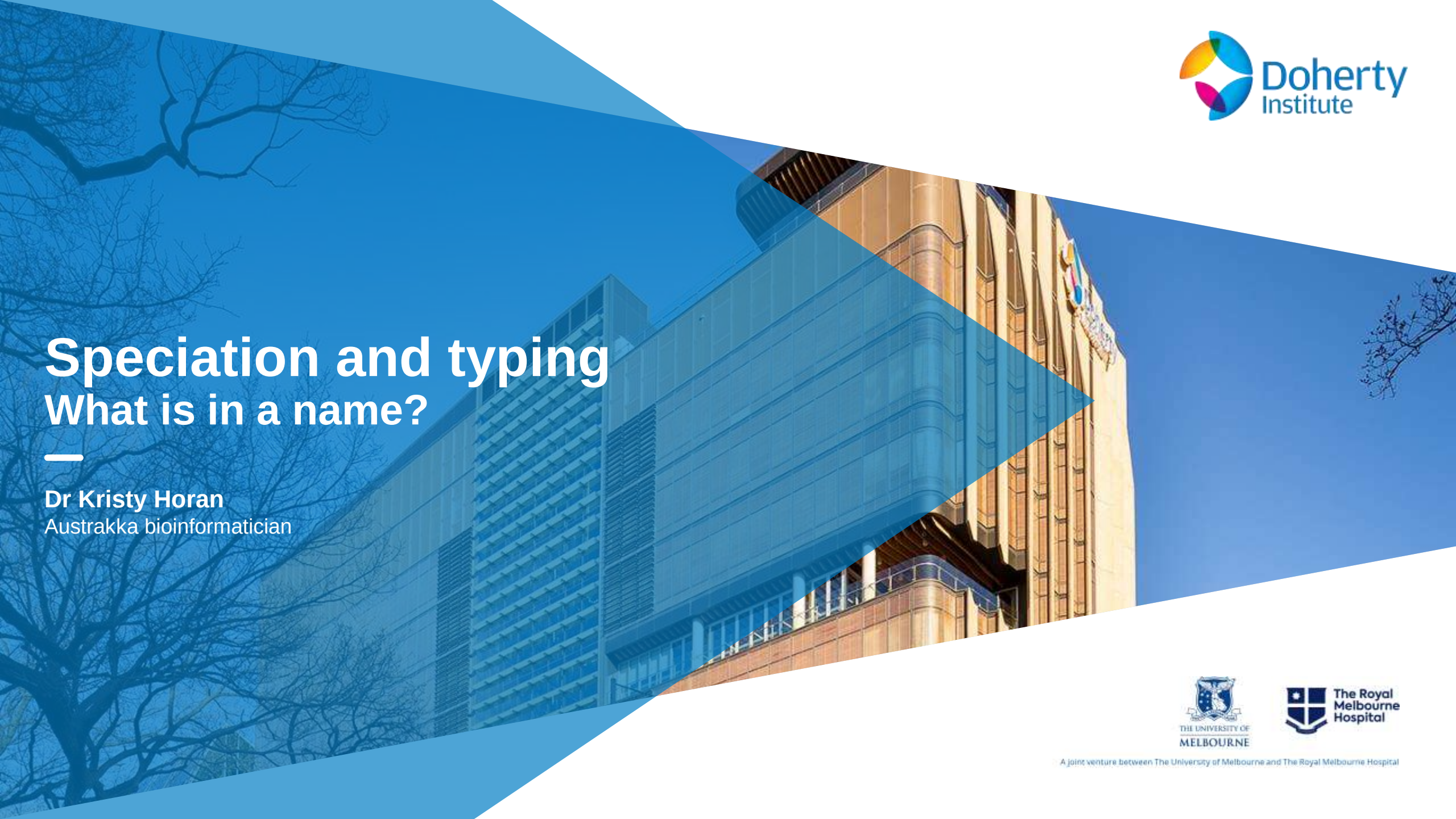
Speciation and typing

What is in a name?

