Session 4.2 – Annotation

Isabel Cuesta

BU-ISCIII

Unidades Comunes Científico Técnicas – SGSAFI-ISCIII

13 al 17 Noviembre 2023

3ª Edición

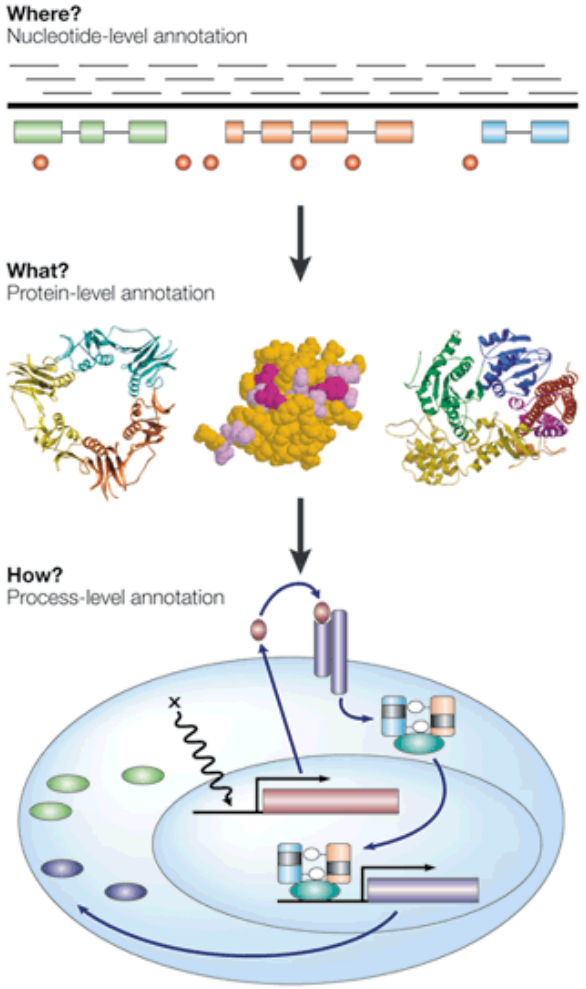
Programa Formación Continua, ISCII

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



Lincoln Stein, 2001. Nat.Review Nature R

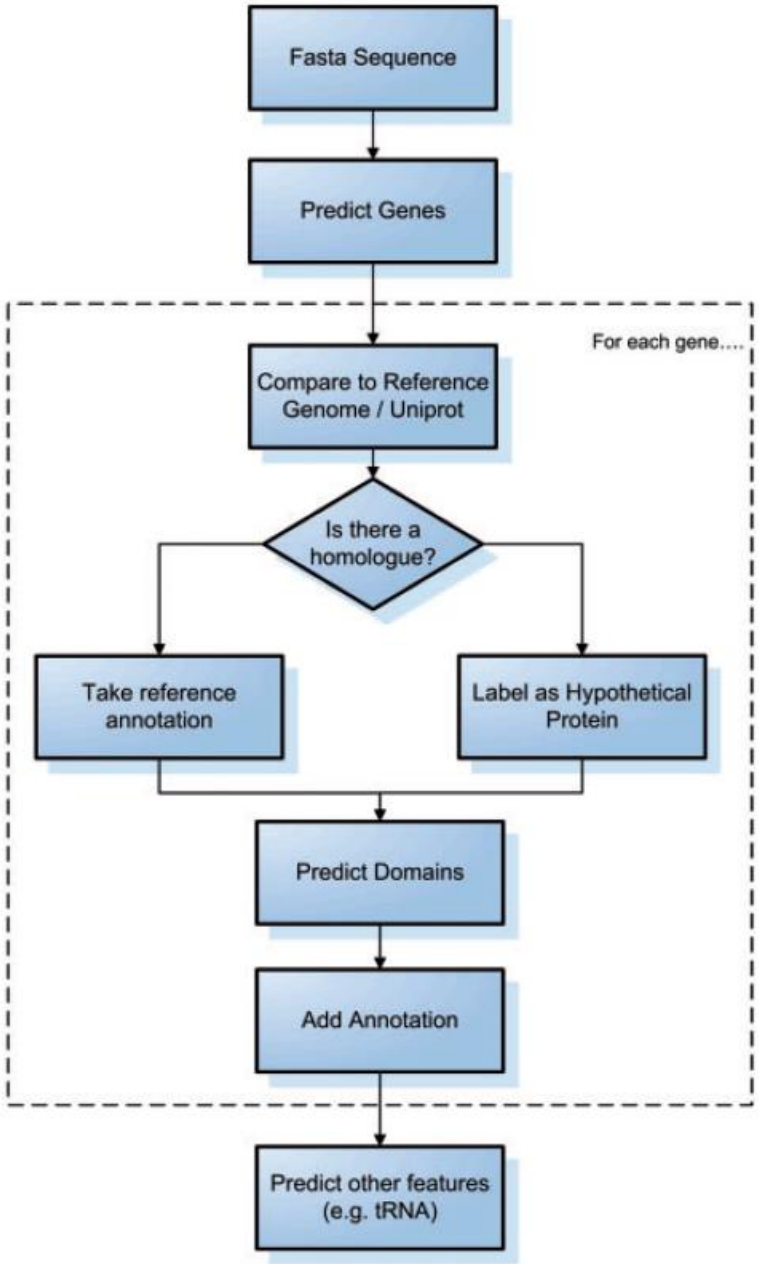
Main categories

- **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

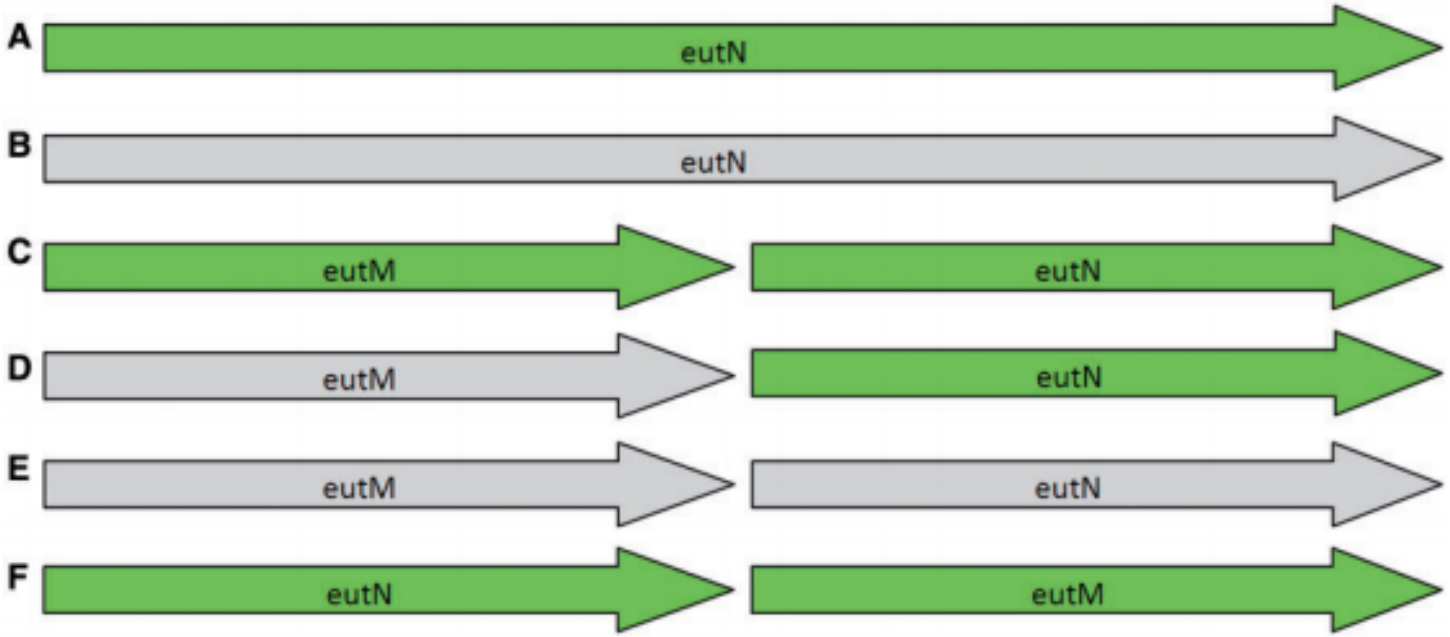
