Genome Annotation

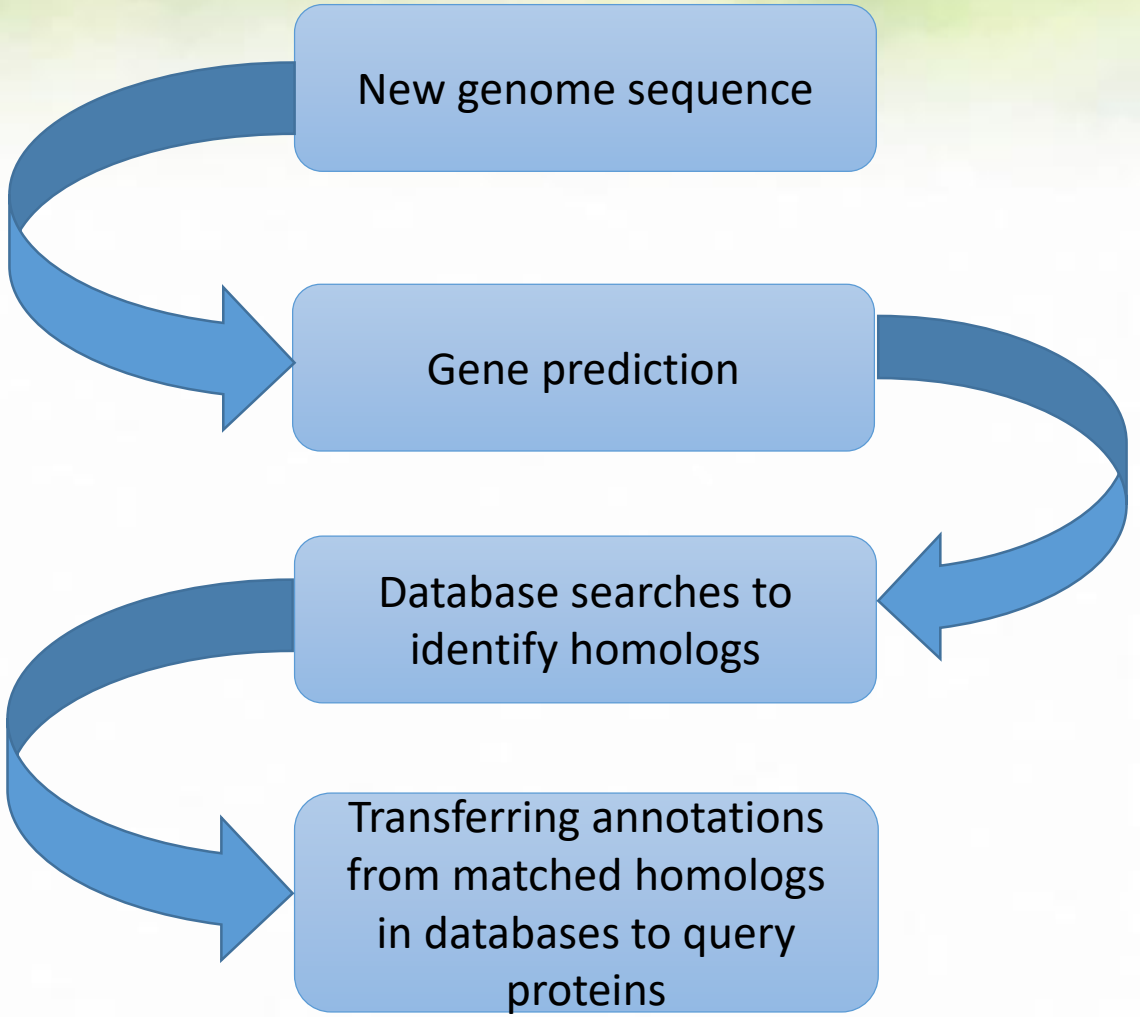
SAGESA AMR Bioinformatics Africa

25 – 27 May 2022

Stan Kwenda, PhD
Lead Data Analyst
National Institute for Communicable Diseases
Johannesburg
South Africa