Dr Kristy Horan
Austrakka bioinformatician



Doherty Institute

THE UNIVERSITY OF MELBOURNE

The Royal Melbourne Hospital

A joint venture between The University of Melbourne and The Royal Melbourne Hospital

Learning objectives

Understand sequence level classification approaches for speciation and subtyping

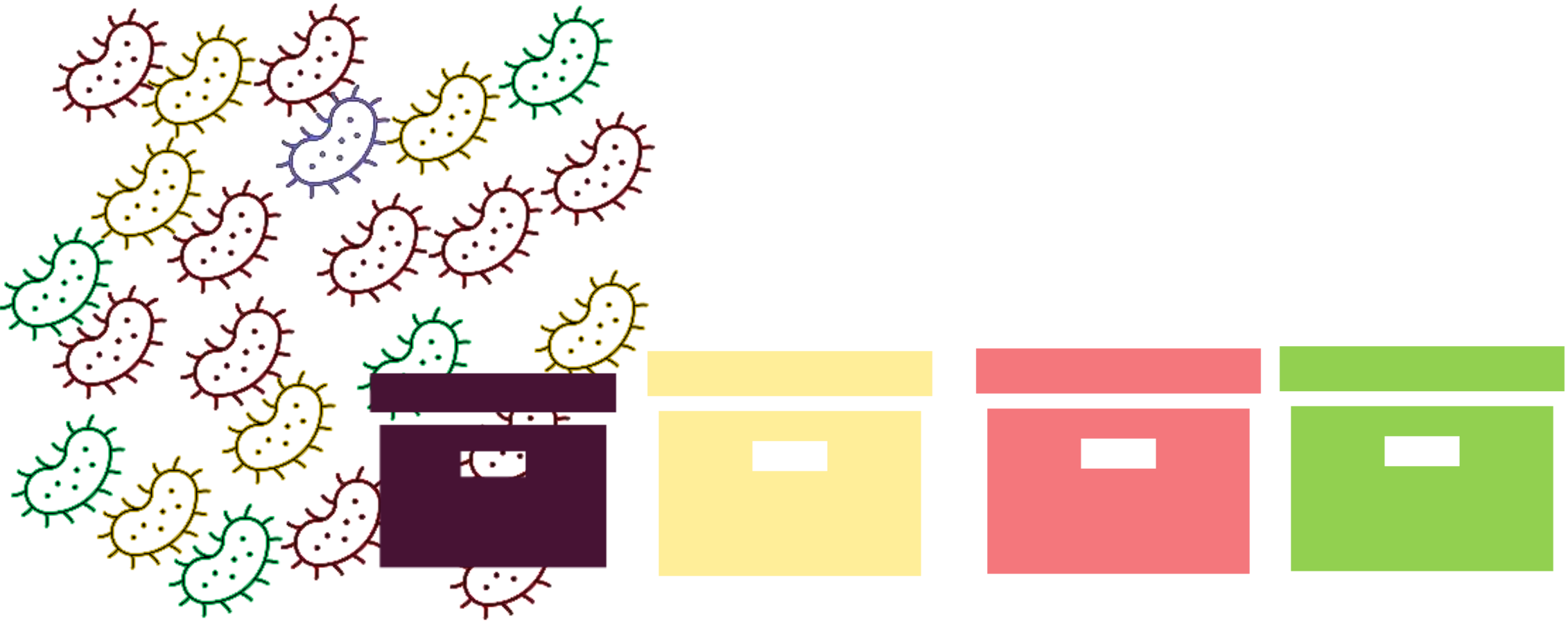
When it is appropriate to applying each of these approaches

Understand interpretation considerations

What is in a name?

