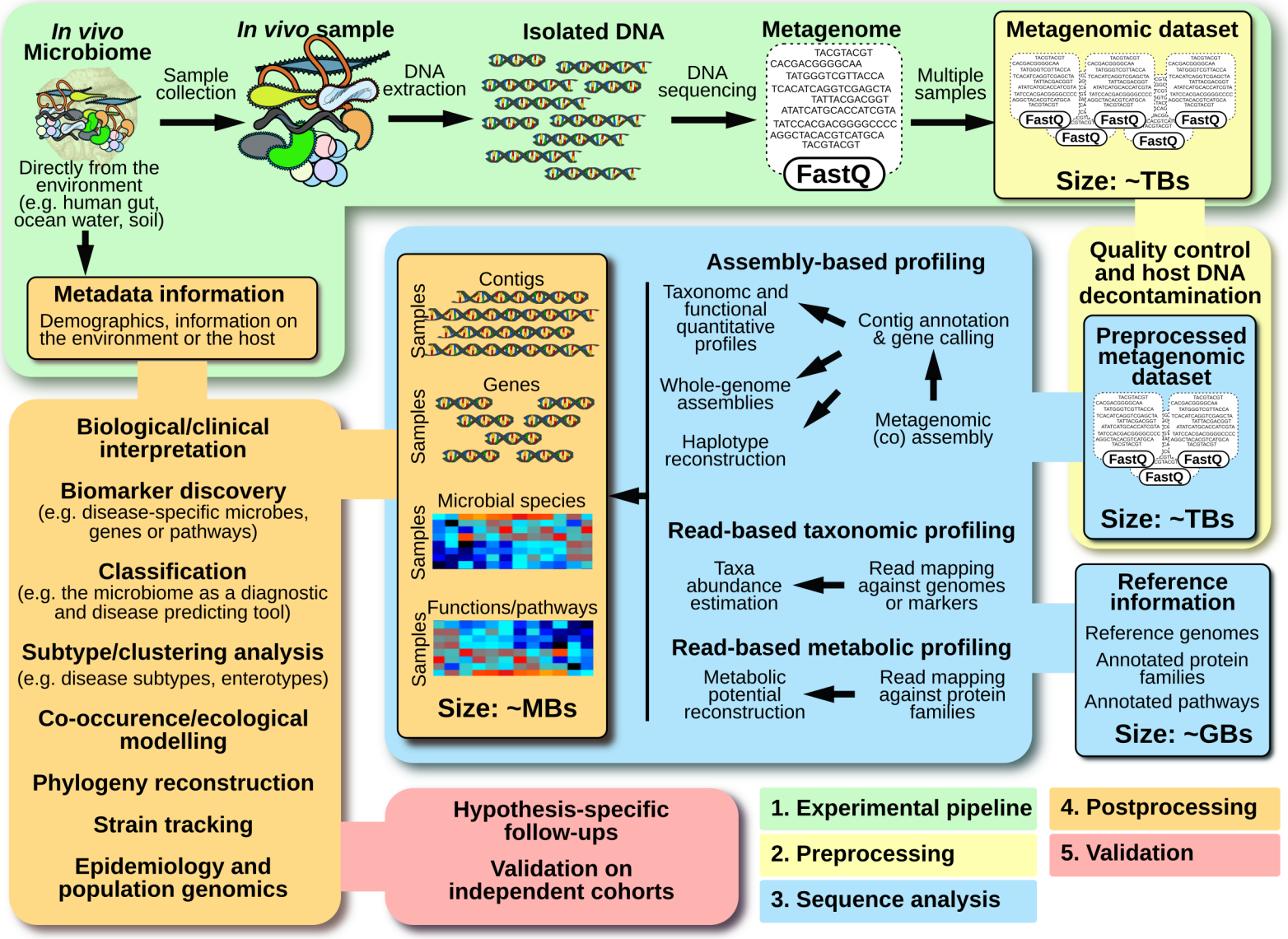


Quince et al Nat. Biotech http://rdcu.be/vMYk

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

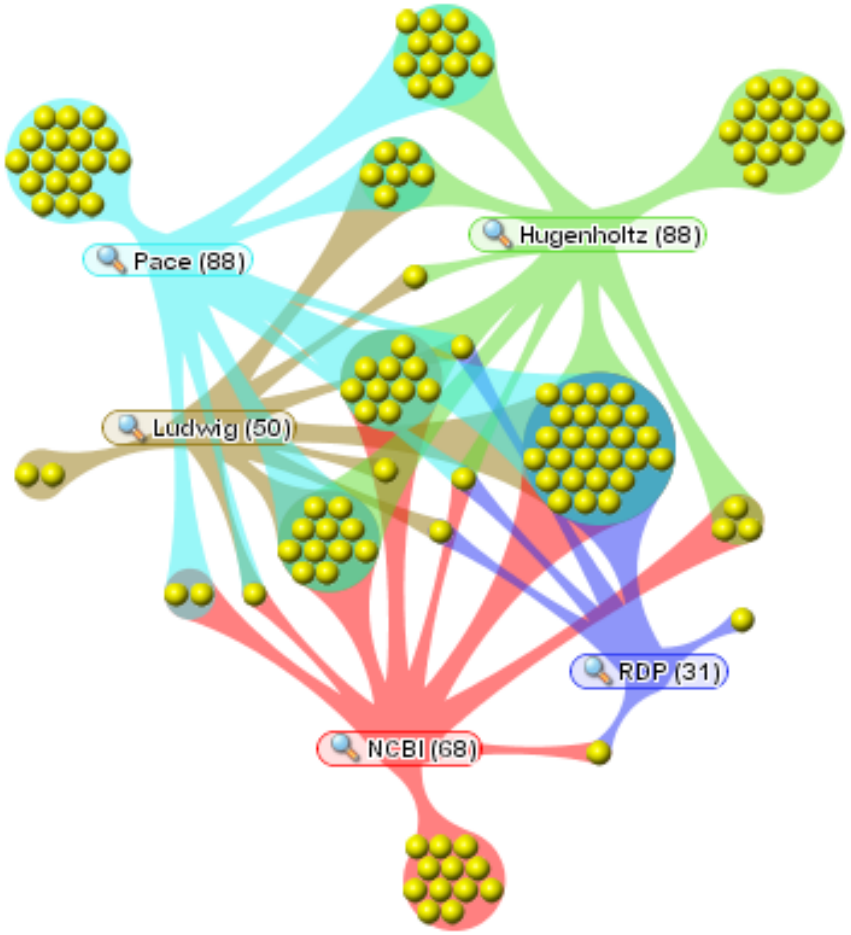
- Classification (supervised learning) problem of identifying to which of a set of categories a new observation belongs
- Requires a training database and an algorithm for comparing against that database
- Database is just a set of sequences with labels
- Query is a sequence too
- Taxonomic classification 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 COGs and the KEGG in a bit more detail…

