BM329 Workshop A: Microbial Identification

Leighton Pritchard

2024-01-25

Table of contents

Preface to the 2023-24 BM329 Block A workshop

Welcome to the BM329 (Biomedical Microbiology) Block A Microbial Identification workshop for 2023-24.

This year is the first presentation of this workshop material, and we would be very grateful to hear feedback by email or through the GitHub repository Issues page.

Overview

The workshop asks you to carry out common microbial identification analyses for a sequenced bacterial isolate, to determine its likely identity. You will be using online bioinformatics services to do this.

> ⚠ Important Note
>
> There is new material in this workshop that is not covered in lectures, and this material is examinable. Please take care to read the text in the expandable callout boxes, as well as that for the workshop, to be sure you have understood the topic and obtain full value from the exercise.
> The workshop material will remain live online for the duration of BM329, and you can revisit it whenever you wish, for practice or revision.

You should be able to complete this workshop in under two hours.

Learning Objectives

By the end of this workshop, students will be able to:

- Obtain an identification for bacterial isolate from its 16S sequence, using public bioinformatics 16S sequence-matching services
- Obtain an identification for a bacterial isolate from its assembled draft genome sequence, using a range of techniques, *via* public bioinformatics services

- Interpret sequence-based bacterial taxonomy assignments in the context of LPSN (the List of Prokaryotic names with Standing in Nomenclature), the arbiter of bacterial nomenclature
- Compare, and make an informed judgement between, alternative taxonomic assignments for the same isolate made using different computational approaches and databases.

Assessment

This workshop activity itself is not formally assessed although, as noted, all the material it contains is examinable. There is a formative assessment in the form of short answer questions on MyPlace, which you should complete as part of the workshop.

1 Introduction

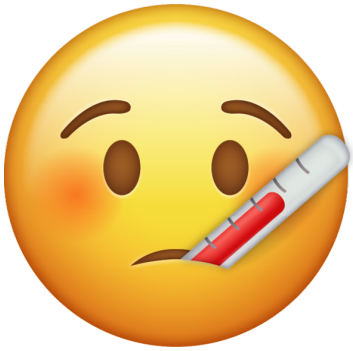
1.1 Say cheese!

Figure 1.1: Ballard Farmer's Market (Wikimedia Commons)

A few weeks ago you went to a food festival and farmer's market, and passed a cheese stall. You love cheese and the smell was making you hungry, so you tried a few samples, and bought the one you found tastiest. It was a delicious blue cheese, made from unpasteurised cow's milk, and with natural rennet.

You ate it on sourdough crackers, with some fig and plum chutney, and it was a savoury taste explosion. You thought to yourself what a great idea it was to go to the farmer's market. And it was even a local producer, so you kept the food miles down and supported the local community. You felt *great*!

1.2 A month later



Now it's about a month later. But you're not feeling great at all. You haven't felt great for a couple of days in fact, and you've got quite a list of signs and symptoms.

- you've had a fever for day or so
- you're sweating, even though it's cool
- you've got a nasty headache you can't shift, even with paracetamol
- you're not hungry and haven't felt like eating at all, since this started
- you're exhausted all the time
- all your joints ache, your muscles feel like you've wrestled an elephant seal, and don't get you started about your back

But you *were* out at the weekend, and caned it a bit. Maybe it's just a really bad hangover?

It's all very nonspecific. It's not 'flu season but even though there's no 'flu going round it feels like the worst 'flu of your life and worse than any hangover you've had so far. So you called 111 and off you popped to the pharmacist like they said, who immmediately suggested you go to your GP, who referred you without delay to Accident and Emergency with suspected bacteremia.

1.3 In the blood

You're in a hospital room now, full of antibiotics and feeling terrible. Your bloods have been taken - so many bloods! - and you've been told that the phlebotomists have found an unexpected bacterium. They're trying to grow it up in agar right now. As you're a microbiologist you're naturally curious, and you ask if they've identified it yet.

It turns out they've been having difficulty finding a medium that the bacteria will grow on. They're trying blood and tryptose broth and, while that's meeting with some success, it's very, very slow-growing.

Figure 1.2: A hospital room (Wikimedia Commons)

They've managed to get a preliminary identification from a small sample by MALDI-TOF, (Barth et al. (2023)) which suggests that it's a *Brucella* strain of some sort.

Your heart sinks. You really, *really* do not want brucellosis.

! Why you don't want brucellosis

Although uncommon the UK, brucellosis is endemic in many developing countries (Laine et al. (2023)). The causative agents, *Brucella* spp. bacteria, grow less quickly in the open environment than in their hosts, and are highly contagious intracellular pathogens readily spread between individuals and by infectious body fluids (Moreno et al. (2023)).

