
ctg123 . exon       1300 1500 . + . Parent=mrna0001
ctg123 . exon       1050 1500 . + . Parent=mrna0001
ctg123 . exon       3000 3902 . + . Parent=mrna0001
ctg123 . exon       5000 5500 . + . Parent=mrna0001
ctg123 . exon       7000 9000 . + . Parent=mrna0001
ctg123 . mRNA       10000 15000 . + . ID=mrna0002;Parent=operon001;Name=subsonicsquirrel
ctg123 . exon       10000 12000 . + . Parent=mrna0002
ctg123 . exon       14000 15000 . + . Parent=mrna0002
```

# Annotation format: gbk

- LOCUS – Annotated sequence
- DEFINITION
- ACCESION
- FEATURES
  - source
  - gene
  - CDS
    - gene
    - function
    - Product
    - protein\_id
    - Translation (sequence)

```

LOCUS      AF068625                200 bp    mRNA    linear    ROD 06-DEC-1999
DEFINITION Mus musculus DNA cytosine-5 methyltransferase 3A (Dnmt3a) mRNA,
            complete cds.
ACCESSION  AF068625 REGION: 1..200
VERSION    AF068625.2 GI:6449467
KEYWORDS   .
SOURCE     Mus musculus (house mouse)
ORGANISM   Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE  1 (bases 1 to 200)
