

## Session 5.1 – Annotation

Isabel Cuesta

BU-ISCIII

Unidades Comunes Científico Técnicas – SGSAFI-ISCIII

16 al 20 octubre 2023, 5ª Edición  
Programa Formación Continua, ISCIII

# Bacterial genome characteristics

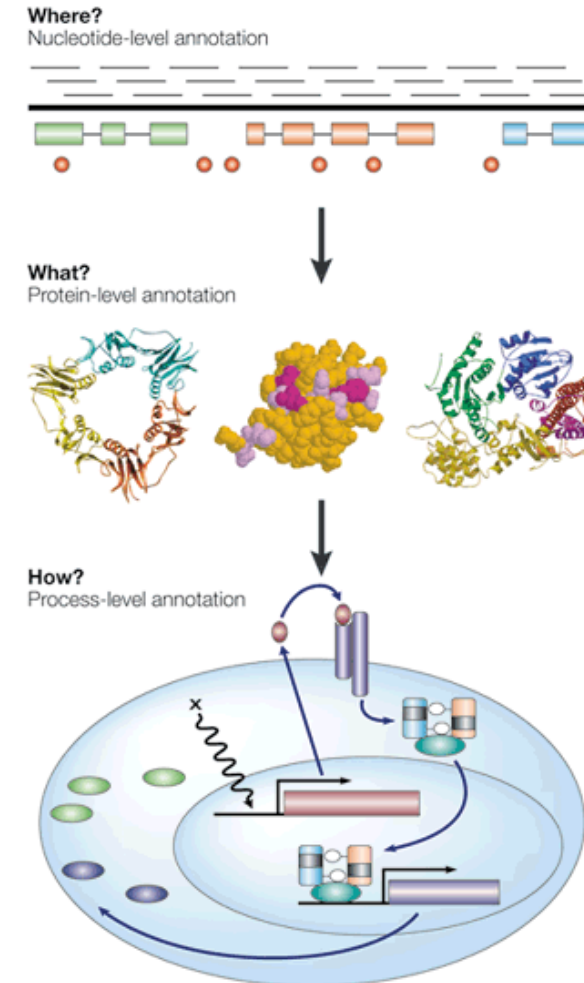
- A bacterial genome is a single "circular" DNA molecule with several million base pairs in size
- Bacteria can contain plasmids (small and circular DNA molecules, that contain (usually) non-essential genes)
- Genomes contain a few thousand genes.
- "Gene density" is much higher than in humans, one million base pairs of bacterial DNA contains about 500 to 1000 genes.
  - bacterial genes have no introns,
  - the average number of codons in bacterial genes is less than in human genes
  - neighboring genes are very close together throughout the genome

# 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>

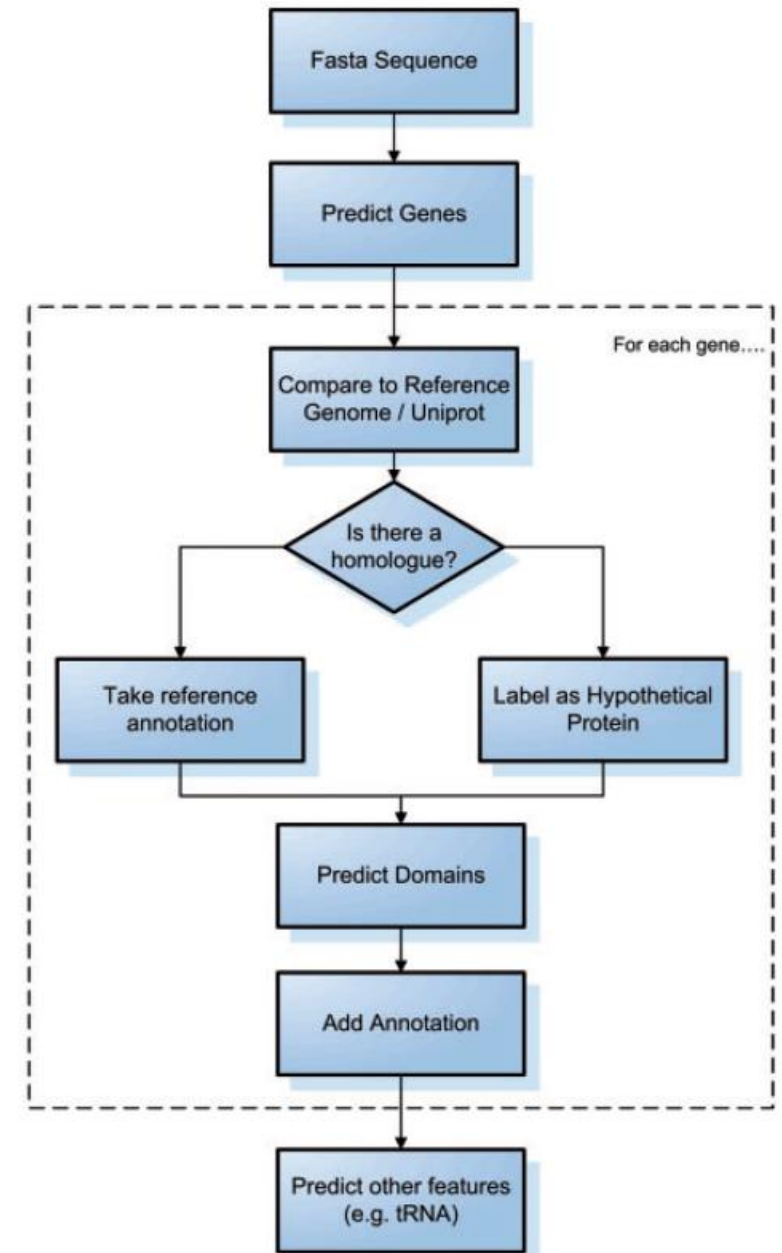


# Main categories

- **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

# Automatic annotation

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



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

# Automatic annotation

## Two strategies for identifying coding genes:

- Sequence alignment to find known protein sequences in the contigs
  - transfer the annotation across
    - will miss proteins not present in your database
    - may miss partial proteins
- Ab initio gene finding o find candidate open reading frames:
  - Build model of ribosome binding sites
  - predict coding regions
    - may choose the incorrect start codon
    - may miss atypical genes, overpredict small genes

# Automatic annotation

- **tRNA:** easy to find and annotate: anti-codon
- **rRNA:** easy to find and annotate: 5s 16s 23s
- **CDS:** straightforward to find candidates
  - false positives are often small ORFs
  - wrong start codon o partial genes
  - Pseudogenes
  - assigning function is the bulk of the workload

