Bacterial gene and genome annotation

13/06/2024

Dr. Luu Phuc Loi (luu.p.loi@googlemail.com)

Content

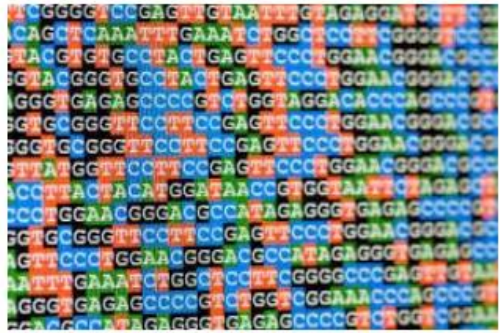
1. Bacterial genomes
2. Bacterial Genes
3. Genome Annotation
4. Prokka
5. Curating genomes
6. Bakta

Bacterial genomes: small genomes

6,000,000,000

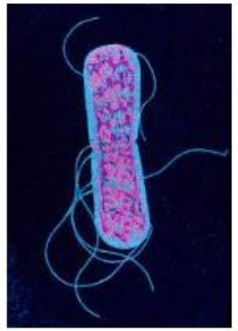
letters

30,000 genes



Genome

A T G C



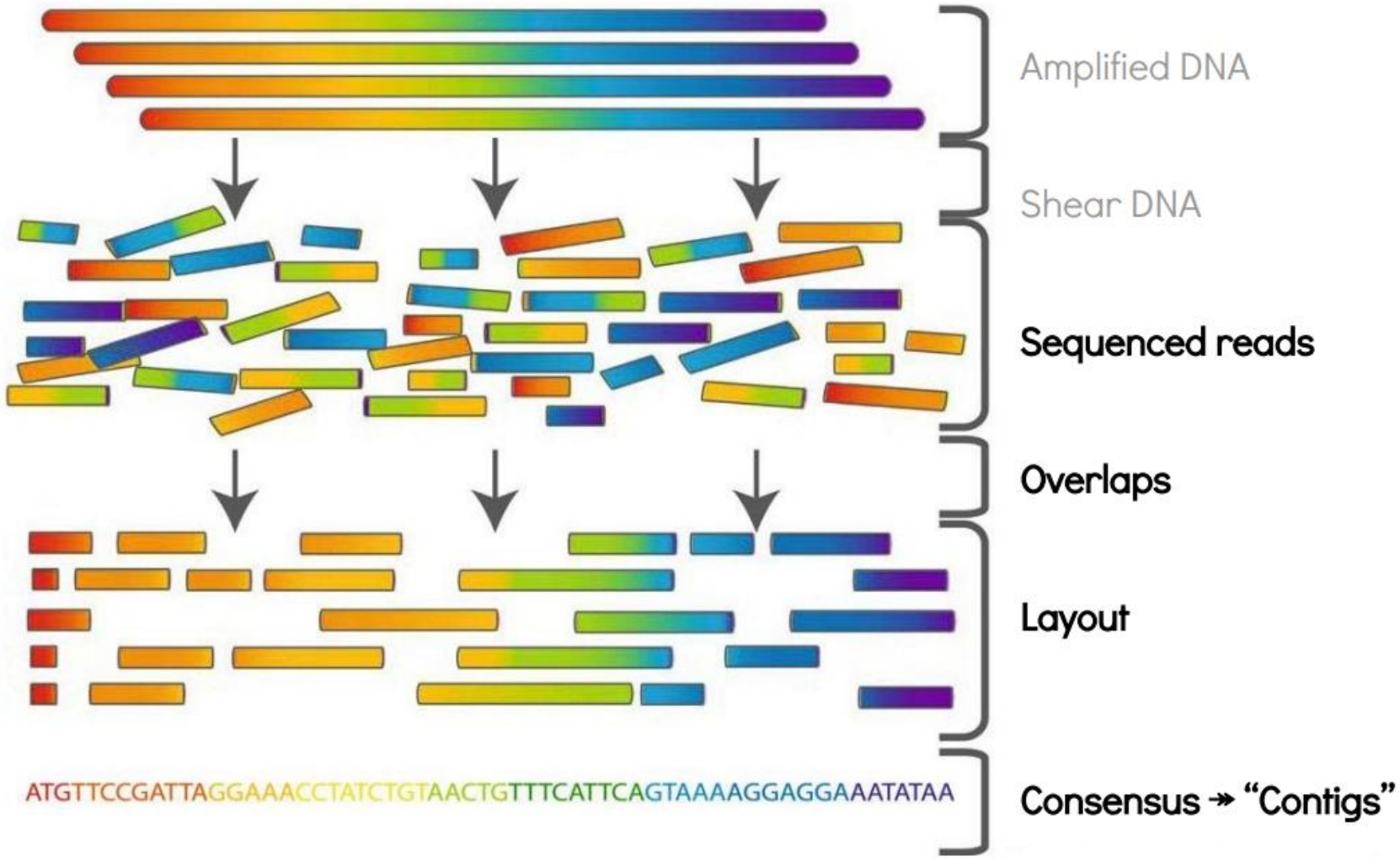
3,000,000

letters

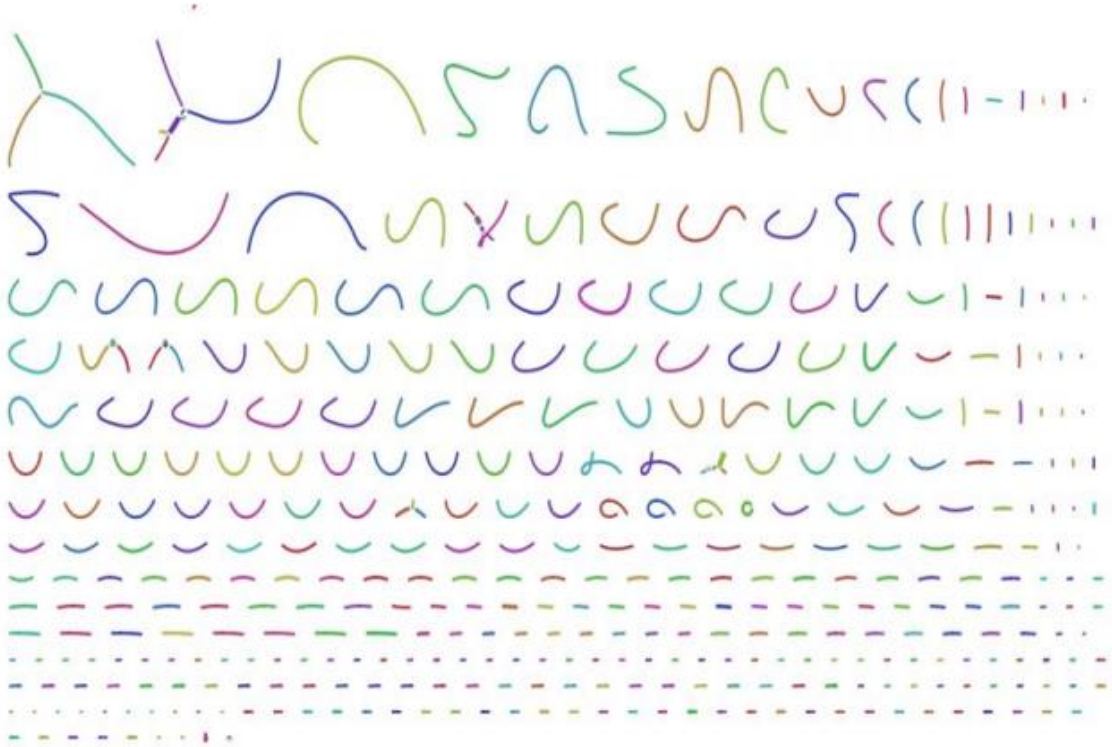
3,000 genes

https://angus.readthedocs.io/en/2016/_static/bacterial_genome_annotation.pdf

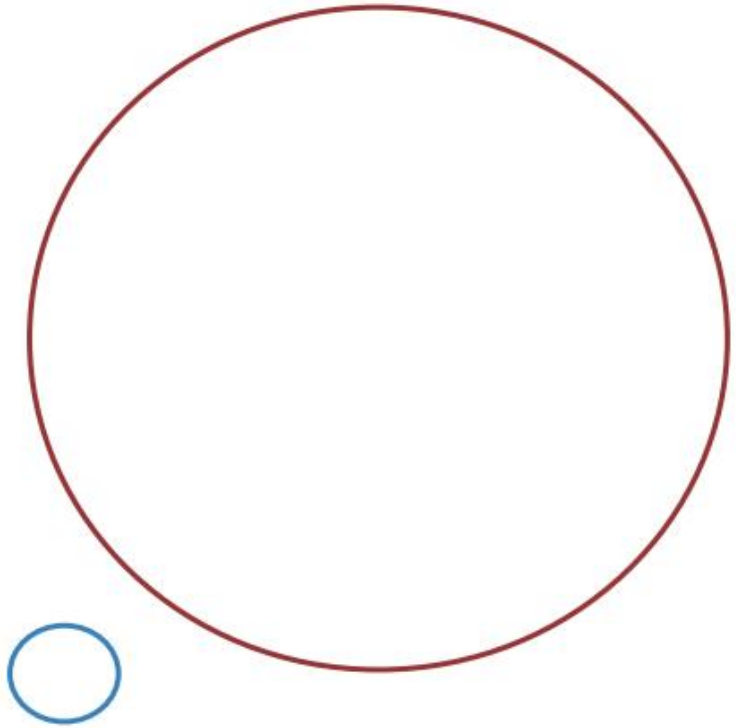
De novo assembly



Draft vs Finished genomes

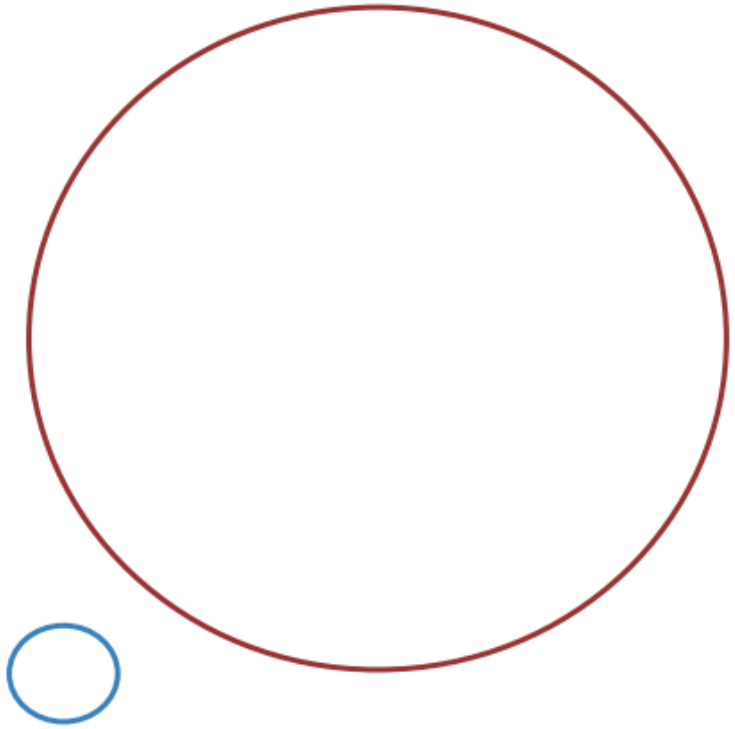


Lots of contigs



One contig per replicon

```
48541 agcccttcaa agaaatgttc tcagcaggca tggagcccag gacttgctcc ctttgggtag
48601 agagccgggt tgaaggtgac tgaagtgaaa tgggacagta gaggcggggg gggtggtgag
48661 ttcctggagg tgggggggtgt gggaacctgc tgtgtactga gatgcacccc tgccagttct
48721 gcctgaagat ttgaggcggg gggcaggggg gcggagtgaa gtcattttac tggtaagtaa
48781 ttttaaacct tttaatatta aagcaaacgt ggatatgtaa tgaatgaaat tcattctgga
48841 atgaaaaatt cacgtgatgt tgaaaaataa cacggggctt cagagaggac tttctggctg
48901 gcagcagact ccagattccc agggcccctg caccctcctc tgcccacagg gcaccttaat
48961 ggagaaggtg tgggaggaga gccaggccgg agtcagagca cactggtgac tccacatttg
49021 cagcgtgccc tgcctctctc ctgaggcttg gcaacgtgca atatgctaag caaactcccc
49081 ctgtccccgt ccagtttctg aggacaagag ccaccacctg tagcaaataa agacccagca
49141 acccttttgac tcatctttgt gagtctctgg aatcagaggg tagccacatc gctgagaggt
49201 ggagtgaagc actcgggtga aaaggtacaa ggaagtcagg gacaggagtg tggggacatc
49261 acctagacaa tgacagagaa gaggggcaca gccgagtgag gggagagggg ccggcagtcc
49321 tacatcccct ggcctgaagc acgctccagg gcagaaggaa aaacactgtc tttggggtcc
49381 aagagacctg agttcaaatt ctggctccac cactgaccac ctgtgtaacc ttgaactgct
49441 gctgcctgaa cctcaggttc cccttctaaa aatagaggag aaaaggatgc atttctcctt
49501 gcccctgtga gaacgaaatg gtgcaagcac caaggagcct cagcaaaggt cgggcctgcc
49561 cccgcctggc caaacctttc ctcttcagga ggccacggca accgtagttt gacagaagag
