

## Session 5.1 – Annotation

**Pedro J. Sola Campoy**

**BU-ISCIII**

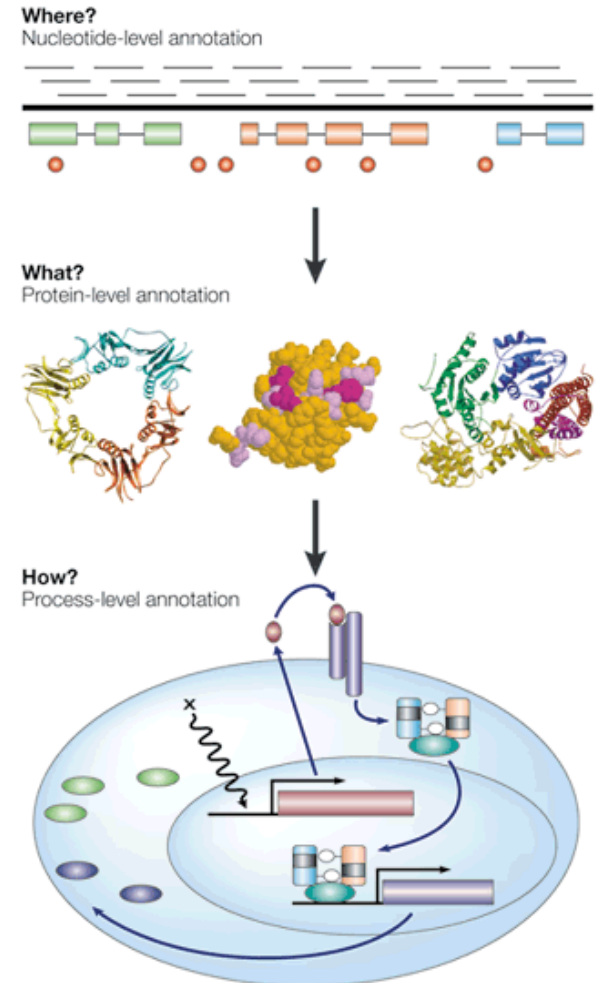
**Unidades Comunes Científico Técnicas – SGSAFI-ISCIII**

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 coding elements on the genome, a process called **gene prediction**
- attaching **biological information** to these elements



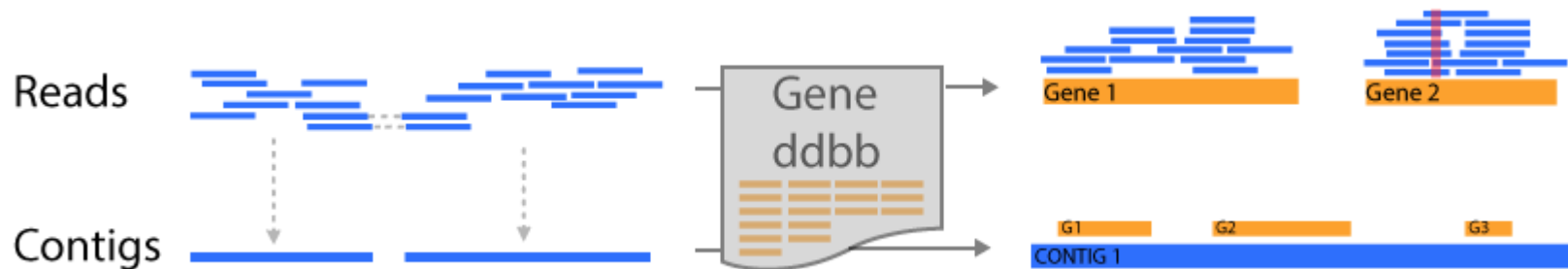
<https://galaxyproject.github.io/training-material/topics/genome-annotation/tutorials/genome-annotation/tutorial.html>