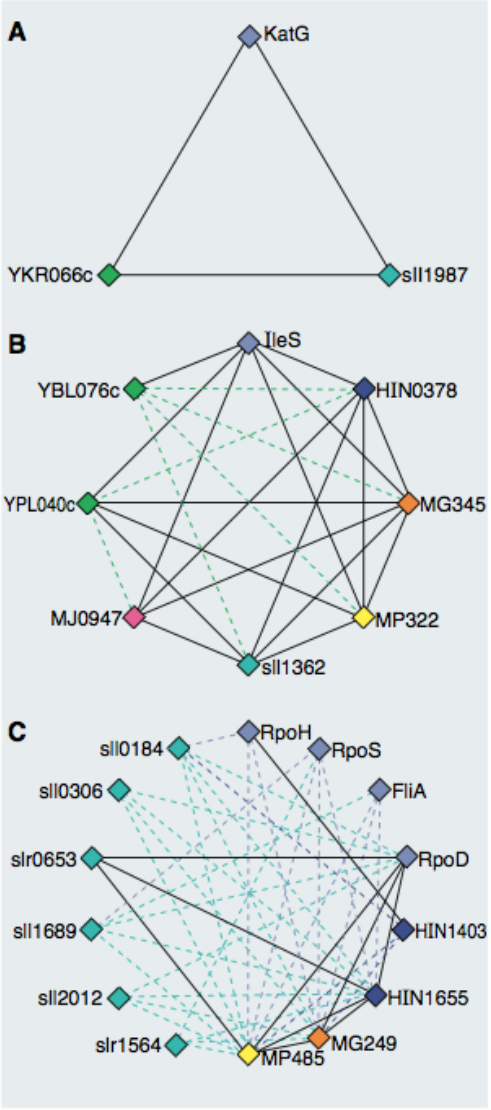
Clusters of orthologous genes (COGs)

http://www.ncbi.nlm.nih.gov/COG/

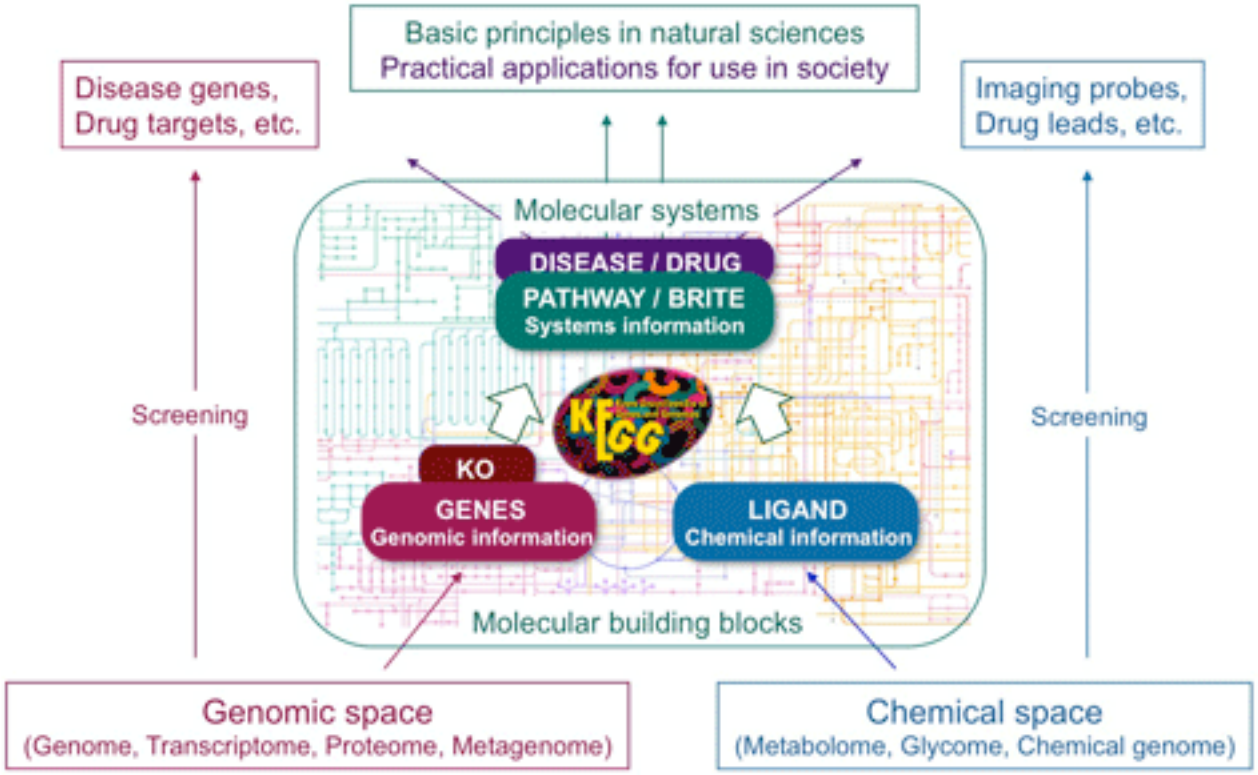
- Each COG consists of individual orthologous genes or orthologous groups of paralogs from three multiple lineages
- Any two proteins from different lineages that belong to the same COG are orthologs
- COGs can be used in phlogenies and are strongly related to function:

  e.g. COG0001 Glutamate-1-semialdehyde aminotransferase
- Currently (2003) there are 4873 COGs from 66 genomes
- Extensions include group specific e.g. archaeal COGs and unsupervised eggNOGS

  http://eggnog.embl.de/version_4.0.beta/



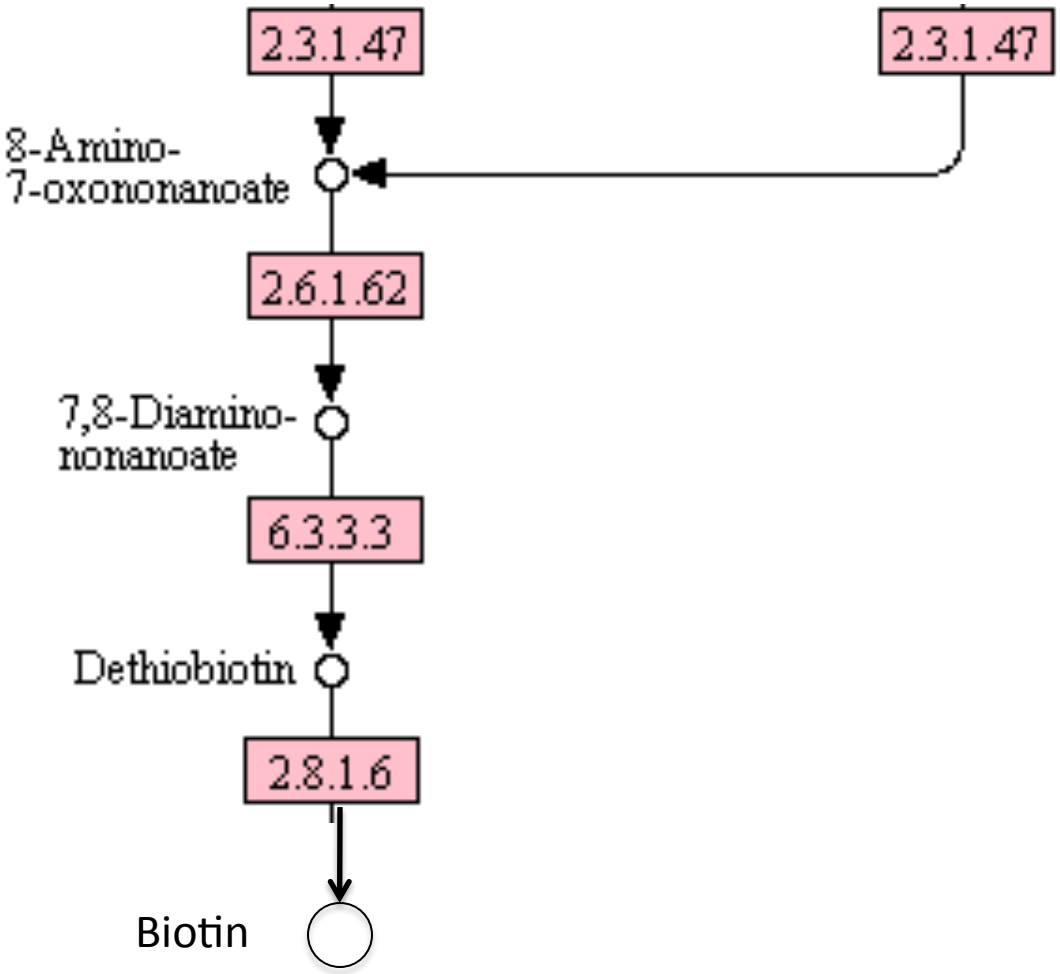
Kyoto Encyclopedia of Genes and Genomes (KEGG) database