# Automatic annotation: limitations

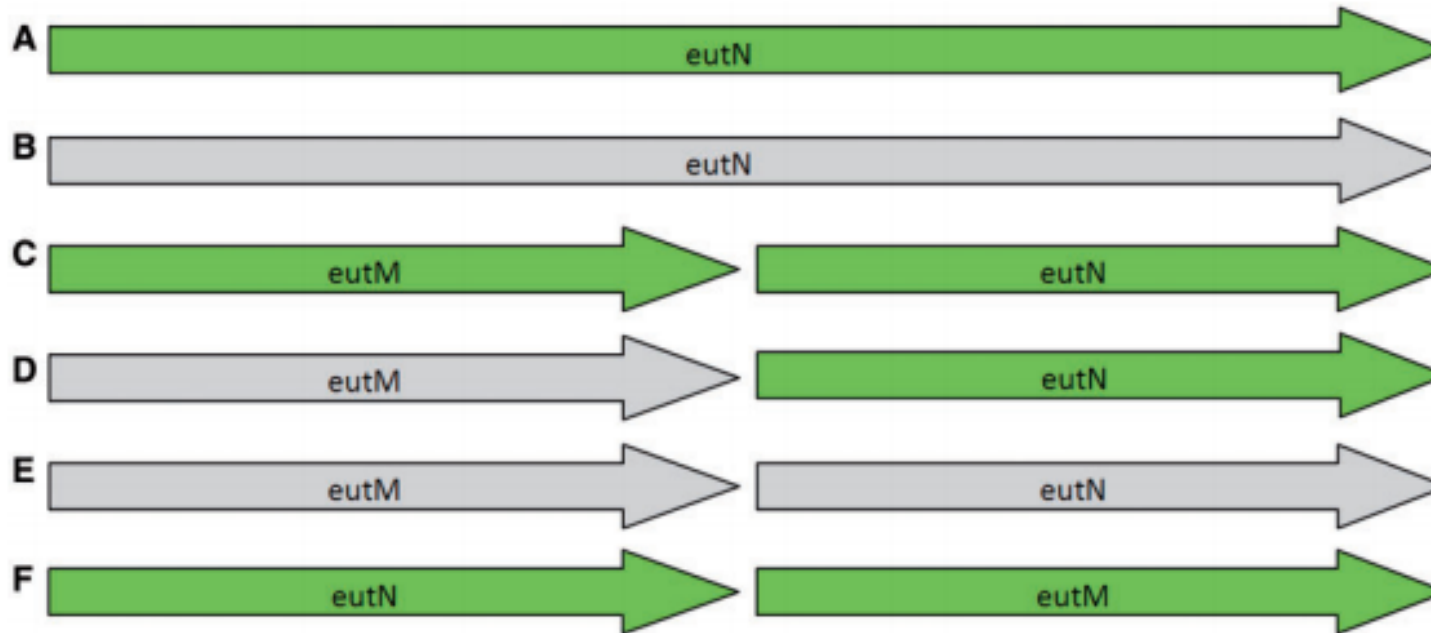
- If sequence homologous are found, may **not be functional** homologous
- If **no homology found**- limited information can be inferred
- Incorrect annotation can be **propagated** when similarity is over part on sequence not used in annotation
  - Multidomain proteins (HMM)
- Inconsistent annotation (**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

*Richardson and Watson. Briefings  
in Bioinformatics. 2012*



# Automatic annotation: limitations

Inconsistent annotation, en un gen descrito evento de fusión genica



*Salmonella typhi* CT18 (NC\_003198) and *Salmonella typhi* Ty2 (NC\_004631) there is a single ORF of 690 bp

**Figure 2:** The six different models present across 17 RefSeq entries for *Salmonella* species for the *eutM/eutN* locus. Green indicates normal gene/CDS features, lighter grey indicates gene features annotated as pseudogenes. (A) A single intact gene of 690 bp; (B) a single pseudogene of 690 bp; (C) two short intact genes ~300 bp in length; (D) one pseudogene and one intact gene, each ~300 bp in length; (E) two pseudogenes, each 300 bp in length; and (F) two intact genes with the order reversed.

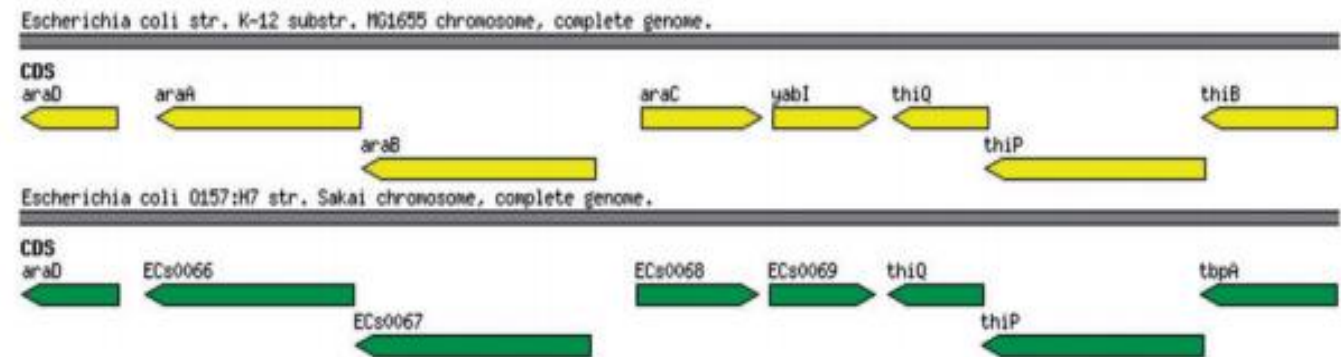
Richardson and Watson. Briefings in Bioinformatics. 2012

# Automatic annotation: limitations

## Inconsistent annotation

These two regions are more than 97% identical at the nucleotide level; however, the annotation differs considerably.

While *E. coli* K12 MG1655 contains features with gene names *araA*, *araB* and *araC*, the equivalent features in *E. coli* O157:H7 Sakai do not have those gene names and have been assigned uninformative locus tags



**Figure 3:** A syntenic block of genes showing inconsistent gene name annotations in *E. coli* K12 MG1655 and *E. coli* O157:H7 Sakai.

Richardson and Watson. *Briefings in Bioinformatics*. 2012

# Automatic annotation: limitations

- **Spelling mistakes**

- There are 128 proteins in UniProt that contain the word ‘syntase’, an incorrect spelling of the word ‘synthase’
- If a user was to visit any of these databases and search for ‘dihydrofolate synthase’ the misspelled entries would be omitted from the search results

*Richardson and Watson. Briefings  
in Bioinformatics. 2012*

# Automatic annotation: limitations

- ‘Same gene name, different product name’
  - The NCBI validation software specifically highlights when this occurs intra-genomically with the description ‘Same gene name, different product name’

**Table 1:** Different product names assigned to features with the gene name ‘int’ across 17 different RefSeq entries for *Salmonella* species

Gene name	Product name	Accession
int	bacteriophage integrase	NC.003198, NC.004631, NC.015761
int	Gifsy-1 prophage Int	NC.006905
int	hypothetical protein	NC.006905
int	Integrase	NC.003198, NC.004631, NC.006511, NC.012125
int	integrase (fragment)	NC.003198
int	phage integrase family site specific recombinase	NC.006905
int	putative cytoplasmic protein	NC.006905
Int	Putative integrase	NC.003384
int	putative integrase protein	NC.006905
int	putative P4-type integrase	NC.006905
int	putative phage integrase protein	NC.006905
int	site-specific recombinase, phage integrase family	NC.012125

Richardson and Watson. Briefings in Bioinformatics. 2012

# Automatic annotation: limitations

## Hypothetical proteins

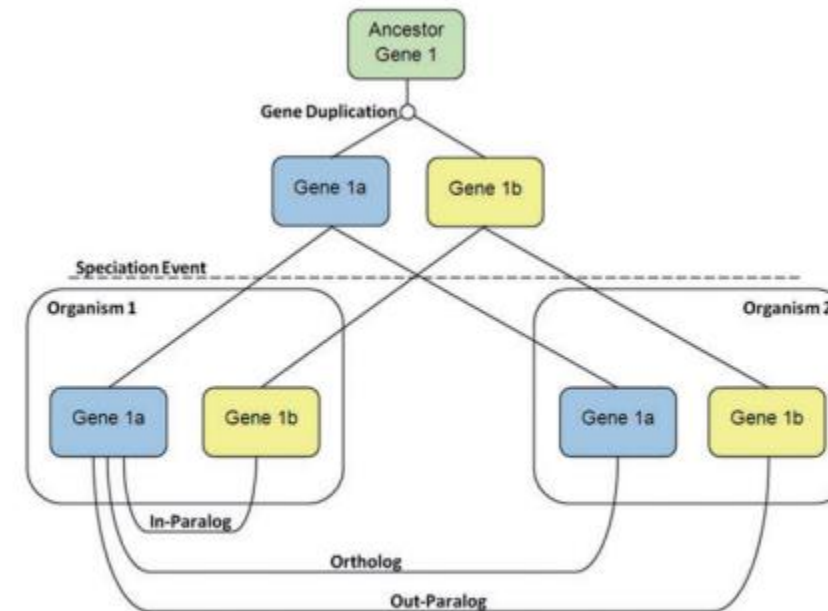
- These may be real genes with no known function or they may be artifacts of the gene prediction process.
- Often there are features which are only orthologous to other hypothetical features and do not contain any domains. These could either be regions with no functionality, a relic of the feature prediction software or the domains present have not been discovered yet
- Whether or not to include them is often a decision made by the annotation team and varies between groups
- As experimental data becomes more ubiquitous evidence tags should play a larger role in annotation.