Genome annotation

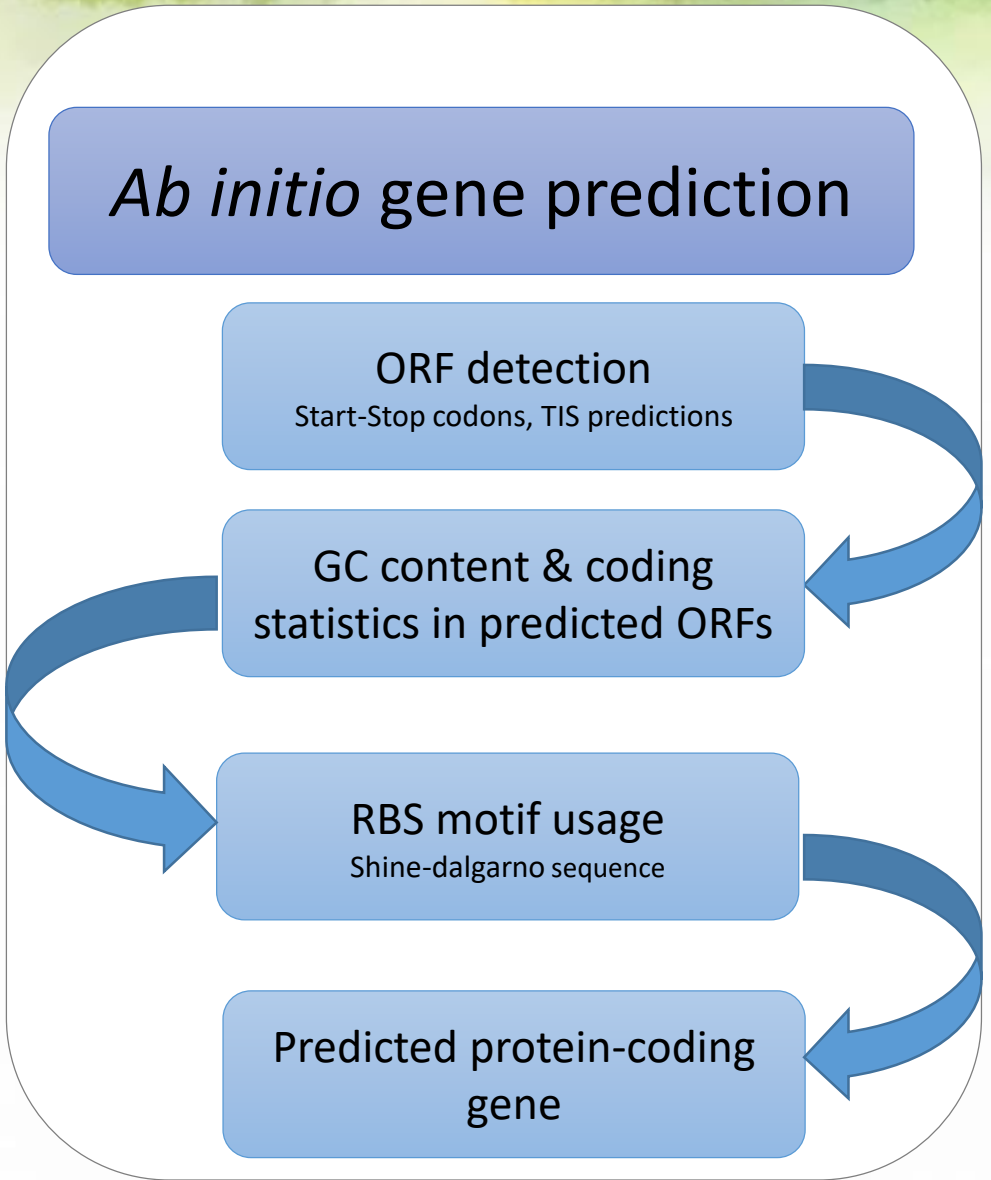
- Extraction of biological information from nucleotide sequencing data
  - Supports our understanding of gene functions and genome evolution
- A successful annotation depends on the quality of the genome assembly
  - The value of the genome is only as good as its annotation



