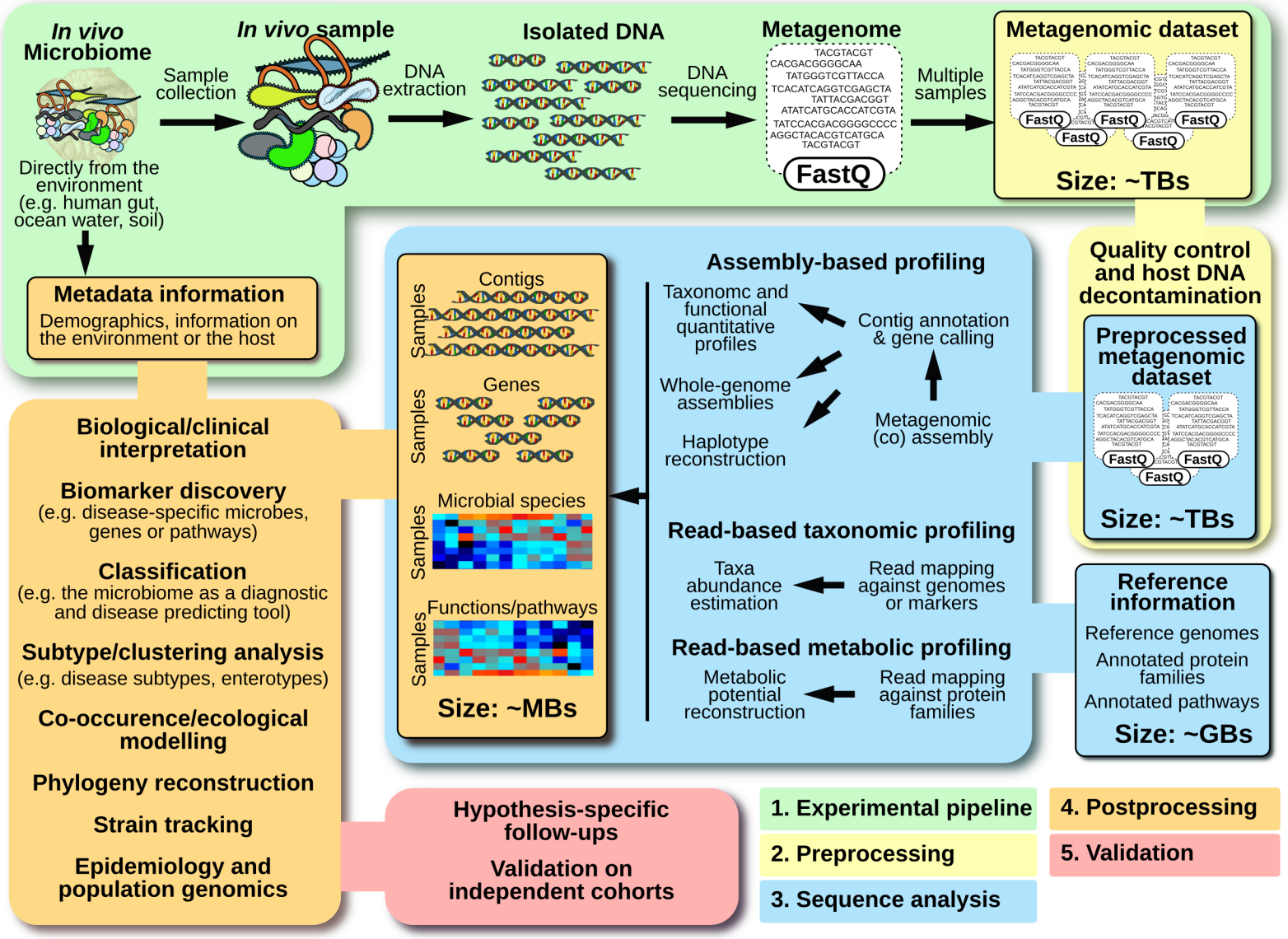


Quince et al Nat. Biotech http://rdcu.be/vMYk

Taxonomic and functional classification of metagenome sequences

Chris Quince

Introduction

- What is in my community and what can they do?
- Most of this talk is focused on read based methods but in principle could be applied to contigs
- There is a distinction between methods that aim to classify every read and those that only aim to profile the community
- Taxonomic and functional classification only differ in the choice of database