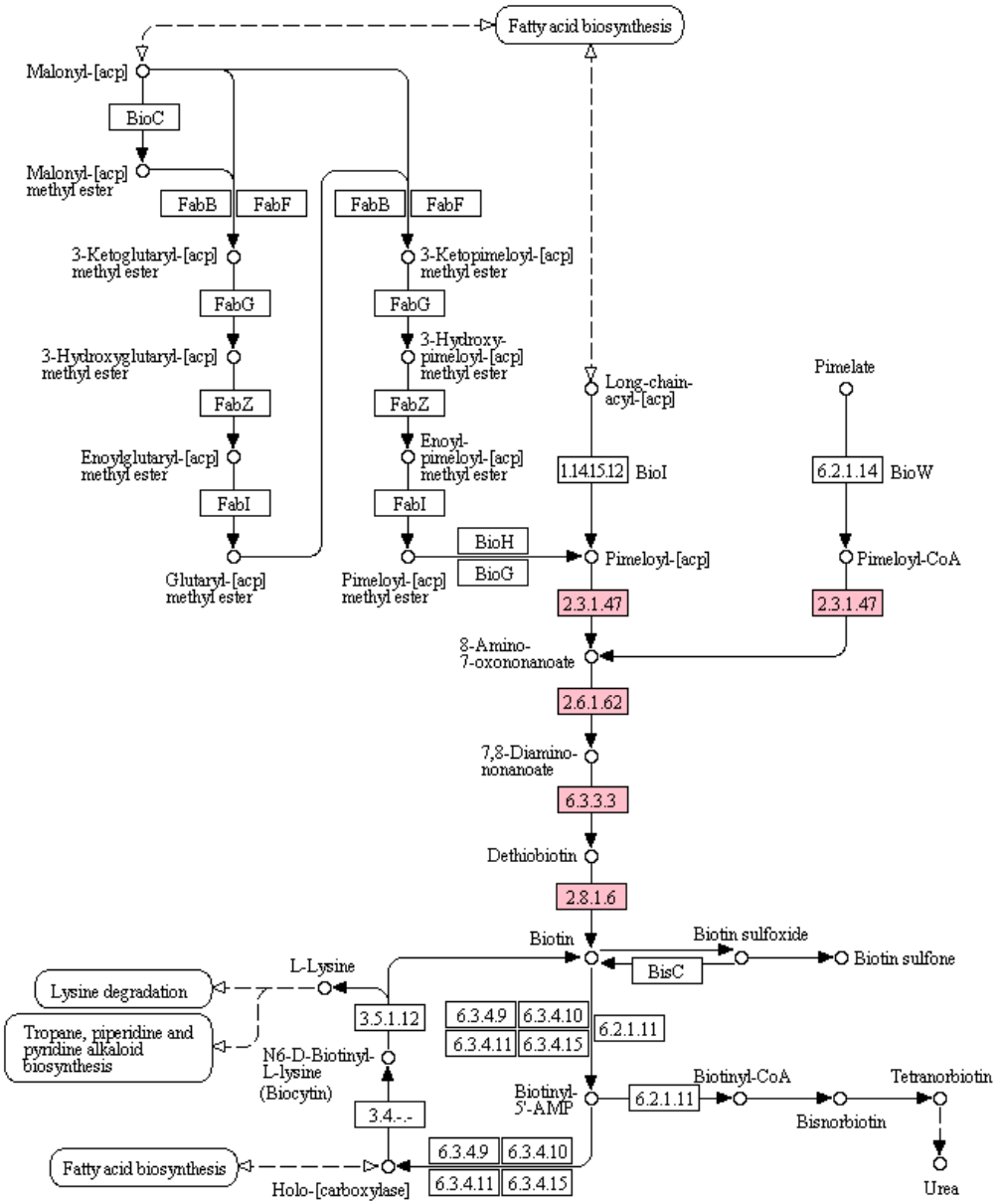
KEGG

Sequence → Gene → KO number → Module → Pathway

Kegg module: M00123 Biotin biosynthesis

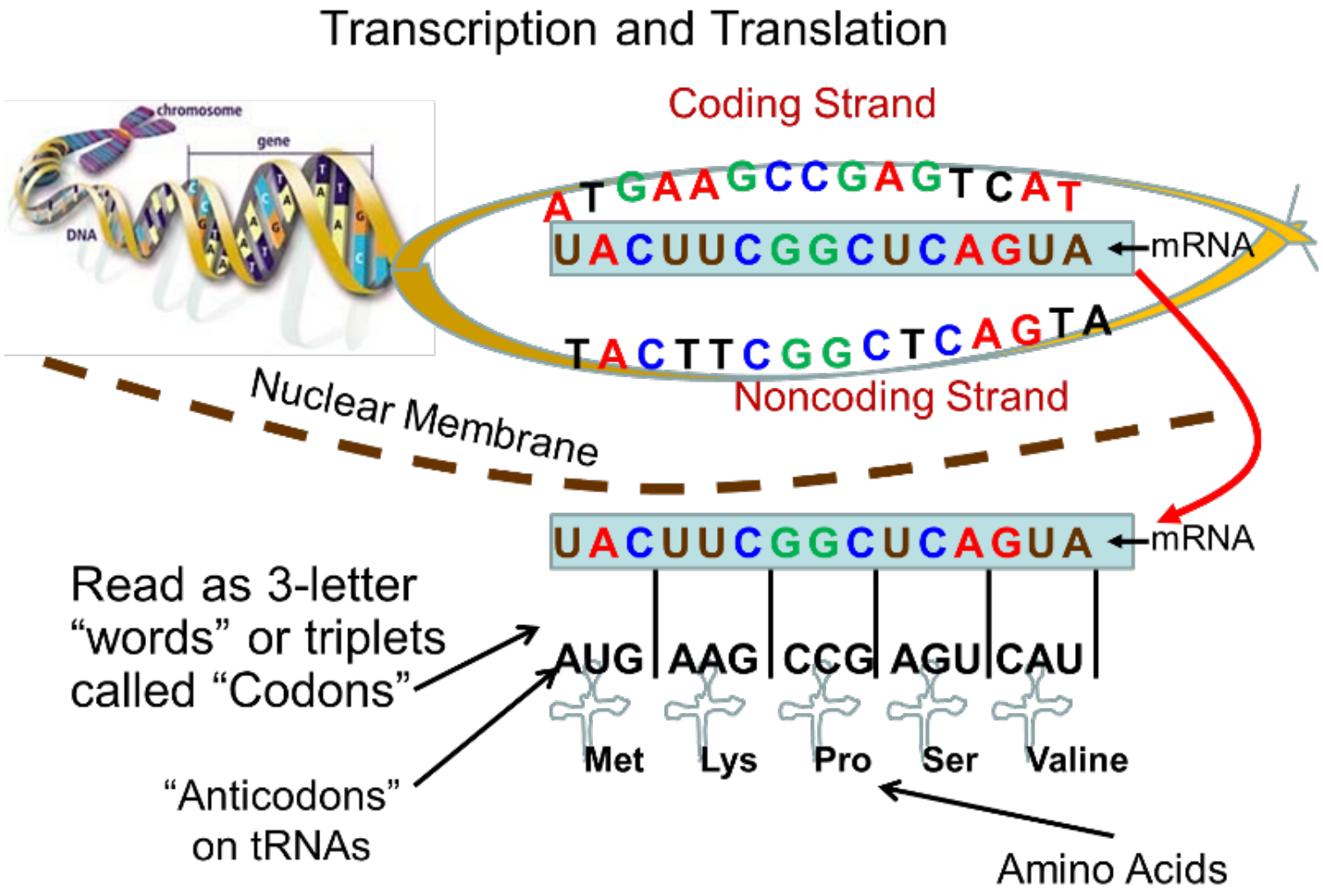


BIOTIN METABOLISM



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 can find 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, vsearch (nucleotide only)