49621 cagcaccttg atttaatgct tcccagcatg tgtccttgag caagtcacct aacctctctg
49681 ggctgcttcc tcattgggaa aatatggctg ccagtaaaac ctgccctgtc cacctcctgg
49741 ggcacttggc aaacagcaaa agagtccaaa tgtgcaggct gggccaggcg cagtggctca
49801 tgcctgtaat cccagcaatt taggaagcca aggtgggcgg atcacctgag gtcaggagtt
49861 tgagaccagc ctggccaaca tggtgaaacc ttgtctctac aaaaatacaa aaattagccg
49921 ggcatgatgg cgggtgcctg taatcccagt tactcgggag gctgaggcaa gagaatcgct
49981 tgaacccgga aggggaaggt tgcagtgagc caagattgtg ccactgcact ccagcctggg
50041 caacagagcg agactctgtc tcaaaaaaaa aaaaaaaaaa aaacaatgca gagctggctg
50101 tgtaaaaaac ctgttccact gcagggccca gtgtccacca ggctggggtg caggcctatg
50161 gggtgggggc ccagcatcag cctctcagca gccctgggag gcggggcgca tcccgtgccc
50221 ctcgtggtct ggatgtgttc tagcccaagt cctaggttac acctgccgtc gcctggcctc
50281 tcaggagagg cccagggtga ggaggagcat ggtaaaggtg aagctgattg ggaagtcagc
50341 tgttgggaaa gcaactcctt gcacattgga ggaaccgaga aagactgacc ccgaggacag
50401 cagccagcat ggcccttcct gggagcccat gttgggggat tcctgctgca gccaaggctc
50461 agcccttgtg gtcgcaggtg ctggtcctgg cctcttcccc tcccatgcag gagcacagga
50521 gagatggctt ctgaggacct gttgcagctg tggccctggg aatagatttg ccagggagct
50581 ttaaagcagc tgagtgtgtc atccagctaa gcctggggaa ggagcttggc tcaggtcctg
50641 acaggtgtga cagggatggg gactgggaag taagagatga aaccctggct ggaggctgtg
50701 agcttccaca gccagcgctg gacagggagg gtccagatat acccactagt gccctcacca
```



One contig per replicon

https://www.ncbi.nlm.nih.gov/nuccore/806904736

Genome Annotation: Adding biological info to sequences

```
48541 agcccttcaa agaaatgttc tcagcaggca tggagcccag gacttgctcc ctttgggtag
48601 agagccgggt tgaaggtgac tgaagtgaaa tgggacagta gaggcggggg gggtggtgag
48661 ttcctggagg tgggggtgt gggaacctgc tgtgtactga gatgcacccc tgccagttct
48721 gcctgaagat ttgaggcggg gggcaggggg gcggagtgaa gtcattttac tggtaagtaa
48781 ttttaaacct tttaatatta aagcaaacgt ggatatgtaa tgaatgaaat tcattctgga
48841 atgaaaaatt cacgtgatgt tgaaaaataa cacggggctt cagagaggac tttctggctg
48901 gcagcagact ccagattccc agggcccctg caccctcctc tgcccacagg gcaccttaat
48961 ggagaaggtg tgggaggaga gccaggccgg agtcagagca cactggtgac tccacatttg