Automatic annotation: limitations

- 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. coliK12MG1655 contains features with gene names araA, araB and araC, the equivalent features in E. coli 0157:H7 Sakai do not have those gene names and have been assigned uninformative locus tags

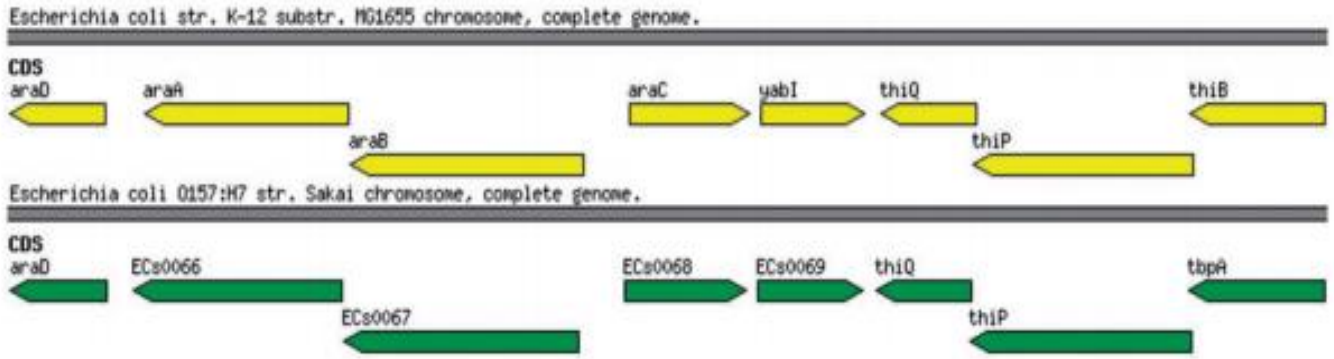


Figure 3: A syntenic block of genes showing inconsistent gene name annotations in *E.coli KI2 MGI655* and *E. coli 0157:H7 Sakai*.