*Richardson and Watson. Briefings  
in Bioinformatics. 2012*

# Automatic annotation: limitations

## Distinguishing orthologs from paralogs

orthologs tend to retain similar functions, whereas paralogs tend to diverge over time to perform different functions



**Figure 4:** A diagram displaying the processes that can lead to, and define, orthologs and paralogs. Gene duplication and speciation events create complex evolutionary relationships between genes.

*Richardson and Watson. Briefings  
in Bioinformatics. 2012*

# 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

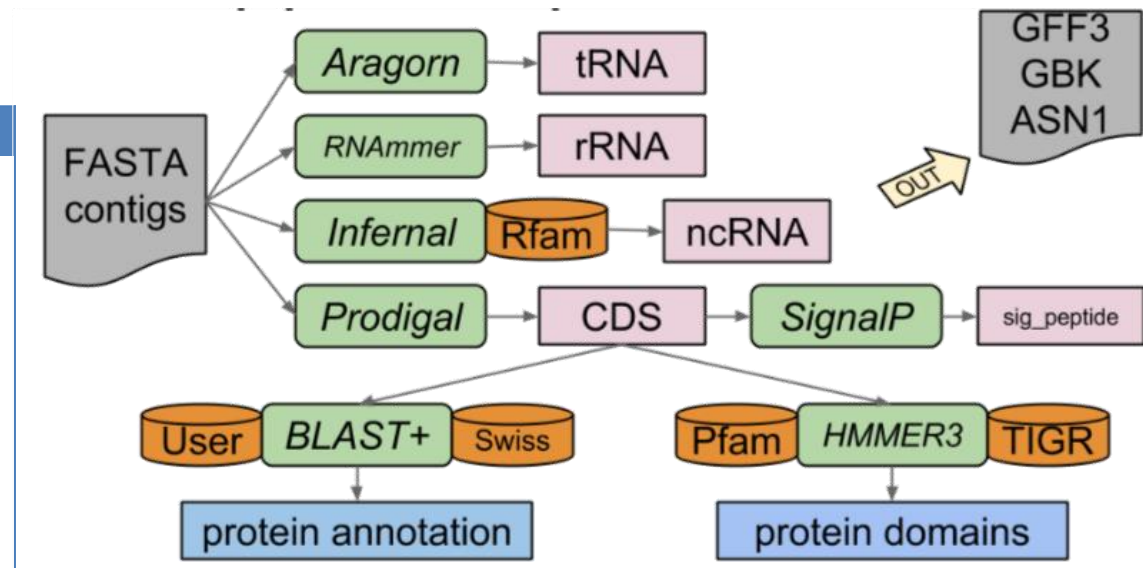
```
beta-lactamase (conceptual)
  class A beta-lactamase (HMM:NF033103)
  metallo-beta-lactamase (HMM:NF012229)
    subclass B1 metallo-beta-lactamase (HMM:NF033088)
      NDM family subclass B1 metallo-beta-lactamase (HMM:NF000259)
        subclass B1 metallo-beta-lactamase NDM-1 (allele)
        subclass 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 (Rapid prokaryotic genome annotation)

Seeman, Bioinformatics 2014

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



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



## Automatic annotation: Prokka

- 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

Prokka uses this method, but in a hierarchical manner, starting with a **smaller trustworthy database**, moving to medium sized but **domain-specific databases**, and finally to **curated models of protein families**

# Automatic annotation: Prokka

- Facts

- searching against smaller databases is faster
- searching against similar sequences is faster

- Idea

- start with small set of close proteins
- advance to larger sets of more distant proteins

- Prokka

- your own custom "trusted" set (optional)
- core bacterial proteome (default)
- genus specific proteome (optional)
- whole protein HMMs: PRK clusters, TIGRfams
- protein domain HMMs: Pfam

Prokka uses this method, but in a hierarchical manner, starting with a **smaller trustworthy database**, moving to medium sized but **domain-specific databases**, and finally to **curated models of protein families**

# Automatic annotation: Prokka

## Core Bacterial proteome

- Many bacterial proteins are conserved
  - experimentally validated
  - good annotations
- Prokka provides this database
  - derived from UniProt-Swissprot
  - only bacterial proteins
  - only accept evidence level 1 (aa) or 2 (RNA)
  - reject "Fragment" entries
  - extract /gene /EC\_number /product /db\_xref •
- First step gets ~50% of the genes
  - BLAST+ blastp, multi-threading to use all CPUs

# Automatic annotation: Prokka

- **Prokka has genus specific databases**
  - aim to capture "genus specific" naming conventions
  - derived from proteins in completed genomes
  - proteins are clustered and majority annotation wins
  - some annotations are rubbish though
- **Custom model databases**
  - I took COG/PRK MSAs and made HMMs
- **Existing model databases**
  - Pfam, TIGRfams are well curated
- **And if all else fails**
  - we always have our friend "hypothetical protein"

# 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
  - 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 gccgcccgc cgcggggcc acacggcaga gccgcctgaa
61 gccacgcgt gaggctgcac ttttcgagg gcttgacatc agggctctatg tttaagtctt
121 agctcttgct tacaagacc acggcaattc cttctctgaa gccctcgag cccacagcgc
181 ccctcgagc cccagctgc
//
  
```

# 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 LPS81_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="MAYQLNINWPEFLEKYWQKQPVLKNAFPDFVDPITPDELAGLA MEPEVDSRLVSLKNGKQASNGPFEHFDGLGETGWSLLAQAVNHNMPAAELVRPFRV LPDWRLLDLMISFSVPGGGVGVPHIDQYDFIQQWIGSRRWRVGDKLPHRQFCPPHALL HVDPFPPIIDEDLQPGDILYIPPGFPHDGIHETALNYSVGFGRGPNRDLISSFADYV LENDLGDEHYSDPDLTCREHPGRVEEYELERLRTHMIDMIRQPEDFKQWFGSFVTTTPR HELDIAPAEPPYEEEEVLDALLGGEKLSRLSGLRVLHIGDSFFVHSEQLDITDAEALD ALCRYTSLGQEELGSGLQNPFAVSELRLINQGYNYFEE"
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="MEQQLTIEMIADAFSYDITGFDGCEALNTFLKEHLKRQHDGQI LRGYALVSGDTPRLLGYITLGGSCFERGMLPSKTQQKKIPYQNPVTLGRLAIDKS VQGGQWGEMLVAHMRVVHMGASKAVGIYGLFVEALNEKAKAFYLRGLFIQLVDENSNL 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"



# Resistance prediction using WGS

Hendrisken et al. *Frontiers in Microbiology*. 2019.

Concordance between phenotypic susceptibility testing and WGS based predicted antimicrobial resistance