Overview

- Databases
- Search algorithms
- Software

Taxonomic/functional classification

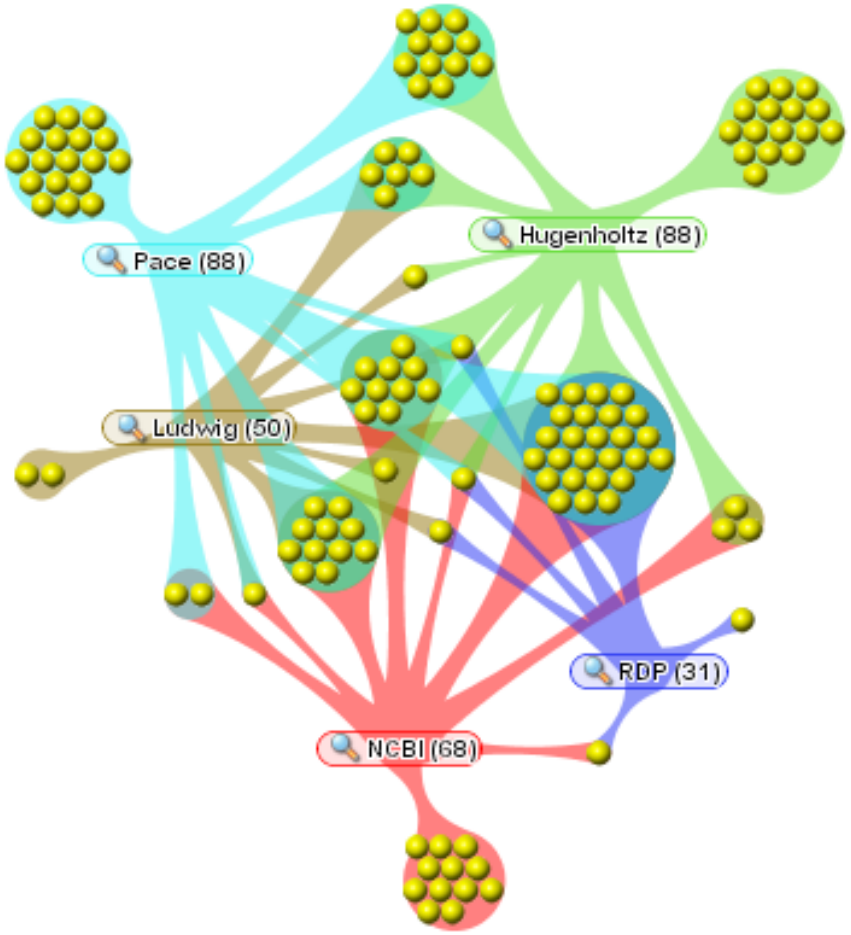
- Classification problem of identifying to which of a set of categories a new observation belongs
- Also known as supervised learning
- Requires a training database and an algorithm for comparing against that database
- Database is just a set of sequences with labels
- Query is a sequence too
- Taxonomic classification just means that database has a hierarchical labelling
- Functional databases are often hierarchical too

What is a Taxonomy?

- Classic hierarchical classification:
  - Kingdom, Phylum, Class, Order, Family, Genera, Species e.g.
    Fungi, Basidiomycota, Agaricomycetes, Agaricomycetidae, Boletales, Boletaceae, Boletus, Boletus edulis

- Taxonomy = system for classification
  Phylogeny = evolutionary history represented as a tree

Taxonomy is an arbitrary labeling..



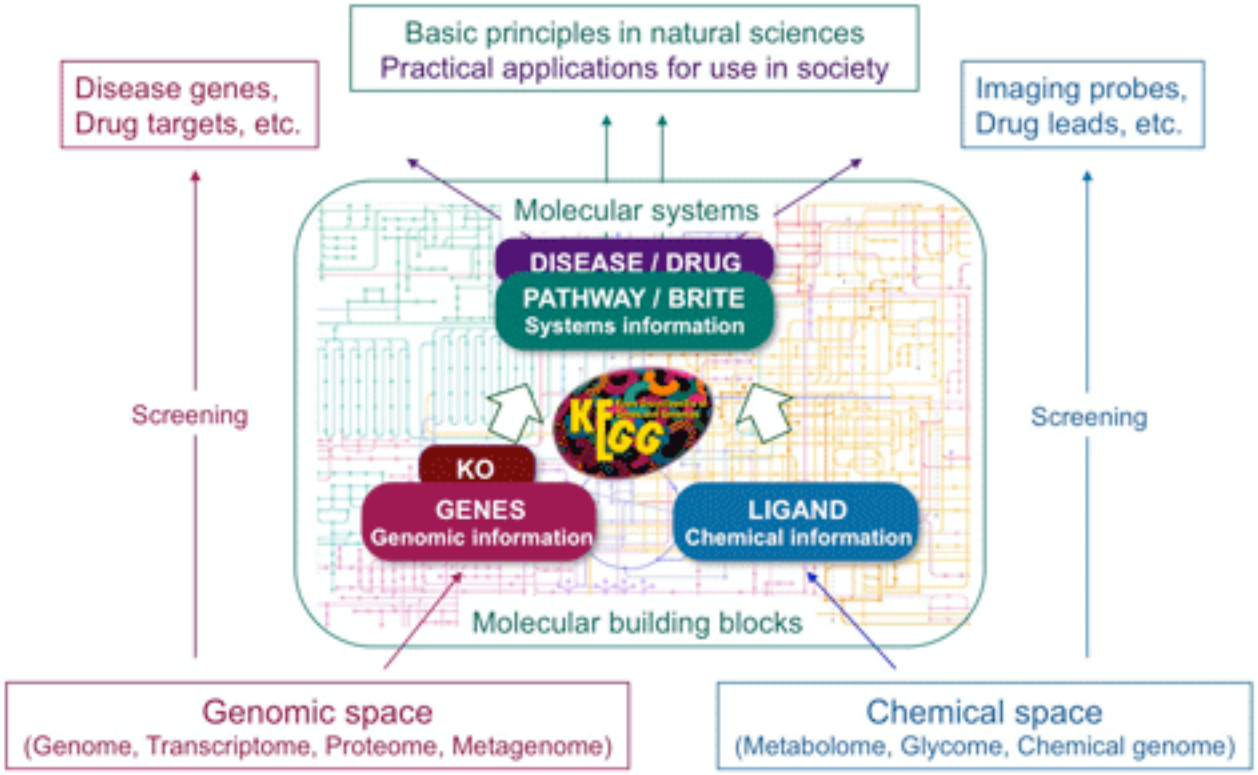
Metagenome taxonomy databases

- Usually based on NCBI Taxonomy possibly with some curation
- Either consists of whole genomes or the whole of the NR/NT
- Trade-off between quality and comprehensivity