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 I: 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

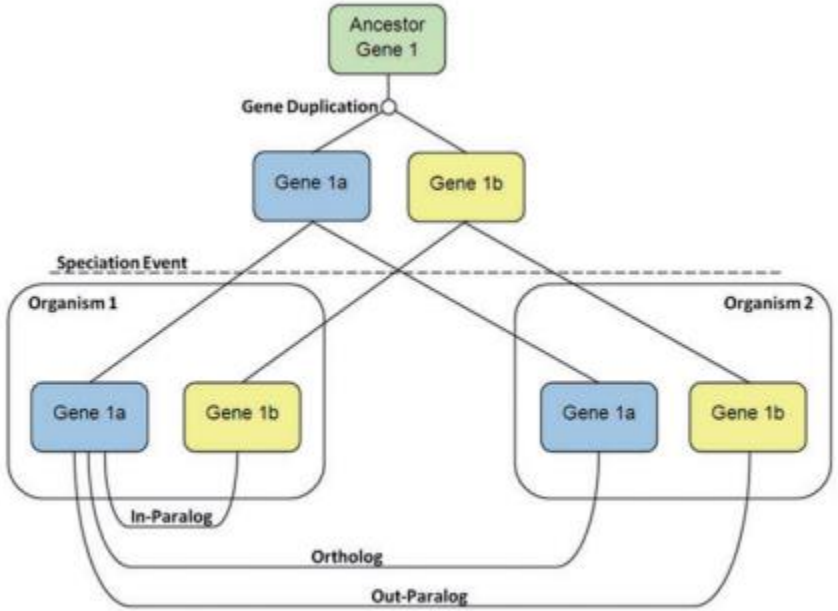
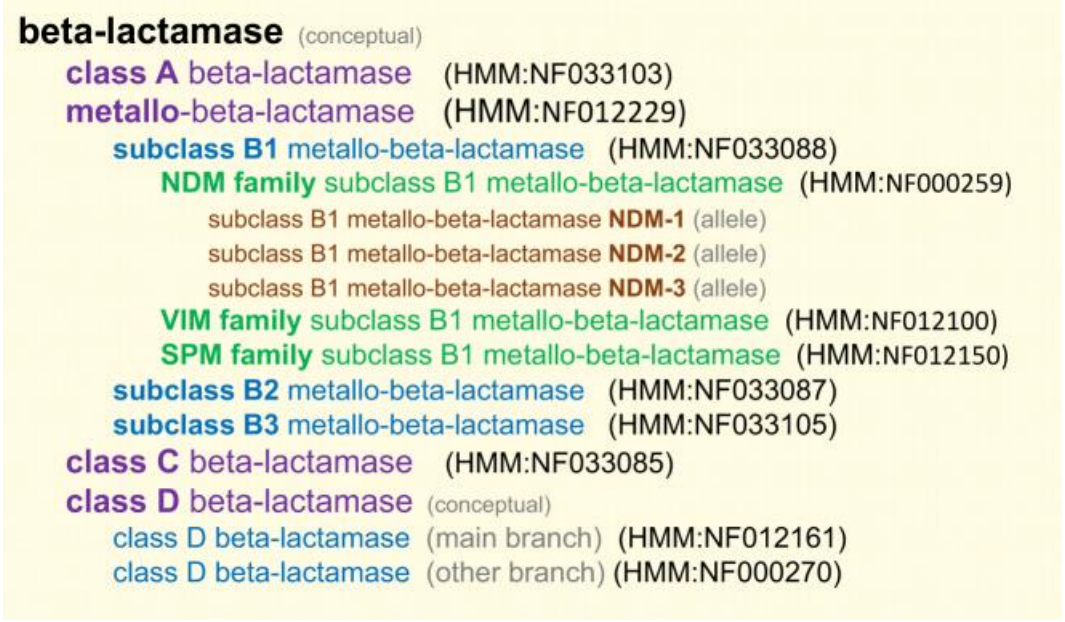


Figure 4: A diagram displaying the processes that can lead to, and define, orthologs and paralogs. Gene duplication and speciation events create complex evolutionary relationships between genes.