	Pathogen	No. of pathogens	AST method	No. of antimicrobials	Bioinformatic tool	Sequencing data	Concordance	Sensitivity	Specificity	Comment	References
2013	<i>S. Typhimurium</i>	49	MIC	17	ResFinder	Assembled, Velvet	99.74%			Disagreement: 7 isolates including 6 <i>E. coli</i> resistant to Spec	(7)
	<i>E. coli</i>	48		14							
	<i>E. faecalis</i>	50									
	<i>E. faecium</i>	50									
2013	<i>E. coli</i> (ESBL)	74	DD	7	BLASTn, selected panel	Assembled, Velvet		96%	97%	VM rate: 1.2%/M rate: 2.1%	(8)
	<i>K. pneumonia</i> (ESBL)	69									
2014	<i>S. aureus</i>	501	DD/MIC (Vitek)	12	BLASTn, selected panel	Assembled, Velvet		97%	99%	VM rate: 0.5%/M rate: 0.7%	(9)
2016	<i>C. jejuni</i>	32	MIC	9	BLASTx	Assembled, CLC-bio	99.2%			Lower concordance to Gen, Azi, Clin, Tel	(10)
	<i>C. coli</i>	82									
2016	<i>S. enterica</i>	104	MIC	14	ResFinder/ ARG-ANNOT/ CARD/BLAST	Assembled, CLC-bio	99.0%	99.2%	99.3%	Lower concordance to aminoglycosides/ $\beta$ -lactams	(11)
		536									
2017	<i>E. coli</i>	31	MIC	4	Custom DB based on ARDB/CARD/ $\beta$ -lactamase alleles			87%	98%	Neg. predictive value: 97% Pos. Predictive value: 91%	(12)
	<i>K. pneumonia</i>	24									
	<i>P. aeruginosa</i>	22									
	<i>E. cloacae</i>	13									
2017	<i>S. enterica</i>	50	MIC	4	ResFinder/ PointFinder	Assembled, SPAdes	98.4%			Disagreement: 2/2 <i>C.jejuni</i> to FQ/ERY 5 <i>E. coli</i> to COL (pmrB)	(13)
	<i>E. coli</i>	50		6							
	<i>C. jejuni</i>	50		4							
2018	<i>E. faecalis</i>	97	MIC	11	ResFinder/NCBI Pathogen DB/BLAST	Assembled, CLC-bio	96.5%				(14)
	<i>E. faecium</i>	100									
2018	<i>S. aureus</i>	501	DD/MIC	12	GeneFinder/ Mykrobe/ Typewriter	FASTQ/assembled, BLAST	98.3%			Disagreements: 0.7% predicted resistant 0.6% predicted susceptible	(15)
		491									
2018	<i>M. tuberculosis</i>	397	MIC	4	Cortex	Assembled	89.5%			97.1%/99.0% predicted R/S 97.5%/98.8% predicted R/S 94.6%/93.6% predicted R/S 91.3%/96.8% predicted R/S	(16)
		10,209									
2019	<i>H. pylori</i>	140	MIC (E-test)	5	ARIBA	FASTQ	99%			Phenotype issues to metronidazole	(17)

1) ESBL: Extended Spectrum Beta-Lactamase, 2) MIC: Minimum Inhibitory Concentration, 3) DD: Disk diffusion, 4) VM: Very Major, 5) M: Major, 6) R/S: Resistant/Susceptible, 7) SPEC: Spectinomycin, 8) GEN: Gentamicin, 9) AZI: Azithromycin, 10) CLIN: Clindamycin, 11) TEL: Telithromycin, 12) FQ: Fluoroquinolone, 13) ERY: Erythromycin, 14) COL: colistin.



# Resistance prediction using WGS

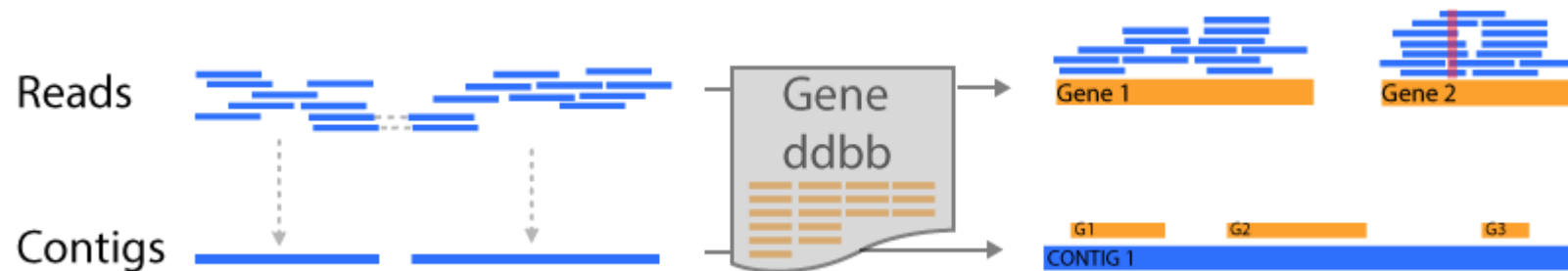
*Hendrisken et al. Frontiers in Microbiology. 2019.*

- **Huge list here:** [https://www.frontiersin.org/files/Articles/478239/fpubh-07-00242-HTML/image\\_m/fpubh-07-00242-t002.jpg](https://www.frontiersin.org/files/Articles/478239/fpubh-07-00242-HTML/image_m/fpubh-07-00242-t002.jpg)

Software	Type
SRST2	Mapping
Ariba	Mapping + assembly
ABRICATE	Assembly
ResFinder	Assembly

# Mapping vs Assembly

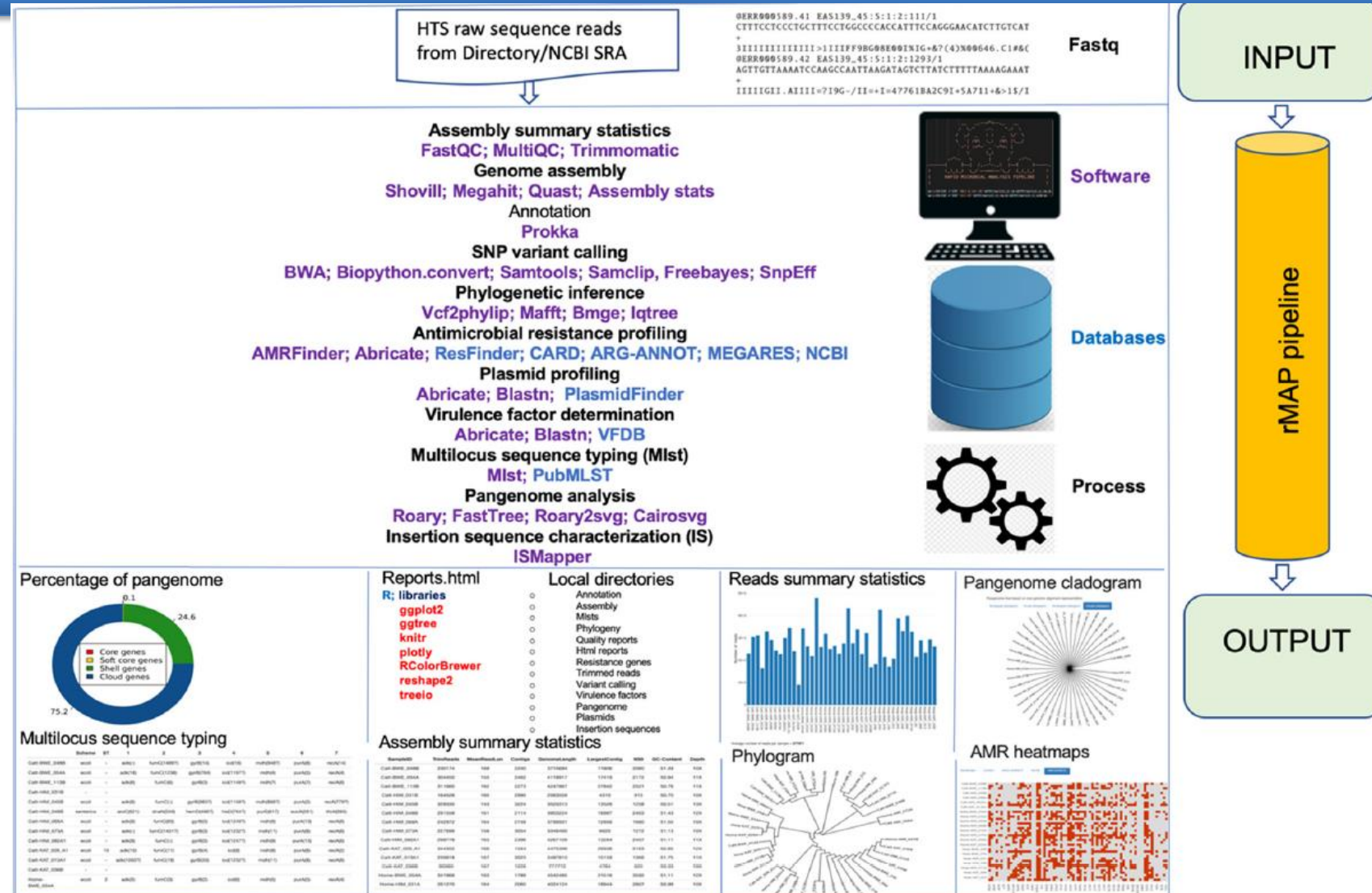
- **Functional annotation based on mapping (srst2)**
  - Pro: more resolute / high quality ddbb
  - Con: Unable to locate genes / no ab initio annotation
- **Functional annotation based on assembly (Resfinder)**
  - Pro: genes are located / related
  - Depend on assembly (close to repetitive regions)



rMAP - rapid microbial analysis pipeline- for ESKAPE bacterial group wgs data

The resistomes of ESKAPE pathogens (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacter species*)