Genome is just a genome

_



Genome is just a genome

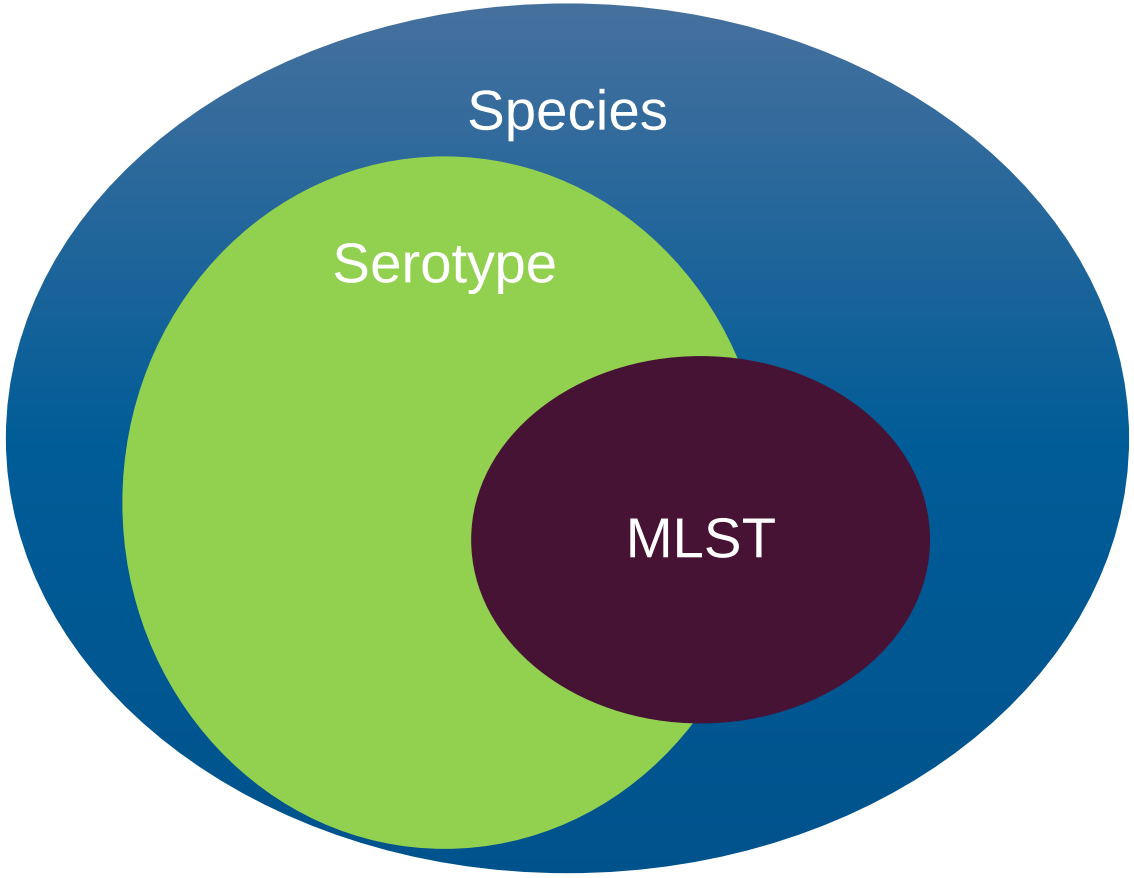
A genome is just a genome

- WGS does allow for more high-resolution genomic data than more ‘traditional’ methods
- Simply very long strings of letters (A,T,C,G)
- Without a mechanism for interpretation or classification it is not useful

Interpretation

1. Sequence level
   - What is it?
   - Is it what we expect?
   - What important information is present?
2. Population level
   - Where does it fit?
   - Are there relationships to things we have seen before?

Approaches to classification of single sequences



Classification

- Recapitulation of ‘traditional’ wet-lab based techniques.
- Different levels of resolution
- Coarse - species level
- Mostly **but not always** as you increase in granularity each group is a subset of the last.
  - Example - MLST subset of serotype which is a subset of species
  - Example - Cases were an MLST will span multiple serotypes
- The genomic mechanisms and approaches that are used to in each classification method are NOT a subset of each other.

Speciation

Taxonomic classification

Taxonomic classification – AKA speciation

Taxonomic classification – AKA speciation

Why?

- MALDI is relatively cheap and very quick and pretty good overall – why bother with speciation from genomic data??

Quality control

- Contamination
- Have we sequenced the correct thing??
  - Mistakes happen in the lab – ensure that the results you are generating are **suitable** for reporting to stakeholders

Unknown

- What happens if ‘traditional’ methods are unable to reliably provide a classification?