49021 cagcgtgccc tgcctctctc ctgaggcttg gcaacgtgca atatgctaag caaactcccc
49081 ctgtccccgt ccagtttctg aggacaagag ccaccacctg tagcaaataa agacccagca
49141 acccttttgac tcatctttgt gagtctctgg aatcagaggg tagccacatc gctgagaggt
49201 ggagtgaagc actcgggtga aaaggtacaa ggaagtcagg gacaggagtg tggggacatc
49261 acctagacaa tgacagagaa gaggggcaca gccgagtgag gggagagggg ccggcagtcc
49321 tacatcccct ggcctgaagc acgctccagg gcagaaggaa aaacactgtc tttggggtcc
49381 aagagacctg agttcaaatt ctggctccac cactgaccac ctgtgtaacc ttgaactgct
49441 gctgcctgaa cctcaggttc cccttctaaa aatagaggag aaaaggatgc atttctcctt
49501 gcccctgtga gaacgaaatg gtgcaagcac caaggagcct cagcaaaggt cgggcctgcc
49561 cccgcctggc caaacctttc ctcttcagga ggccacggca accgtagttt gacagaagag
49621 cagcaccttg atttaatgct tcccagcatg tgtccttgag caagtcacct aacctctctg
49681 ggctgcttcc tcattgggaa aatatggctg ccagtaaaac ctgccctgtc cacctcctgg
49741 ggcacttggc aaacagcaaa agagtccaaa tgtgcaggct gggccaggcg cagtggctca
49801 tgcctgtaat cccagcaatt taggaagcca aggtgggcgg atcacctgag gtcaggagtt
49861 tgagaccagc ctggccaaca tggtgaaacc ttgtctctac aaaaatacaa aaattagccg
49921 ggcatgatgg cgggtgcctg taatcccagt tactcgggag gctgaggcaa gagaatcgct
49981 tgaacccgga aggggaaggt tgcagtgagc caagattgtg ccactgcact ccagcctggg
50041 caacagagcg agactctgtc tcaaaaaaaa aaaaaaaaaa aaacaatgca gagctggctg
50101 tgtaaaaaac ctgttccact gcagggccca gtgtccacca ggctggggtg caggcctatg
50161 gggtgggggc ccagcatcag cctctcagca gccctgggag gcggggcgca tcccgtgccc
50221 ctcgtggtct ggatgtgttc tagcccaagt cctaggttac acctgccgtc gcctggcctc
50281 tcaggagagg cccagggtga ggaggagcat ggtaaaggtg aagctgattg ggaagtcagc
50341 tgttgggaaa gcaactcctt gcacattgga ggaaccgaga aagactgacc ccgaggacag
50401 cagccagcat ggcccttcct gggagcccat gttgggggat tcctgctgca gccaaggctc
50461 agcccttgtg gtcgcaggtg ctggtcctgg cctcttcccc tcccatgcag gagcacagga
50521 gagatggctt ctgaggacct gttgcagctg tggccctggg aatagatttg ccagggagct
50581 ttaaagcagc tgagtgtgtc atccagctaa gcctggggaa ggagcttggc tcaggtcctg
50641 acaggtgtga cagggatggg gactgggaag taagagatga aaccctggct ggaggctgtg
50701 agcttccaca gccagcgctg gacagggagg gtccagatat acccactagt gccctcacca
```