# 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?

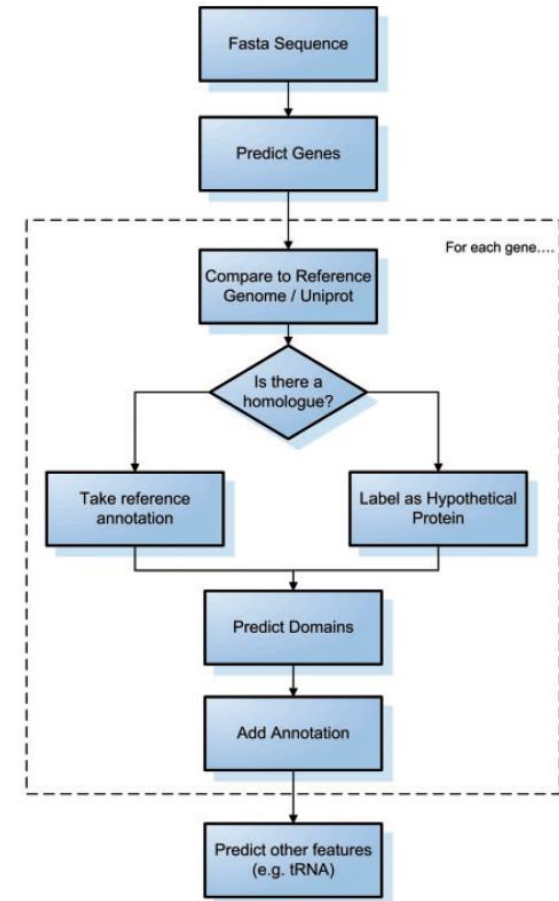
# Mapping vs Assembly

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



# Automatic annotation: limitations

- If sequence homologues are found, may **not be functional** homologues
  - Not truncated
- 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
  - 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