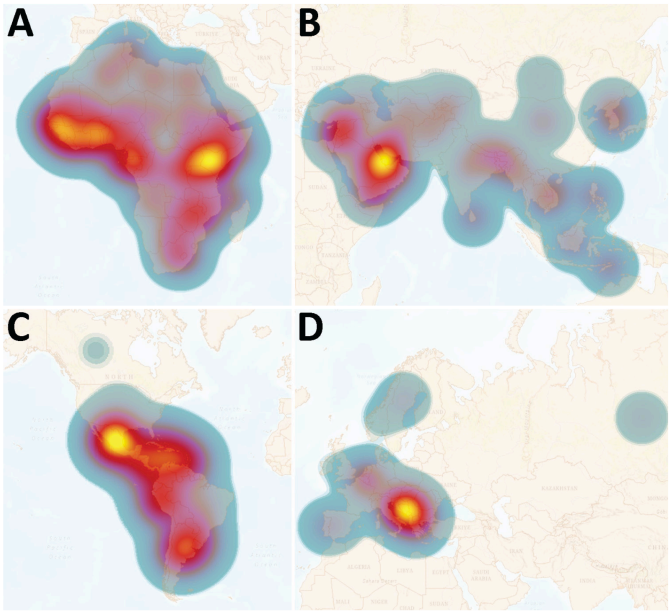


Figure 5. Heatmaps of regional annual incidence of human brucellosis estimated per 1 million population at risk. Each region has a different scale for incidence per 1 million population at risk. Heatmaps are intended to represent transnational zones that require priority control or surveillance initiative, not to represent the risk of individual countries. The heat scale shows high risk to low risk; yellow to blue. A) Africa: average risk is ≈750 new cases per million; high is >3,000. B) Asia: average risk is ≈500 new cases per million; high is ≥4,000. C) Americas: average risk is ≈20 new cases per million; high is ≥75. D) Europe: average risk is ≈10 new cases per million; high ≥100.

Figure 1.3: Figure 5 from (Laine et al. (2023)) illustrating estimated incidence of human brucellosis, worldwide.

In humans brucellosis is not ofen fatal, but it can be severely disabling. Brucellosis presents in both infectious and non-infectious forms and can be highly persistent, causing chronic debilitation and granulomas within organs. The most common presentations involve greatly enlarged liver and spleen, but other complications - including neurological - may be seen (Franco et al. (2007)). The presentation can be extremely variable, to the extent that "in endemic areas, everything can be due to brucellosis until proven different" (Bosilkovski, Keramat, and Arapović (2021)).

Blood and bone marrow culture remains the gold standard for diagnosis of brucellosis, but *Brucella* spp. grow slowly, and it's not unusual for a culture to take a week or more to reach a size suitable for diagnosis. ELISA is popular but less specific than than many

other diagnostic tests. Molecular methods such as PCR and MALDI-TOF have not been as widely validated (Franco et al. (2007)).
Treatment of brucellosis is unpleasant for the patient, and has hardly changed this century. Treatment requires simultaneous administration of multiple antibiotics for an extended period (e.g. 200mg doxycycline plus 900mg rifampicin, and perhaps a fluoroquinoline in support, for 6-8 weeks). Relapse rates are high (up to 25%) with oral delivery alone but lower with parenteral delivery (up to about 8%) (Franco et al. (2007), Bosilkovski, Keramat, and Arapović (2021)).
Hospital staff potentially exposed to the infectious agent undergo the same treatment as prophylaxis.

At the lab they've managed to sequence the bacterium that's in your blood, using an Oxford Nanopore flongle they had lying around. The assembled draft genome has come back, but the NHS is chronically underfunded and the staff are massively overworked, so your consultant hasn't had time to learn how to classify bacteria from genome sequence data.

"I'll have a go, if you like," you say. Then the light starts to hurt and you have a nap.