Functional databases

- These comprise generic databases that attempt to contain every functional protein family e.g. Pfam http://pfam.xfam.org/
- Or they may be curated for a specific class of functions e.g. CAZy http://www.cazy.org/
- Discuss one the KEGG in a bit more detail…

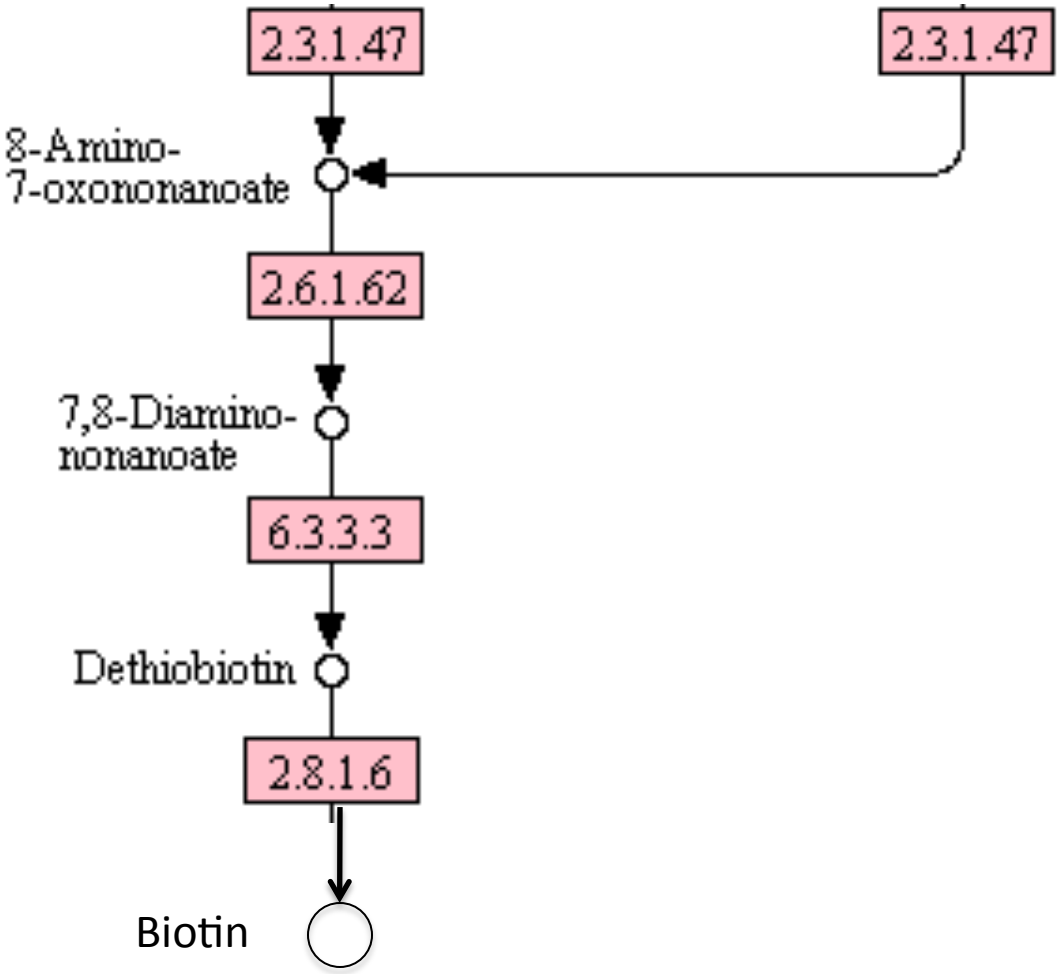
Kyoto Encyclopedia of Genes and Genomes (KEGG) database