Richardson and Watson. Briefings in Bioinformatics. 2012

Automatic annotation: limitations

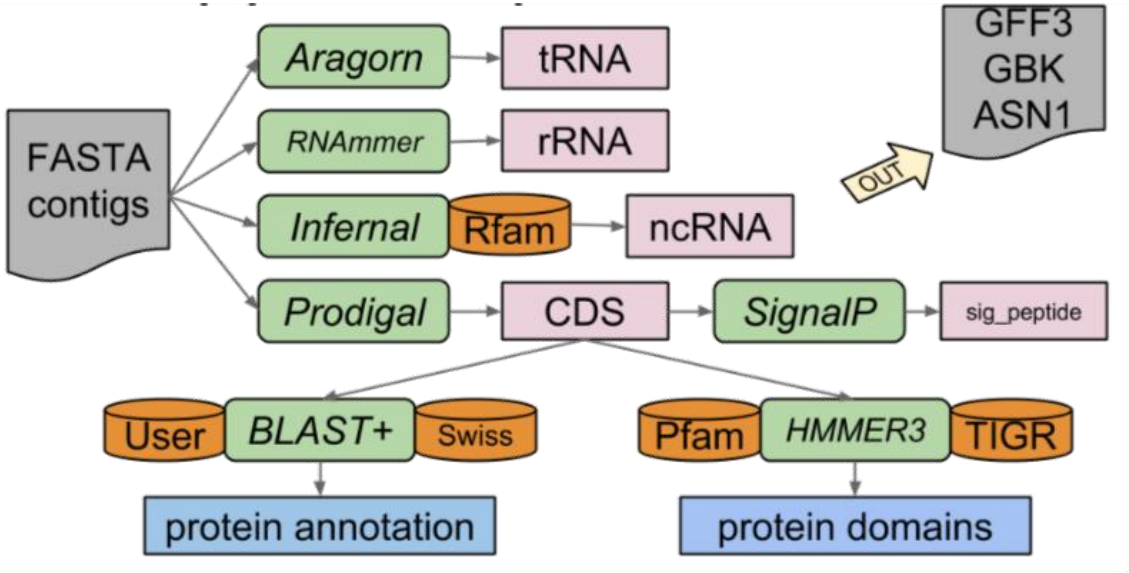
- **RefSeq is one attempt to standardize and improve the quality of genome annotation**
  - WP_ prefix. All identical proteins regardless of species
  - Standard classification



Automatic annotation: Prokka (Rapid prokaryotic genome annotation)

Seeman, Bioinformatics 2014

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



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**

Viral genome annotation

PROPERTIES

- DNA, ssDNA, dsDNA, RNA, ssRNA, fragmented RNA
- Non-coding ORF
- Coding ORF
- Overlapping reading frames
- Non-standard nomenclature for viral gene products
- RNA editing (the RNA polymerase co-transcriptionally adds one or two nucleotides that are not on the template, including multiple proteins in a single gene. Annotated protein sequence does not match the expected translated nucleotide sequence)
- Ribosome slippage (Allow viruses to produce two proteins from a single mRNA transcript by having the ribosome 'slip' one or two nucleotides along the mRNA transcript, thus changing the reading frame.)
- Viral sequence variability

Viral genome annotation

APPROACHES

- Identification hallmark genes conserved within known virus families
- Detection of short nucleotide sequences believed to be enriched in viruses (DeepVirFinder: reference-free and alignment-free machine learning method, for identifying viral sequences in metagenomic data using deep learning. Ren et al., Quan Biol 2020)
- Tools for specific virus (i.e. Influenza)

Viral genome annotation

LIMITATIONS

- Pitfalls that can lead to false-positives or false negatives
- Some tools are limited by minimum sequence length
- Detection of a limited range of virus families.
- High diversity of DNA and RNA viruses presents a challenge for development of a universal annotator

VAPiD: a lightweight cross-platform viral annotation pipeline and identification tool to facilitate virus genome submissions to NCBI GenBank

- Users can provide a specified reference from which to annotate all viruses
- Provide their own BLASTn database
- Force VAPiD to search NCBI's NT database