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
- Restricted to nucleotide comparisons?
- Precise but not sensitive

Kmers

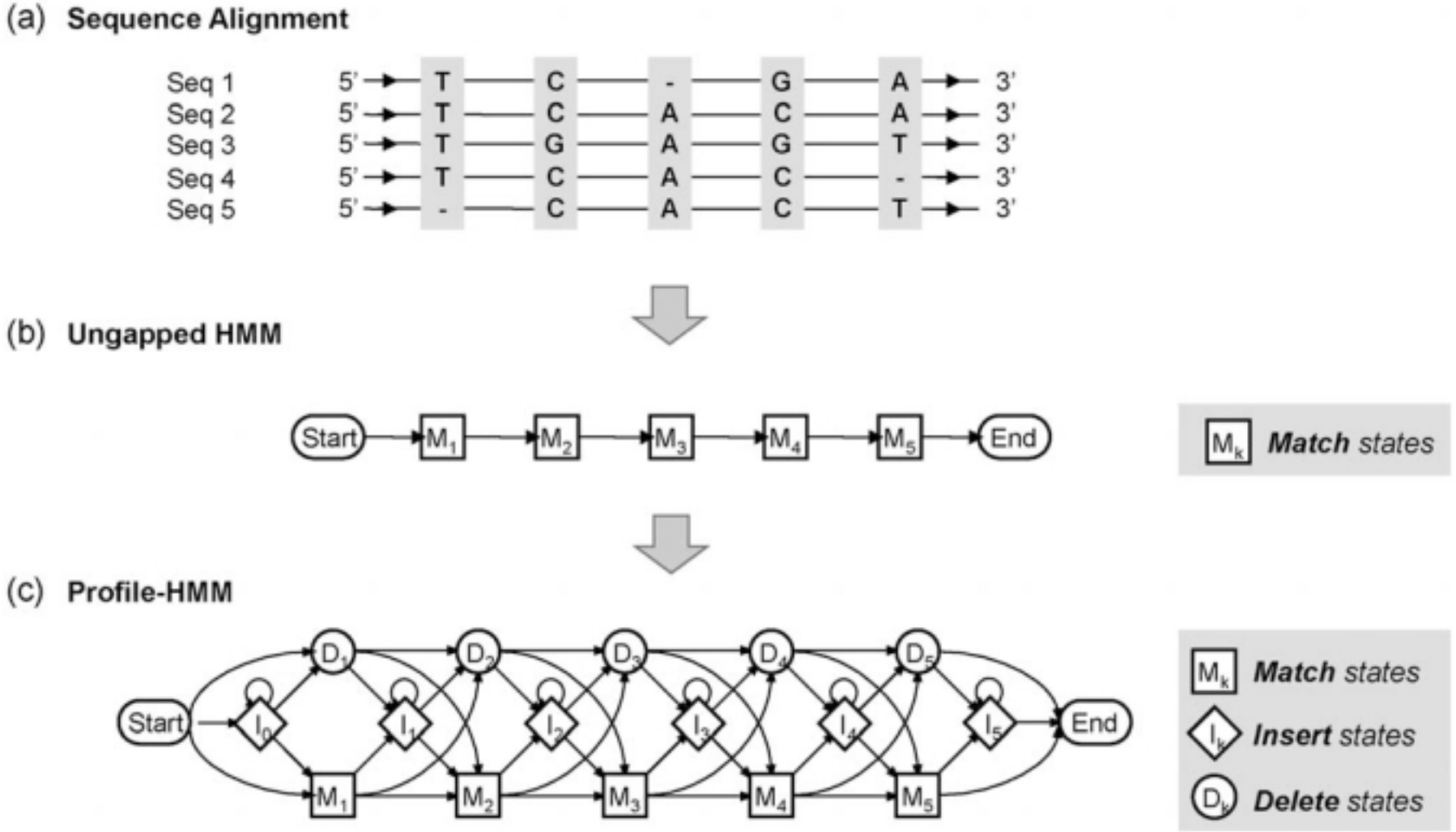
- A "k-mer" is a word of DNA that is k long

