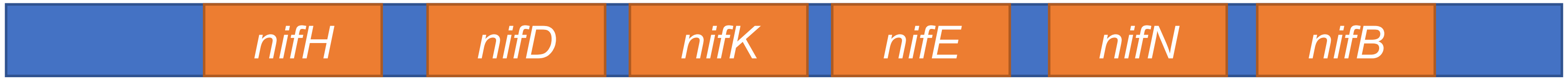
Genome annotation

MDBP-105

Gene annotation

- Adding biological information to sequences
- There is a gene X in contig Y on location Z
  - Size of the gene
  - Name of the gene
  - Function of the gene (protein / RNA gene)

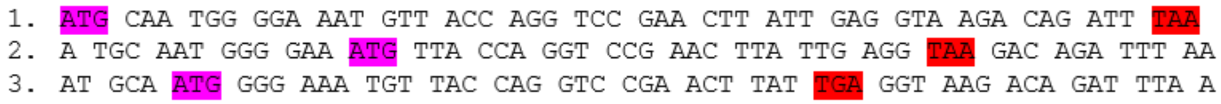


Contig Y = 20 035 bp

Ways to identify protein coding genes

- Sequence alignments
  - E.g. BLAST
  - Search contigs against a database
  - Computationally (and manually) intensive

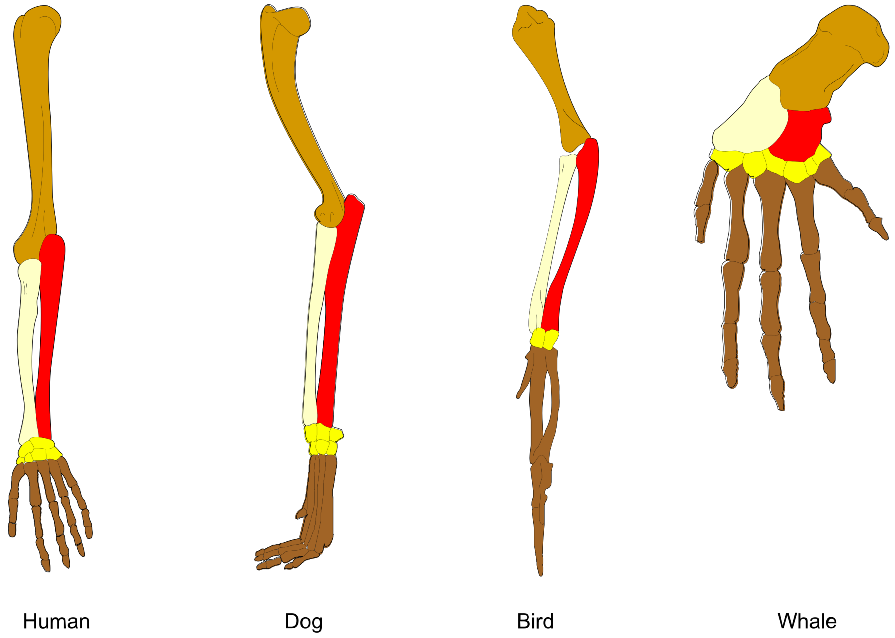
- Gene finding
  - Start codon (ATG)
  - Open reading frame (ORF)
  - Stop codon (TAA, TAG, TGA)



```
1.  ATG CAA TGG GGA AAT GTT ACC AGG TCC GAA CTT ATT GAG GTA AGA CAG ATT TAA
2.  A TGC AAT GGG GAA ATG TTA CCA GGT CCG AAC TTA TTG AGG TAA GAC AGA TTT AA
3.  AT GCA ATG GGG AAA TGT TAC CAG GTC CGA ACT TAT TGA GGT AAG ACA GAT TTA A
```

