Session 5.1 – Annotation

Pedro J. Sola Campoy

BU-ISCIII

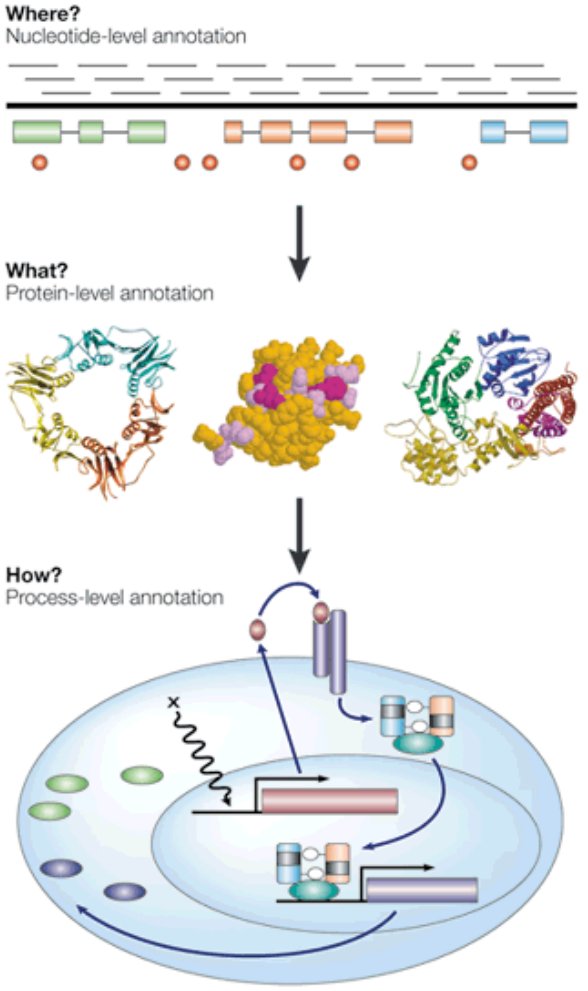
Unidades Comunes Científico Técnicas – SGSAFI-ISCIII

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



Lincoln Stein, 2001. Nat.Review Nature R

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

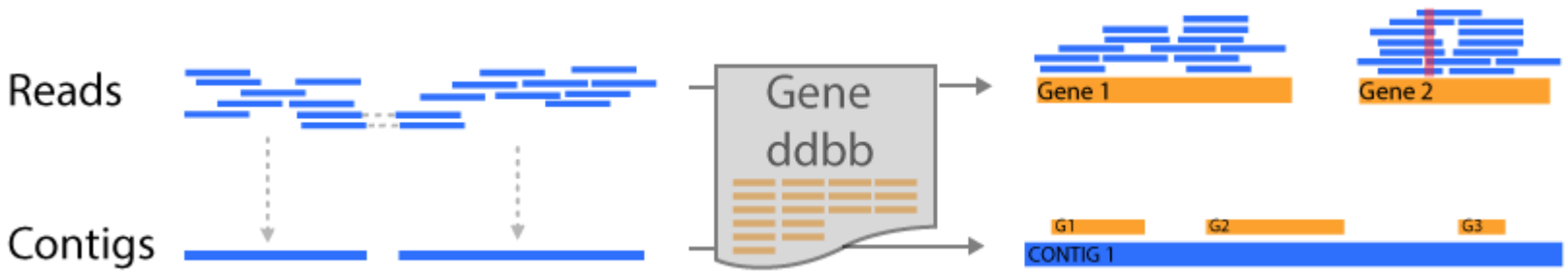
- **Functional annotation** – Elements are used in database searches to attach biologically relevant information to whole sequence and individual objects

- Do they depend on each other?