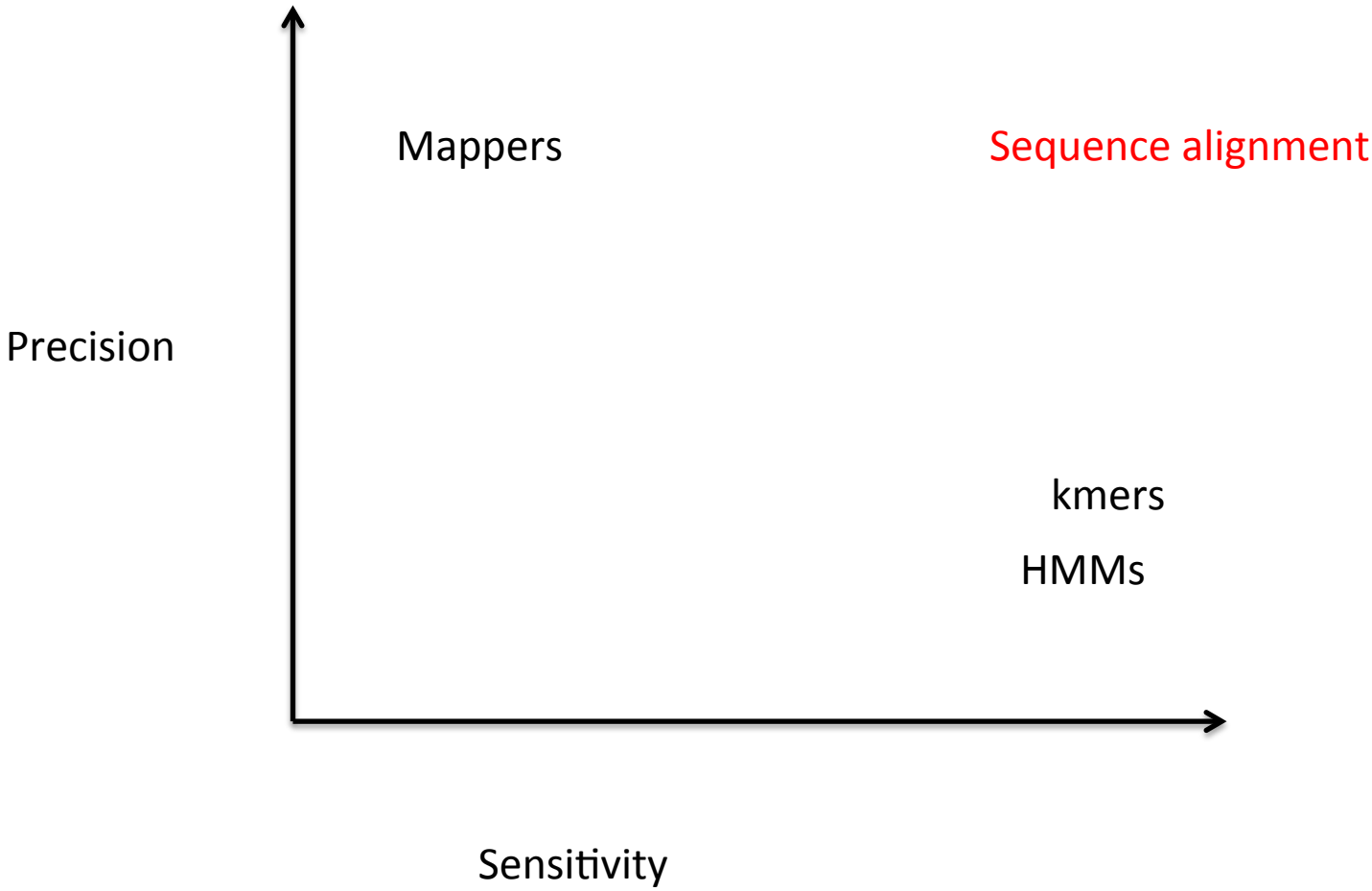
$$sequence = \left(CTGGCTTGA\right)$$

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

- 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

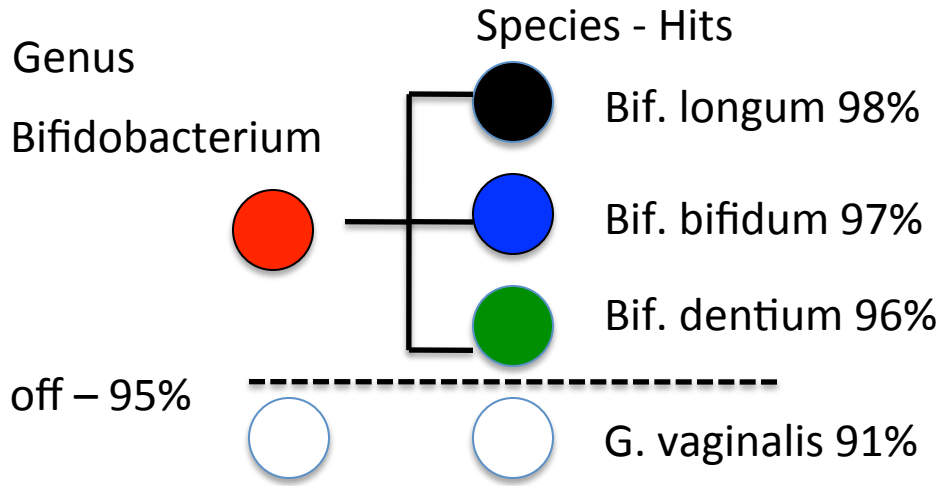
- All software for metagenome read classification consists of some variant on one of these algorithms and a database:
  - Homology search: MEGAN
  - Kmer based: Kraken
  - FM - index: Centrifuge
- Profilers same principle but use database of marker genes e.g. MetaPhlan2 or mOTU hence cannot classify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

MEGAN – Metagenome analyser

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

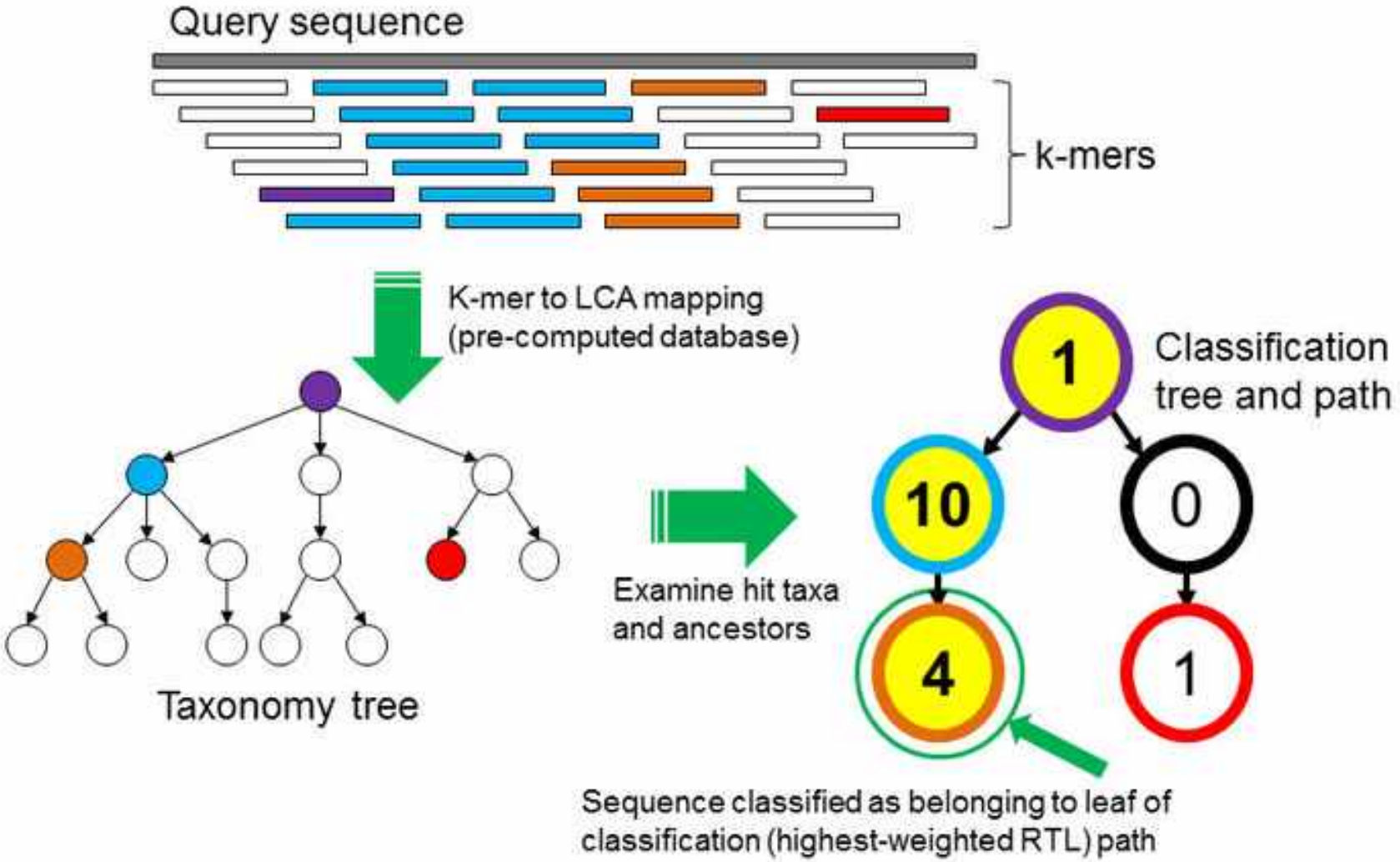
Kingdom Phylum … Family

Bacteria Actinobacteria … Bifidobacteriaceae



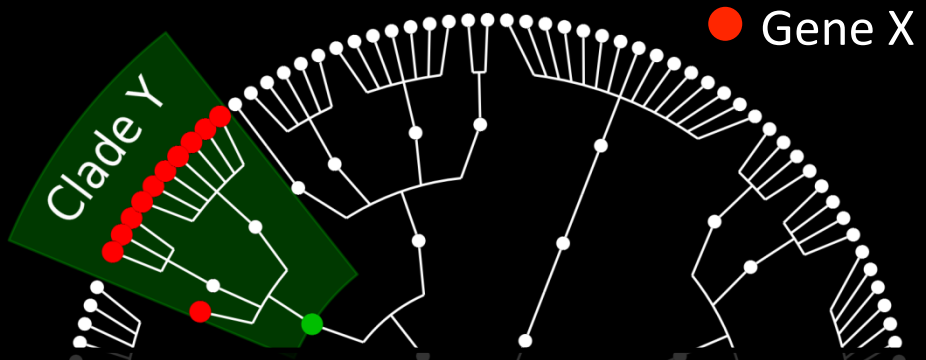
Kraken

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



MetaPhlAn: the basic idea

X is a unique marker gene for clade Y



THE INPUT

- ~25.000 genomes (bacteria, archara, fungi, viruses)
- ~1/10 of genomes are final, ~9/10 draft
- ~7,100 species (excludes incomplete annotations, spp., etc.)