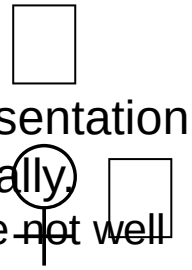
Improve bioinformatics analysis

- Not all tools or methods are suitable for all species
  - May get unexpected behavior or result when using a tool designed for a specific species on the wrong species.
- Can improve the quality of comparative analysis methods

Taxonomic classification – AKA speciation

This is hard

- What actually defines a species??
- Historically phenotypic or clinical presentation that may not be easily identified genomically.
  - Mechanisms which determine these are ~~not well~~ understood.
  - Increased resolution available through WGS means we have needed to re-consider some ideas around what a species IS?
- Databases can be opinionated and dirty
  - Contamination of databases with incorrectly annotated sequences
  - Curator has opinion which leaks into the database annotation – not wrong, may just be not appropriate to your use case.



Example

- *E. coli* vs *Shigella* spp.
  - “They are the same thing!!”
  - There are clinical manifestations and phenotypic characteristics that mean that we can’t stop there…
  - Different species of of Shigella have evolved from both divergent AND convergent evolution
    - There are complex genomic mechanisms which distinguish these species
  - *Shigella* have lost some of their genome compared to *E. coli* and gained a plasmid with virulence genes.
    - Complex genomic features which can be hard to distinguish.

Taxonomic classification - tools

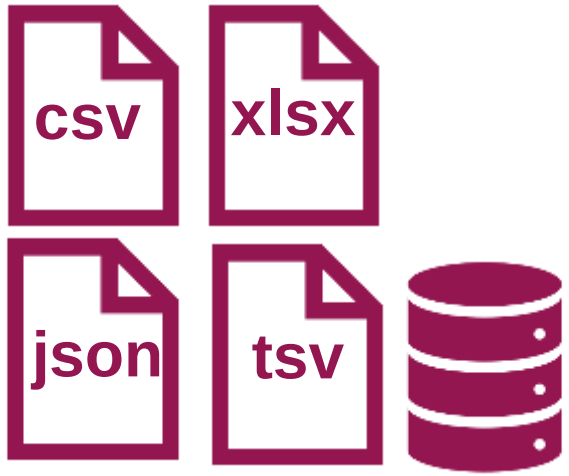
_

- **Kmer Identification (kraken2)**
  - Can take fastq and fasta as input
  - Quick comparison
  - Highly accurate for well studied organisms
    - Databases have improved dramatically
- **Relatedness**
  - Not so quick
    - Average nucleotide identity (ANI)
    - Genetic distance
  - New approaches using sketching improves performance – making these approaches more feasible
    - skani - %ANI
    - Mash – can provide distances and context with tree visualization

A word on “databases”

Knowledge base

- **Almost all bioinformatics typing strategies will rely on some sort of knowledge base**
  - csv/tsv
  - Actual database (sql, postgres)
  - Combination
- **Your choice of knowledgebase WILL IMPACT YOUR RESULTS**
  - Different collections of data may have been developed for different purposes
    - Researcher/developer opinions
    - Different focus
      - Whole genes vs variants
      - Bacteria vs fungi
  - Maintenance and curation
    - How is the collection curated?
    - Frequency?
    - Quality?