Sserwadda & Mboowa,  
Microbiology, 2021



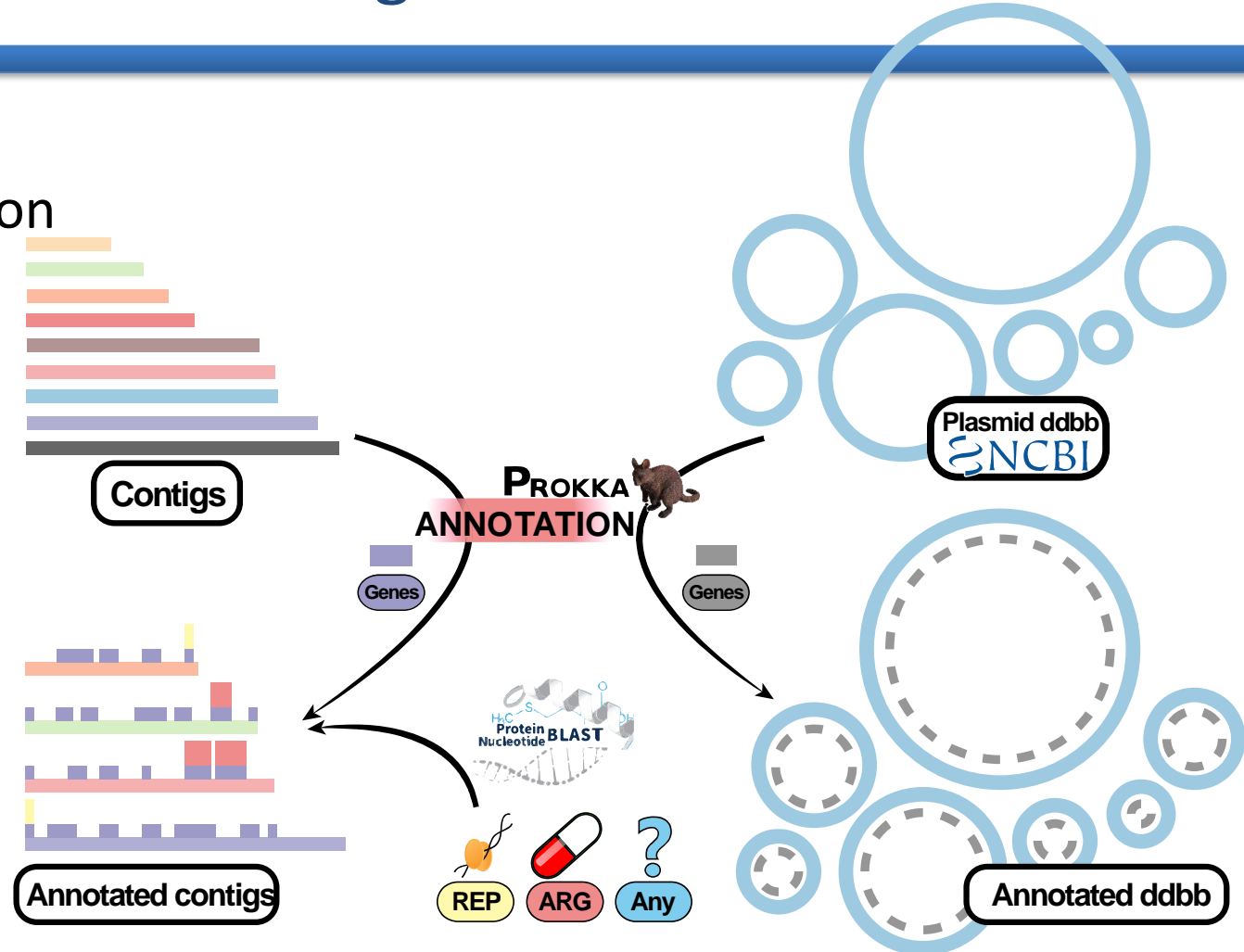
# Annotation visualization using PlasmidID