Mapping vs Assembly

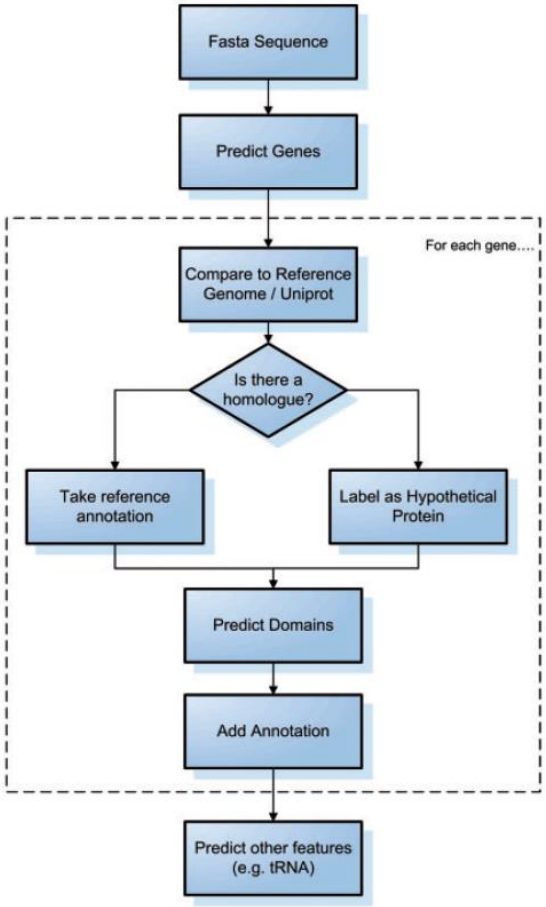
- **Functional annotation based on mapping (srst2)**
  - Pro: more resolutive / high quality ddbb
  - Con: Unable to locate genes / no ab initio annotation

- **Functional annotation based on assembly (prokka)**
  - Pro: genes are located / related
  - Depend on assembly (close to repetitive regions)



Automatic annotation

- **Exponntial 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



https://www.ncbi.nlm.nih.gov/books/NBK20253/#!po=3.12500

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

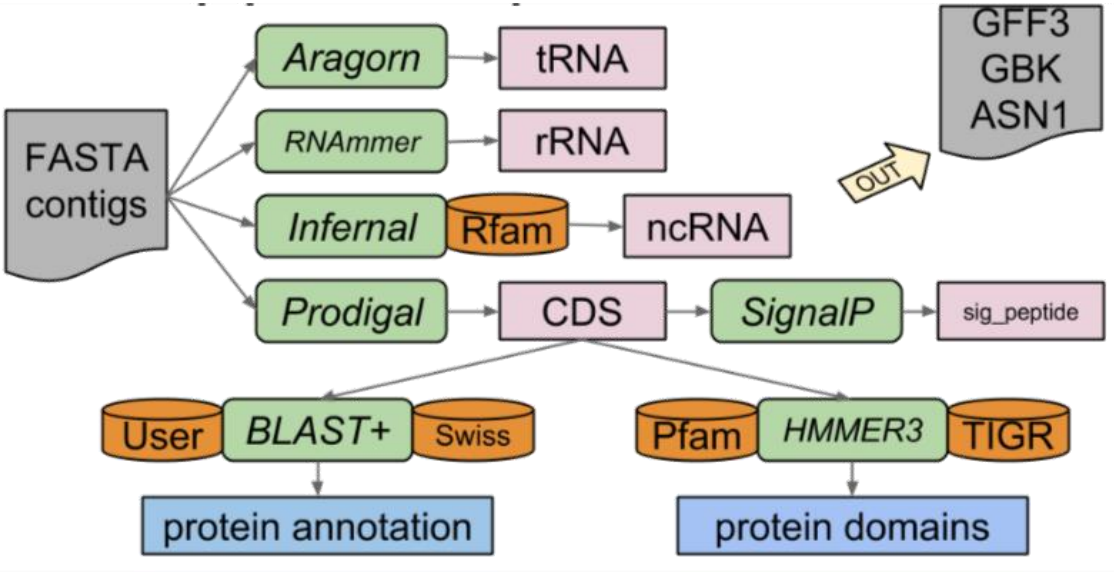
Automatic annotation: limitations

- **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)
RNAmmer (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 *et al.* , 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	**Sequin** 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
   - Name
   - Alias
   - Parent
   - Target
   - Gap
   - Derives_from
   - Note
   - Dbxref
   - Ontology_term

```
##gff-version 3.2.1
##sequence-region ctg123 1 1497228
ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
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
ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             5000  5500  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
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 gccgcccgac ccgccgggcc acacggcaga gccgcctgaa
       61 gcccagcgct gaggctgcac ttttccgagg gcttgacatc agggtctatg tttaagtctt
      121 agctcttgct tacaaagacc acggcaattc cttctctgaa gccctcgcag ccccacagcg
      181 ccctcgcagc cccagcctgc
//
```