Back to Speciation

Kraken2

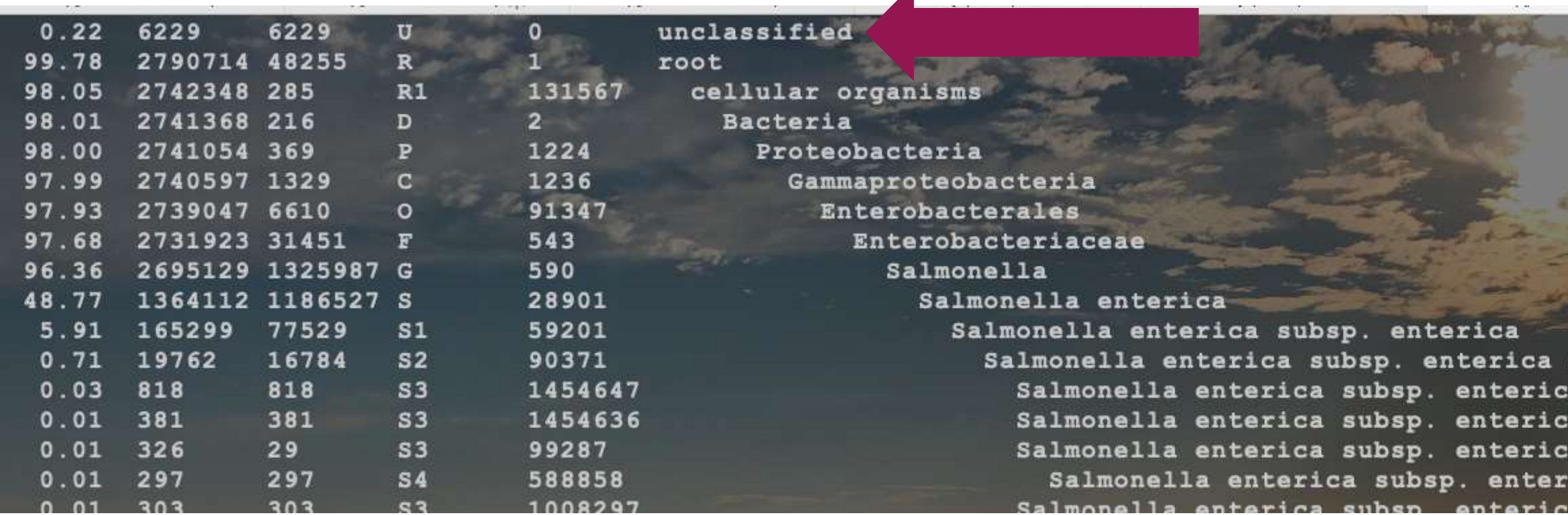
Kraken2 - databases

_

- **Kraken2 is the tool**
- **Kraken2 database format is specific**
- Can’t just use any collection of raw files
  - Many different databases are publicly available (https://github.com/BenLangmead/aws-indexes/tree/master/docs)
  - Build your own (https://github.com/DerrickWood/kraken2/wiki/Manual#build)
- **Not all kraken2 databases are equal**
- Different sized databases will have different representation
  - May result in different proportions or slightly different results
- Different sources (eg: RefSeq vs GTDB) can also result in differences in results due to some differences in how the taxonomy is annotated
  - RefSeq in general is based on submitter supplied information – although many of these are supported by NCBI curation it is not perfect
  - GTDB based on the GTDB-TK taxonomy (https://ecogenomics.github.io/GTDBTk/)

Kraken 2 – key points

–



0.22	6229	6229	U	0	unclassified
99.78	2790714	48255	R	1	root
98.05	2742348	285	R1	131567	cellular organisms
98.01	2741368	216	D	2	Bacteria
98.00	2741054	369	P	1224	Proteobacteria
97.99	2740597	1329	C	1236	Gammaproteobacteria
97.93	2739047	6610	O	91347	Enterobacterales
97.68	2731923	31451	F	543	Enterobacteriaceae
96.36	2695129	1325987	G	590	Salmonella
48.77	1364112	1186527	S	28901	Salmonella enterica
5.91	165299	77529	S1	59201	Salmonella enterica subsp. enterica
0.71	19762	16784	S2	90371	Salmonella enterica subsp. enterica
0.03	818	818	S3	1454647	Salmonella enterica subsp. enteric
0.01	381	381	S3	1454636	Salmonella enterica subsp. enteric
0.01	326	29	S3	99287	Salmonella enterica subsp. enteric
0.01	297	297	S4	588858	Salmonella enterica subsp. enter
0.01	303	303	S3	1008297	Salmonella enterica subsp. enteric