https://github.com/rcs333/VAPiD

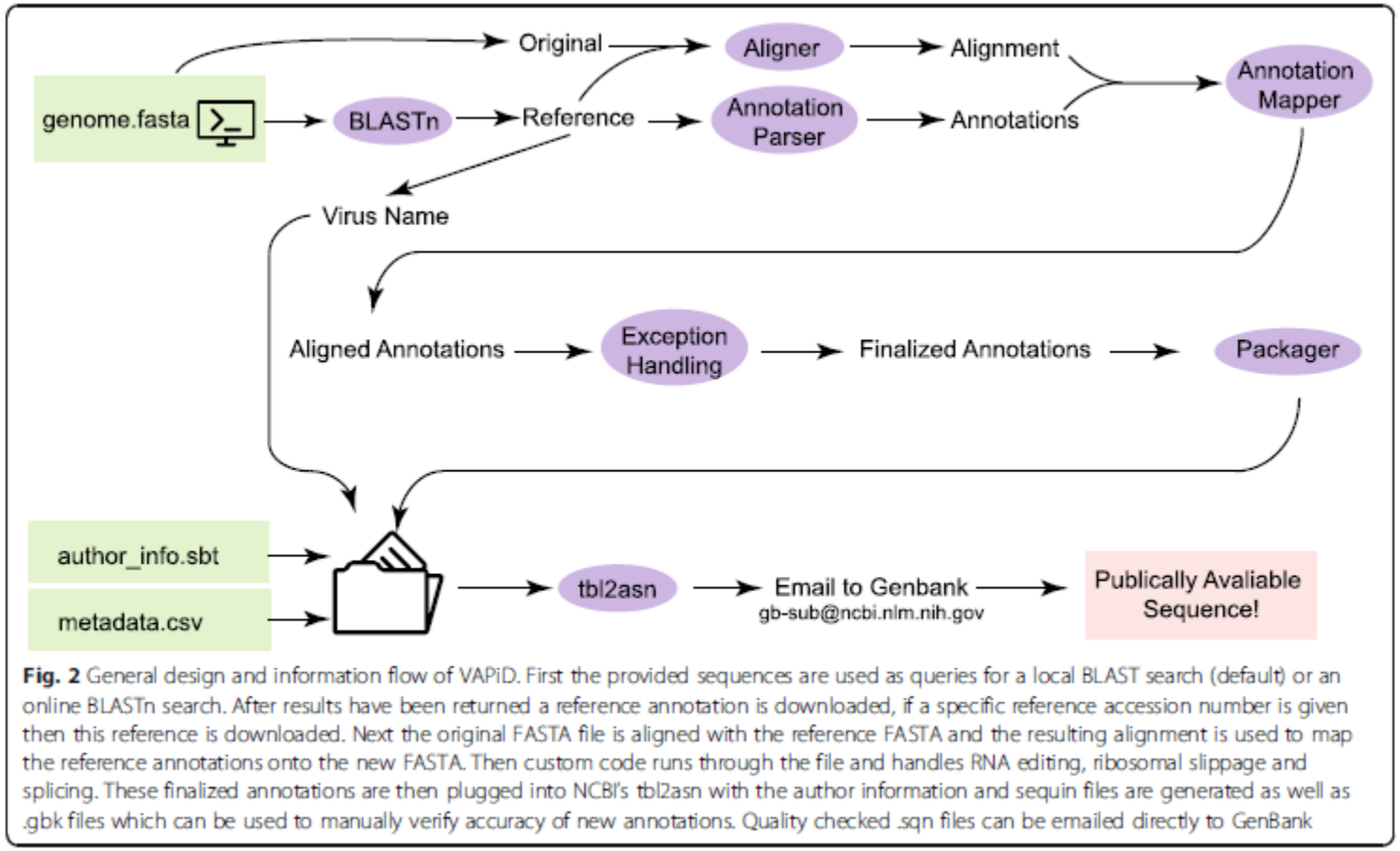


Fig. 2 General design and information flow of VAPiD. First the provided sequences are used as queries for a local BLAST search (default) or an online BLASTn search. After results have been returned a reference annotation is downloaded, if a specific reference accession number is given then this reference is downloaded. Next the original FASTA file is aligned with the reference FASTA and the resulting alignment is used to map the reference annotations onto the new FASTA. Then custom code runs through the file and handles RNA editing, ribosomal slippage and splicing. These finalized annotations are then plugged into NCBI's tbl2asn with the author information and sequin files are generated as well as .gbk files which can be used to manually verify accuracy of new annotations. Quality checked .sqn files can be emailed directly to GenBank

Shean BMC Bioinformatics 2019

VAPiD: a lightweight cross-platform viral annotation pipeline and identification tool to facilitate virus genome submissions to NCBI GenBank

ALGORITHM STEPS:

1. Find the correct reference sequence.
2. Gene locations are stripped from the reference
3. Pairwise nucleotide alignment between the reference and the submitted sequence is generated using MAFFT
4. The relative locations of the genes on the reference sequence are then mapped onto the new sequence
5. Gene names are taken from the annotated reference sequence
6. Spellchecking
7. RNA editing
8. Ribosome slippage
9. Genbank file generation

