Genome annotation

Sequences

- Sequencing
- Mapping
- Assembly

Annotation

- Identifying the locations of genes and all of the coding regions
- Determining their functions

Annotation steps

- Identify non-coding part of genome
- Identify genome elements
- Assign functional information

Genome elements

- Coding
  - genes
  - Transcription and translation
- Non-coding
  - Structural DNA (not transcribed)
  - Functional RNA (not translated)
  - Introns (removed before translation)

Non-coding

- Structural
  - Telomeres, centromeres, repetetives
- Functional
  - tRNA, rRNA
- Introns
  - Removed from mRNA

Structural annotation

- Open Reading Frames
- Gene structure
- Coding regions
- Regulatory motifs

ORFs

- Part of genome between start and stop codons
- 6 frames for one sequence
- 1, 2, 3, -1, -2, -3

Search potential genes

- BLAST+, HMM search, KRAKEN
- Comparing with known set of genes
- Score the similarity

Search potential product

- Compare translated product with known database
- DIAMOND

Identification ab initio

- Search for patterns
- AI to derive function
- GLIMMER, GeneScan

Databases

- ENCODE
- Entrez Gene
- Ensembl
- GENCODE
- Gene Ontology Consortium
- RefSeq
- Uniprot
- Vega
- ...

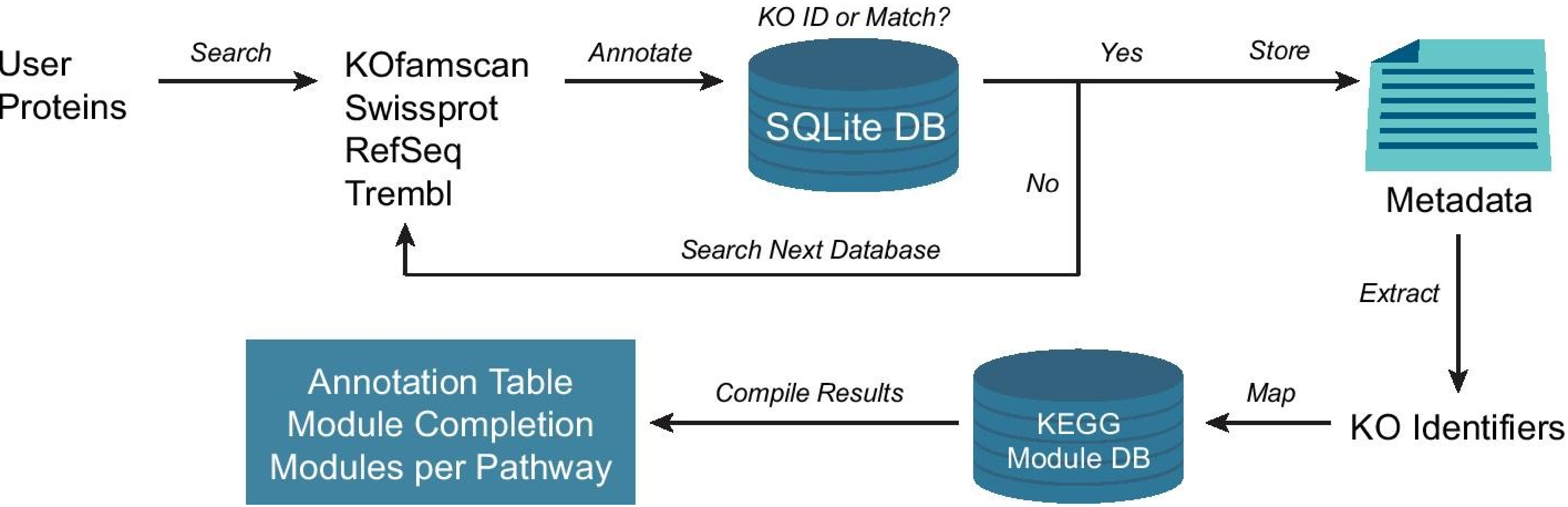
Useful tools

- Prokka
- MicrobeAnnotator
- NCBI Prokaryotic Genome Annotation Pipeline

Prokka

- Illumina BaseSpace app
- Vendor lock

MicrobeAnnotator



User Proteins

Search

KOfamscan
Swissprot
RefSeq
Trembl

Annotate

KO ID or Match?