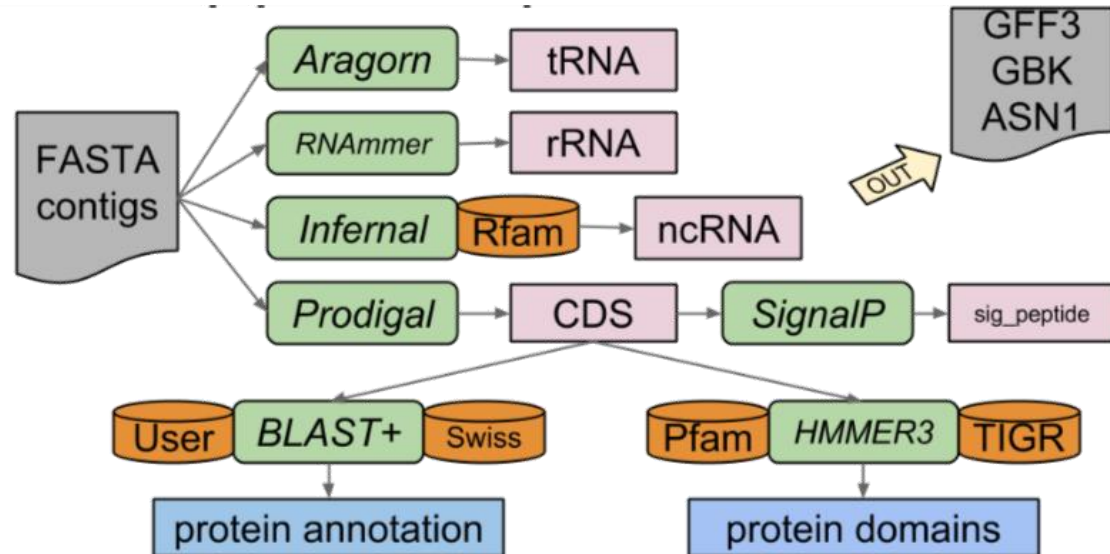
# Automatic annotation: limitations

- 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<br>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                             | ctg123 . gene            | 1000 | 9000 | . | + | . | ID=gene00001;Name=EDEN                            |
| 2. | Source - program                         | ctg123 . TF_binding_site | 1000 | 1012 | . | + | . | ID=tfbs00001;Parent=gene00001                     |
| 3. | Type - term or SOFA<br>sequence ontology | ctg123 . mRNA            | 1050 | 9000 | . | + | . | ID=mRNA00001;Parent=gene00001;Name=EDEN.1         |
| 4. | Start                                    | ctg123 . mRNA            | 1050 | 9000 | . | + | . | ID=mRNA00002;Parent=gene00001;Name=EDEN.2         |
| 5. | End                                      | ctg123 . mRNA            | 1300 | 9000 | . | + | . | ID=mRNA00003;Parent=gene00001;Name=EDEN.3         |
| 6. | Score                                    | ctg123 . exon            | 1300 | 1500 | . | + | . | ID=exon00001;Parent=mRNA00003                     |
| 7. | Strand - (+/-)                           | ctg123 . exon            | 1050 | 1500 | . | + | . | ID=exon00002;Parent=mRNA00001,mRNA00002           |
| 8. | Phase - (0/1/2)                          | ctg123 . exon            | 3000 | 3902 | . | + | . | ID=exon00003;Parent=mRNA00001,mRNA00003           |
| 9. | Attributes                               | ctg123 . exon            | 5000 | 5500 | . | + | . | ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003 |
|    |  | ctg123 . exon            | 7000 | 9000 | . | + | . | ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003 |
|    |  | ctg123 . CDS             | 1201 | 1500 | . | + | 0 | ID=cds00001;Parent=mRNA00001;Name=edenprotein.1   |
|    | - Name                                   | ctg123 . CDS             | 3000 | 3902 | . | + | 0 | ID=cds00001;Parent=mRNA00001;Name=edenprotein.1   |
|    | - Alias                                  | ctg123 . CDS             | 5000 | 5500 | . | + | 0 | ID=cds00001;Parent=mRNA00001;Name=edenprotein.1   |
|    | - Parent                                 | ctg123 . CDS             | 7000 | 7600 | . | + | 0 | ID=cds00001;Parent=mRNA00001;Name=edenprotein.1   |
|    | - Target                                 | ctg123 . CDS             | 1201 | 1500 | . | + | 0 | ID=cds00002;Parent=mRNA00002;Name=edenprotein.2   |
|    | - Gap                                    | ctg123 . CDS             | 5000 | 5500 | . | + | 0 | ID=cds00002;Parent=mRNA00002;Name=edenprotein.2   |
|    | - Derives_from                           | ctg123 . CDS             | 7000 | 7600 | . | + | 0 | ID=cds00002;Parent=mRNA00002;Name=edenprotein.2   |
|    | - Note                                   | ctg123 . CDS             | 3301 | 3902 | . | + | 0 | ID=cds00003;Parent=mRNA00003;Name=edenprotein.3   |
|    | - Dbxref                                 | ctg123 . CDS             | 5000 | 5500 | . | + | 1 | ID=cds00003;Parent=mRNA00003;Name=edenprotein.3   |
|    | - Ontology_term                          | ctg123 . CDS             | 7000 | 7600 | . | + | 1 | ID=cds00003;Parent=mRNA00003;Name=edenprotein.3   |
|    |  | ctg123 . CDS             | 3391 | 3902 | . | + | 0 | ID=cds00004;Parent=mRNA00003;Name=edenprotein.4   |
|    |  | ctg123 . CDS             | 5000 | 5500 | . | + | 1 | ID=cds00004;Parent=mRNA00003;Name=edenprotein.4   |
|    |  | ctg123 . CDS             | 7000 | 7600 | . | + | 1 | ID=cds00004;Parent=mRNA00003;Name=edenprotein.4   |

# Annotation format: gbk

- LOCUS – Annotated sequence
- DEFINITION
- ACCESION
- FEATURES
  - source
  - gene
  - CDS
    - Locus tag
    - 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 agggctctatg ttttaagtctt
121 agctcttgct tacaaagacc acggcaattc cttctctgaa gccctcgag cccacagcg
181 ccctcgagc cccagcctgc
//
    