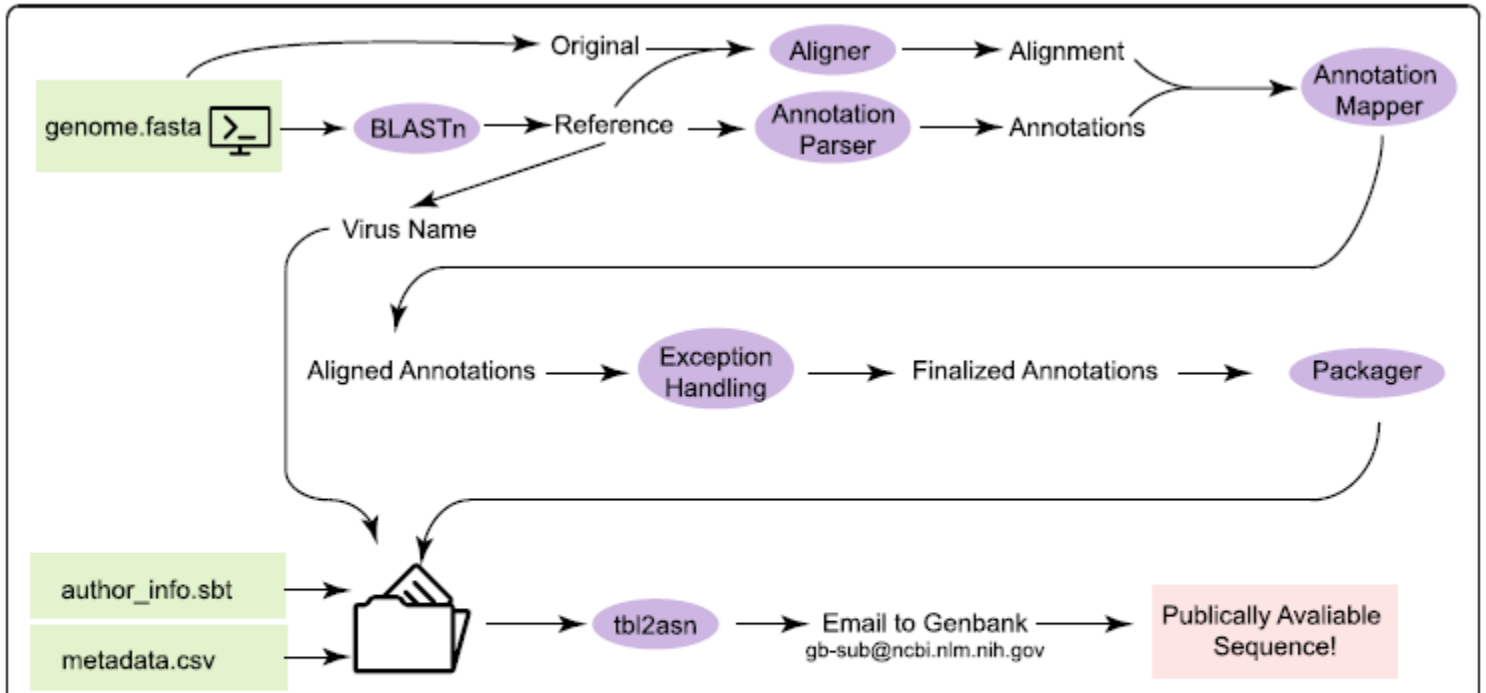


Fig. 2 General design and information flow of VAPiD. First the provided sequences are used as queries for a local BLAST search (default) or an online BLASTn search. After results have been returned a reference annotation is downloaded, if a specific reference accession number is given then this reference is downloaded. Next the original FASTA file is aligned with the reference FASTA and the resulting alignment is used to map the reference annotations onto the new FASTA. Then custom code runs through the file and handles RNA editing, ribosomal slippage and splicing. These finalized annotations are then plugged into NCBI's tbl2asn with the author information and sequin files are generated as well as .gbk files which can be used to manually verify accuracy of new annotations. Quality checked .sqn files can be emailed directly to GenBank

Shean BMC Bioinformatics 2019

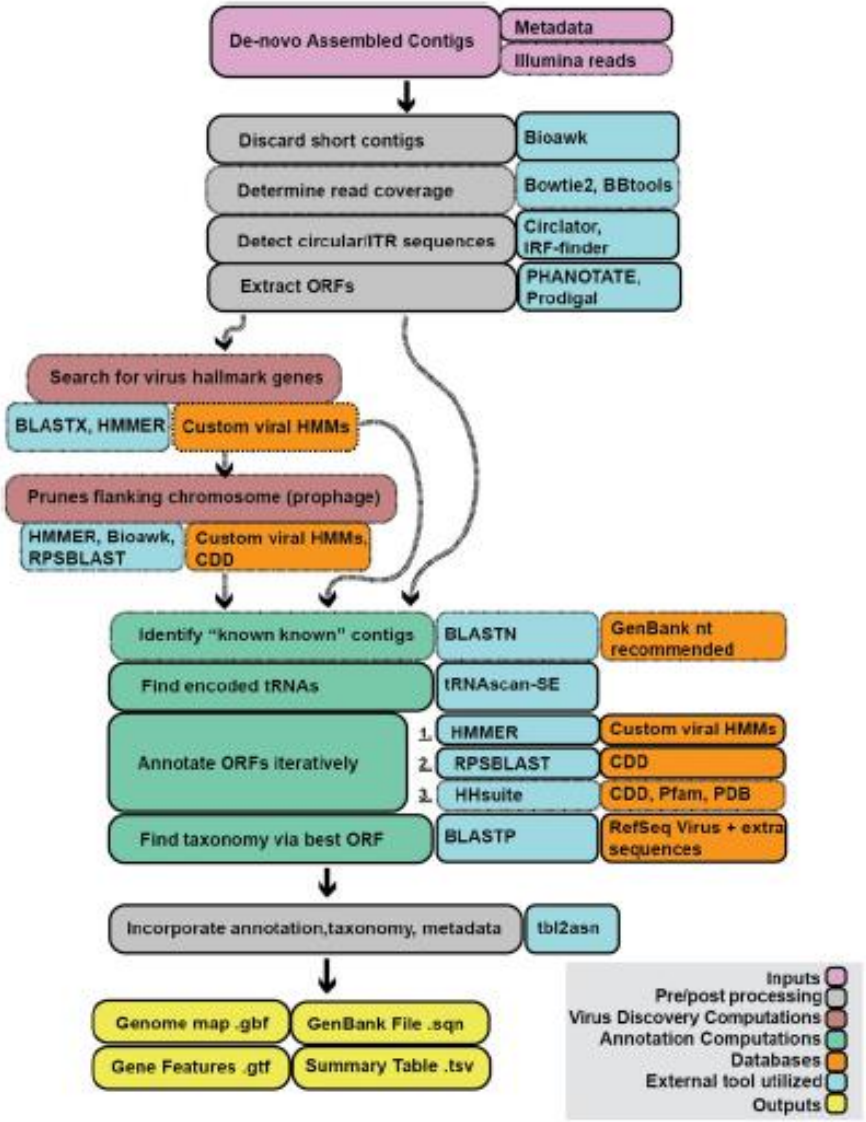
VAPiD: a lightweight cross-platform viral annotation pipeline and identification tool to facilitate virus genome submissions to NCBI GenBank