- Automatic annotation

- Prokka
  - DDBB plasmid
  - Contigs
- Gff to bed

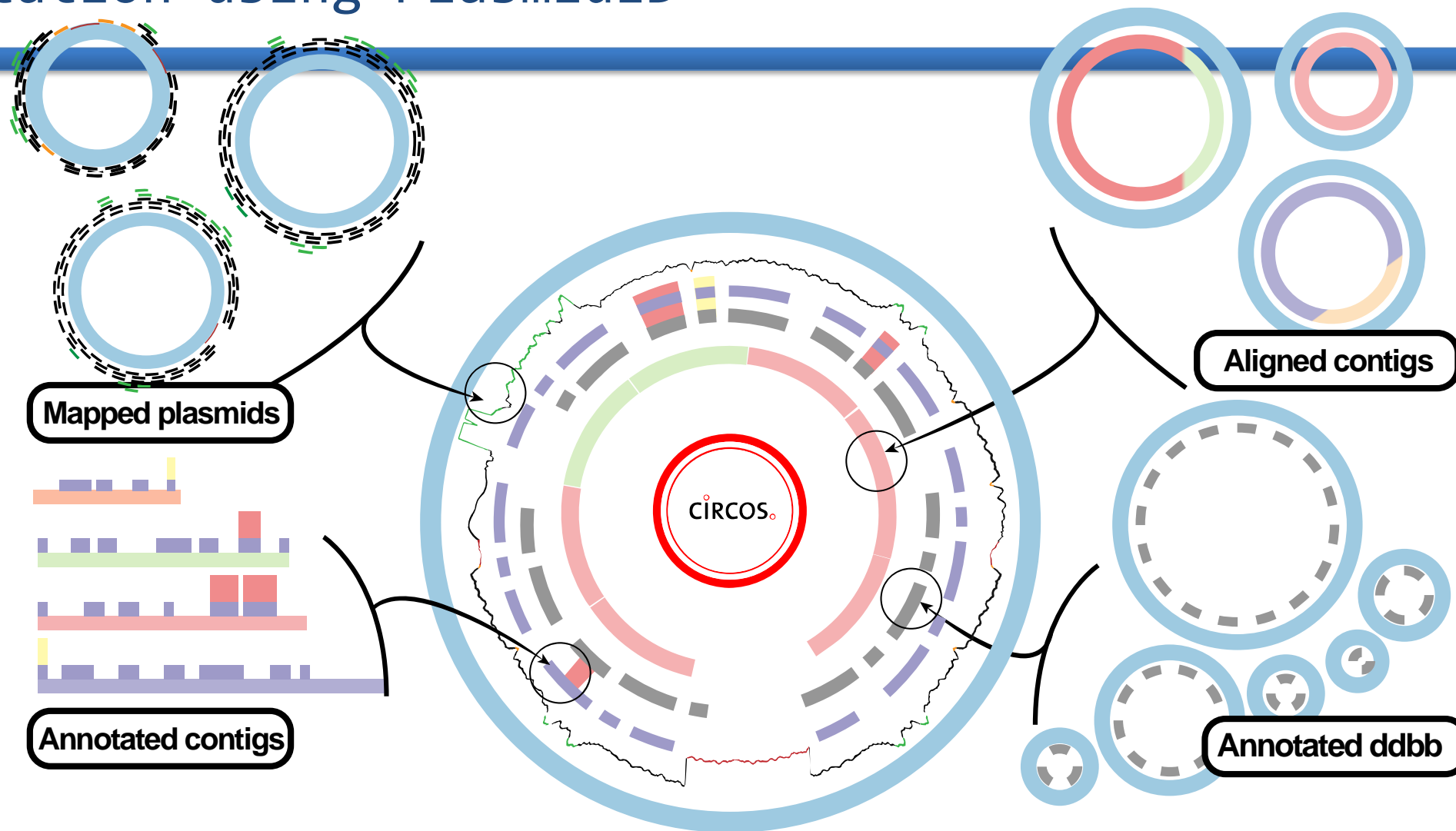
- Specific annotation

- BLAST+
- ABR & REP
- User input FASTA

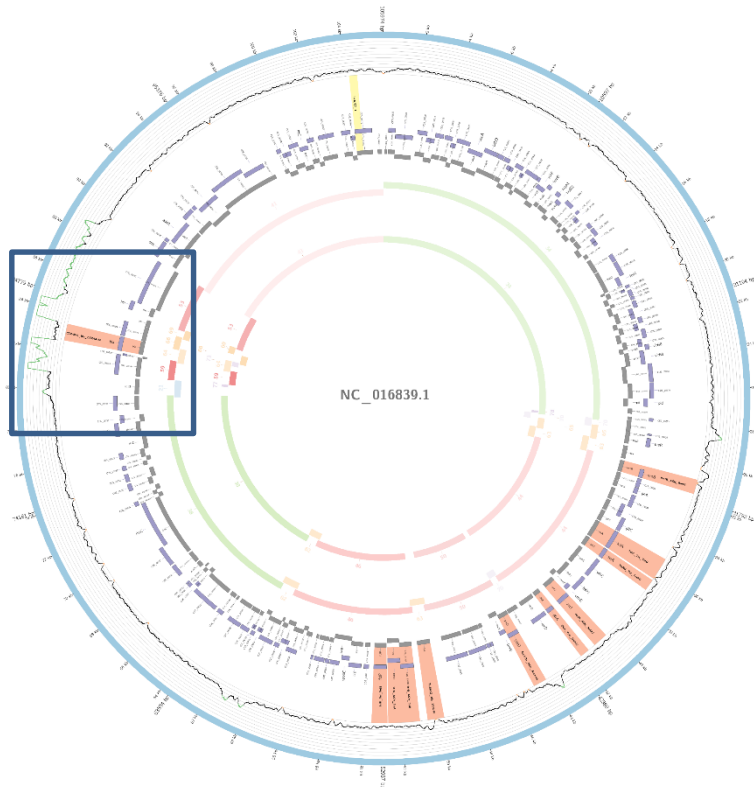




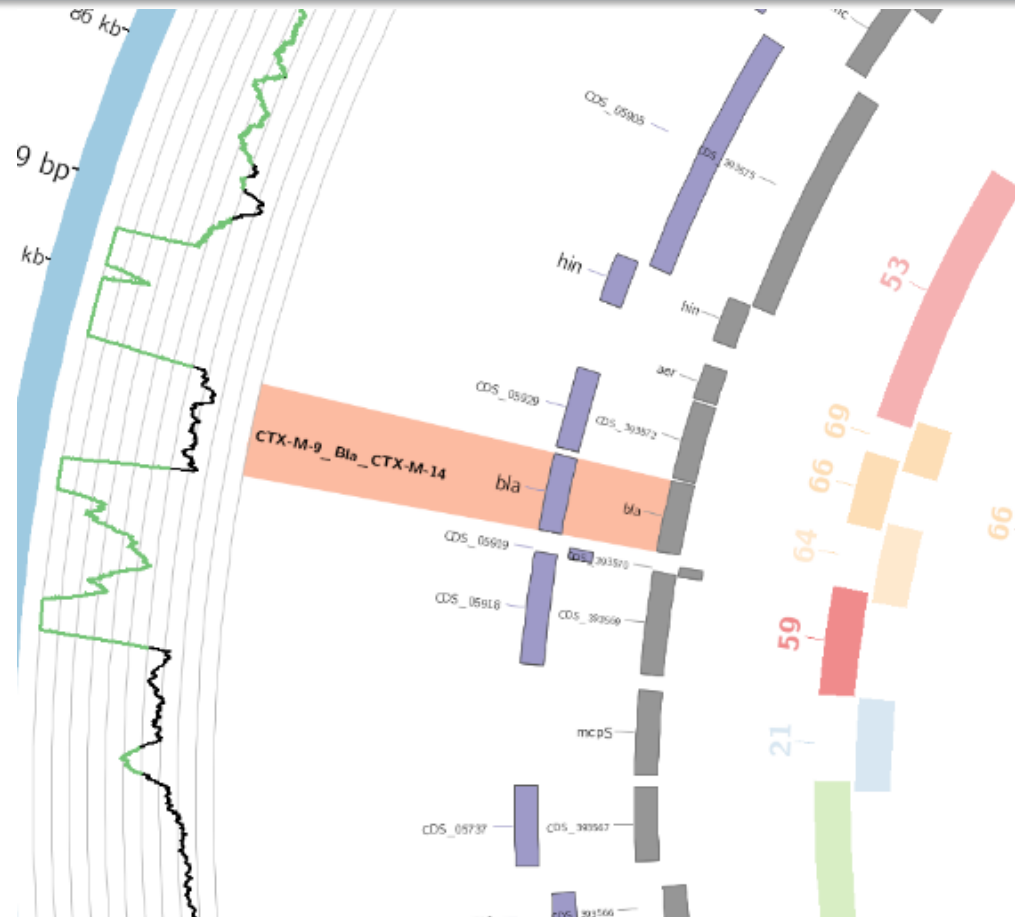
# Annotation using PlasmidID



# Annotation using PlasmidID



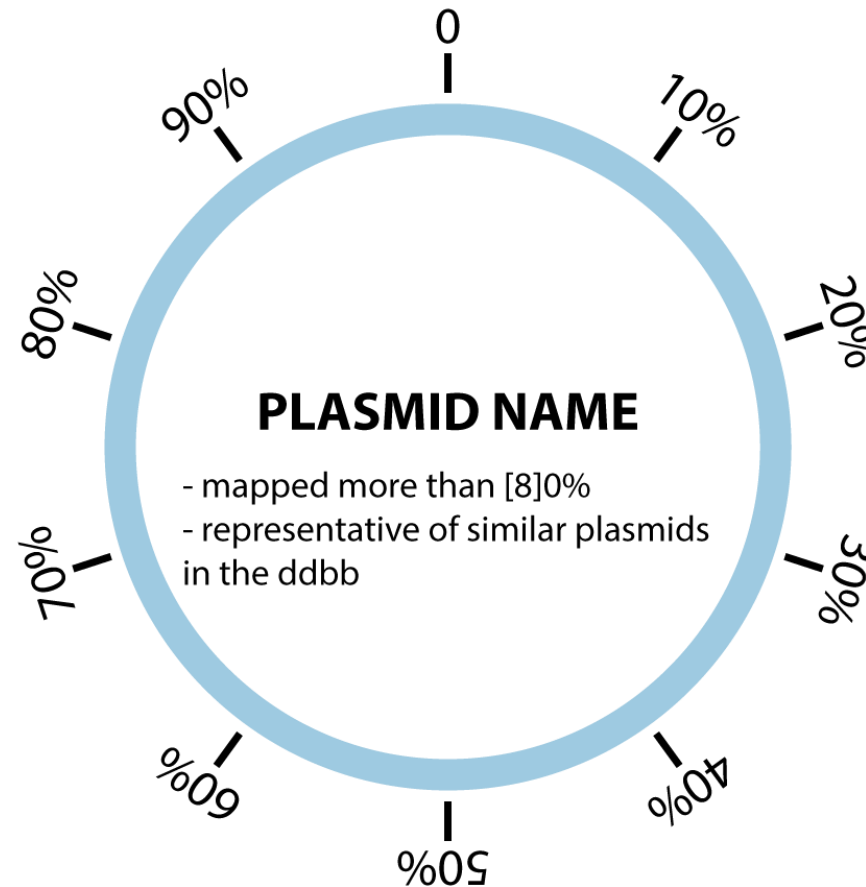