SQLite DB

Yes

Store

Metadata

No

Search Next Database

Extract

KO Identifiers

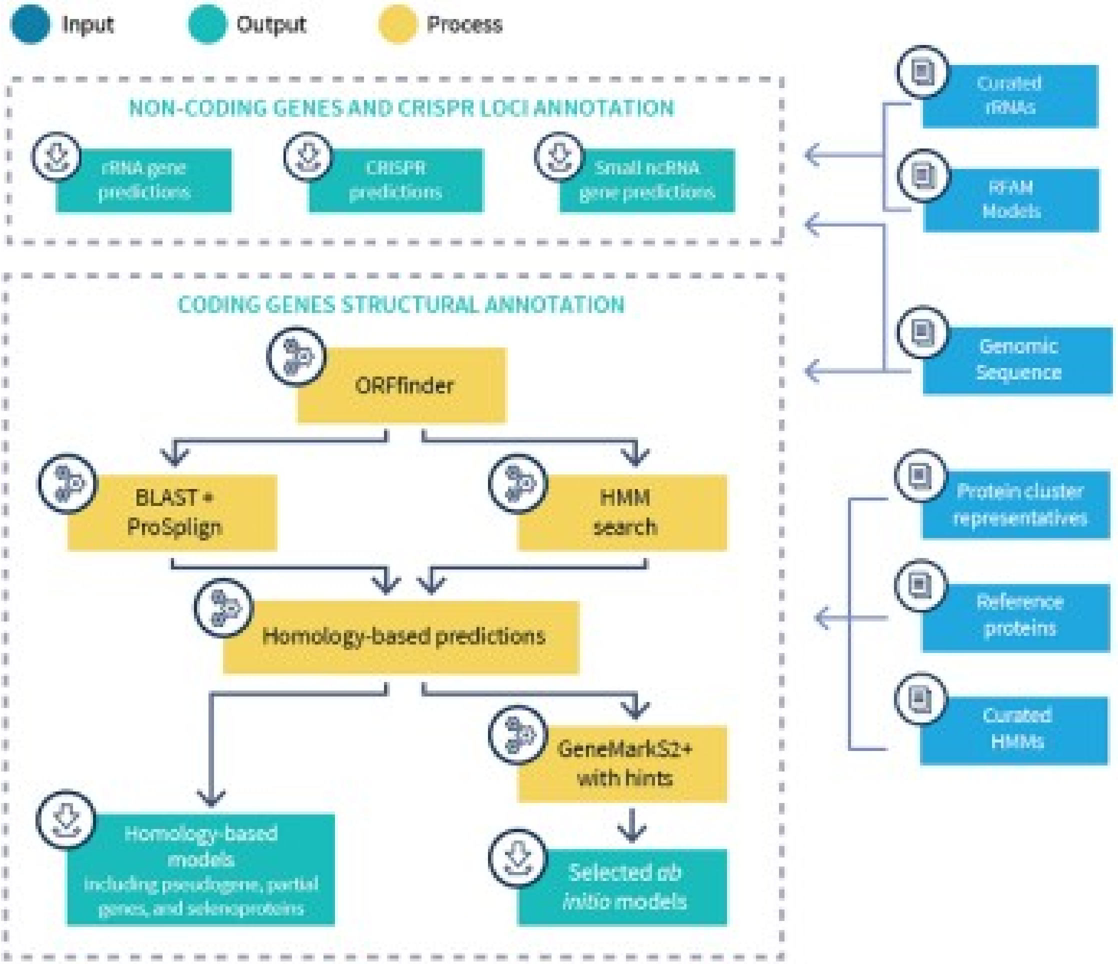
Map

KEGG Module DB

Compile Results

Annotation Table
Module Completion
Modules per Pathway

NCBI Annotation pipeline



● Input ● Output ● Process

NON-CODING GENES AND CRISPR LOCI ANNOTATION

rRNA gene predictions

CRISPR predictions

Small ncRNA gene predictions

Curated rRNAs

RFAM Models

Genomic Sequence

CODING GENES STRUCTURAL ANNOTATION

ORFfinder

BLAST + ProSplign

HMM search

Homology-based predictions

GeneMarkS2+ with hints