LIMITATIONS

- VAPiD is not the preferred annotation tool for novel or extremely divergent viral species
- Not perform ab initio gene annotation
- Any errors that are in the downloaded reference will be transferred to the new genome (i.e. misspelling
- VAPiD performs best on high-quality and accurate reference sequences

Shean BMC Bioinformatics 2019

Cenote-Taker 2



Tisza et al., Virus Evolution 2021

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
//
```

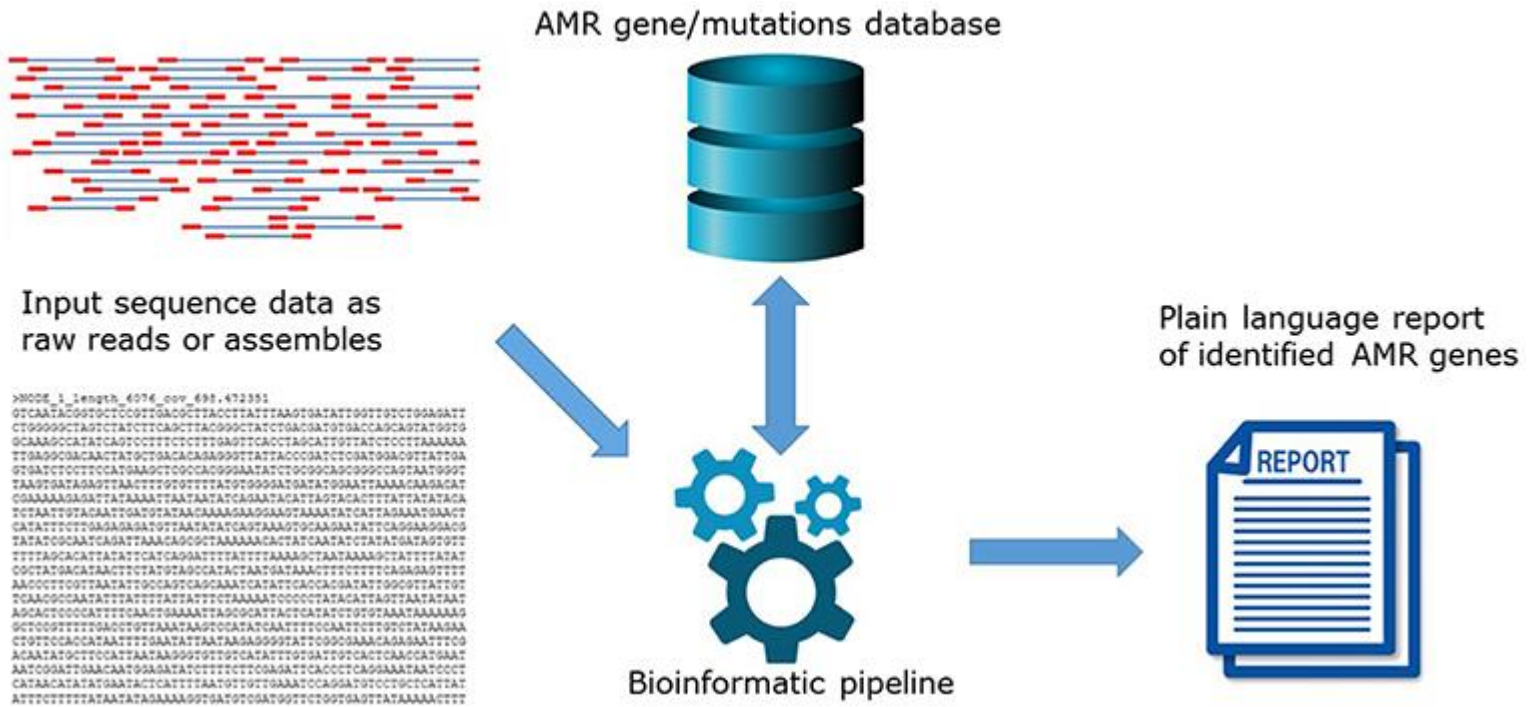
Annotation format: gbk

- 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"
```

Resistance prediction using WGS

Hendrisken et al. Frontiers in Microbiology. 2019.