AUTHORS   Okano,M., Xie,S. and Li,E.
TITLE     Cloning and characterization of a family of novel mammalian DNA
            (cytosine-5) methyltransferases
JOURNAL    Nat. Genet. 19 (3), 219-220 (1998)
PUBMED     9662389
REFERENCE  2 (bases 1 to 200)
AUTHORS   Xie,S., Okano,M. and Li,E.
TITLE     Direct Submission
JOURNAL    Submitted (28-MAY-1998) CVRC, Mass. Gen. Hospital, 149 13th Street,
            Charlestown, MA 02129, USA
REFERENCE  3 (bases 1 to 200)
AUTHORS   Okano,M., Chijiwa,T., Sasaki,H. and Li,E.
TITLE     Direct Submission
JOURNAL    Submitted (04-NOV-1999) CVRC, Mass. Gen. Hospital, 149 13th Street,
            Charlestown, MA 02129, USA
REMARK     Sequence update by submitter
COMMENT     On Nov 18, 1999 this sequence version replaced gi:3327977.
FEATURES   Location/Qualifiers
            source
                1..200
                /organism="Mus musculus"
                /mol_type="mRNA"
                /db_xref="taxon:10090"
                /chromosome="12"
                /map="4.0 cM"
            gene
                1..>200
                /gene="Dnmt3a"
ORIGIN
1 gaattccggc ctgctgccgg gccgccgac ccgccgggcc acacggcaga gccgcctgaa
61 gccacgcgct gaggctgcac ttttcgagg gcttgacatc agggctcatg ttttaagtctt
121 agctcttgct tacaaagacc acggcaattc cttctctgaa gccctgcgag cccacagcgc
181 ccctcgacgc ccagcctgc
//
    
```

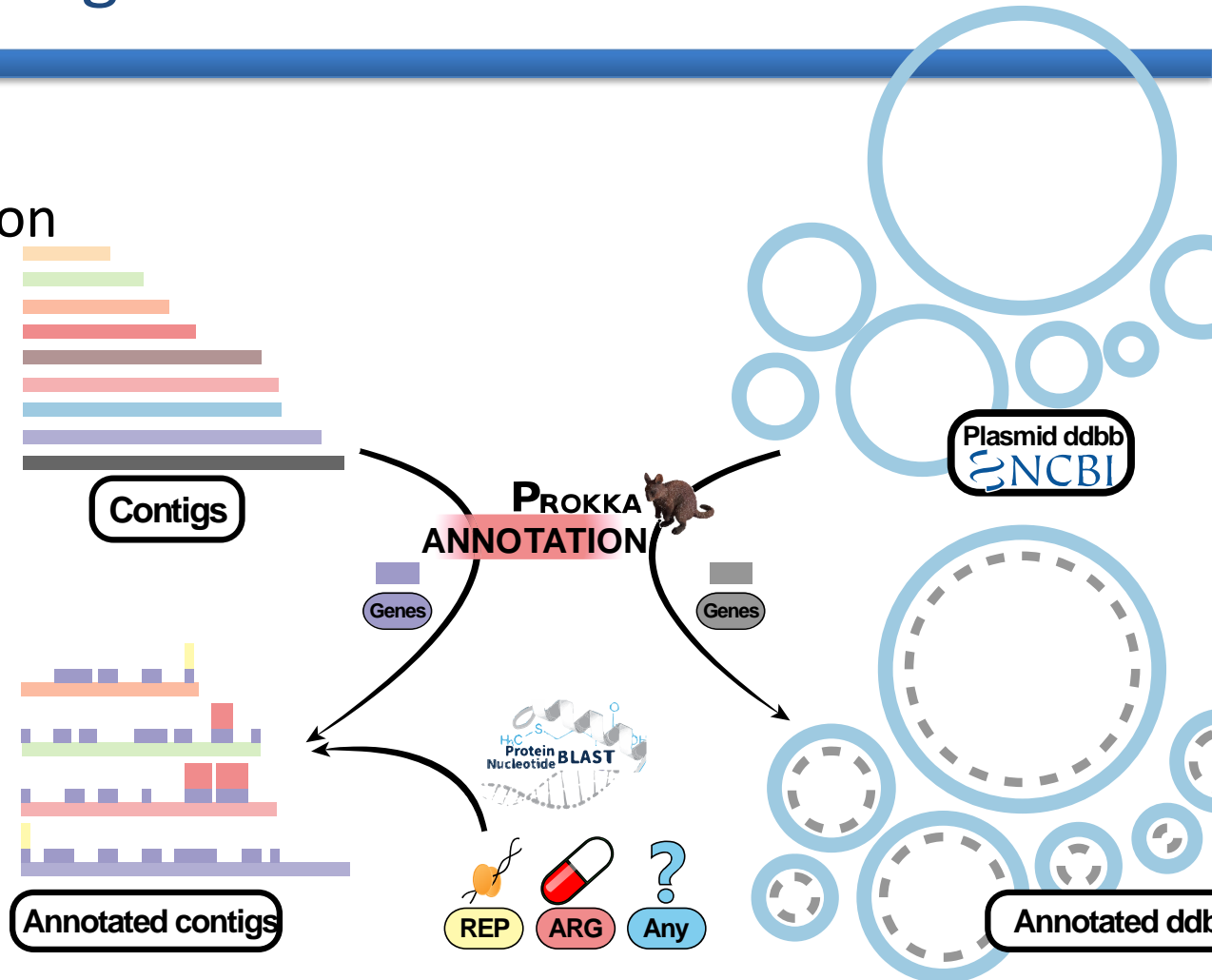
# Annotation using PlasmidID

- Automatic annotation

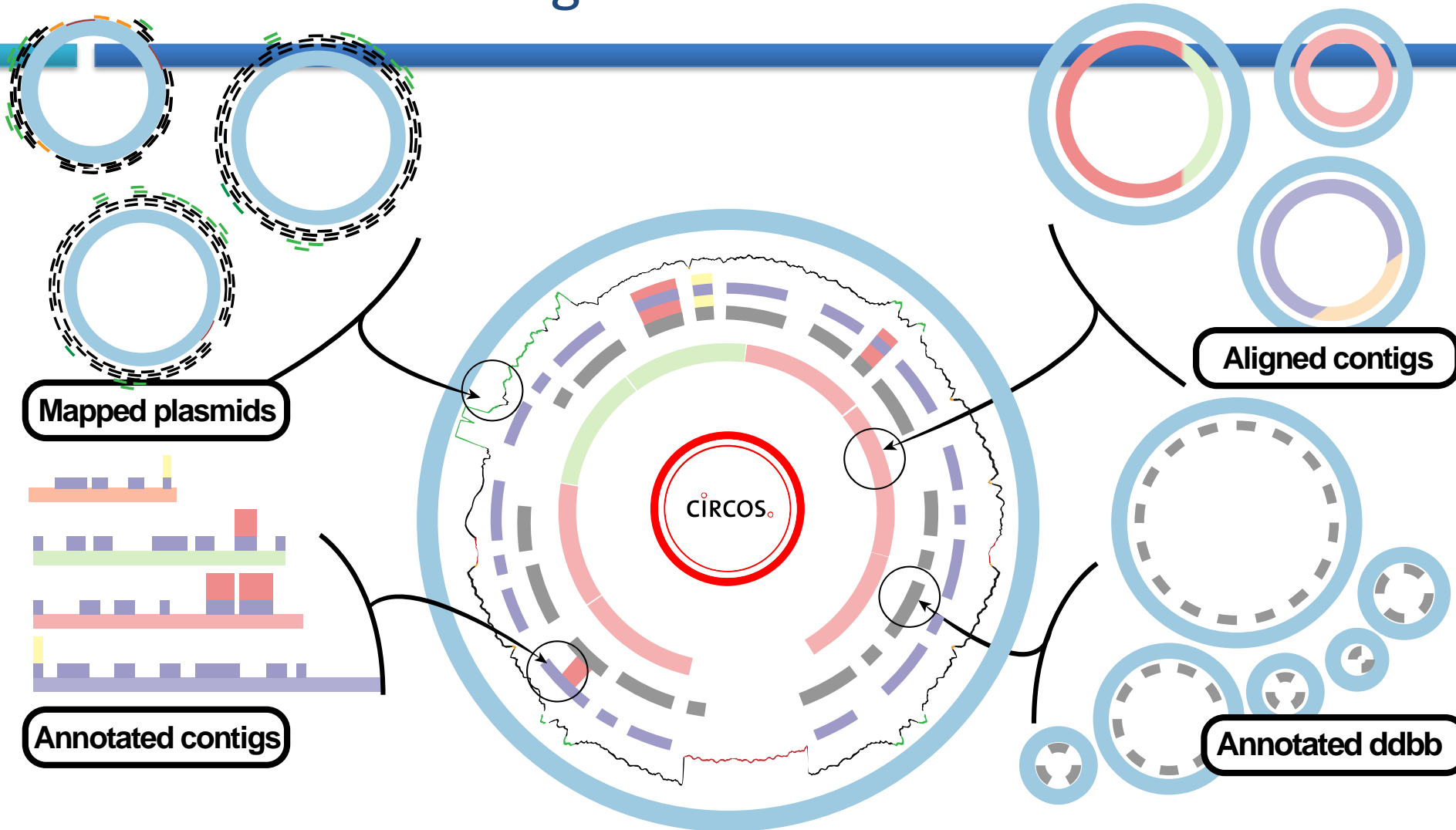
- Prokka
  - DDBB plasmid
  - Contigs
- Gff to bed

- Specific annotation

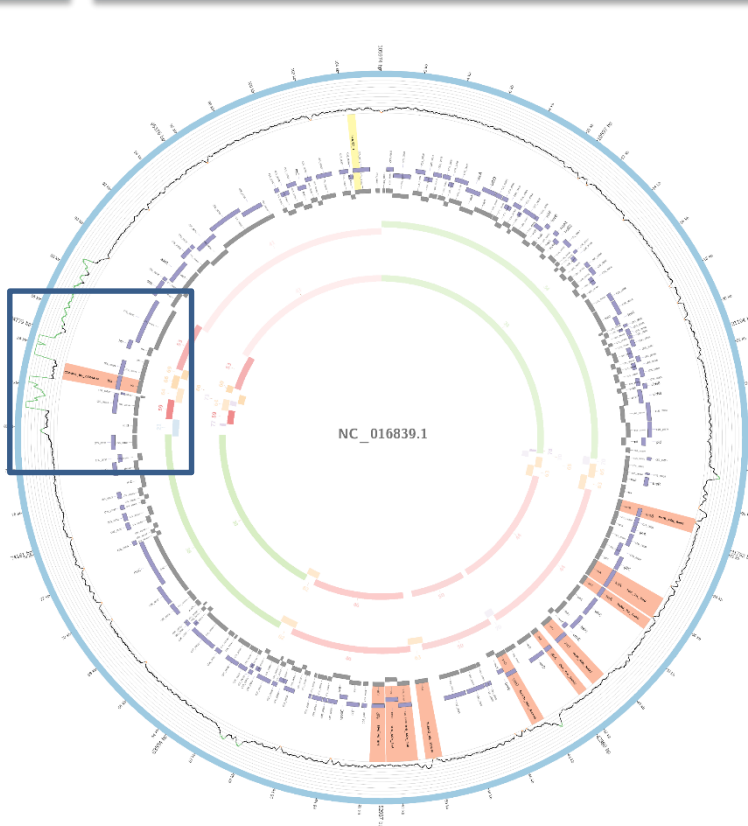
- BLAST+
- ABR & REP
- User input FASTA



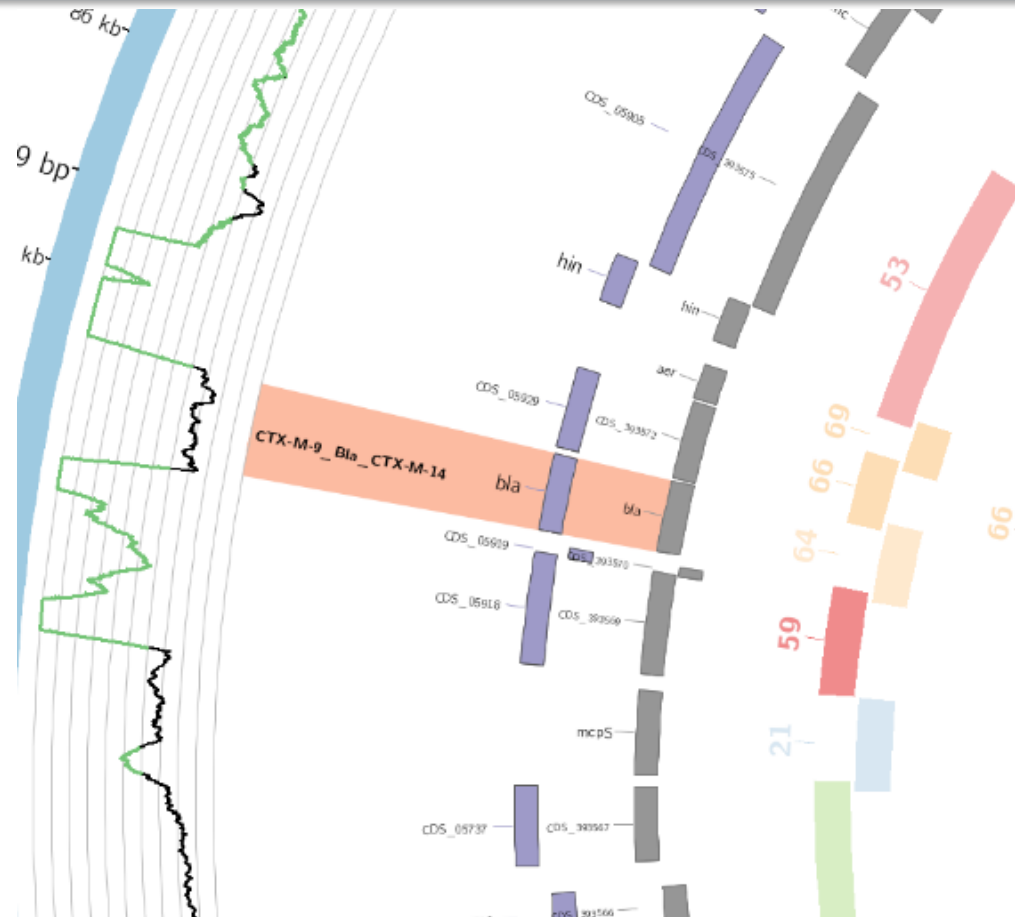
# Annotation using PlasmidID



# Annotation using PlasmidID



Annotation on short contigs



# Manual annotation: Artemis

Artemis is a DNA sequence viewer and annotation tool that allows visualisation of sequence features and the results of analyses within the context of the sequence, and its six-frame translation.

1. 
2. 
3. 