1.4 Controversy

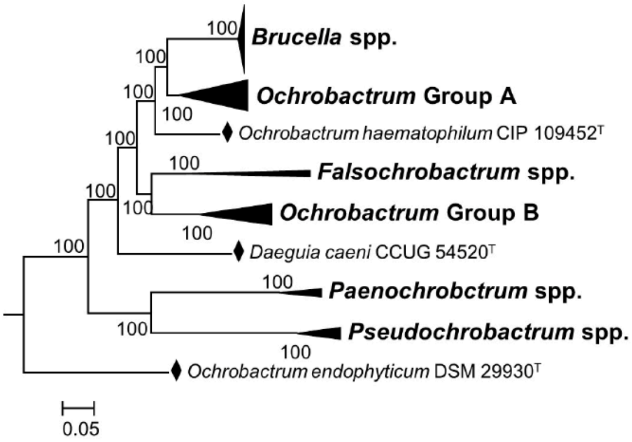


Figure 1. Phylogenetic relationships among genera of the family *Brucellaceae* based on whole sequence genome analysis (a black diamond denotes type strains). Node labels give percentage bootstrap support. *Rhizobium etli* (CFN42T) was used to root the phylogenetic tree (not shown). The tree was constructed using the maximum likelihood method, based on the general time-reversible model, as described by Ashford et al. [30] (adapted from Figure 4 of this reference).

Figure 1.4: Figure 1 from (Moreno et al. (2022)), showing the relationship between *Brucella* and *Ochrobactrum* species. *Brucella* is a group of bacteria nested within *Ochrobactrum*.

As an expert biomedical microbiologist, *you* know that there has recently been a lot of controversy around identification of *Brucella*, and that preliminary molecular and genomic identifications of *Brucella* spp. shouldn't necessarily be taken at face value.

You are aware that recent publications in microbial nomenclature renamed a whole genus of much less worrying bacteria, belonging to the *Ochrobactrum* genus, to *Brucella* (Hördt et al. (2020)), even though *Ochrobactrum* spp. are only opportunistic pathogens and not particularly virulent (Hagiya et al. (2013)). Many microbiologists who work in clinical settings, or with *Brucella*, object to the renaming (Moreno et al. (2022), Moreno et al. (2023)). Even so, some commercial diagnostic equipment has updated its databases and now calls bacteria *Brucella* that used to be called as *Ochrobactrum*.

! The *Brucella*/*Ochrobactrum* controversy

In 2020, a manuscript was published applying a bioinformatics methodology to over 1,000 genomes of *Alphaproteobacteria* type strains. *Alphaproteobacteria* is the taxonomic Class that contains both *Brucella* and *Ochrobactrum* genera (Hördt et al. (2020)).
In phylogenetic trees, *Brucella* and *Ochrobactrum* are closely-related. *Ochrobactrum* form a group called a clade (they share a single common ancestor), but *Brucella* appears as a coherent subgroup with multiple *derived characters* within this clade.
The bioinformatics method applied in Hördt et al. (2020) reproduced this relationship, and a proposal was made that, because *Brucella* was nested within *Ochrobactrum*, the entire group of bacteria - including all *Ochrobactrum* - should be named *Brucella*. It was argued in this paper that, even though *Brucella* and *Ochrobactrum* are not as virulent as each other, they should be considered to belong to different "risk groups," but also to be fundamentally similar organisms (Hördt et al. (2020)).
Clinical microbiologists and other scientists rejected this proposal, arguing that the names of bacteria do more than indicate where they lie in a phylogenetic tree. They argue that the long-standing use of brucellosis to refer to the serious disease caused only by *Brucella* ends up being confused - potentially dangerously - if relatively low-risk bacteria such as *Ochrobactrum* are renamed as *Brucella*. They also argue that "typology," taking into consideration more than phylogenetic placement, is necessary for meaningful naming and subdivision of bacteria, and that bioinformatics methods alone are not sufficient to define bacterial taxa. They consider that *Brucella* and *Ochrobactrum* strains show quite different phenotypes, and should therefore remain known by different names (Moreno et al. (2022), Moreno et al. (2023)).

Table 1. Comparison between the *Brucella* and *Ochrobactrum* genus.