```

# Annotation format: gbk

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

| FEATURES | Location/Qualifiers  |
|----------|--|
| source   | 1..381113<br>/organism="Klebsiella pneumoniae subsp. pneumoniae SA1"<br>/mol_type="genomic DNA"<br>/strain="SA1"<br>/sub_species="pneumoniae"<br>/db_xref="taxon:1379688"<br>/note="contig LP581_2557_Contig_49"   |
| gene     | 415..1536<br>/locus_tag="KPST86_490001"  |
| CDS      | 415..1536<br>/locus_tag="KPST86_490001"<br>/inference="ab initio prediction:AMIGene:2.0"<br>/note="Evidence 4:Homologs of previously reported genes of unknown function"<br>/codon_start=1<br>/transl_table=11<br>/product="conserved hypothetical protein"<br>/protein_id="CDI25656.1"<br>/translation="MAYQLNINWPEFLEKYWQKQPVLKNAFPDFVDPITPDELAGLA<br>MEPEVDSRLVSLKNGKMQASNGPFEHFDLGETGWSLLAQAVNHMMPAELVRPFRV<br>LPDWRLLDLMISFSVPGGGVGPHIDQYDFIIGWIGSRHRVGDGLPHRQFCPPHALL<br>HVDPPPIIDEDLQPGDILYIPPGPHDGIHETALNYSVGFPGPNRDLISSFADYV<br>LENDLGDEHYSDPDLTCREHPGRVEEYELERLRTHMIDMIRQPEDFKQWFGSFVTTPR<br>HELDIAPAEPPYEEEEVLDALLGGEKLSRLSGLRVLHIGDSFFVHSEQLDITDDEALD<br>ALCRYTSLGQEELGSGLNPAFVSELTRLINQGYNYFEE" |
| gene     | complement(1584..2117)<br>/locus_tag="KPST86_490002"   |
| CDS      | complement(1584..2117)<br>/locus_tag="KPST86_490002"<br>/inference="ab initio