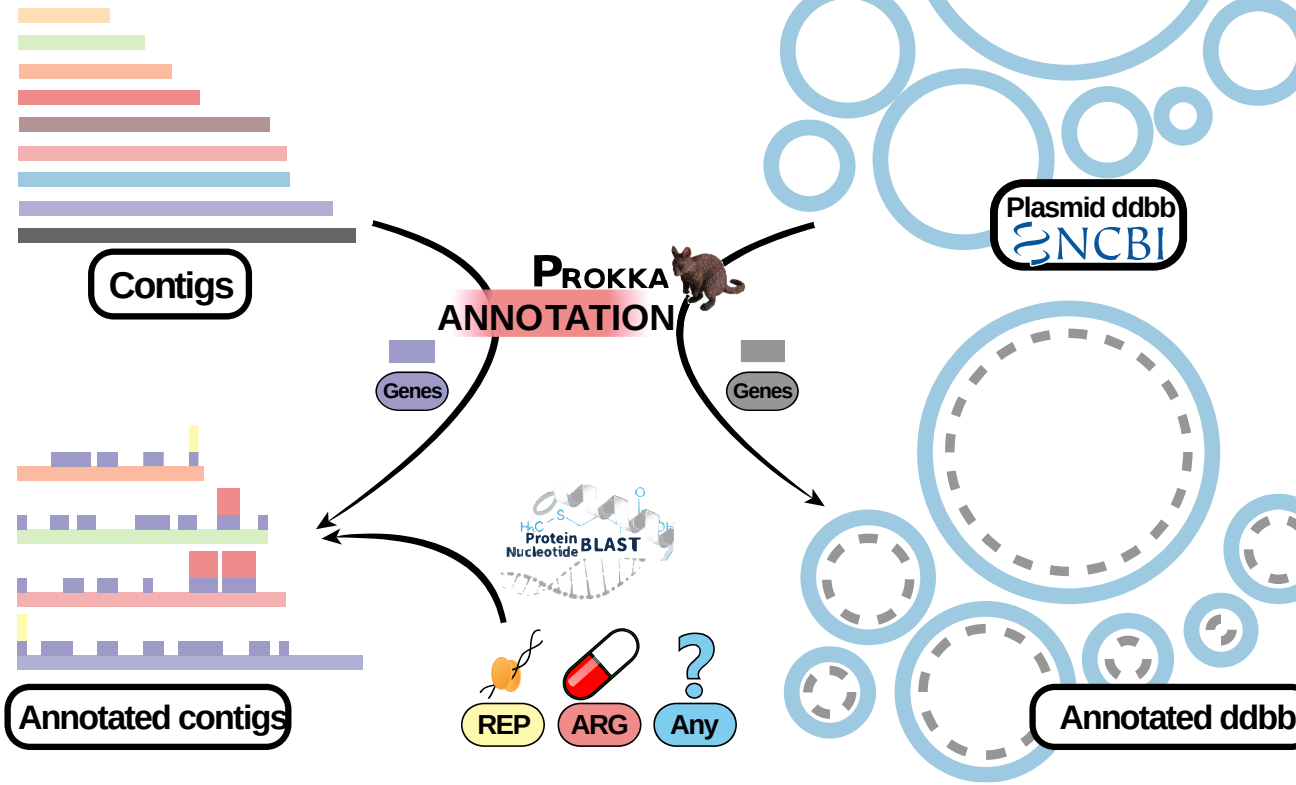
Annotation format: gbk

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

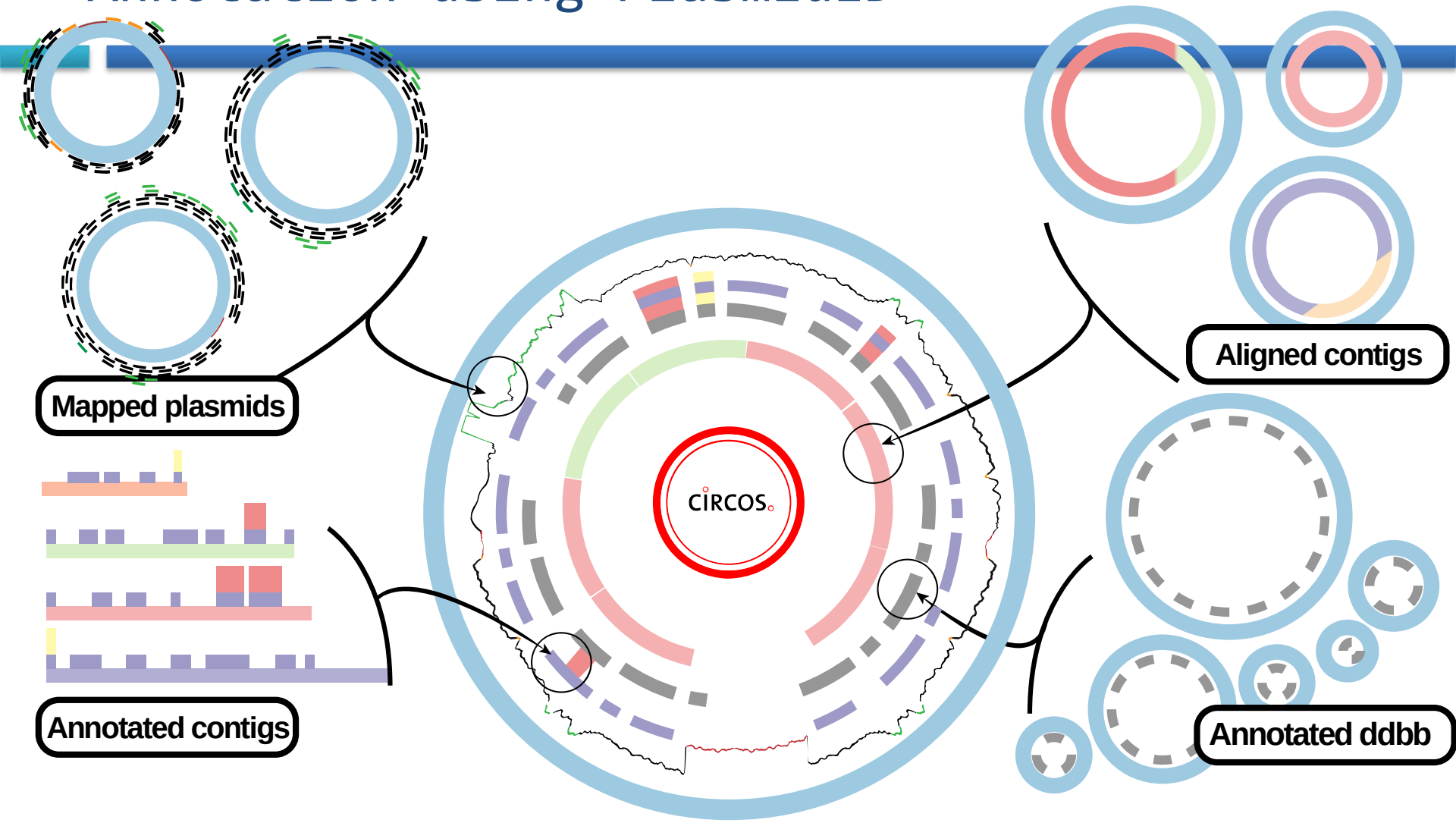
```
FEATURES             Location/Qualifiers
     source          1..381113
                     /organism="Klebsiella pneumoniae subsp. pneumoniae SA1"
                     /mol_type="genomic DNA"
                     /strain="SA1"
                     /sub_species="pneumoniae"
                     /db_xref="taxon:1379688"
                     /note="contig LPSB1_2557_Contig_49"
     gene            415..1536
                     /locus_tag="KPST86_490001"
     CDS             415..1536
                     /locus_tag="KPST86_490001"
                     /inference="ab initio prediction:AMIGene:2.0"
                     /note="Evidence 4:Homologs of previously reported genes of
                     unknown function"
                     /codon_start=1
                     /transl_table=11
                     /product="conserved hypothetical protein"
                     /protein_id="CDI25656.1"
                     /translation="MAYQLNINWPEFLEKYWQKQPVVLKNAFPDFVDPITPDELAGLA
                     MEPEVDSRLVSLKNGKWQASNGPFEHFDGLGETGWSLLAQAVNHWHMPAAELVRPFRV
                     LPDWRLDDLMISFSVPGGGVGPHIDQYDVFIIQGMGSRRWRVGDKLPMRQFCPHPALL
                     HVDPFPPIIDEDLQPGDILYIPPGFPHDGITHETALNYSVGFRGPNGRDLISSFADYV