Divergent Properties	*Brucella*	*Ochrobactrum*	References
Genome Size	3.1–3.4 Mb	4.7–8.3 Mb	[20,21]
Pangenome	Closed	Open	[20,22,23,24,25,26,27]
Plasmid	No	Variable (up to 6)	[20,21,23,25,28,29]
Phylogeny	Monophyletic	Polyphyletic	[8,30]
Active Phages	No	>4	[29,31]
Lateral gene transfer	Absent	Present	[29]
Speciation type	Allopatric	Sympatric	[32,33,34]
Cell envelope permeability	Permeable to hydrophobic probes and resistant to destabilization by polycationic peptides	Impermeable to hydrophobic probes and sensitive to polycationic peptides	[35,36]
Metabolic redundancy	Low	High	[22,37]
Degradation of complex molecules	No	A large variety of such molecules	[38,39,40,41]
Removing toxic metals	No	Yes (some species/strains)	[42,43]
Capable to root nodulation	No	Yes (some species/strains)	[44,45]
Life style	Pathogen (class 3)	Saprophyte	[6,44,46]
Natural habitat	Intracellular	Soil and root plant surfaces	[24,35,44]
Transmission	Host-host interaction/animal products	Mostly iatrogenic	[47,48]
Virulence	Finely tuned	Fortuitus/opportunistic	[6,46,47]
Virulence mechanisms	Escape from the immune response/deviation of the intracellular trafficking	No true ones and virulence depending on host immune status	[35,46,49,50]
Infection dynamics	Long-lasting infection and low proinflammatory response	Acute proinflammatory/pyogenic; self-limiting in immunocompetent hosts	[46,47,51]
Animal disease	Very important	Seldom	[46,48,52,53,54]
Human health	Very important	Negligible	[22,46,47,48,52]
Diagnosis	Well-standardized serological methods	No serological tests are available or necessary	[52,55]
Treatment	WHO recommended long bi-therapy in uncomplicated cases	Based on antibiotic resistance/short monotherapy	[47,56,57]
Antibiotic resistance	Seldom and well-defined	High	[25,47,56,57,58,59,60,61]
Vaccine	Available (domestic ruminants) and critically important to control disease	Unnecessary	[62,63]
WHO/OIE/FAO regulations	Very important	Null	[57]

Figure 1.5: Table 1 from (Moreno et al. (2022))

Since the nomenclature change, there have been instances of bacteria formerly known as *Ochrobactrum* being identified as *Brucella* by commercial MALDI-TOF instruments in a clinical context, leading to avoidable prophylactic treatment of hospital staff (Moreno et al. (2023)).
Now that the name *Brucella* has been validly published in the literature for bacteria formerly known as *Ochrobactrum*, and is considered the "correct" name for the group, it is highly unlikely it would be withdrawn, under the the rules of the International Code of Nomenclature of Prokaryotes (ICNP, the Prokaryotic Code). The controversy continues.

You know that if you make a positive identification of this bacterium as something other than *Brucella*, you might avoid eight weeks of unpleasant multiple antibiotic treatment, and so can staff who were potentially exposed to *Brucella* through your sample. If this infection does turn out to be *Brucella* they would also need to undergo prophylactic treatment.

Your identification could save a lot of people - not least you! - quite a bit of unpleasantness.

1.5 Your challenge

Identify the infectious bacterium using bioinformatics methods

You will be provided with the assembled draft genome sequence of the organism that was isolated from your blood and sequenced on the flongle and, for convenience, some additional information (e.g. 16S sequence data) that may help with identification.
In the sections that follow, you will be guided through identification of the organism using a number of different bioinformatics approaches and online resources. Once you have conducted these analyses and identified your organism, you should complete the formative questions on MyPlace. These will ask you about your identification, and your reasons for making that identification.

Let's get started

Begin your analysis by clicking on the link to `Data` (here, or below)

2 Data

2.1 Your datasets

The hospital laboratory have provided you with the following data from the bacterium that was isolated from your blood:

- Draft genome sequence: isolate_genome.fasta
- 16S sequence from the draft genome: isolate_16S.fasta

Downloading data files

You can download your data files using the links above. Clicking on the link may open the file in your browser. If this is the case, then you can use the `File -> Save As` menu option to save the file.
Alternatively, right-click (or `Ctrl`-click) on the link, and choose `Save file as…` (or similar) to save the file.
These data files are also available from the BM329 MyPlace page.

FASTA file format

The sequences you have been given are in FASTA format. This is a very common standard format for representing biological sequences, and looks like this:

```
>R431BS_isolate_from_bloods 16S sequence obtained from full genome
CAACTTGAGAGTTTGATCCTGGCTCAGAACGAACGCTGGCGGCAGGCTTAACACATGCAAGTCGAGCGCC
CCGCAAGGGGAGCGGCAGACGGGTGAGTAACGCGTGGGAATCTACCTTTTGCTACGGAATAACTCAGGGA
AACTTGTGCTAATACCGTATGTGCCCTTCGGGGGAAAGATTTATCGGCAAAGGATGAGCCCGCGTTGGAT
TAGCTAGTTGGTGAGGTAAAGGCTCACCAAGGCGACGATCCATAGCTGGTCTGAGAGGATGATCAGCCAC
ACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGCAAG
```

The general format is

```
>sequence_identifier sequence_description
[symbols representing the biological sequence]
>sequence_identifier sequence_description
[symbols representing the biological sequence]
[...]
```

Each new sequence in the file starts with a right angled bracket (`>`) to indicate a *header line*, which is immediately followed by a unique sequence identifier. This is typically, but not always, an accession number that uniquely identifies the sequence in a database.
If there are any space characters on the *header line*, all text after the space is taken to be a free-text description of the sequence itself.
The lines following the *header line* then contain the symbols (`A`, `C`, `G`, `T` for DNA, the protein alphabet for protein) that represent the biological sequence itself.