Multi-step process involving structural and functional annotations

Nucleotide level genome annotation

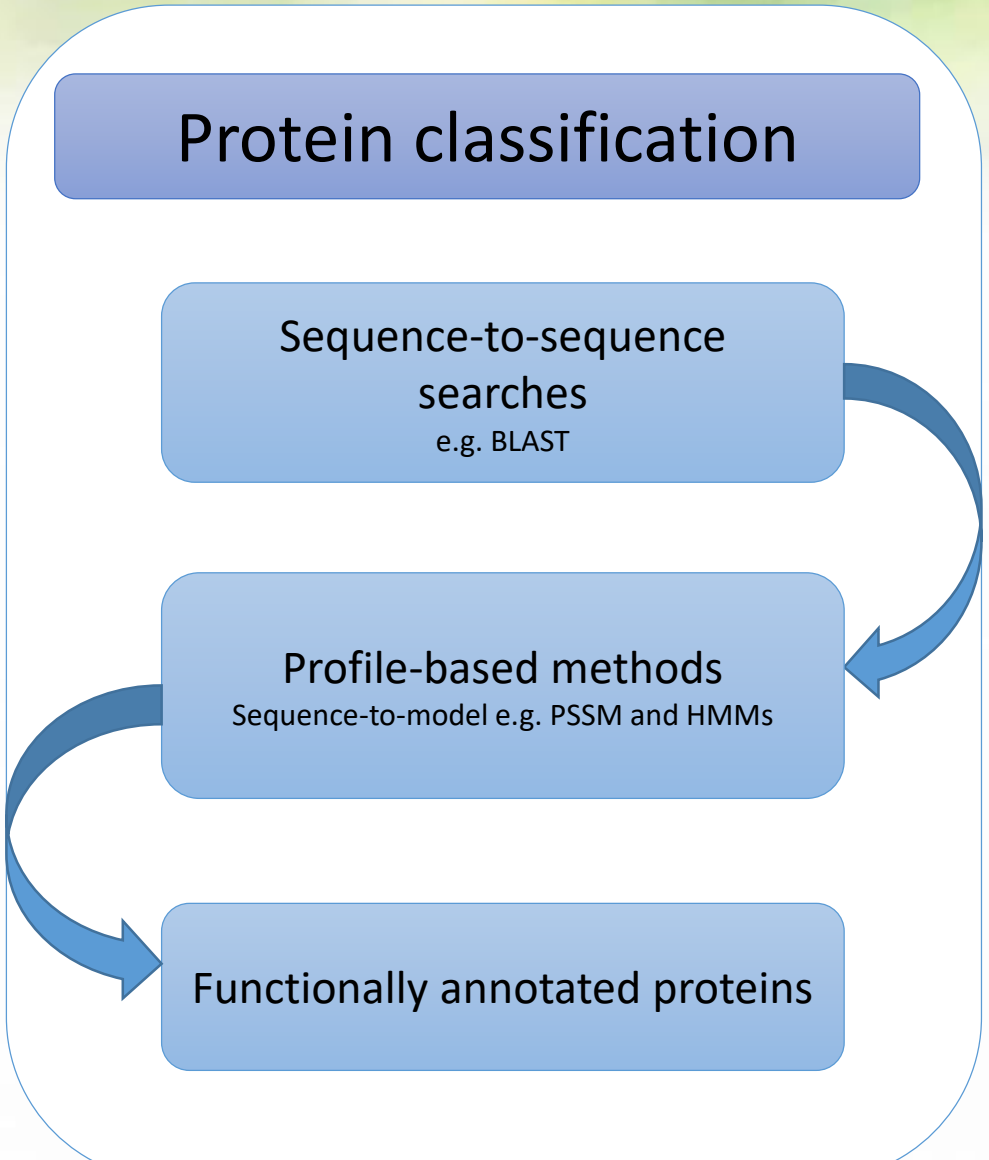
- Structural annotation
  - Gene location and structure and functions of regulatory regions
    - Where in the genome are the genes and genetic markers?
  - Repetitive elements
  - Gene duplications e.g. presence of paralogs
  - Detecting variations e.g. SNPs
  - Noncoding RNA and regulatory elements
- High GC content in transcribed regions
- Statistical properties of the CDS
- Most bacteria use genetic code 11
  - AUG most common start codon. Alternatives include UUG, GUG, AUU
  - 3 stop codons: TAA, TGA, and TAG
  - Some use genetic code 4 and 25
- Gene prediction tools:
  - Prodigal, Glimmer, Genemark