Kraken2 - unclassified

_

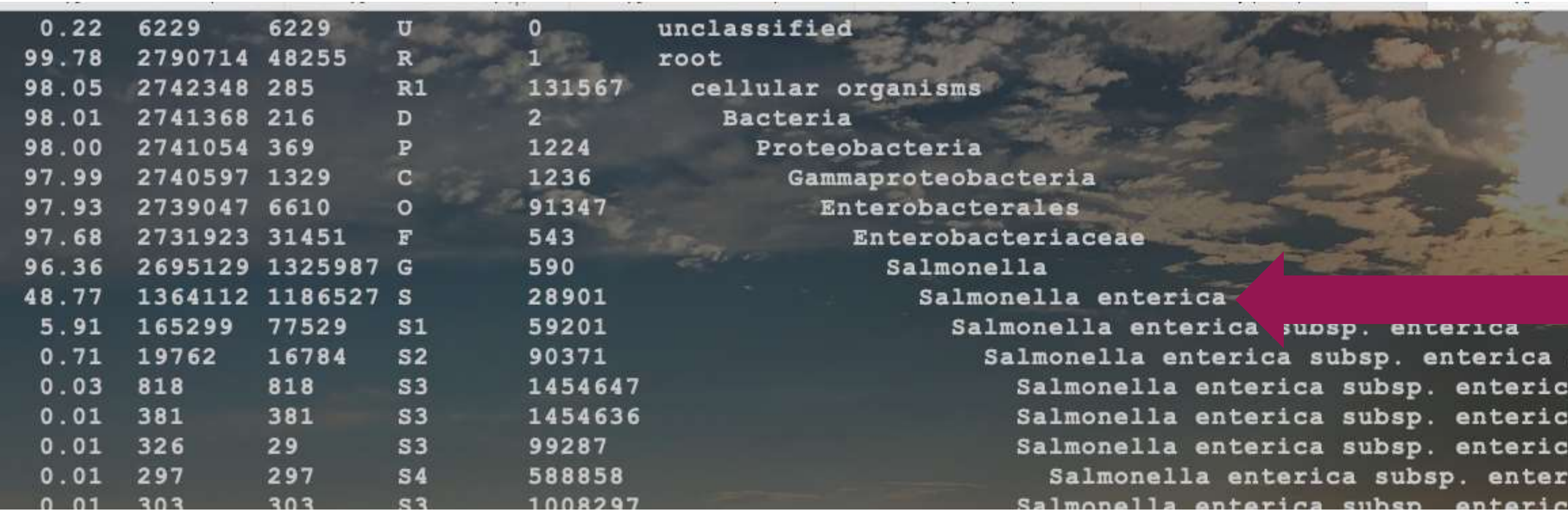
- **The kraken report is ordered in a tree structure by the most common domain first, then below that kingdom, family, genus, species and sub-species.**
- **The first line in a kraken2 report is the proportion of unclassified reads**
- Unclassified is the the amount of data in your sequence which can not be classified using the database that you have provided
- **This number should be very low (< 10%)**
- **High proportion of unclassified sequence (> 20 – 50%)**
- Indicates that the species present in the sample that was sequenced is

  NOT WELL REPRESENTED IN THE DATABASE

  - The database is inappropriate or incomplete
  - There is contamination of your sequence with something that is not in your DB ie human/fungi/virus
  - Novel strain or species (NOT common – other more mundane causes should be thoroughly investigated)
- The species represented in the sequence can not be identified using this combination of tool and database

Taxonomic classification – key points

–



0.22	6229	6229	U	0	unclassified
99.78	2790714	48255	R	1	root
98.05	2742348	285	R1	131567	cellular organisms
98.01	2741368	216	D	2	Bacteria
98.00	2741054	369	P	1224	Proteobacteria
97.99	2740597	1329	C	1236	Gammaproteobacteria
97.93	2739047	6610	O	91347	Enterobacterales
97.68	2731923	31451	F	543	Enterobacteriaceae
96.36	2695129	1325987	G	590	Salmonella
48.77	1364112	1186527	S	28901	Salmonella enterica
5.91	165299	77529	S1	59201	Salmonella enterica subsp. enterica
0.71	19762	16784	S2	90371	Salmonella enterica subsp. enterica
0.03	818	818	S3	1454647	Salmonella enterica subsp. enteric
0.01	381	381	S3	1454636	Salmonella enterica subsp. enteric
0.01	326	29	S3	99287	Salmonella enterica subsp. enteric
0.01	297	297	S4	588858	Salmonella enterica subsp. enter
0.01	303	303	S3	1008297	Salmonella enterica subsp. enteric