ribosome binding site

delta toxin
PubMed: 15353161

ACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGA
AAAGCAGCCTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTC
CCAGGCCAGTGCCGGGCCCCTCATAGGAGAGGAAGCTCGGGAGGTG
GCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGG
ACAGAATGCCCTGCAGGAACTTCTTCTAGAAGACCTTCTCCTCCTG
CAAATAAAACCTCACCCATGAATGCTCACGCAAGTTTAATTACAGA
CCTGAAACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCT
CTCCGTCCGTCCGTGGGCCACGGCCACCGCTTTTTTTTTTGCC

transfer RNA
Leu-(UUR)

tandem repeat
CCGT x 3

homopolymer
10 x T

What's in an annotation?

- Location
  - which sequence? *chromosome 2*
  - where on the sequence? 100..659
  - what strand? -ve
- Feature type
  - what is it? *protein coding gene*
- Attributes
  - protein product? *alcohol dehydrogenase*
  - enzyme code? *EC:1.1.1.1*
  - subcellular location? *cytoplasm*
  - note? *beer processing*

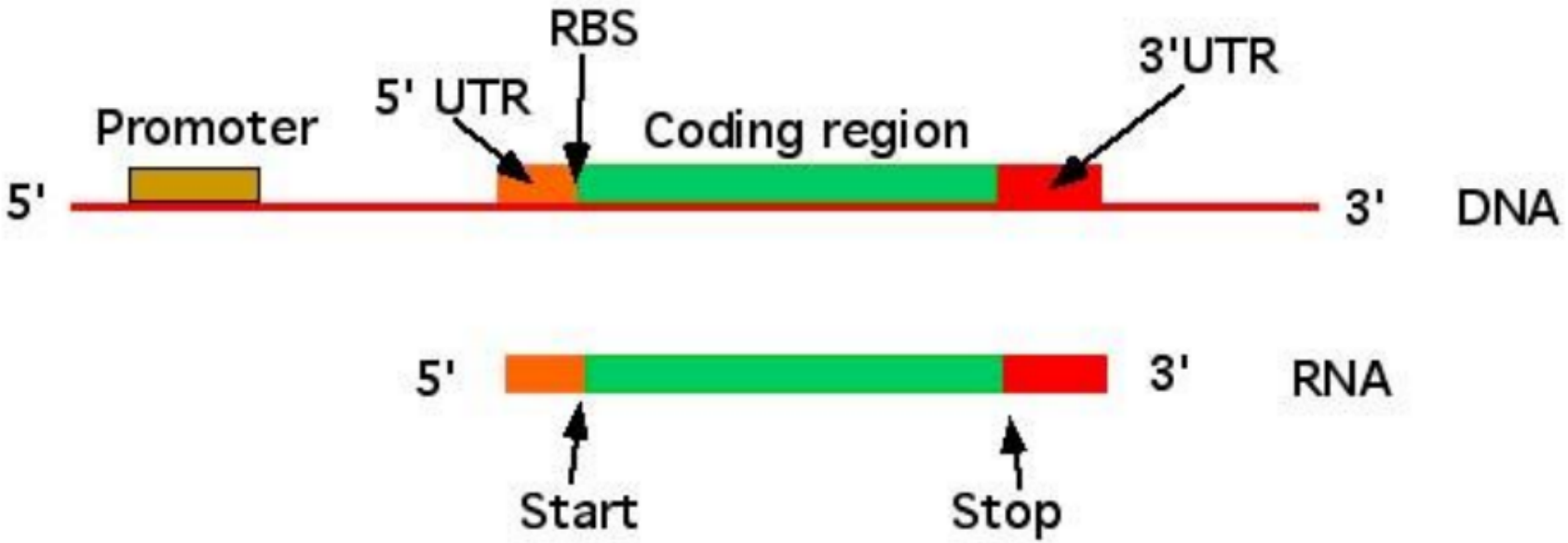
Bacterial feature types

- protein coding genes
  - promoter (-10, -35)
  - ribosome binding site (RBS)
  - coding sequence (CDS)
    - signal peptide, protein domains, structure
  - terminator
- non coding genes
  - transfer RNA (tRNA)
  - ribosomal RNA (rRNA)
  - non-coding RNA (ncRNA)
- other
  - repeat patterns, operons, origin of replication, ...

Key bacterial features

- 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
  - partial genes, remnants
  - pseudogenes
  - assigning function is the bulk of the workload

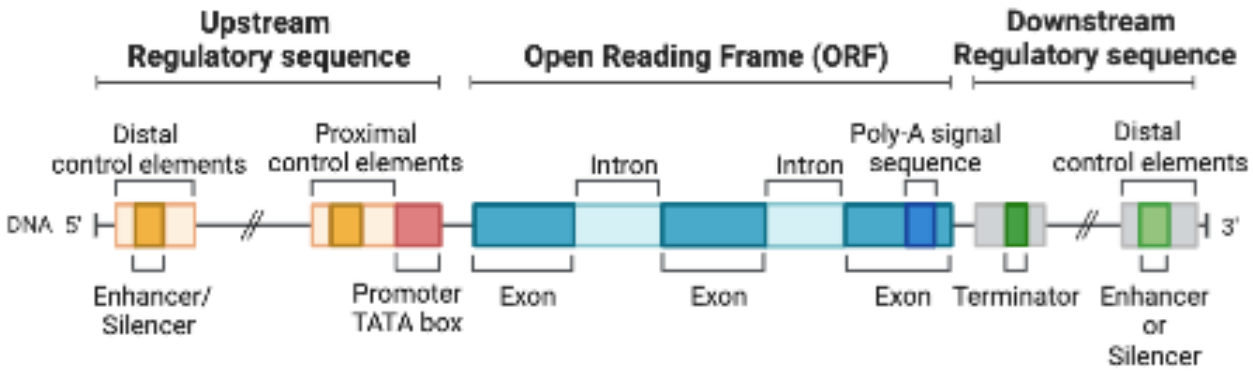
Bacterial genes: no introns!