4. 

WUBLAST misc feature

LTR

800 1600 2400 3200 4000 4800

foo;bar;baz;zap;foobar;ldggwdgjiwe

CDS
5. 

D L L Y L Q H I L R T Y T T S Q A I N C M K L I E L K S F (

I C F I C N T Y \* G L T Q H H K Q S T V \* N L S N \* K A F

GTATGCTTATTTGCAACACATATGAGGACTTACACAAACATCACAAAGCAATCAACGTATGAAACTTATCGAACTGAAAAGCTTTC

20 40 60 80

CTAGACGAAATAAACGTTGTGTATACTCCTGAATGTGTGTAGTGTTCGTTAGTTGACATACCTTGAATAGCTTGACTTTCGAAAG

I Q K I Q L V Y Q P S V C C \* L C D V T H F K D F Q F A K I

D A K N A V C I S S K C L M V L L \* S Y S V # R V S F S E

. R S # K C C M N L V # V V D C A I L Q I F S I S S F L K \*
6. 

source	1	41173	
CDS	1	353	SPBC16A3.01, spn3, septin homolog spn3, len:117aa, i der
WUBLASTN HIT	3	353	
misc_feature	498	539	
CDS	784	1821	c MLCB458.06, fas, probable type I fatty acid synthase, le
misc_feature	790	1788	c Pfam match to entry adh_zinc PF00107, Zinc-binding d ehy
LTR	2383	2740	tfl like LTR
CDS	3838	5811	c SPBC16A3.03c
CDS	7720	8379	SPBC16A3.04, unknown, len:220aa, similar eg. to YIL0 93C
CDS	8379	10153	