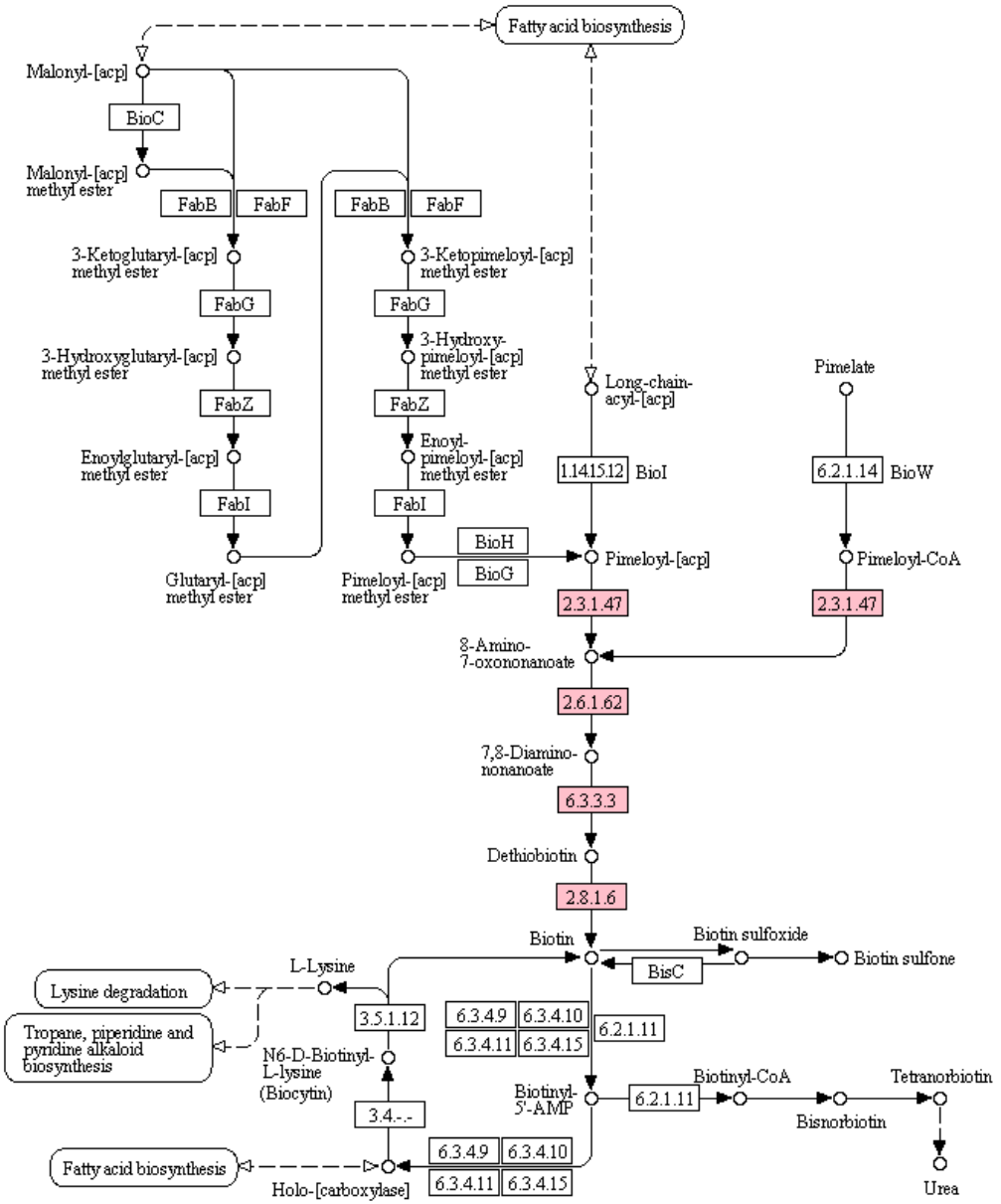
KEGG

Sequence → Gene → KO number → Module → Pathway

Kegg module: M00123 Biotin biosynthesis

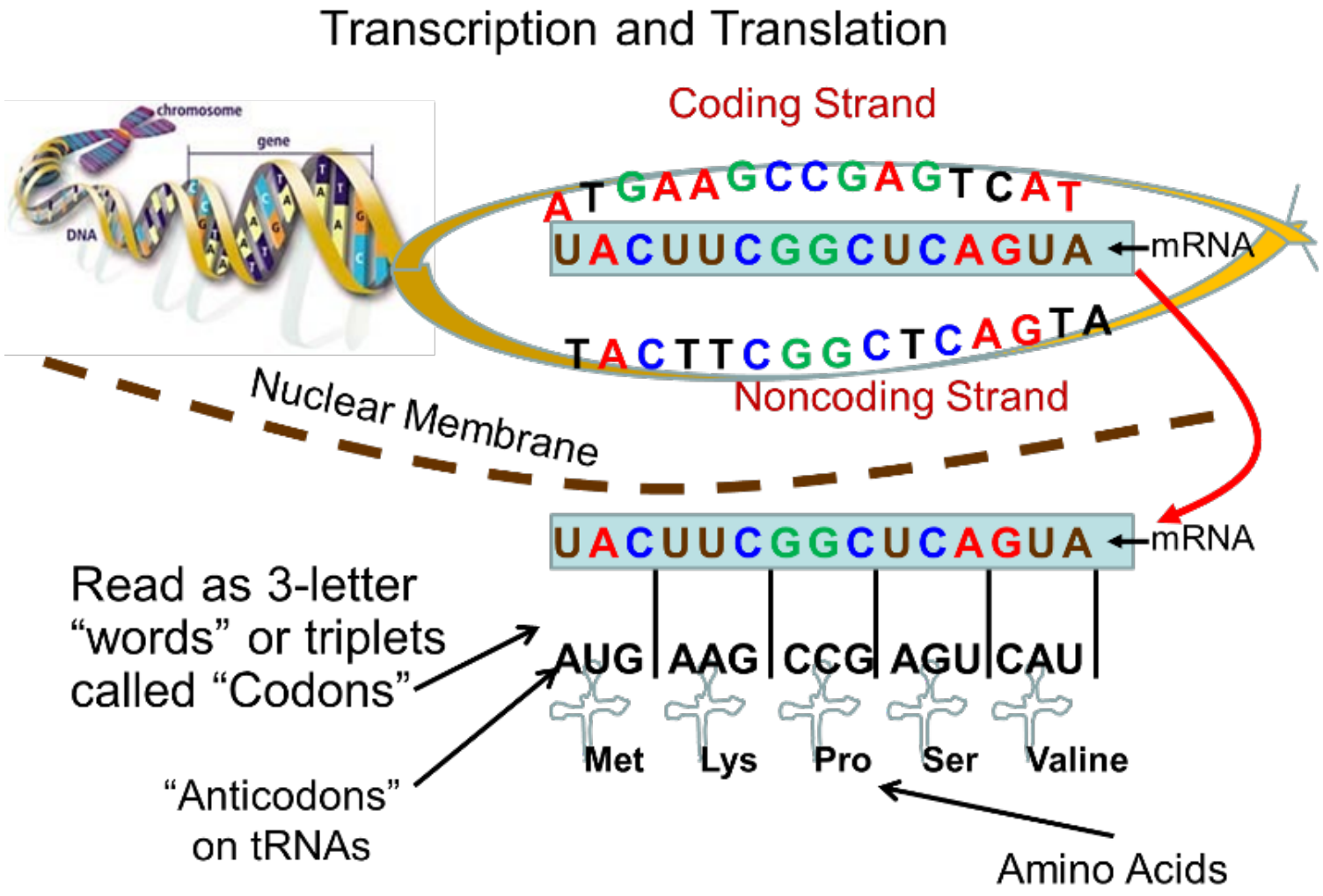


BIOTIN METABOLISM



00780 1/8/14

Searches are performed in either nucleotide or amino acid space…



Sequence search algorithms

- Four basic approaches:
  - Alignment or Sequence homology
  - Mapping
  - Kmers
  - Hidden Markov models (HMMs)

Local Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| ||||||||||||||||
Query Sequence 5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Global Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   ||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

Query Sequence

- Sequence alignment is sensitive and can find very distant relationships between query and target
- Also precise can distinguish between two very similar targets
- Drawback alignment is slow!
- Examples of alignment search algorithms are Blast, Diamond, Rapsearch