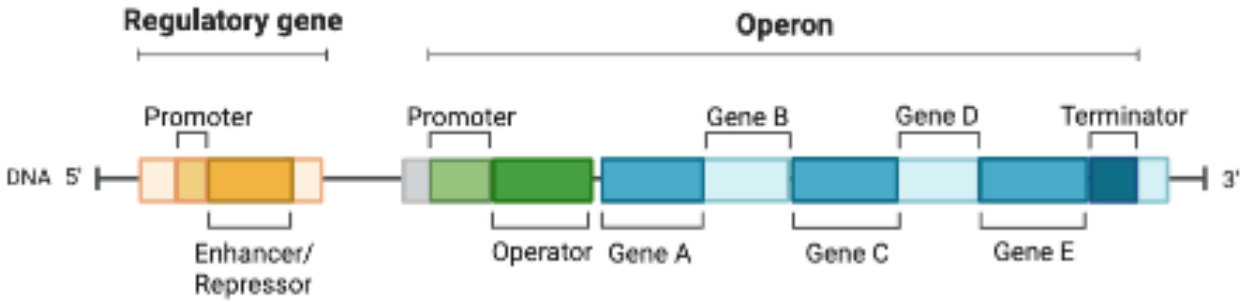
- have >= 3 potential start codons (species dependent)
- haploid, but lots of horizontal gene transfer
- methylation used as primitive immune system
  - restriction modification system against phage

Prokaryotic vs eukaryotic Genes

Eukaryotic Gene Structure

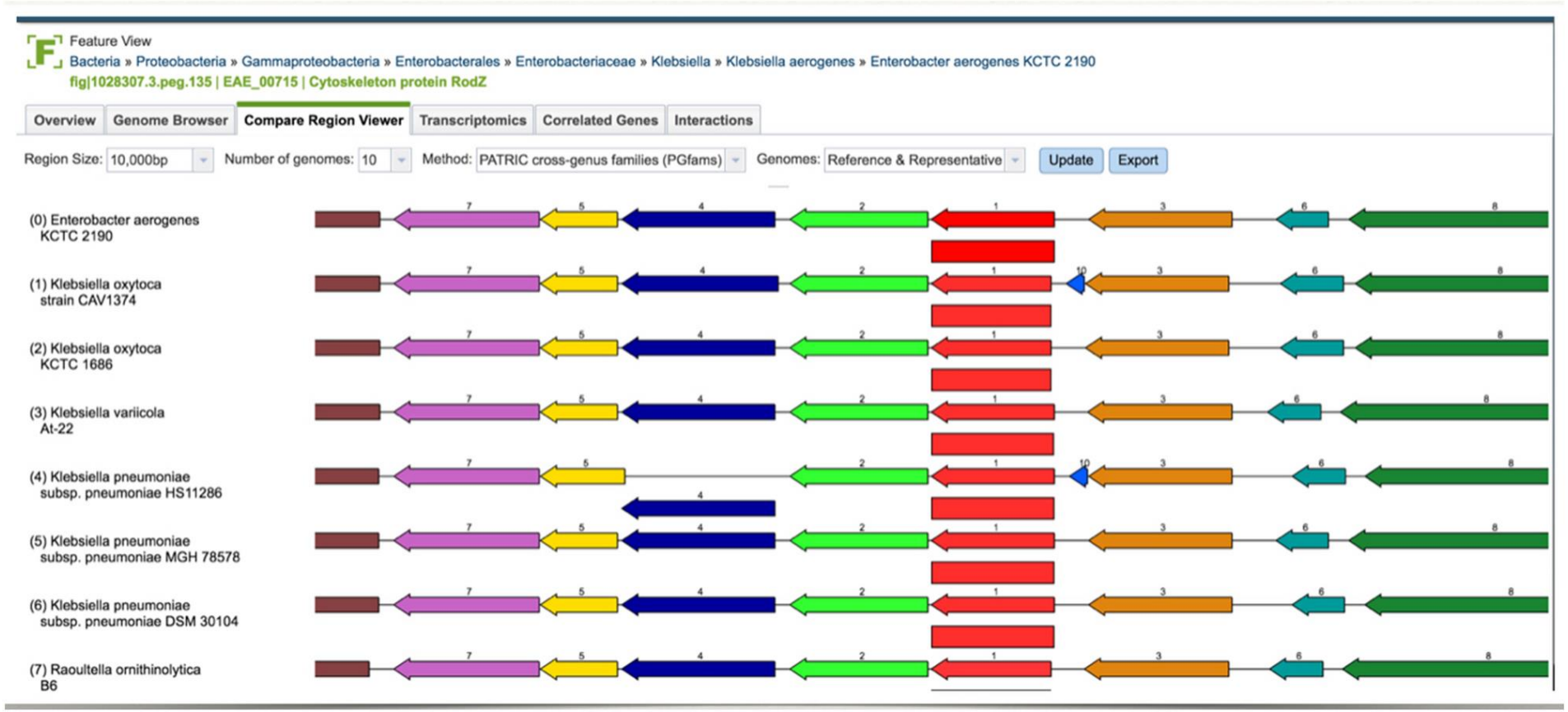


Prokaryotic Gene Structure



https://study.com/skill/practice/contrasting-the-regulation-of-gene-expression-in-prokaryotic-eukaryotic-organisms-questions.html

Detecting operons in bacterial genomes via visual representation learning



Snapshot of the Compare Region Viewer service provided by PATRIC (https://www.patricbrc.org). The image shows a genomic region of the query genome (first row) aligned against a set of other genomes, anchored at the focus gene (represented as a red arrow). The service starts with finding other genes that are of the same family as the focus gene, and then aligns their flanking regions accordingly.

https://www.nature.com/articles/s41598-021-81169-9

Automatic Genome Annotation

Two strategies for identifying coding genes:

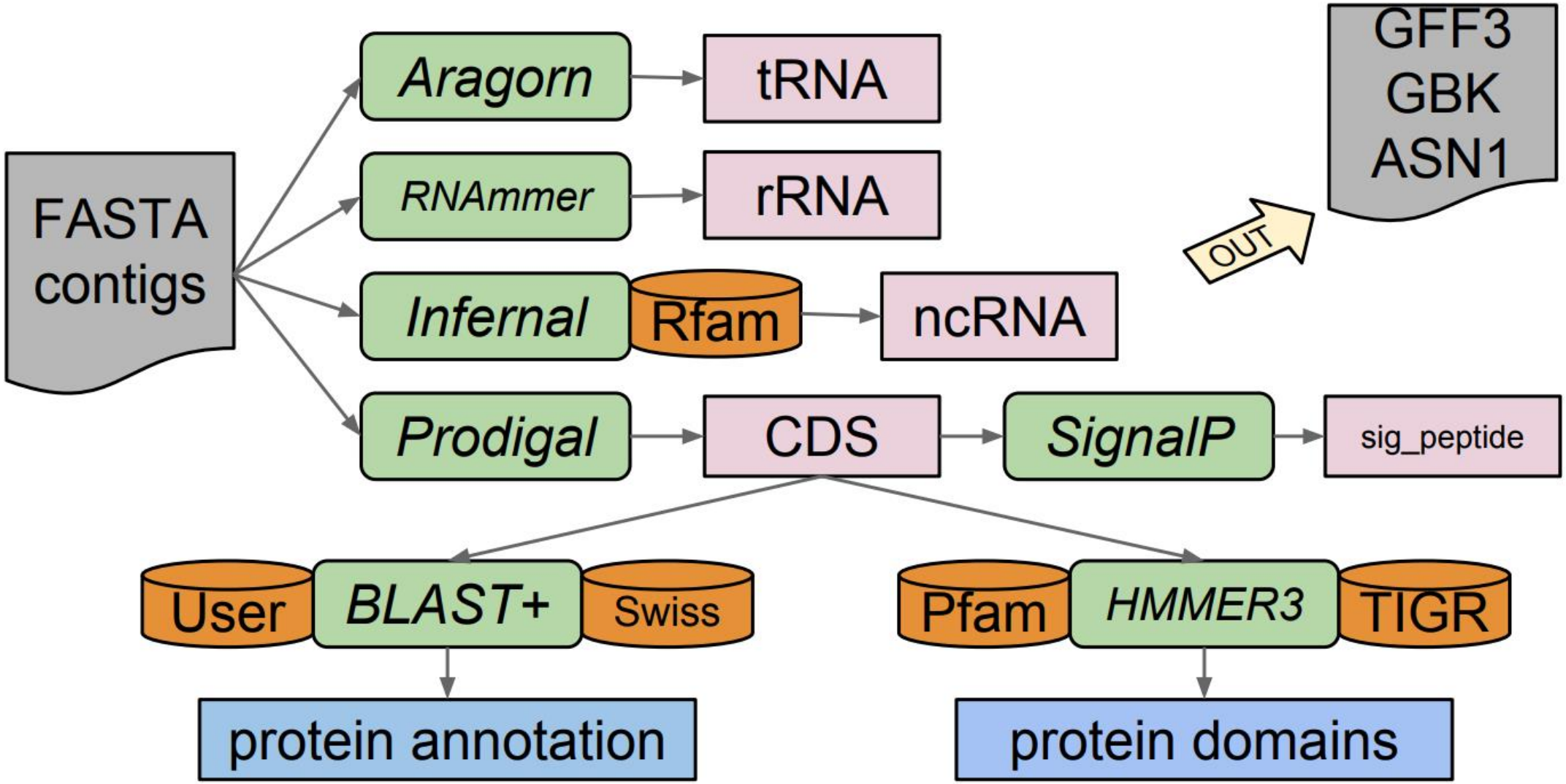
- **sequence alignment**
  - find known protein sequences in the contigs
    - transfer the annotation across
  - will miss proteins not in your database
  - may miss partial proteins

- ***ab initio* gene finding**
  - 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

Some good existing tools

Software	*ab initio*	align-ment	Availability	Speed
RAST	yes	yes	web only	12-24 hours
xBASE	yes	no	web only	>4 hours
BG7	no	yes	standalone	>10 hours
PGAAP (NCBI)	yes	yes	email / we	>1 month

Prokka pipeline (simplified)



Prokka pipeline (simplified)

Bioinformatics Advance Access published March 18, 2014

Genome Analysis

Prokka: rapid prokaryotic genome annotation

Torsten Seemann[1,2,*]

[1] Victorian Bioinformatics Consortium, Monash University, Melbourne, Australia.

[2] Life Sciences Computation Centre, Victorian Life Sciences Computation Initiative, Melbourne, Australia.

Associate Editor: Prof. Alfonso Valencia

ABSTRACT

Summary: The multiplex capability and high yield of current day DNA sequencing instruments has made bacterial whole genome sequencing a routine affair. The subsequent *de novo* assembly of reads into contigs has been well addressed. The final step of annotating all relevant genomic features on those contig can be achieved slowly using existing web and email-based systems, but these are not applicable for sensitive data or integrating into computational pipelines. Here we introduce Prokka, a command line software tool to fully annotate a draft bacterial genome in about ten minutes on a typical desktop computer. It produces standards-compliant output files for further analysis or viewing in genome browsers.

Availability and Implementation: Prokka is implemented in Perl and is freely available under an open source GPLv2 license from http://vicbioinformatics.com/.

Contact: torsten.seemann@monash.edu

2 DESCRIPTION

2.1 Input

Prokka expects pre-assembled genomic DNA sequences in FASTA format. Finished sequences without gaps are the ideal input, but it is expected that the typical input will be a set of scaffold sequences produced by *de novo* assembly software. This sequence file is the only mandatory parameter to the software.

2.2 Annotation

Prokka relies on external feature prediction tools to identify the coordinates of genomic features within contigs. These tools are listed in Table 1, and all of them, except for Prodigal, provide co-

Example from Prokka

Feature Type:

tRNA

Location:

contig000341 @ 655..730 +

Attributes:

/gene="tRNA-Leu(UUR)"

/anticodon=(pos:678..680,aa:Leu)

/product="transfer RNA-Leu(UUR)"

/inference="profile:Aragorn:1.2"

Provenance

Recording where an annotation came from

Prokka uses Genbank "evidence qualifier" tags:

Wet lab

```
/experiment="EXISTENCE:Northern blot"
```

Dry lab

```
/inference="similar to DNA sequence:INSD:AACN010222672.1"
/inference="profile:tRNAscan:2.1"
/inference="protein motif:InterPro:IPR001900"
/inference="ab initio prediction:Glimmer:3.0"
```