THE RESULTING DATABASE

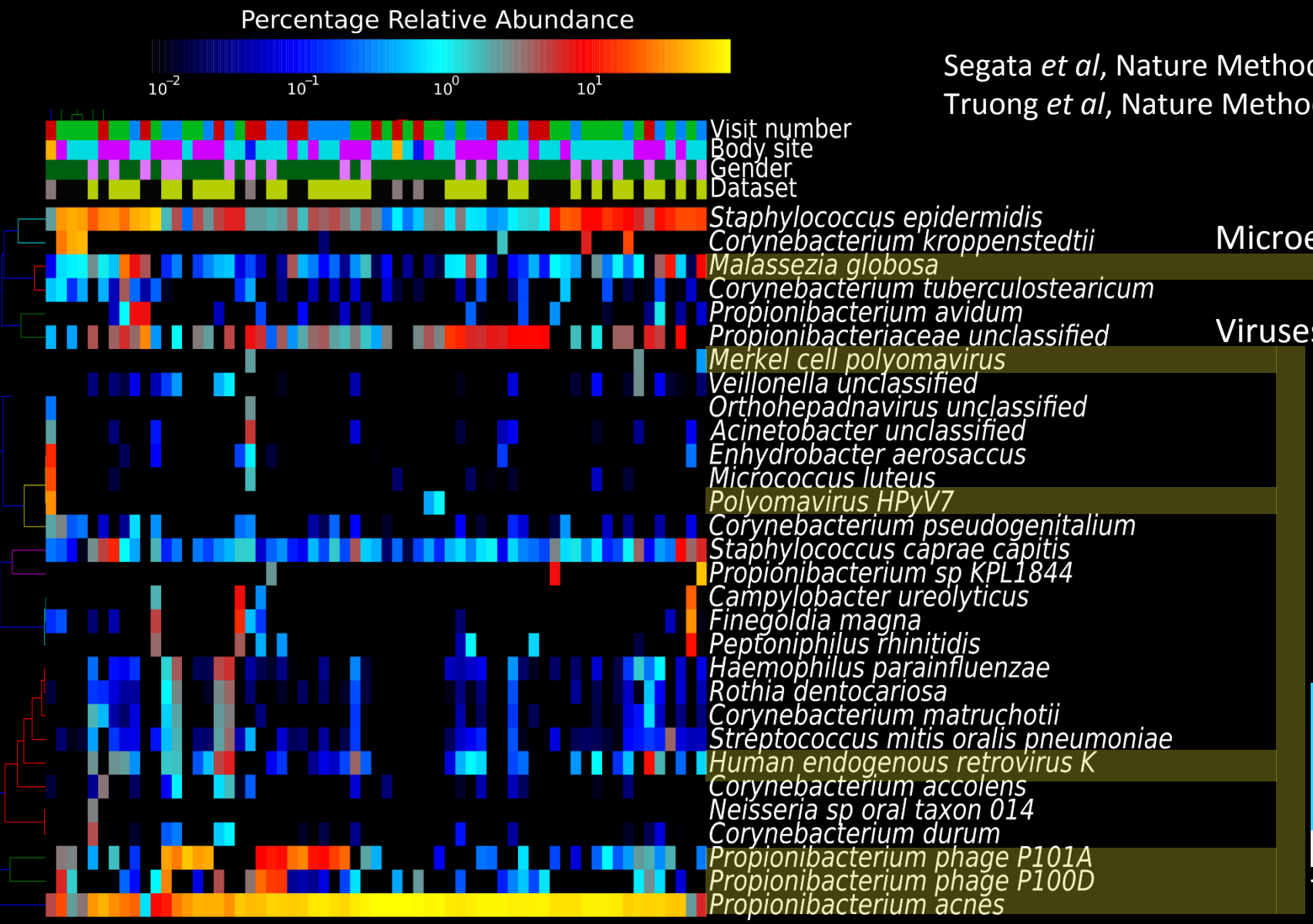
- ~15M total unique marker genes
- ~1M "most representative" unique marker genes
- 180±45 markers per species (200 fixed max)
- Quasi-markers used to resolve ambiguity in postprocessing