Homology distances

- Once sequences are aligned metrics are calculated to indicate similarity to reference sequences
  - Edit distance
  - E value

ACTGCTTTAGGGGG -> database
ACT- CTTAAGGGGT -> query

Edit distance = 3

describes the number of hits one can "expect" to see by chance when searching a database of a particular size

- Typically top N hits are returned

Mapping

- Realisation that we do not always need searches that can find distant relationships
- If we restrict the search to a everything within a threshold of similarity and only return hits better than that then we can use efficient algorithms e.g. Burrows-Wheeler transform
- Examples of mappers are:
  - BWA
  - Bowtie2
- Vsearch and usearch are based on the same principle but follow fast map with slow alignment
- Restricted to nucleotide comparisons?
- Precise but not sensitive

Kmers

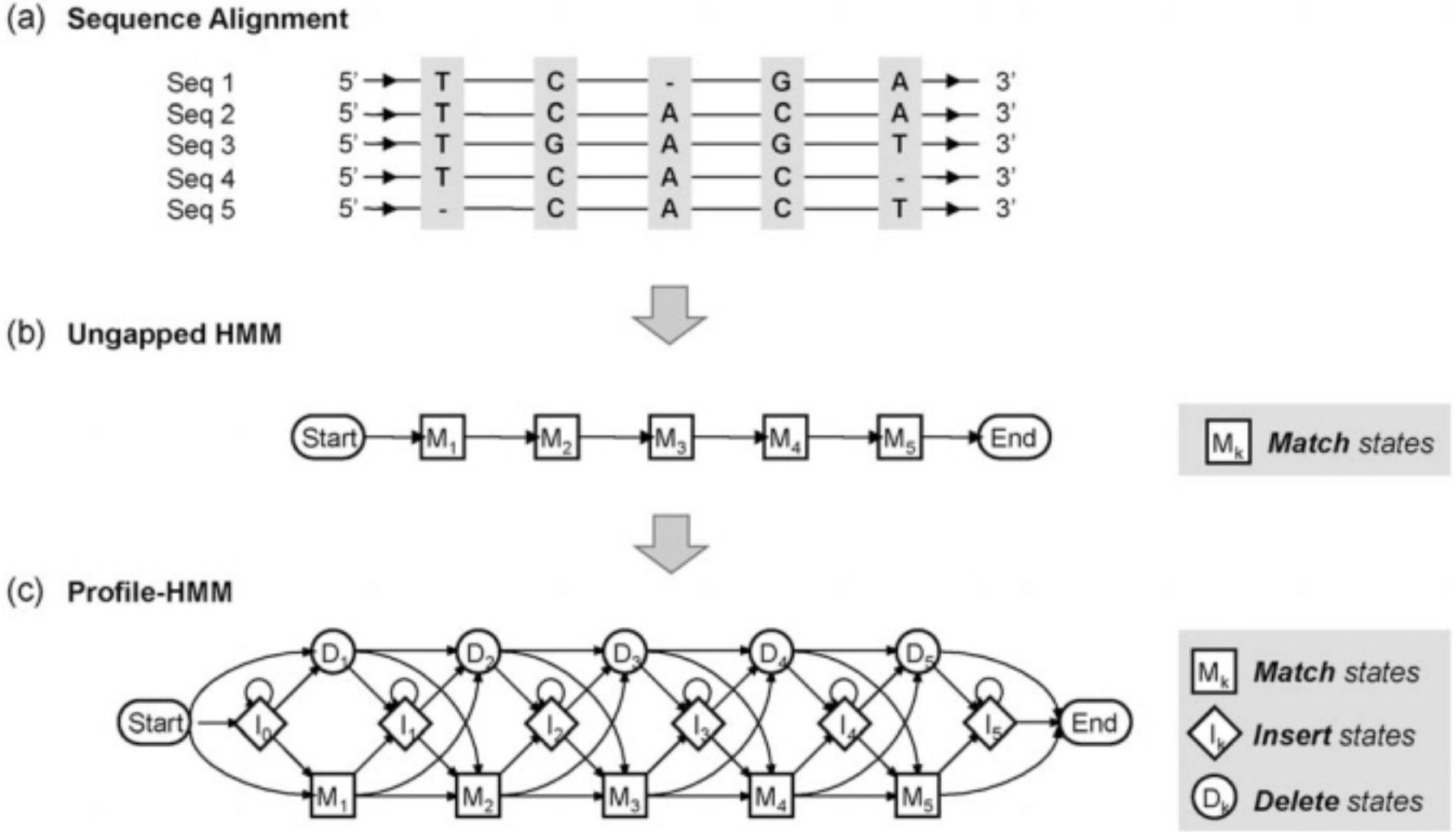
- A "k-mer" is a word of DNA that is k long