Taxonomic classification – Top species –

- **Depending on the species (and database) you will see some results that may be confusing.**
- 96% sequence data has been classified to the Salmonella genus level only 48% of that is classified to the species Salmonella enterica. What happened to the rest of the sequence?
  - There are lots of representatives of Salmonella enterica in the database – can lead to a ‘dilution’ effect.
  - There are not many representatives – not enough information to classify a large proportion of sequence
- In other cases where there are closely related species – eg *N. meningiditis* vs *N. gonorrhoeae* which often share parts of their genome (recombination) can also cause

Serotyping and MLST

In silico serotyping

- **In most cases this is a gene detection exercise**
- **Many serotyping tests based on a phenotypic or biochemical observation that has been deemed to have microbiological relevance**
- The mechanism leading to the phenotypic or biochemical result may be complex involving many mechanisms which needs to be well understood.
- The relationship between genes and their ‘meaning’ in the context of serotype are often capture in a database or knowledgebase.
- **Well established mechanisms may still pose a challenge**
- Poor sequence quality
- Sequence bias
- Missing genes

In silico serotyping - Salmonella

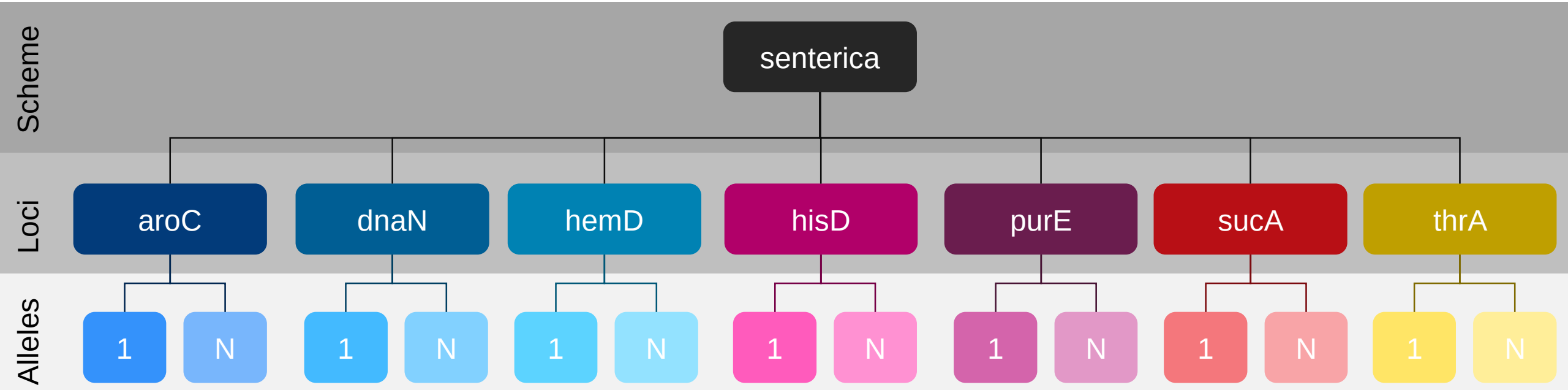
- **Two commonly used tools are sistr (https://github.com/phac-nml/sistr_cmd) and seqsero2 (https://github.com/denglab/SeqSero2)**
- **sistr uses a multi-pronged approach to return a serotype**
- Takes fasta as input type
- Detection of ‘traditional’ antigenic combinations
- Distance based approaches (mash and cgMLST) to representative reference sequences
- Combination of these tests results in a serotype

In silico serotyping - Others

- **Serotyping can be difficult**
- Species specific
- Genomic mechanisms are unclear
- Regions of the genome can be hard to sequence – resulting in gene drop outs
- Biological reasons
- **Neisseria spp.**
- Well defined mechanism but gene transfer is prolific and genes are hard to sequence
- **Shigella spp.**
- Serotype vs biotype
- Evolution of Shigella – means that not all ‘observations’ may have the same mechanism
  - Mechanisms not completely understood