https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi#SG11

Protein-level genome annotation

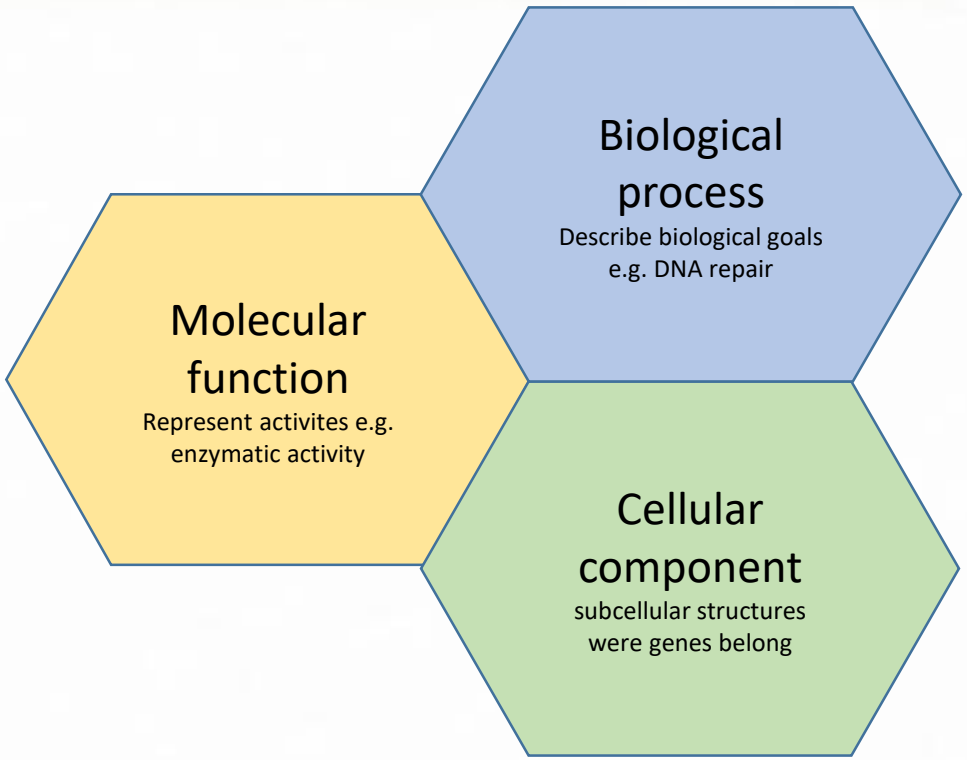
- Cataloging and naming proteins and assigning functions
  - Homology-based searches
    - Primarily reliant on homology detection between newly identified genes/proteins and previously annotated sequences
    - Sequence-to-sequence or sequence-to-model searches
- Not all proteins are well characterized
  - Proteins of unknown function
  - Protein families arising from duplication events and divergence
- Functional annotation
  - Comparison of proteins between species
    - Orthologs due to common ancestry tend to have similar sequences and often similar functions
    - Classify predicted proteins based on functional domains, folds and motifs
- Precise ordering of nucleotides and amino acids confer specific functional properties to biological sequences



Process level genome annotation

- Relating the genome to biological processes
- Gene function descriptions e.g. Gene Ontology terms
- Annotation of genes to various levels of specificity
- Integrating information at the level of regulatory and metabolic networks and protein-protein interactions
- Variant annotation

Gene ontology



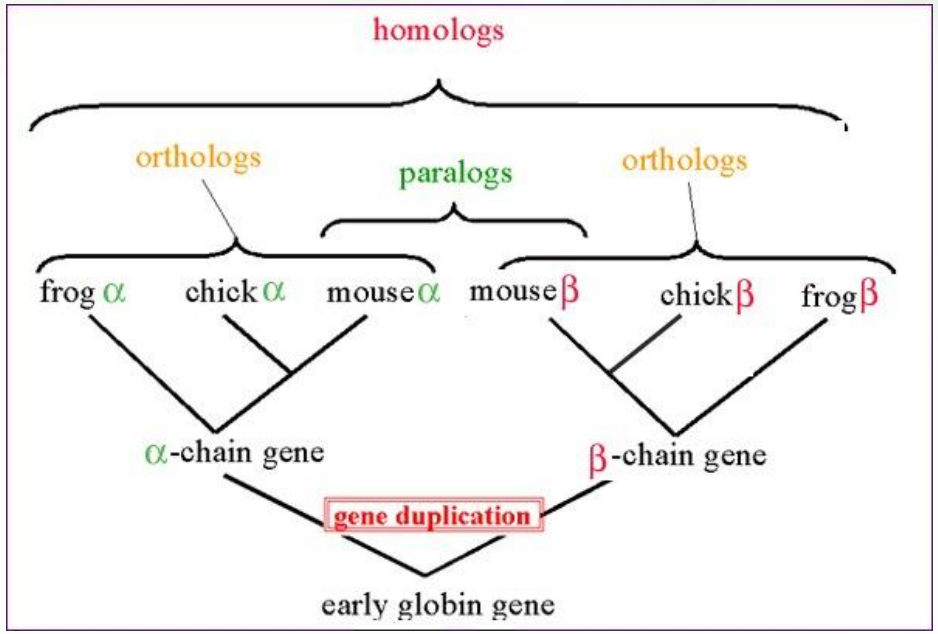
Factors affecting annotation completeness