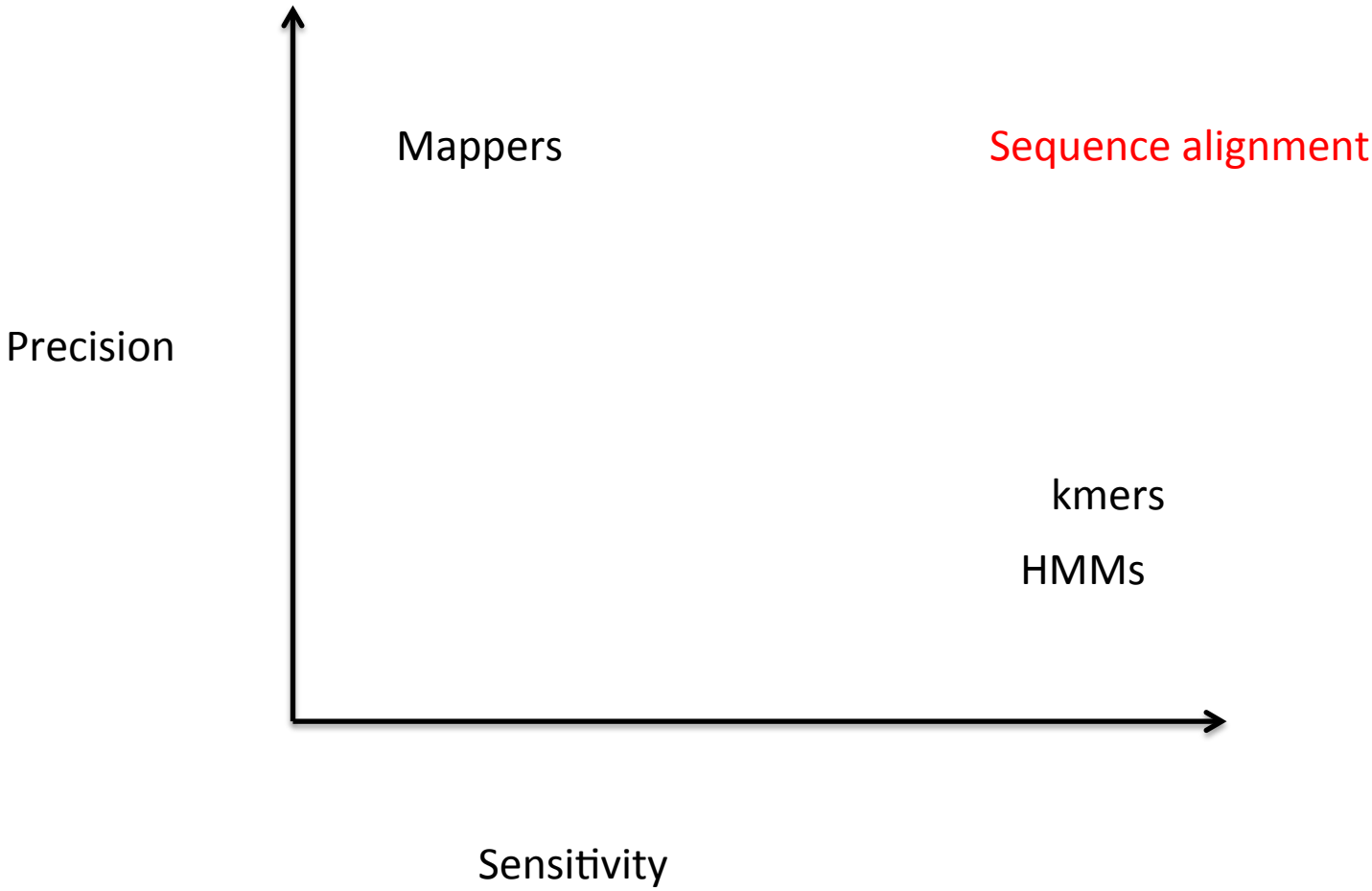
$$sequence = \left(CTGGCTTGA\right)$$

$$4-mers: 2\times CTGG, TGGC, GGCT, GCTT, TTGA$$

- Homologous sequences share kmers
- Comparing kmer composition is a fast way to search queries against a database
- Drawback lacks precision