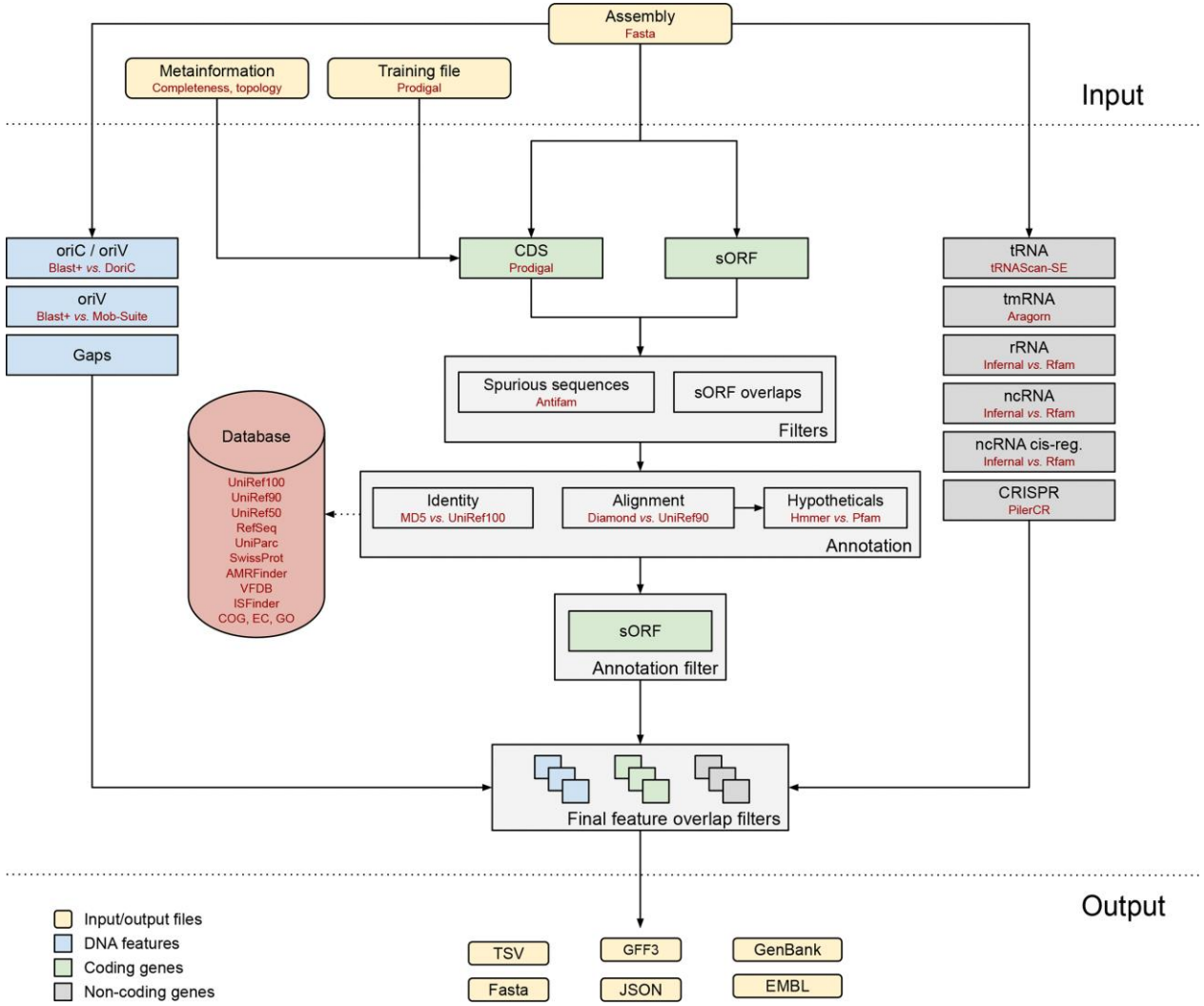
Curating genomes: Improving annotations

- Some annotations are wrong
  - False annotation
  - Missing annotation
  - Partially wrong annotation
- Curation
  - Manual effort to improve annotations
  - Community curation

Web Apollo: a web-based genomic annotation editing platform

https://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-8-r93

Bakta is a tool for the rapid & standardized annotation of bacterial genomes and plasmids from both isolates and metagenome-assembled genomes (MAGs)



Overview of the Bakta annotation workflow (Schwengers *et al.* 2021)

Further structural annotation

- **PlasmidFinder** (Carattoli and Hasman 2020), a tool for the identification and typing of plasmid sequences in Whole-Genome Sequencing.

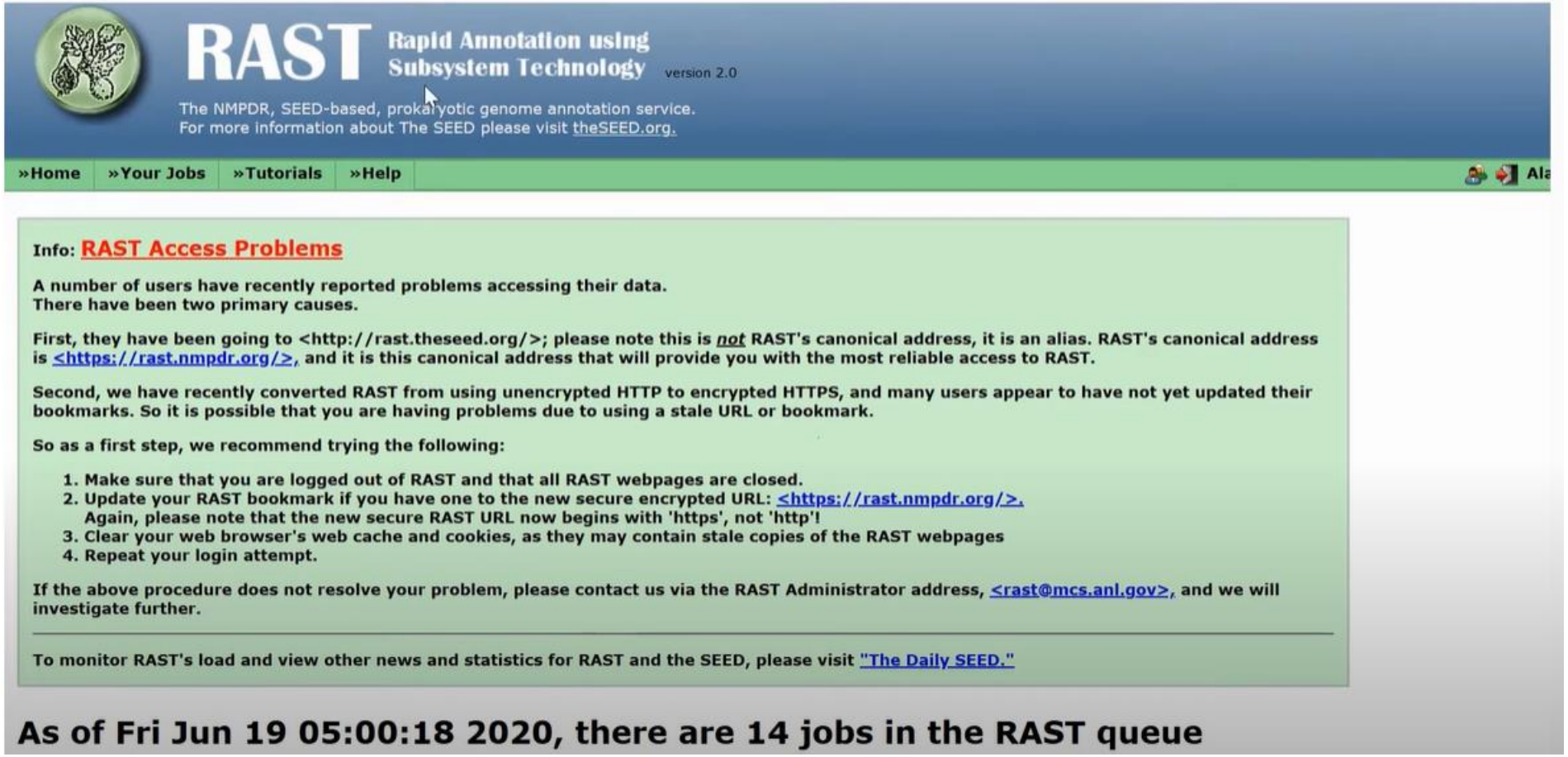
- **IntegronFinder** (Néron *et al.* 2022), a tool for detecting integrons