Resistance prediction using WGS

Hendrisken et al. Frontiers in Microbiology. 2019.

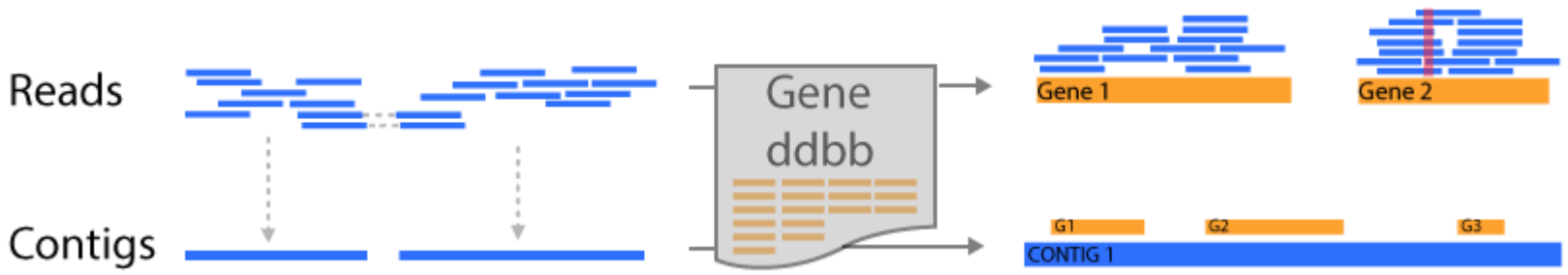
- **Huge list here: https://www.frontiersin.org/files/Articles/478239/fpubh-07-00242-HTML/image_m/fpubh-07-00242-t002.jpg**

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

Mapping vs Assembly

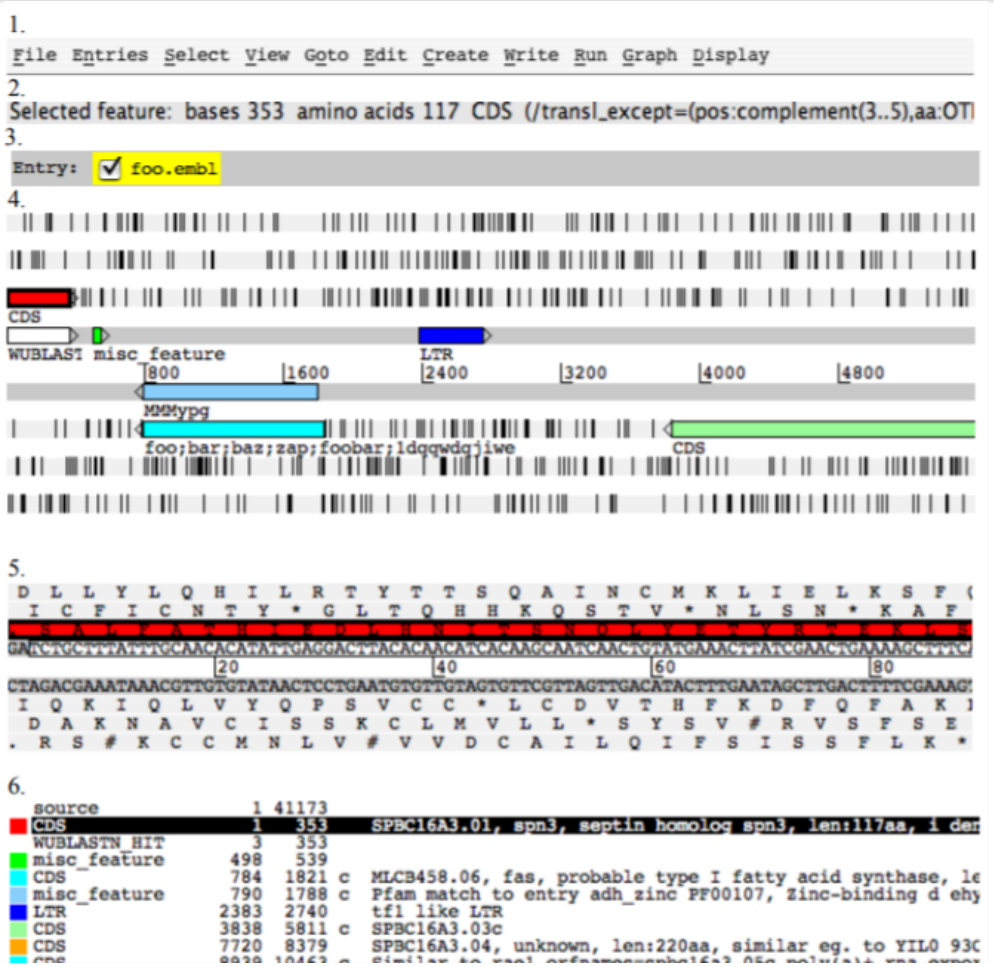
- **Functional annotation based on mapping (srst2)**
  - Pro: more resolutive / high quality ddbb
  - Con: Unable to locate genes / no ab initio annotation

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



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.



Thanks for your attention!