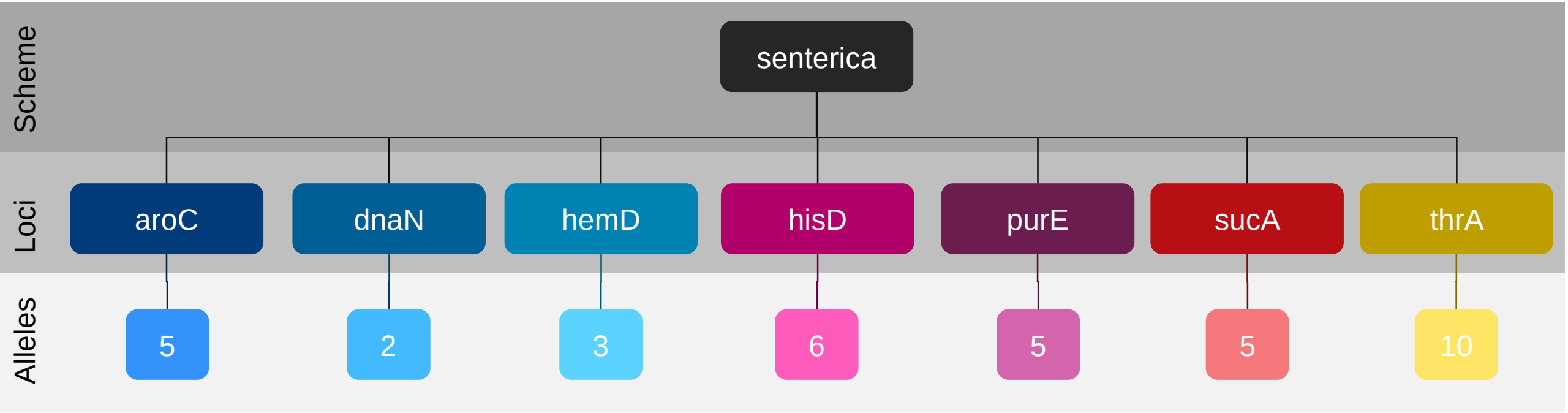
MLST

- **Multi-locus serotyping**
- A method which relies on 7 – 10 species specific housekeeping genes
- Genes are defined on a per ‘species’ level
  - Can be genus level or sub-species level



MLST

aroC(5);dnaN(2);hemD(3);hisD(6);purE(5);sucA(5);thrA(10) ➔ ST10



MLST

- **Schemes can be developed for any genus, species or sub-species**
- Often multiple schemes for the same genus OR species
  - koxytoca, klebsiella, kpneumonia
  - ecoli_acthmann and ecoli_Pasteur
- **Public schemes are maintained and curated**
- https://pubmlst.org/
- https://bigsdb.pasteur.fr/
- New alleles can be added to schemes and profiles designated by domain experts
- **Tools**
- pubMLST has a graphical user interface for upload of assemblies (https://pubmlst.org/organisms)
- mlst is an open source CLI tool https://github.com/tseemann/mlst
  - Comes packaged with the pubMLST database (2022)

Tying it all together

Interpretations typing

_

- **Often species is not enough to determine how concerned we should be about a particular pathogen – not all Salmonella are the same**
- **Recapitulation of traditional wet-lab typing can be a useful first step in characterizing a genome.**
- Although there are many layers to further characterize genomes.
- **Often using a combination of species, serotype and MLST can be extremely useful for triaging the urgency/importance/relevance of a finding**
- *Salmonella enterica* species
- Serotype Typhimurium
- MLST 313

None != absent

The absence of detection does not mean the absence of a gene

- **Genes can be missing for many reasons**
- Poor sequence quality
- Low coverage due to AT bias
- Method of assembly
- Method of detection
- Quality of the database you are using to compare the sequence to
- **Troubleshooting**
- Alternative assembly method
- Different database
- Resequence

Summary

- **Speciation and typing of genomes can be a great first step towards characterisation of a genome**
- **The choice of database is extremely important in the quality of the result observed**
- **Serotyping is a species (or subspecies) specific result**
- Incorrect application of a serotyping test may result in unexpected and uninterpretable results.
- **MLST is a specific agnostic tool, using species specific schemes**
- MLST schemes are generated on a per species or genus level
- Public schemes are maintained, curated and updated at pubMLST
- **Combinations of typing results can be useful to triage sequences for further analysis**
- **Absence of detection does not mean the absence of a gene.**

That was a lot

Thoughts or questions?

Feel free to reach out

kristy.horan@unimelb.edu.au