Homology-based models including pseudogene, partial genes, and selenoproteins

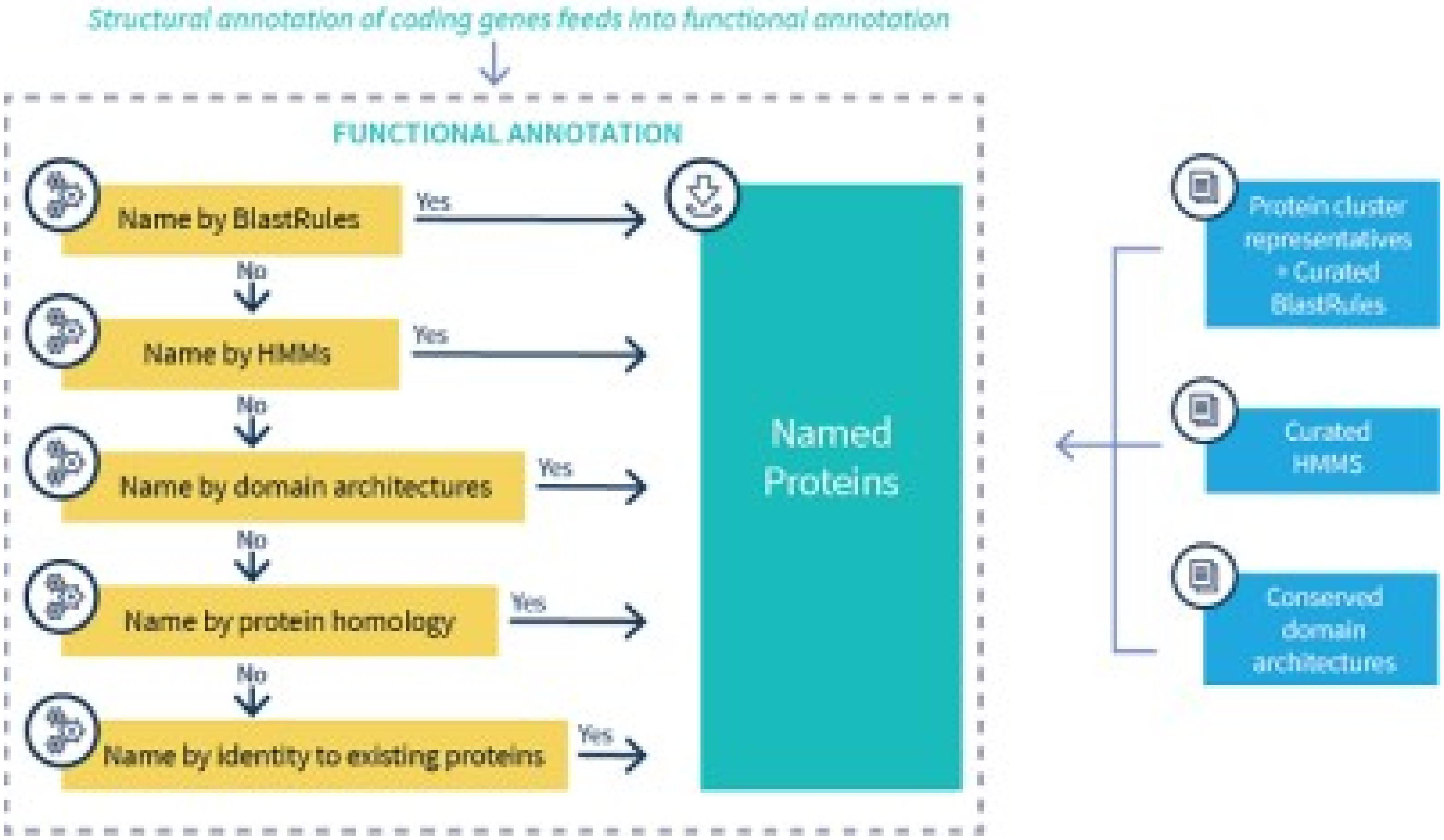
Selected *ab initio* models

Protein cluster representatives

Reference proteins

Curated HMMs

NCBI Annotation pipeline


Structural annotation of coding genes feeds into functional annotation
FUNCTIONAL ANNOTATION
Name by BlastRules
Yes
No
Name by HMMs
Yes
No
Name by domain architectures
Yes
No
Name by protein homology
Yes
No
Name by identity to existing proteins
Yes
Named Proteins
Protein cluster representatives + Curated BlastRules
Curated HMMS
Conserved domain architectures