- Annotation method
  - Annotation coverage can be as high as 98% in some species.
- Taxonomy
- Genome size
  - Tend to be more uncharacterized proteins in the accessory proteomes
- Research bias
  - Early model organisms tend to have higher annotation completeness levels
- Publication date

Lobb *et al.*, (2020) *Microbial Genomics*

Key definitions in annotation

- Paralogs
  - Genes within the same genome that have arisen from duplications
- Orthologs
  - Genes that have originated from a single ancestral gene in the last common ancestor of the compared genomes
- Identity
  - Is the extent to which two sequences are invariant
- Percentage identify
  - Provides a statistic for an alignment of two sequences
- Motif
  - A protein sequence motif is a set of conserved amino acid residues that are important for function
- Domain
  - A structurally compact, independently folding unit forming a stable three dimensional structure
  - Typically contains one or more motifs



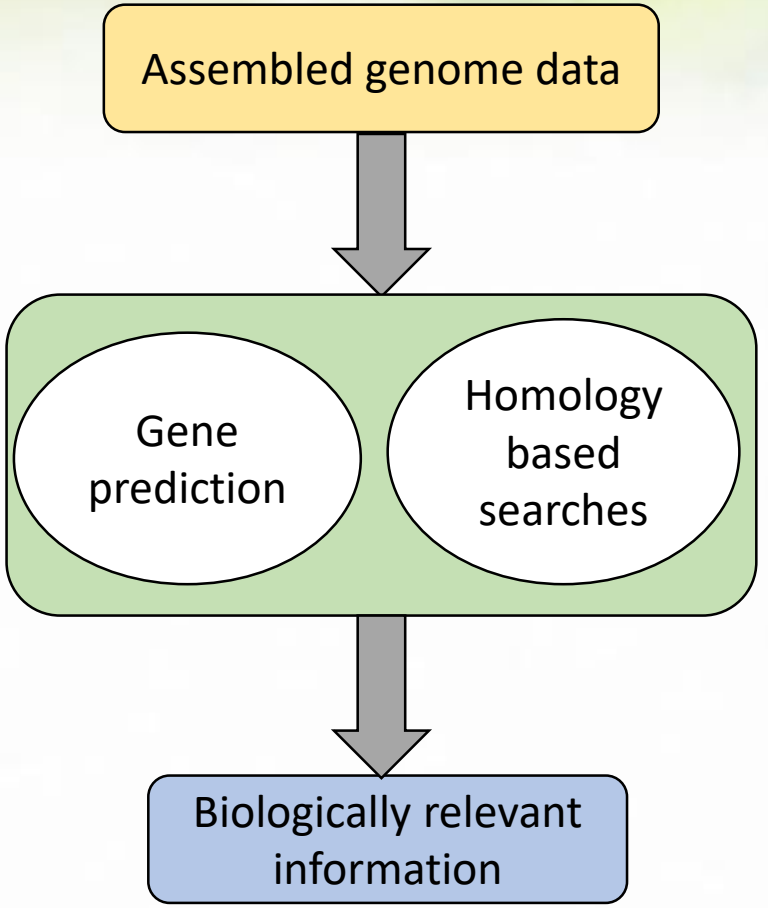
https://www.ncbi.nlm.nih.gov/books/NBK62051/

Curated data repositories for genome annotation

- UniProt/SwissProt
  - (https://www.uniprot.org/downloads)
- RefSeq
  - (https://www.ncbi.nlm.nih.gov/refseq/)
  - Non-redundant protein or nucleotide databases
- Other databases specialized databases:
  - Pfam (https://pfam.xfam.org/)
  - Rfam (https://rfam.xfam.org/)
  - KEGG (https://www.genome.jp/kegg/)
  - BioCyc (https://biocyc.org/)
  - STRING