- An **integron** is minimally composed of:
  - ❑ a gene encoding for a site-specific recombinase (intI)
  - ❑ a proximal recombination site (attI), which is recognized by the integrase and at which gene cassettes may be inserted
  - ❑ a promoter (Pc) which directs transcription of cassette-encoded genes

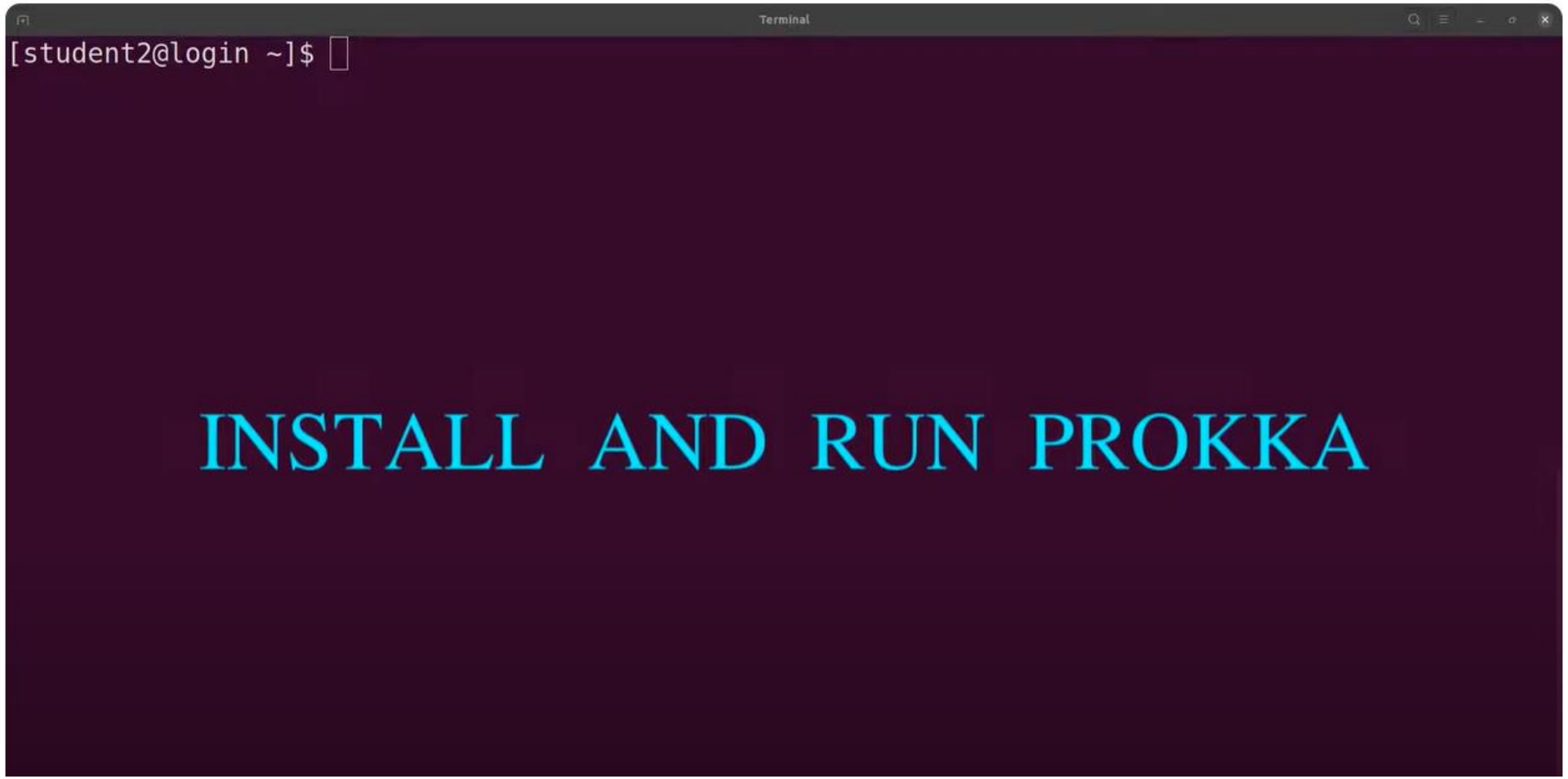
- **IS (Insertion Sequence)** elements: a short DNA sequence that acts as a simple transposable element

- **ISEScan** (Xie and Tang 2017) for detection of IS

Exercise 1: Using RAST to annotate bacterial genome sequence as following https://www.youtube.com/watch?v=pICJ4Gmqq1U



Exercise 2: Using PROKKA to annotate bacterial genome sequence as following
https://www.youtube.com/watch?v=NYbYxGf-8NM&list=PLe1-kjuYBZ04YmORKfXBCOqGDDYJtudVA&index=8



Thank you for your attention!