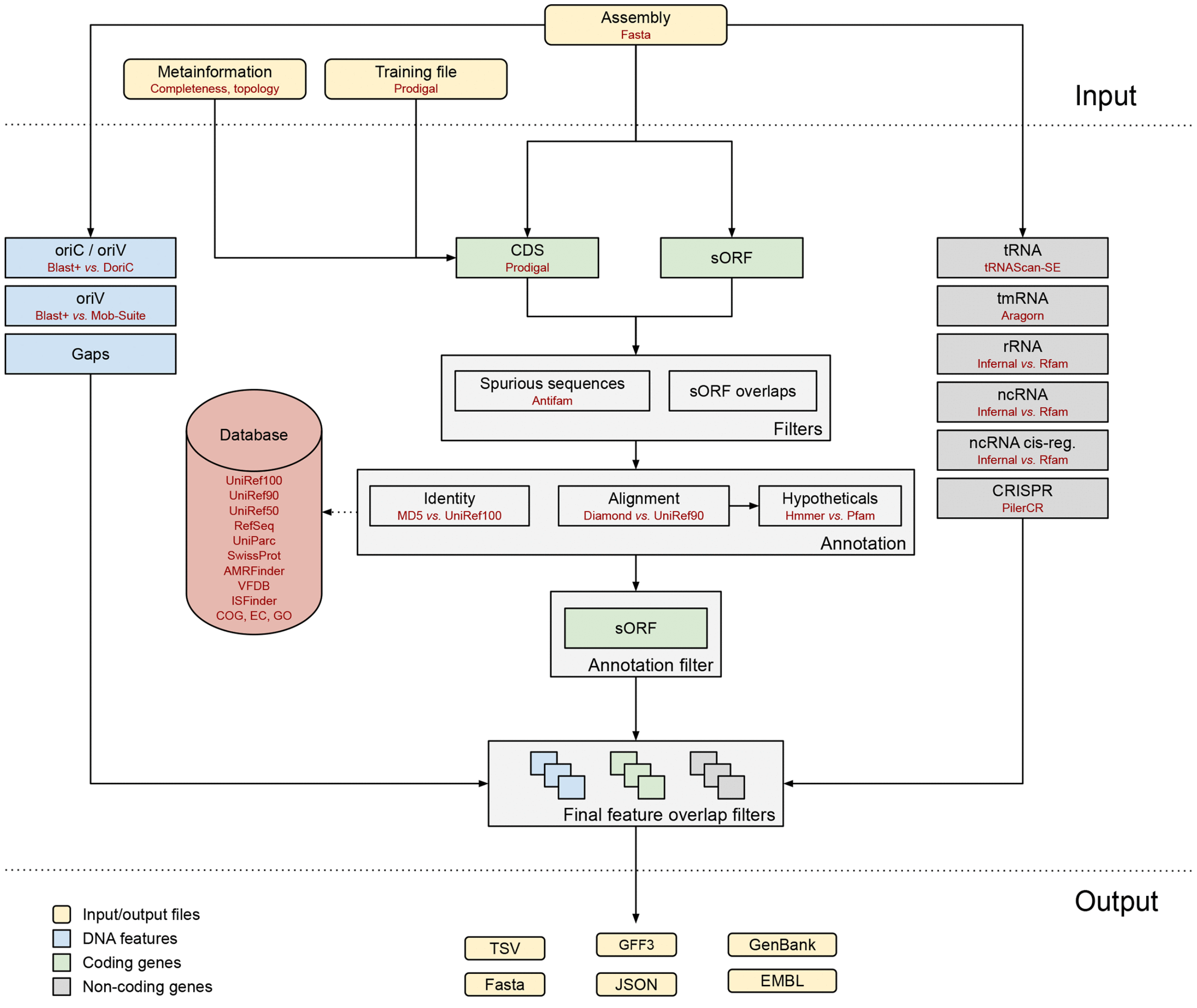
Functions to genes

- Homology
- Statictical modelling of protein families/domains
- Annotation databases
    - NCBI
    - KEGG
    - COG
    - SEED
    - GO
    - UNIPROT
    - INTERPRO
    - PFAM
    - TIGR
    - …



BAKTA

- rapid and standardized annotation of bacterial genomes via alignment-free sequence identification



https://doi.org/10.1099/mgen.0.000685

Taxonomy and completeness of your genome

CheckM2

- Predicts genome completeness and contamination based on ML model
- Designed for metagenome-assembled genomes (MAGs)

GTDB-Tk

- The Genome Taxonomy Database Toolkit
- Taxonomic assignment based on GTDB
- Domain-specific concatenated protein reference trees