prediction:AMIGene:2.0"<br>/note="Evidence 4:Homologs of previously reported genes of unknown function"<br>/codon_start=1<br>/transl_table=11<br>/product="conserved hypothetical protein"<br>/protein_id="CDI25658.1"<br>/translation="MEQQLTIEMIADAFSYDITGFDCGEEALNTLKEHLKRQHDGQI<br>LRGYALVSGDTPRLLGYTTLGSGCFERGLPSKTQQKKIPYQNPVTLGRLAIDKS<br>VQGGWQGEMLVAHMRVVMGASKAVGIYGLFVEALNEKAKAFYLRGLFIQLVDENSNL<br>LFYPTKISIEQLFTDDES"  |
| gene     | complement(2128..2394)<br>/locus_tag="KPST86_490003"   |
| CDS      | complement(2128..2394)<br>/locus_tag="KPST86_490003"<br>/inference="ab initio prediction:AMIGene:2.0"<br>/note="Evidence 4:Homologs of previously reported genes of unknown function"  |

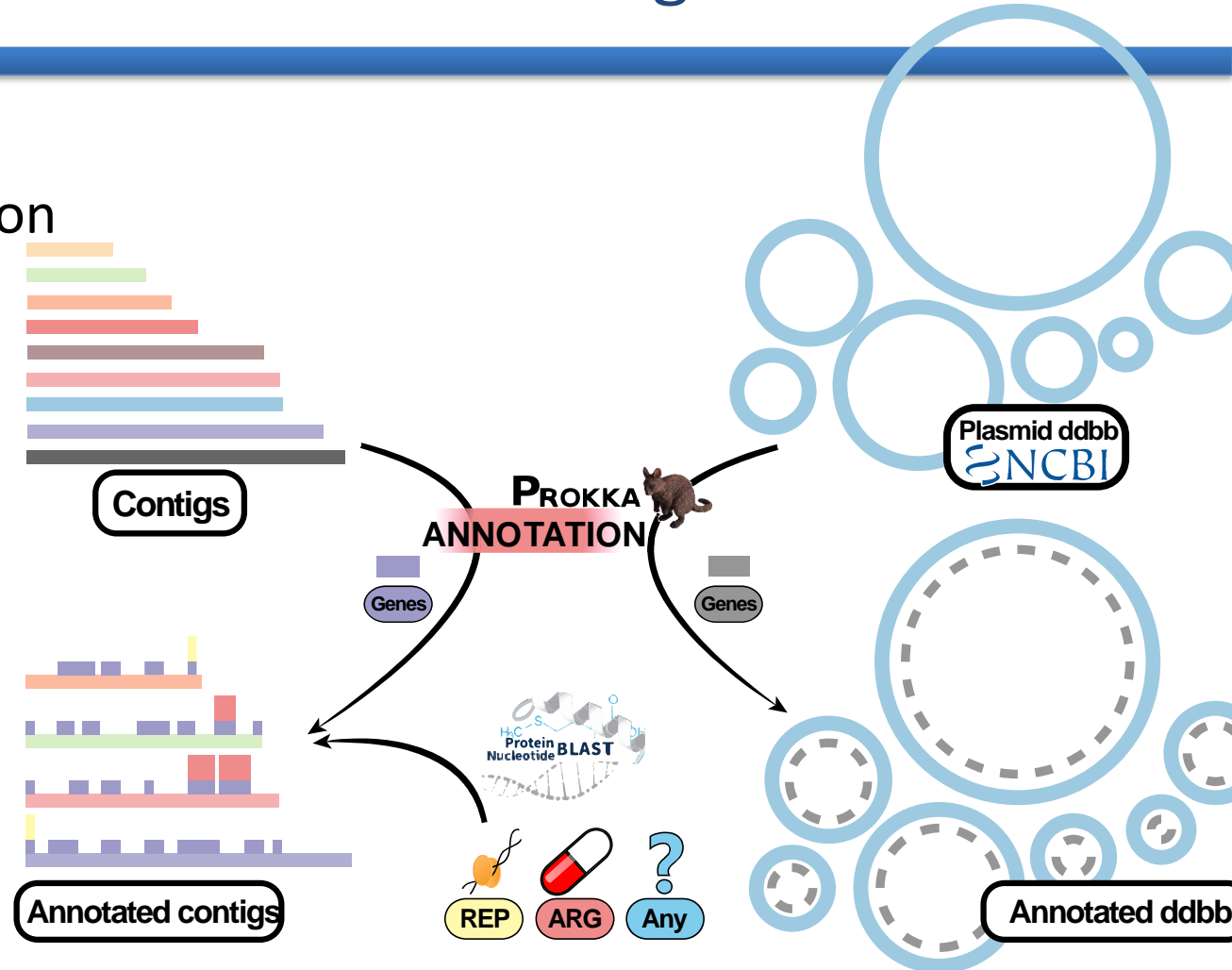
# Annotation visualization using PlasmidID

- Automatic annotation

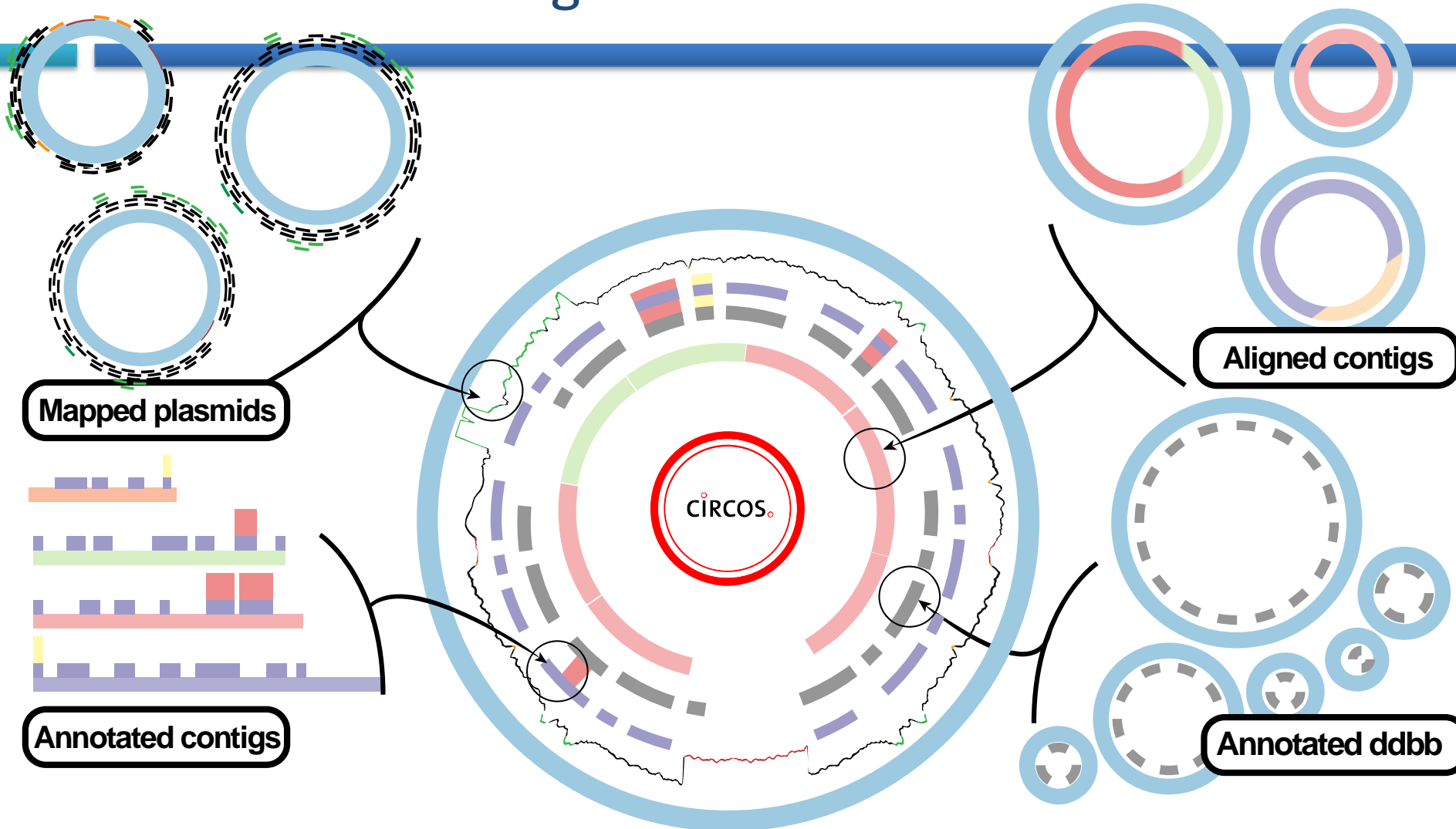
- Prokka
  - DDBB plasmid
  - Contigs
- Gff to bed

- Specific annotation

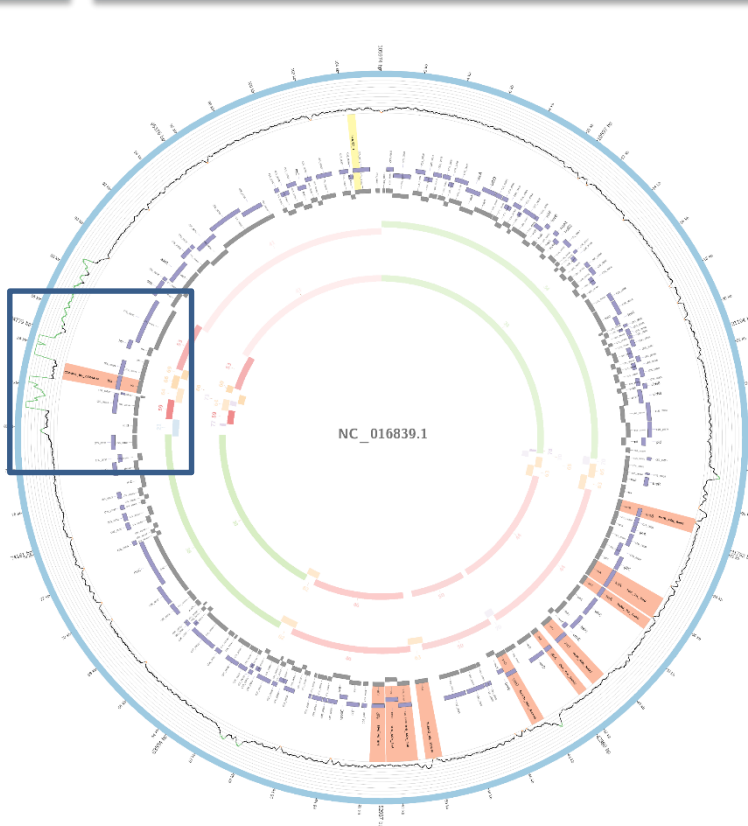
- BLAST+
- ABR & REP
- User input FASTA



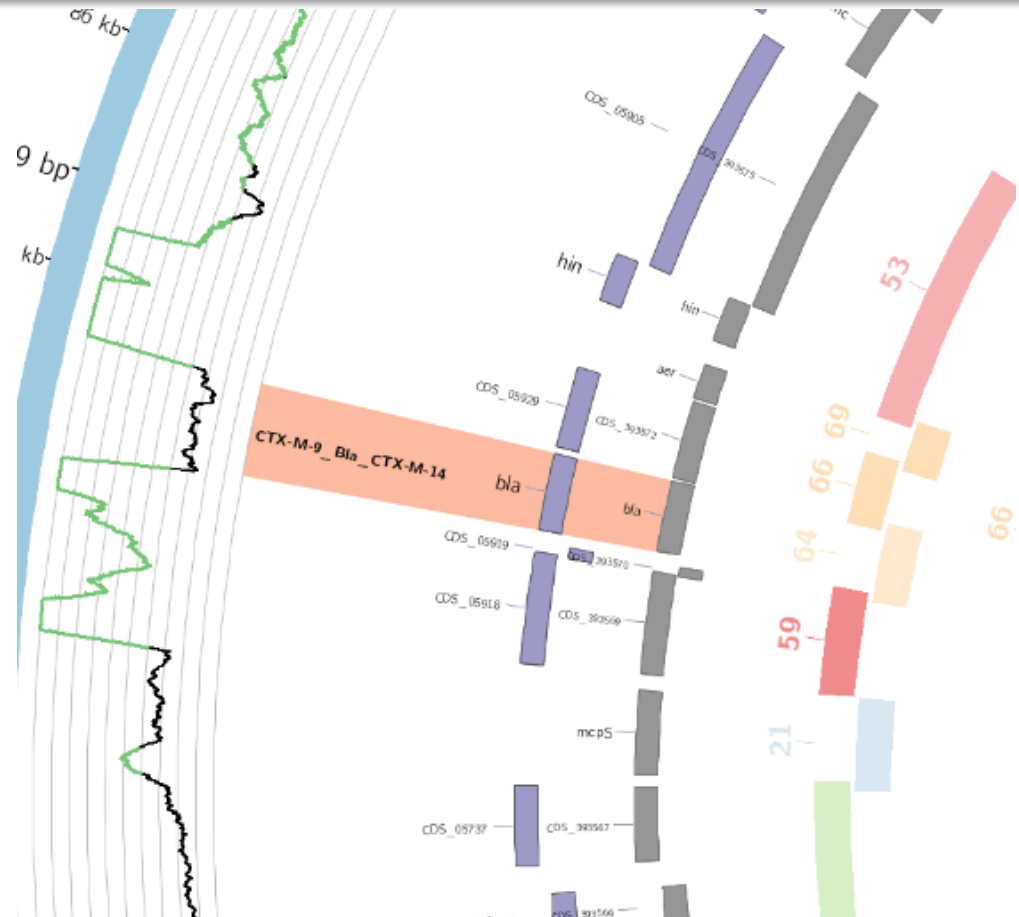
# Annotation using PlasmidID



# 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. Selected feature: bases 353 amino acids 117 CDS (/transl\_except=(pos:complement(3..5),aa:OT)
3. Entry: ☒ foo.embl
- 4.
- 5.
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 |  |