                     LENDLGDEHYSDPDLTCREHPGRVEEYELERLRTMMIDMIRQPEDFKQWFGSFVTTPR
                     HELDIAPAEPPYEEEEVLDALLGGEKLSRLSGLRVLHIGDSFFVHSEQLDTTDAEALD
                     ALCRYTSLGQEELGSGLQNPAFVSELTRLINQGYWYFEE"
     gene            complement(1584..2117)
                     /locus_tag="KPST86_490002"
     CDS             complement(1584..2117)
                     /locus_tag="KPST86_490002"
                     /inference="ab initio prediction:AMIGene:2.0"
                     /note="Evidence 4:Homologs of previously reported genes of
                     unknown function"
                     /codon_start=1
                     /transl_table=11
                     /product="conserved hypothetical protein"
                     /protein_id="CDI25658.1"
                     /translation="MEQQLTIEMIADAFSYDITGFDCGEEALNTFLKEHLKRQHDGQI
                     LRGYALVSGDTVPRLLGYYTLSGSCFERGMLPSKTQQKKIPYQNAPSVTLGRLAIDKS
                     VQGQGWGEMLVAHAMRVVWGASKAVGIYGLFVEALNEKAKAFYLRLGFIQLVDENSNL
                     LFYPTKSIEQLFTDDES"
     gene            complement(2128..2394)
                     /locus_tag="KPST86_490003"
     CDS             complement(2128..2394)
                     /locus_tag="KPST86_490003"
                     /inference="ab initio prediction:AMIGene:2.0"
                     /note="Evidence 4:Homologs of previously reported genes of
                     unknown function"
```