Hidden Markov Models (HMMs)




Mappers
Sequence alignment
Precision
kmers
HMMs
Sensitivity

Taxonomic classification/profiling software

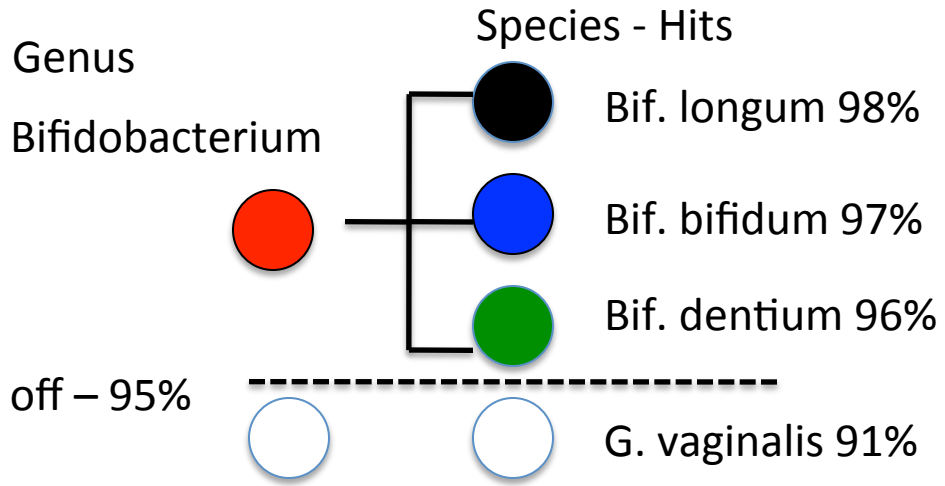
- All software for metagenome read classification consists of some variant on one of these algorithms and a database:
  - Homology search: MEGAN
  - Kmer based: Kraken
  - FM - index: Centrifuge
- Profilers same principle but use database of marker genes e.g. MetaPhlan2 or mOTU hence cannot classify every read

MEGAN – Metagenome analyser

- Matches against the NR using Diamond blastx
- Lowest common ancestor (LCA) algorithm based on NCBI taxonomy