Genome annotation pipelines

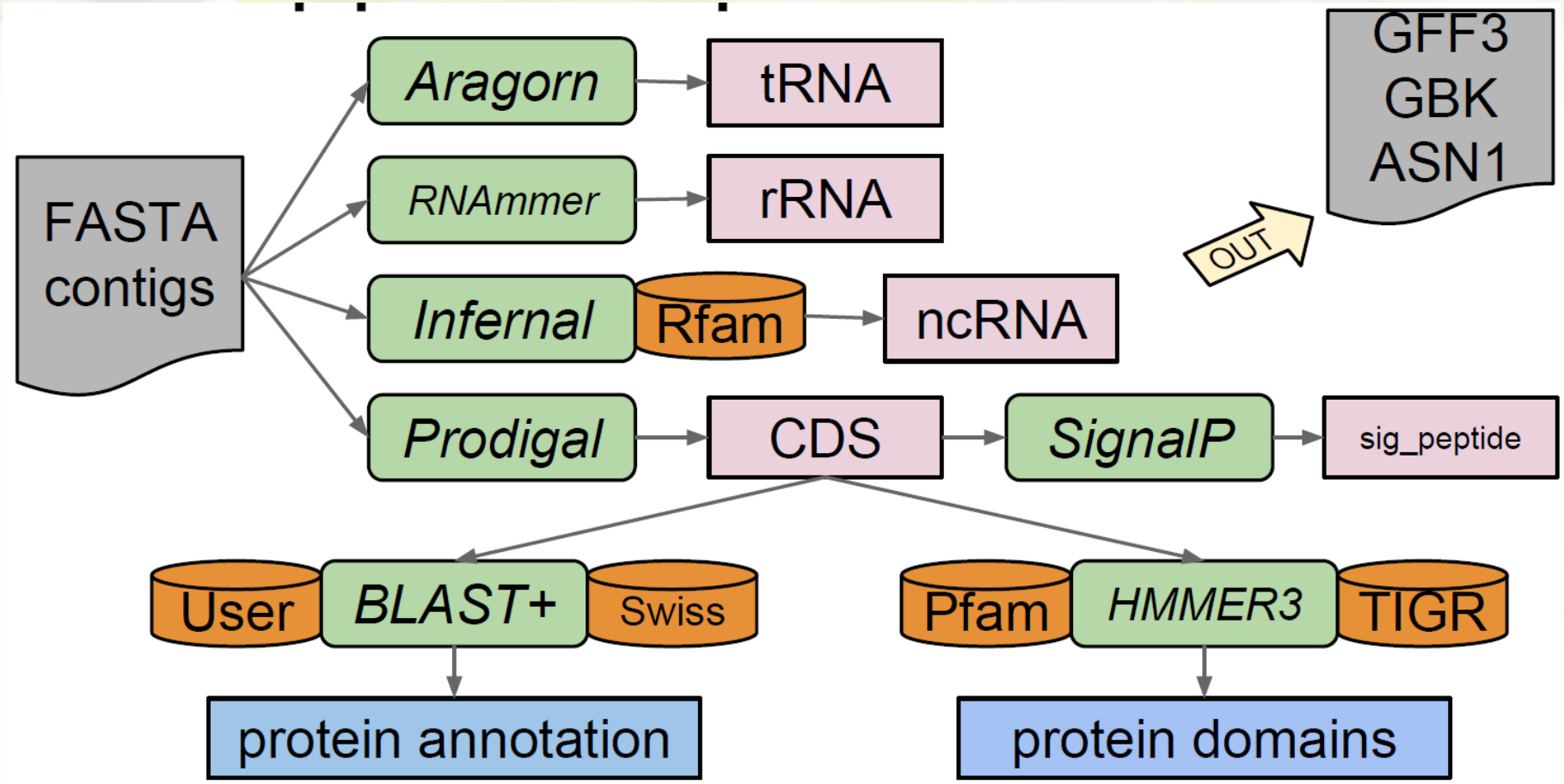
- Prokka
  - (https://github.com/tseemann/prokka)
- RAST
  - (https://rast.nmpdr.org/)
- NCBI PGAP
  - (https://www.ncbi.nlm.nih.gov/genome/annotation_prok/)
- Bakta
  - (https://github.com/oschwengers/bakta)



Summary

- Understanding the structure of a gene is a crucial step in comprehending its function and the significance of variations
- Errors can be easily propagated downstream e.g. erroneous annotations in a reference genome can be easily passed to query sequences in homology-based searches
- Genome annotation success rates higher in species phylogenetically closer to model organisms
- Re-annotation necessary to correct faulty annotations or to update older annotations
- Gene-finding relatively straightforward in bacterial genomes
- Challenges around automated assignment of gene function leading to high numbers of hypothetical sequences of unknown function
- Experimentally derived functional annotations largely obtained from early model organisms such as *E. coli*

Prokka pipeline overview



https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/annotation-with-prokka/slides-plain.html

AGESA

CAMBRIDGE AFRICA

Thank you

wellcome connecting science

NATIONAL INSTITUTE FOR COMMUNICABLE DISEASES

Division of the National Health Laboratory Service