Annotation visualization using PlasmidID

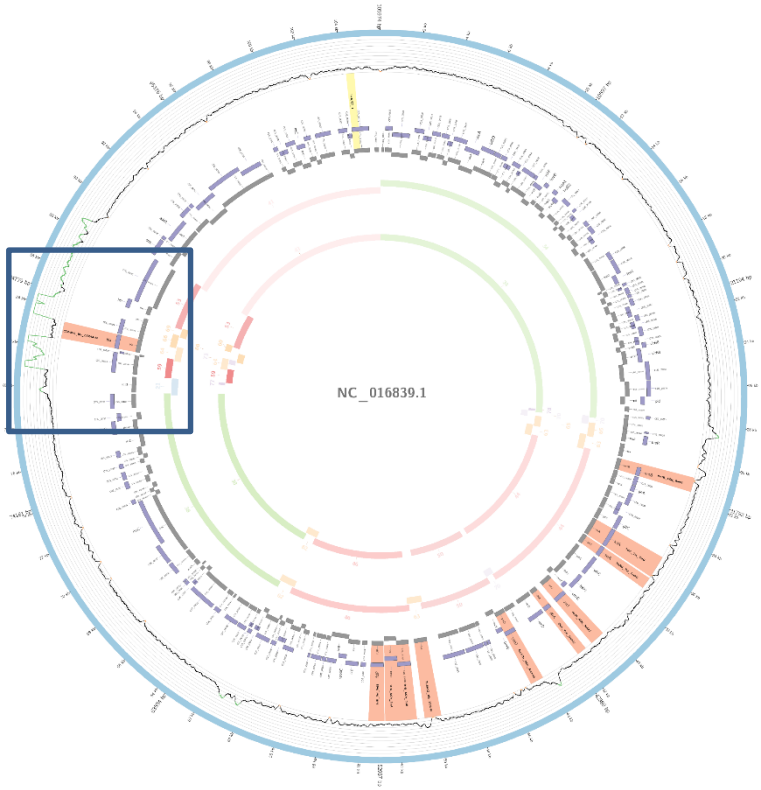
- Automatic annotation
  - Prokka
    - DDBB plasmid
    - Contigs
  - Gff to bed
- Specific annotation
  - BLAST+
  - ABR & REP
  - User input FASTA



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.