Annotation on short contigs





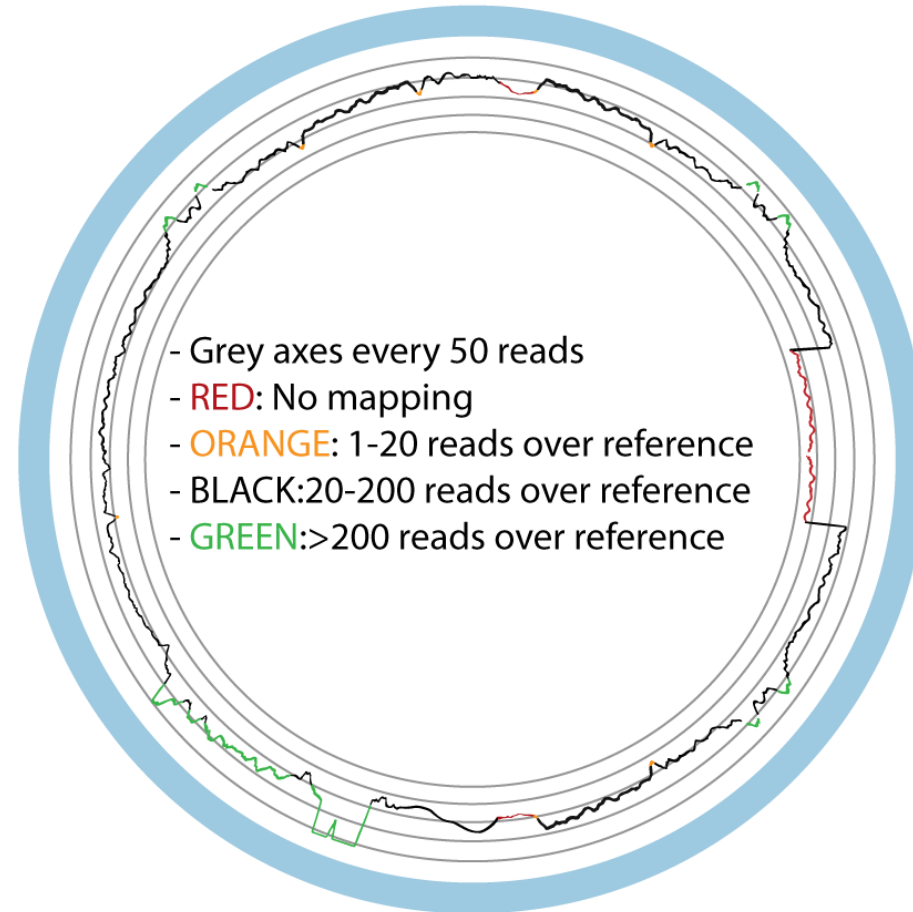
# Understanding the image: track by track

Plasmid Track



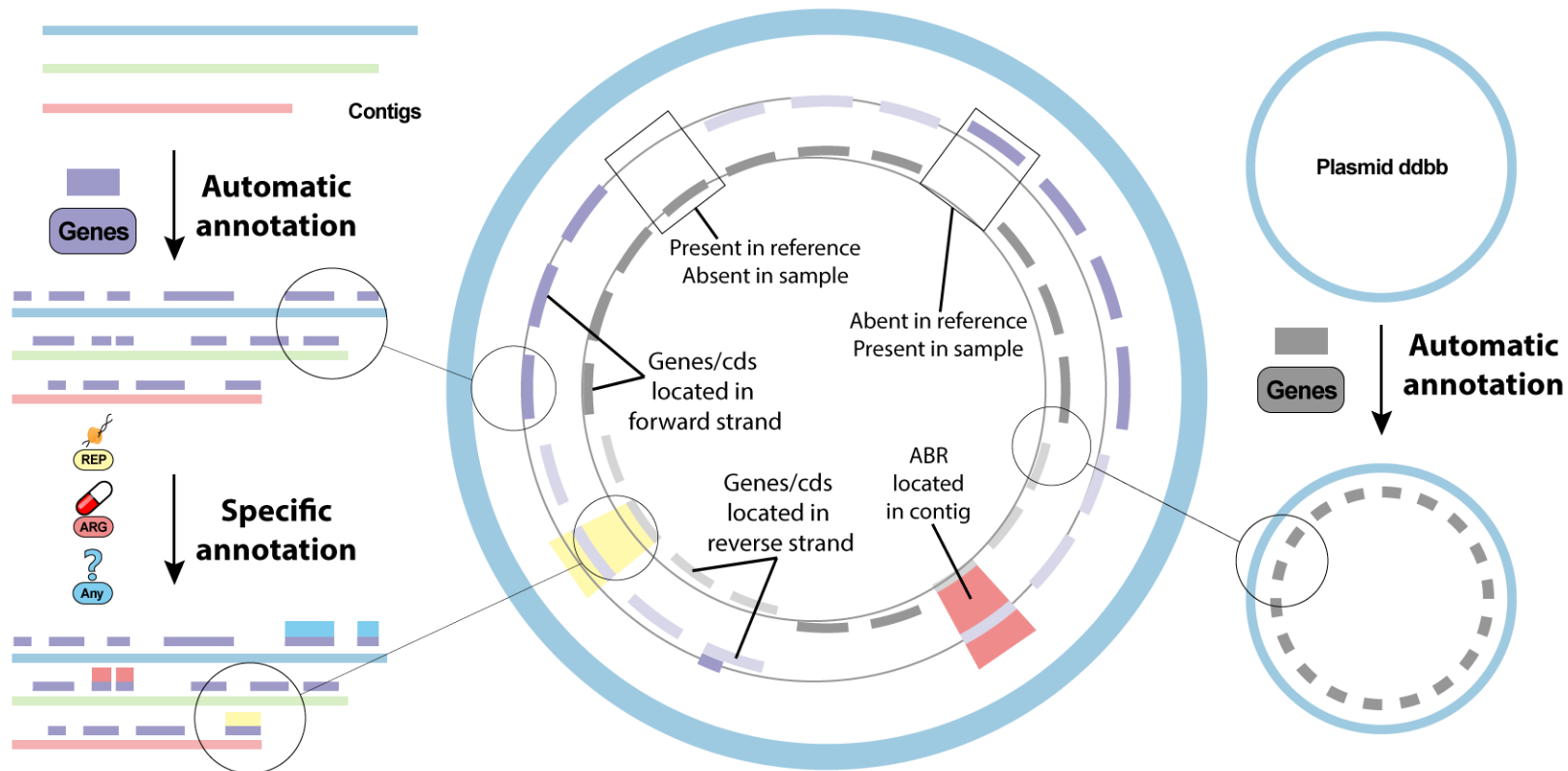
# Understanding the image: track by track

## Coverage Track



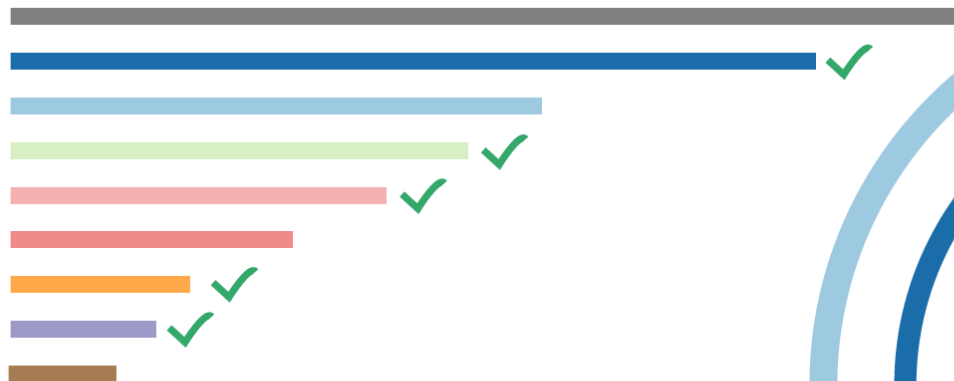
# Understanding the image: track by track