IDEA

1. Pre-identify markers from reference genomes
2. Use markers as proxy for taxonomic clades in shotgun metagenomics

Method: ChocoPhlAn

MetaPhlAn2: trans-kingdom profiling using marker genes



Segata *et al*, Nature Methods 2012
Truong *et al*, Nature Methods, 2015

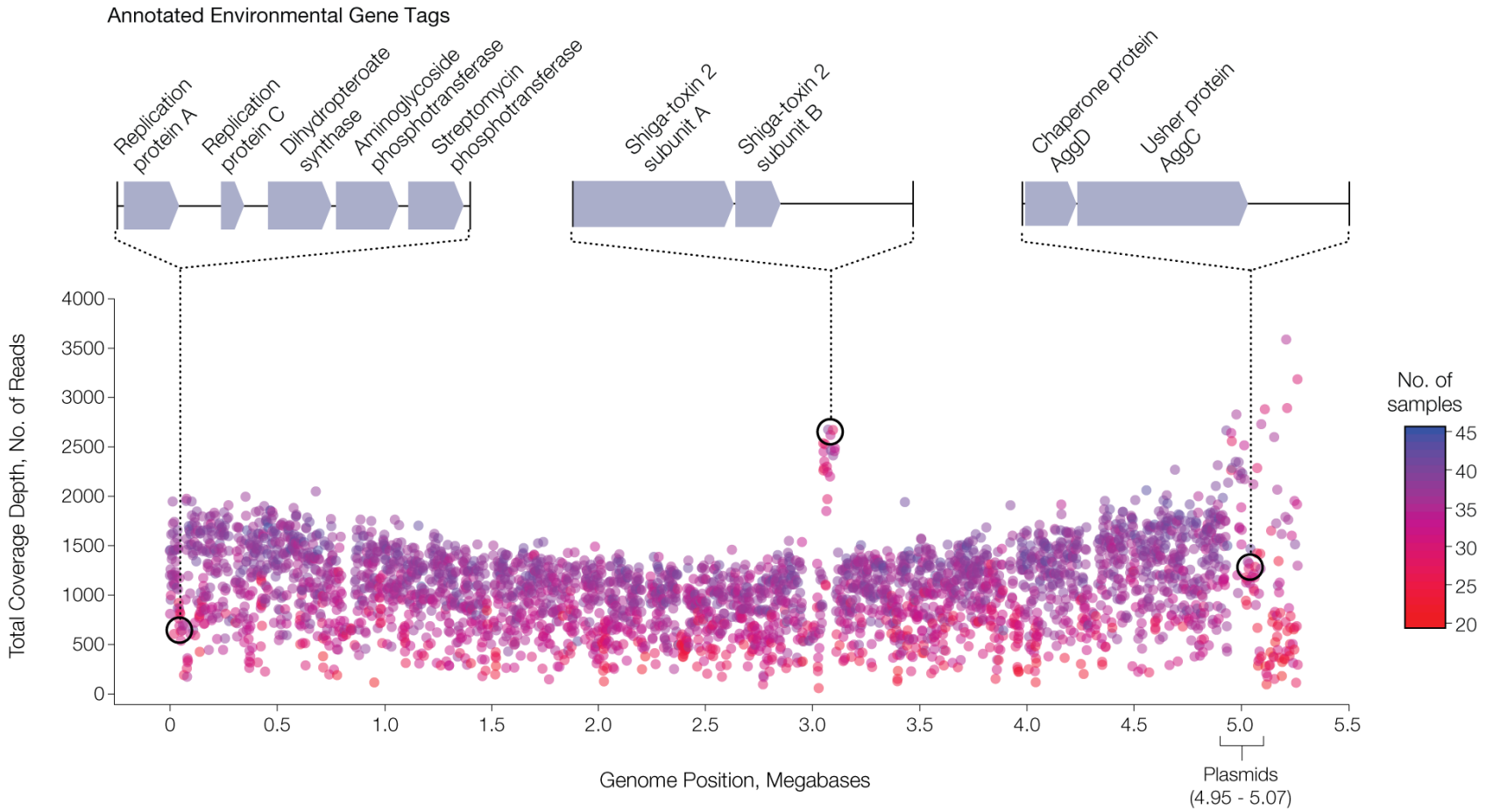
Tin Truong

False positives

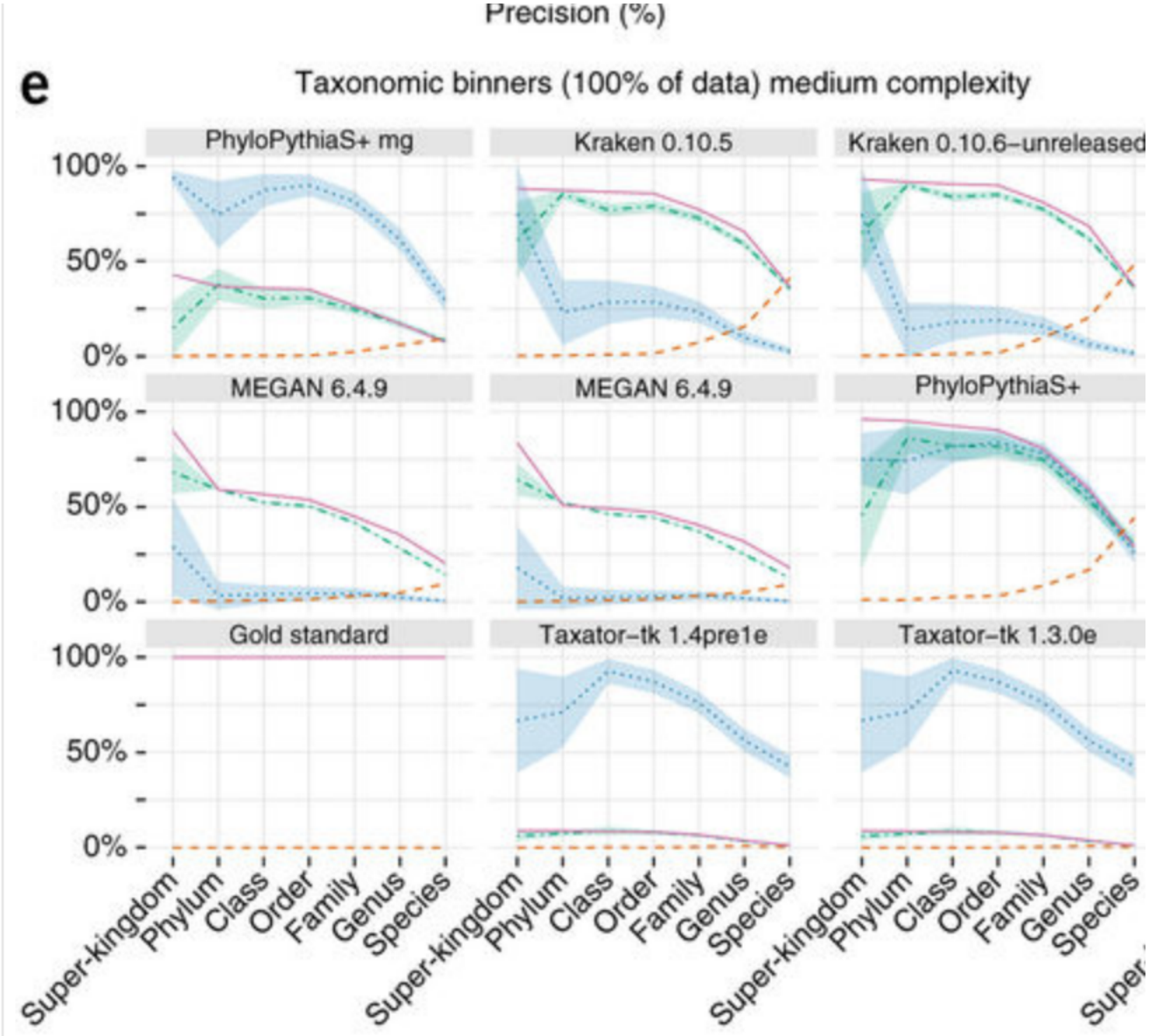
- The challenge with taxonomic labelling to strain/ species level of reads is false positives due to shared genes
- Classic example plague on the New York subway! (Afshinekoo et al. Cell 2015):

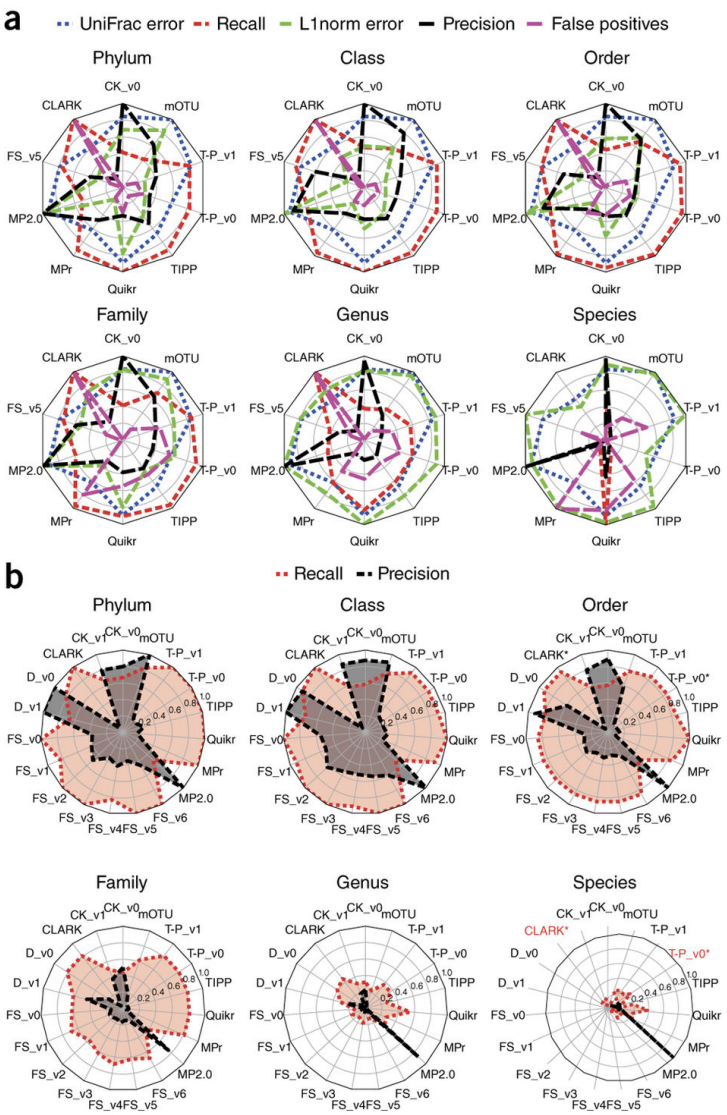
"A smaller proportion (12%) of the detected taxa with species-level identification were known pathogens, including Yersinia pestis (Bubonic plague) and Bacillus anthracis (anthrax)."