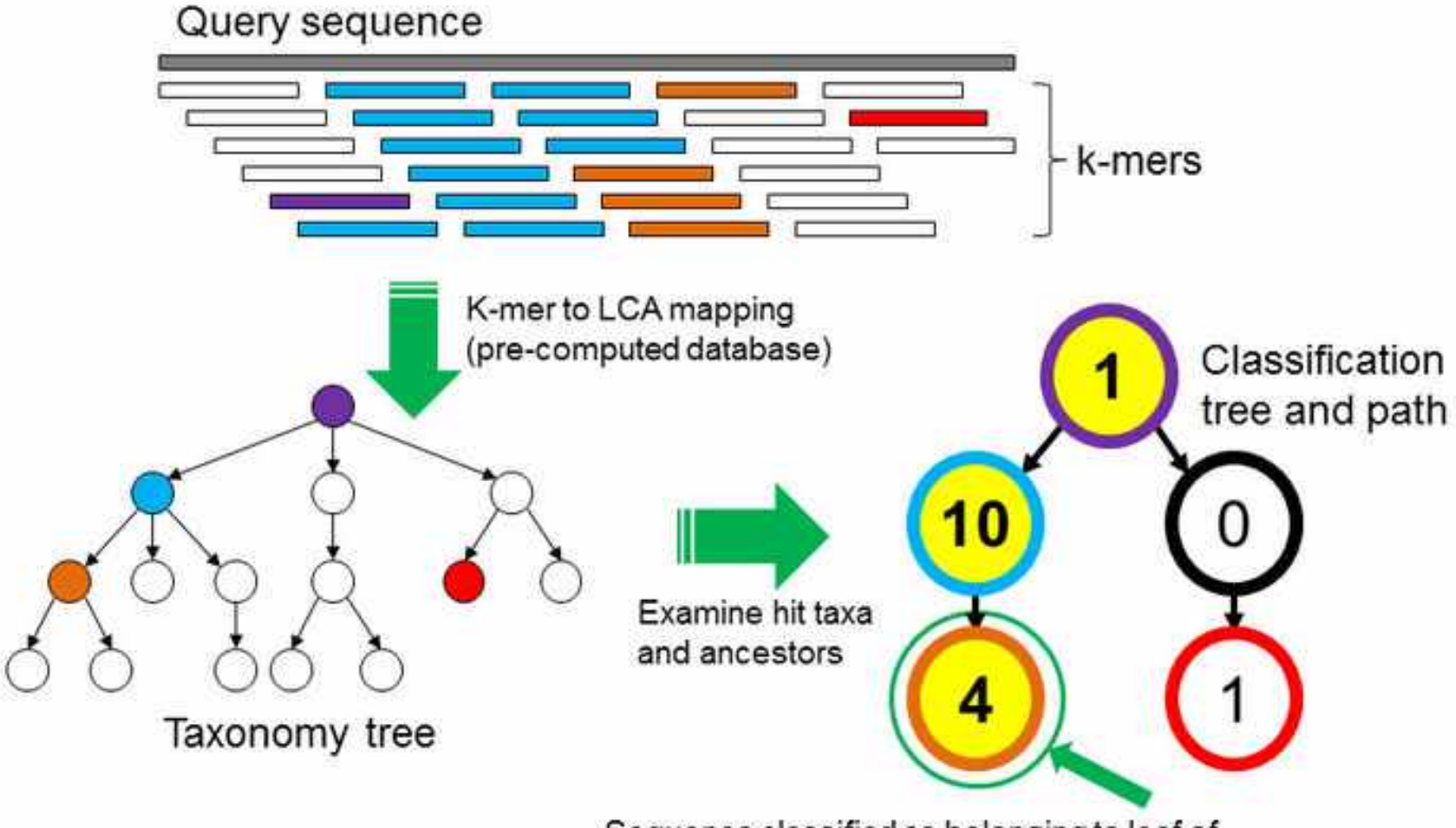
Kingdom Phylum … Family

Bacteria Actinobacteria … Bifidobacteriaceae



Kraken

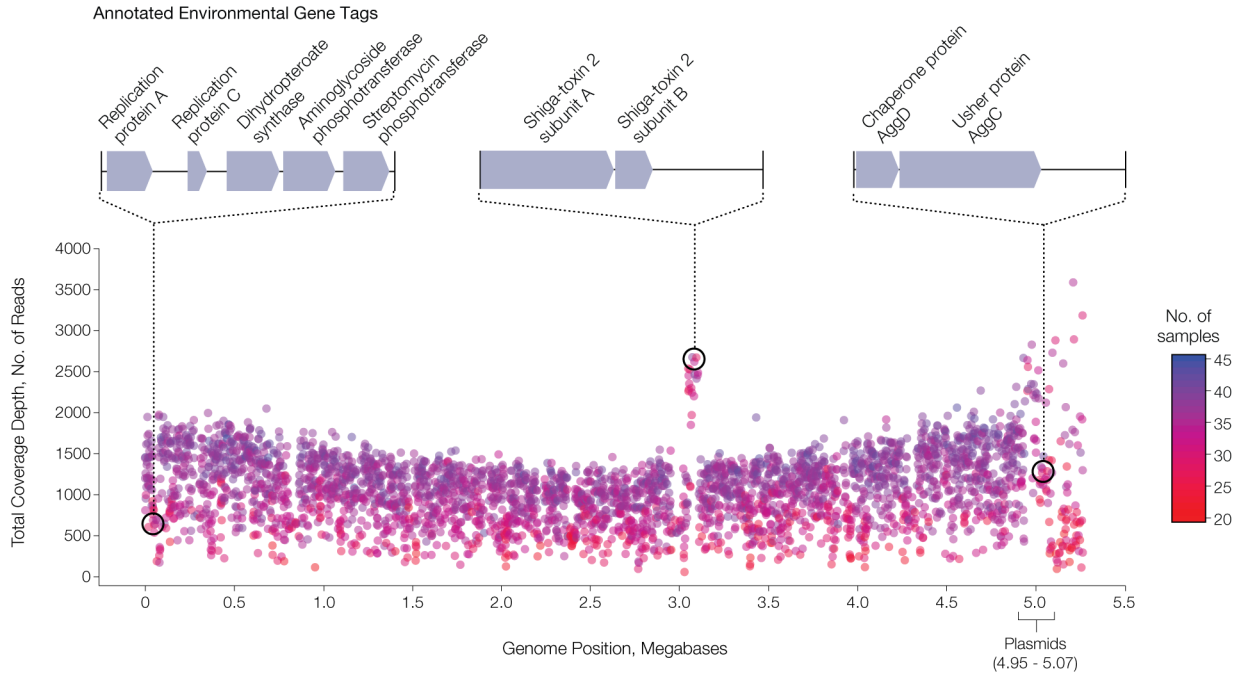
- Kmer based default 31bp
- Default database comprises RefSeq 2014



Functional profiling

- Same principles apply to functional profiling communities except now the searches have to be distant homology or HMM and in amino acid space
- HUMAnN: The HMP Unified Metabolic Analysis Network (Abubucker et al. PLoS Comp Biol 2012)

Summary


Annotated Environmental Gene Tags
Replication protein A
Replication protein C
Dihydropteroate synthase
Aminoglycoside phosphotransferase
Streptomycin phosphotransferase
Shiga-toxin 2 subunit A
Shiga-toxin 2 subunit B
Chaperone protein AggD
Usher protein AggC
Total Coverage Depth, No. of Reads
4000
3500
3000
2500
2000
1500
1000
500
0
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
Genome Position, Megabases
Plasmids
(4.95 - 5.07)
No. of samples
45
40
35
30
25
20