## Annotation Track



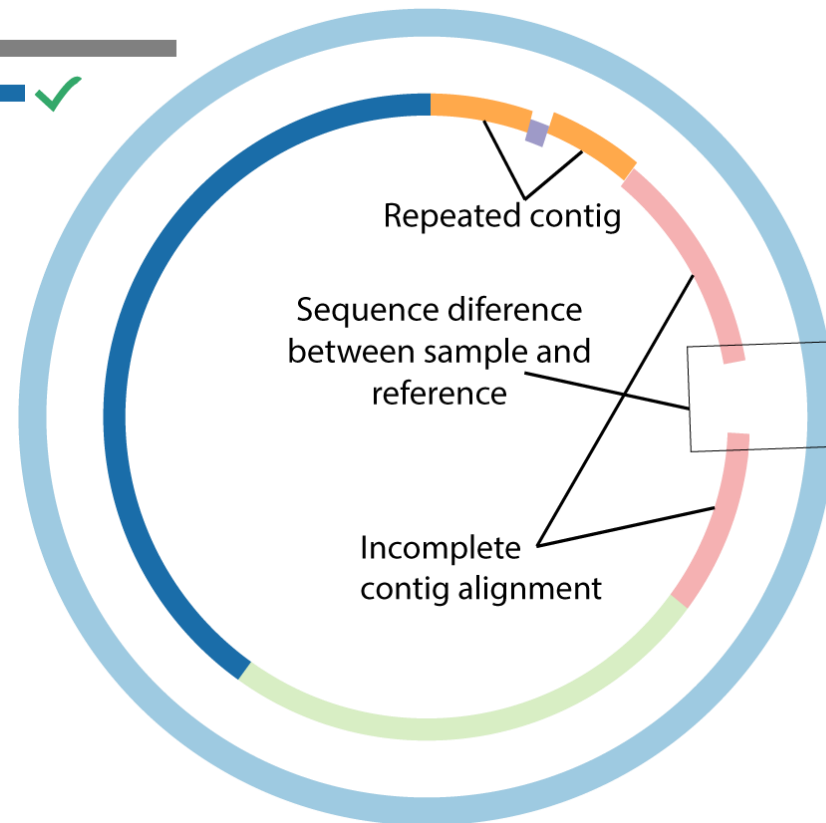
# Understanding the image: track by track

## Contig Track



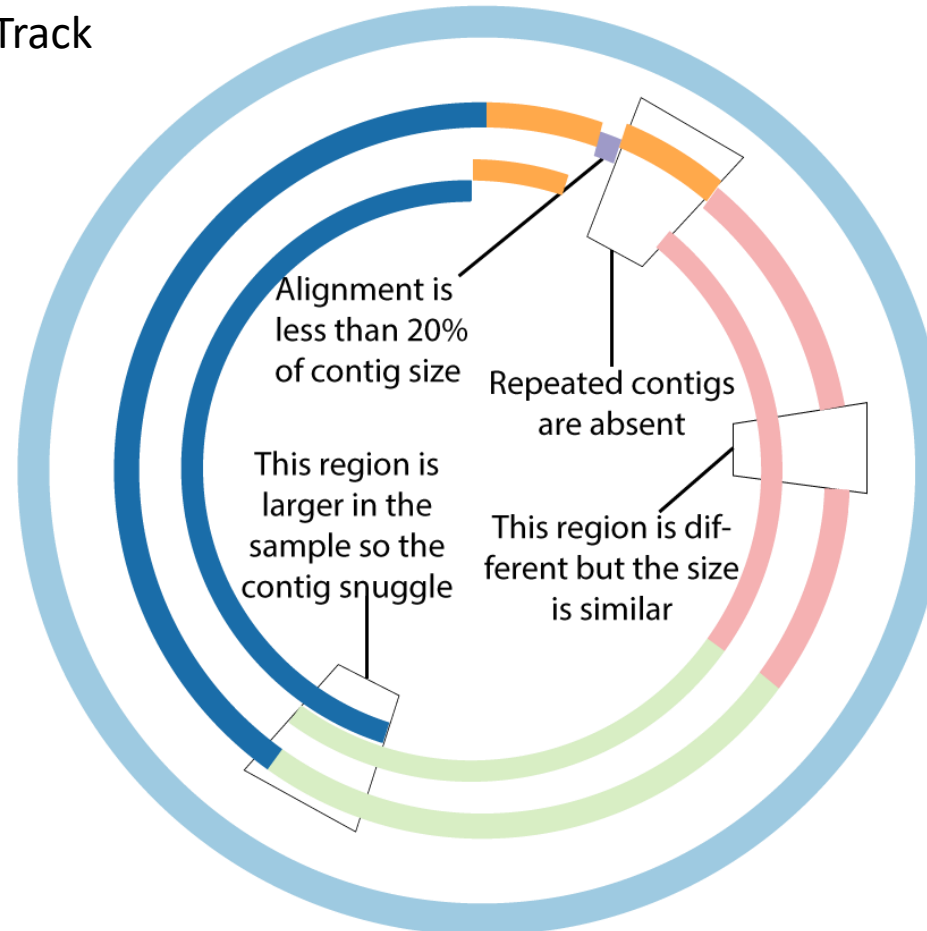
Contigs

- Represent **ONLY LOCAL** alignment
- Does not specify if the contig is inverted neither the total length of the contig



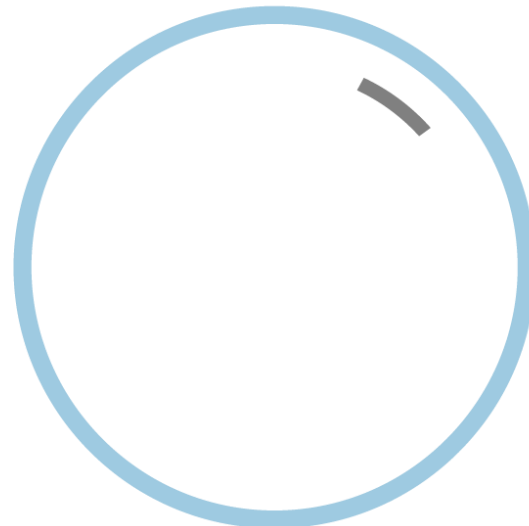
# Understanding the image: track by track

## Complete contig Track

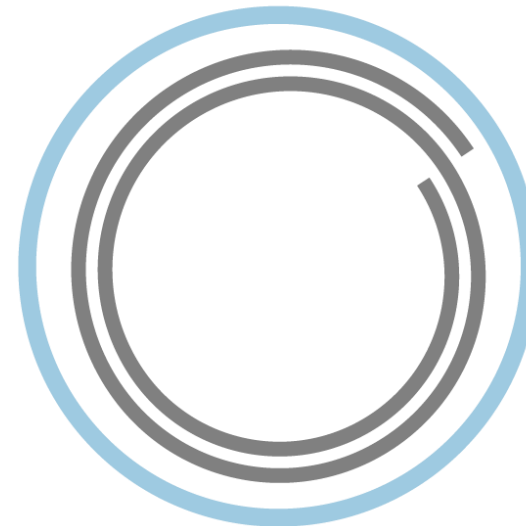


# Understanding the image: track by track

## Complete contig Track



Contig track



Complete contig track

# 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.

- File Entries Select View Goto Edit Create Write Run Graph Display
- Selected feature: bases 353 amino acids 117 CDS (/transl\_except=(pos:complement(3..5),aa:OT)
- Entry: ☒ foo.embl
- 

WUBLAST misc feature LTR

800 1600 2400 3200 4000 4800

foo;bar;baz;zap;foobar;ldggwdgjiwe
- ```

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
G A C G C T T A T T G C A C A C A T T G A G G A C T T A C A C A A C C A A C A A C G T A T G A A C T T A T G A A C T G A A A G C T T C
20 40 60 80
C T A G A C G A A A T A A C G T T G T A T A C T C C T G A A T G T G T T G T A G T T C G T T A G T T G A C A T A C T T T G A A T A G C T T G A C T T T C G A A A G
I O 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 *
    
```
- | source       | 1    | 41173 |
|--------------|------|-------|
| CDS          | 1    | 353   |
| WUBLASTN HIT | 3    | 353   |
| misc_feature | 498  | 539   |
| CDS          | 784  | 1821  |
| misc_feature | 790  | 1788  |
| LTR          | 2383 | 2740  |
| CDS          | 3838 | 5811  |
| CDS          | 7720 | 8379  |
| CDS          | 8379 | 10453 |

SPBC16A3.01, spn3, septin homolog spn3, len:117aa, i der

MLCB458.06, fas, probable type I fatty acid synthase, le

Pfam match to entry adh\_zinc PF00107, Zinc-binding d ehy

tfl like LTR

SPBC16A3.03c

SPBC16A3.04, unknown, len:220aa, similar eg. to YIL0 93C

# Thanks for your attention!

---