- Map reads onto genomes and look for consistent coverage – integrated into SPARSE (unpublished Zhemin Zhou)



Loman et al. JAMA 2013 “A Culture-Independent Sequence-Based Metagenomics Approach to the Investigation of an Outbreak of Shiga-Toxigenic Escherichia coli O104:H4”


Precision (%)
e
Taxonomic binners (100% of data) medium complexity
PhyloPythiaS+ mg
Kraken 0.10.5
Kraken 0.10.6–unreleased
MEGAN 6.4.9
MEGAN 6.4.9
PhyloPythiaS+
Gold standard
Taxator–tk 1.4pre1e
Taxator–tk 1.3.0e
100%
50%
0%
Super-kingdom
Phylum
Class
Order
Family
Genus
Species



c

	Best method (score)	2nd best method (score)	3rd best method (score)
Recall	Quikr (35)	CLARK (43)	TIPP (46)
Precision	MetaPhlAn 2.0 (16)	Common Kmers v0 (25)	mOTU (41)
L1 Norm	MetaPhyler (28)	FOCUS v5 (45)	TIPP (66)
UniFrac	MetaPhyler (4)	Taxy-Pro v0 (4)	CLARK (5)

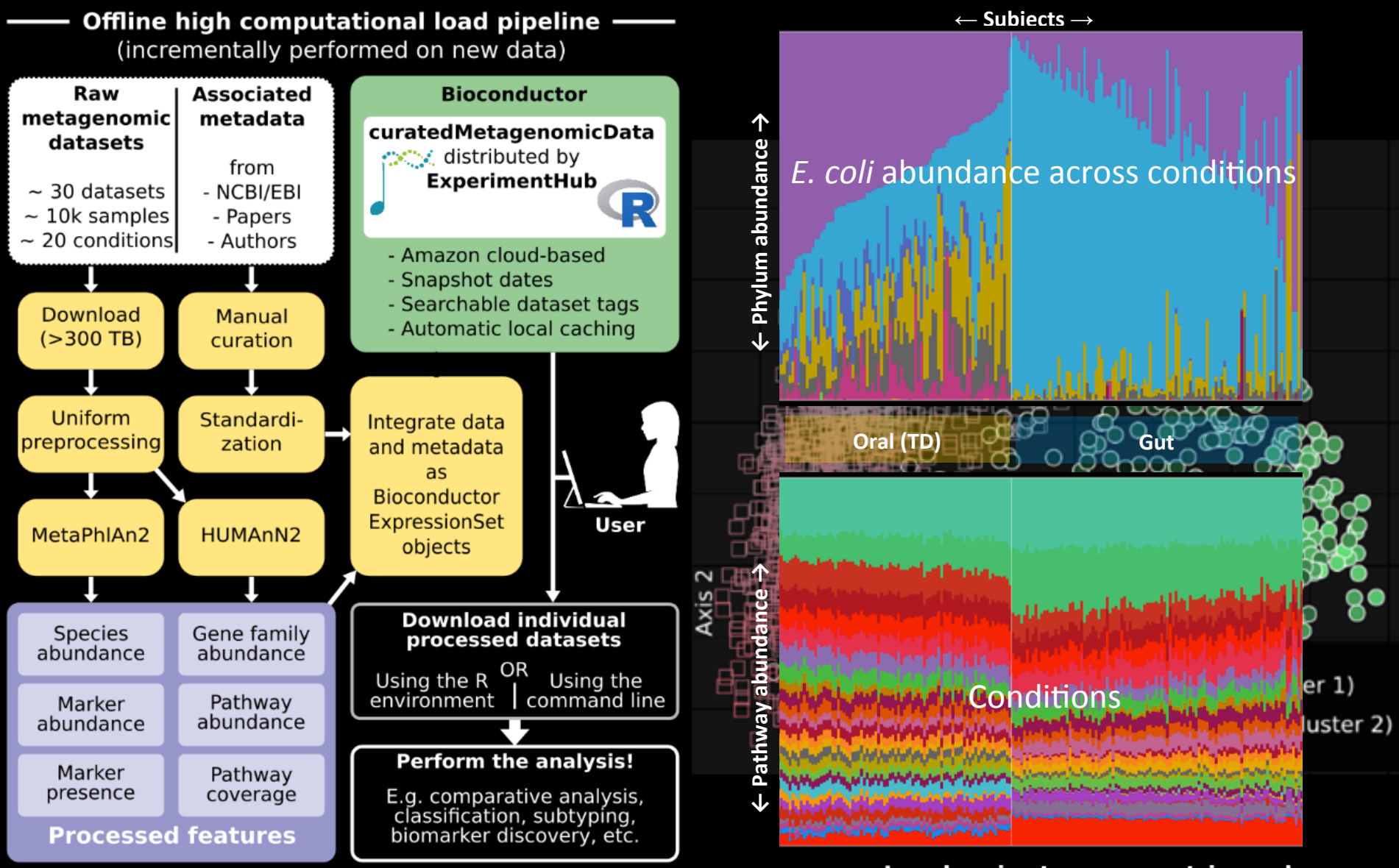
Functional profiling

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

How to find a COG?

- We use RPS-BLAST against NCBI COG database
- RPS Blast: Reversed Position Specific Blast
  - searches a query sequence against a database of profiles
- What is a BLAST profile?
- PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool) derives a position-specific scoring matrix (PSSM) or profile from the multiple sequence alignment of sequences
- The PSSM captures the conservation pattern in alignment and stores it as a matrix of scores for each position in the alignment-highly conserved positions receive high scores and weakly conserved positions receive scores near zero
- Capable of detecting distant sequence similarities

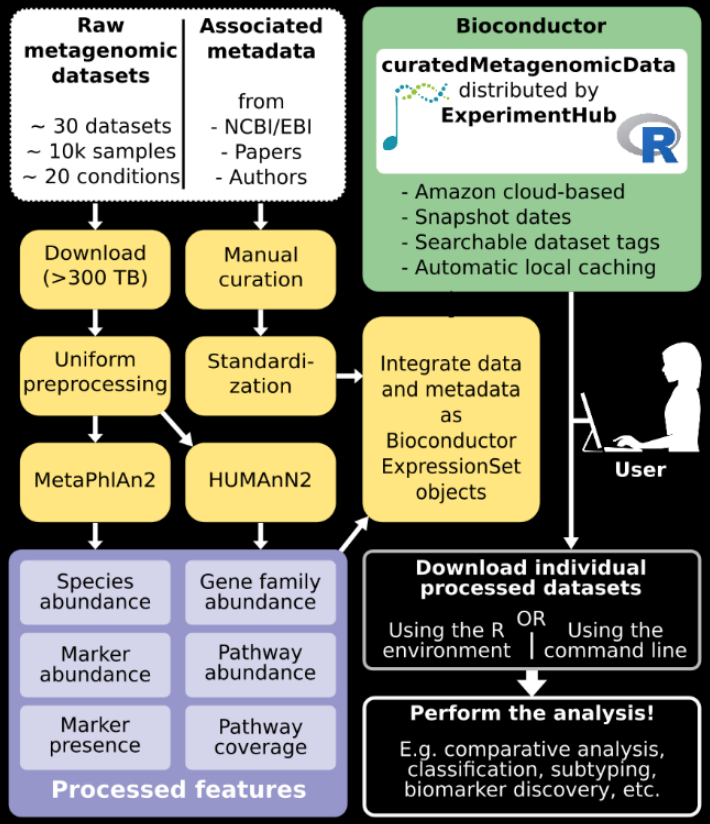
The curatedMetagenomicData resource



Pasolli et al, *Nature Methods, in press*

https://waldronlab.github.io/curatedMetagenomicData/

An integrative resource: curatedMetagenomicData



Edoardo Pasolli

Levi Waldron

- Mandatory metadata fields:

  sampleID, subjectID, body_site, antibiotics_current_use, study_condition, disease, age_category, gender, country, non_westernized, sequencing_platform, DNA_extraction_kit, PMID, number_reads, number_bases, minimum_read_length, median_read_length, NCBI_accession

- Optional metadata fields: all the available ones!

Preprint: http://biorxiv.org/content/early/2017/01/27/103085
Main page: https://waldronlab.github.io/curatedMetagenomicData/
Repository: https://github.com/waldronlab/curatedMetagenomicData
incl. data, scripts, tutorials, examples

- Currently 6,000 samples
- Will continue adding datasets
  - 7,000 new samples prioritized for addition
  - 3,000 additional samples without minimal info
- Will add new analyses

Pasolli et al, *Nature Methods, in press*

Summary

- Profiling effective way to summarise community structure and compare communities
- Does not provide high resolution picture linking organism to function (requires assembly…)
- Always be cautious of results:
  - Presence of gene does not imply expression
  - Correlation does not imply causation