Tip

FASTA files are a *plain text* format. You can open them in an editor like `notepad` (Windows), `emacs` (Linux) or `TextEdit` (macOS) to see their contents.

Warning

Avoid opening FASTA files in Word or other word-processing software. This software can insert hidden characters which corrupt the data and make it unusable for analysis. If you save your sequence as a Word file it will be unreadable by bioinformatics programs.

2.2 Draft genome sequence

Your isolate's genome is not entirely complete, and is not in one single contiguous piece - therefore it is a *draft* genome. Instead, it is in 41 *contigs* (contiguous sequences) - sections of the genome. The contigs are not in order which means that, on the real genome, the second contig may not follow the first (and so on), and the first contig may not be where the genome "starts".

The contigs are described in the `isolate_genome.fasta` file: one sequence per contig. Taken together, the total sequenced genome length is 5,116,355 bases, and the genome has a GC% content of 53.5%.

2.3 16S sequence

16S ribosomal RNA (16S rRNA) is an evolutionarily highly-optimised, essential component of bacterial SSU (small subunit) ribosomes. Because it is essential, and its function in protein synthesis is central to correct operation of the cell, its biological sequence is highly constrained (it is unlikely that a change in the sequence will enhance or be neutral to function). This has two main implications that make it useful for evolutionary analysis and microbial identification.

- The 16S sequence is similar enough to be recognisable in all prokaryotes
- The 16S sequence is usually most similar in organisms that share a recent common ancestor, and less similar in organisms that share a more distant common ancestor.

These properties enable the 16S rRNA sequence to be used to reconstruct evolutionary histories of prokaryotes, including the landmark paper that defined Archaea as a new domain of life (Woese, Kandler, and Wheelis (1990)). Several online reference databases with curated 16S sequence data are commonly used to enable bioinformatic identification of organisms.

Note

The properties noted above also make 16S rRNA useful for identifying which bacteria are present in complex communities: microbiome analyses (Johnson et al. (2019))).
The similarity of 16S sequences across all bacteria mean that a single set of primers (*universal primers*) can be used to amplify most 16S bacterial genes in a sample using PCR.
Sequencing the amplified 16S genes with high-throughput sequencing then allows the distinct identities of many bacteria in a sample to be determined simultaneously.

Disadvantages of 16S sequence data for identification and classification

16S sequence data has more recently fallen out of favour for bacterial classification and identification. This is in part due to the increased availability of high-throughput whole-genome sequencing; with this technology the whole genome can be obtained at the same time as the 16S gene sequence, and the extra information allows for more precise identification. But there are also inherent disadvantages of the approach.

- Whole-genome sequencing is practical, inexpensive, and provides more information, including the 16S sequence.
  - Amplifying a variable subregion of 16S provides about 300bp of information, sequencing the full 16S gene about 1500bp, and sequencing a genome between 1,500,000bp and 12,000,000bp (depending on organism)
- Some distinct species share identical 16S sequences (R. C. Edgar (2018a))

- Some organisms/species possess multiple distinct 16S sequences (R. C. Edgar (2018a))
- There is no universal level of similarity (percentage identity) between 16S sequences that always corresponds to a taxonomic division (R. C. Edgar (2018b))
- 16S databases contain many sequences that are incomplete, contain errors, or are annotated with the wrong taxonomic identity (R. Edgar (2018))

The 16S sequence for your isolate has been identified from the genome sequence and is described in the `isolate_16S.fasta` FASTA file.

Begin your identification

You should begin the identification of your isolate by submitting the 16S sequence in `isolate_16S.fasta` to some of these public resources. click on the link to `16S` (here, or below), to get started.

3 16S

3.1 16S identification of your isolate

In this part of the workshop, you will use publicly available 16S sequence databases and resources to obtain an identification for your isolate.

> **!** Important
>
> Please ensure that you have downloaded the 16S sequence file for your isolate (`isolate_16S.fasta`) to a suitable location on your computer.

3.2 NCBI Blast

BLAST is a bioinformatics tool developed by NCBI (the National Center for Biotechnology Information) that takes as input a *query* sequence and searches for similar sequences in a reference database. BLAST ranks matches from most (first) to least similar. The tool can be downloaded and run on your own machine, or it can be used through the NCBI-BLAST webservice.

- NCBI-BLAST webservice
- NCBI-BLAST help pages
- How BLAST works

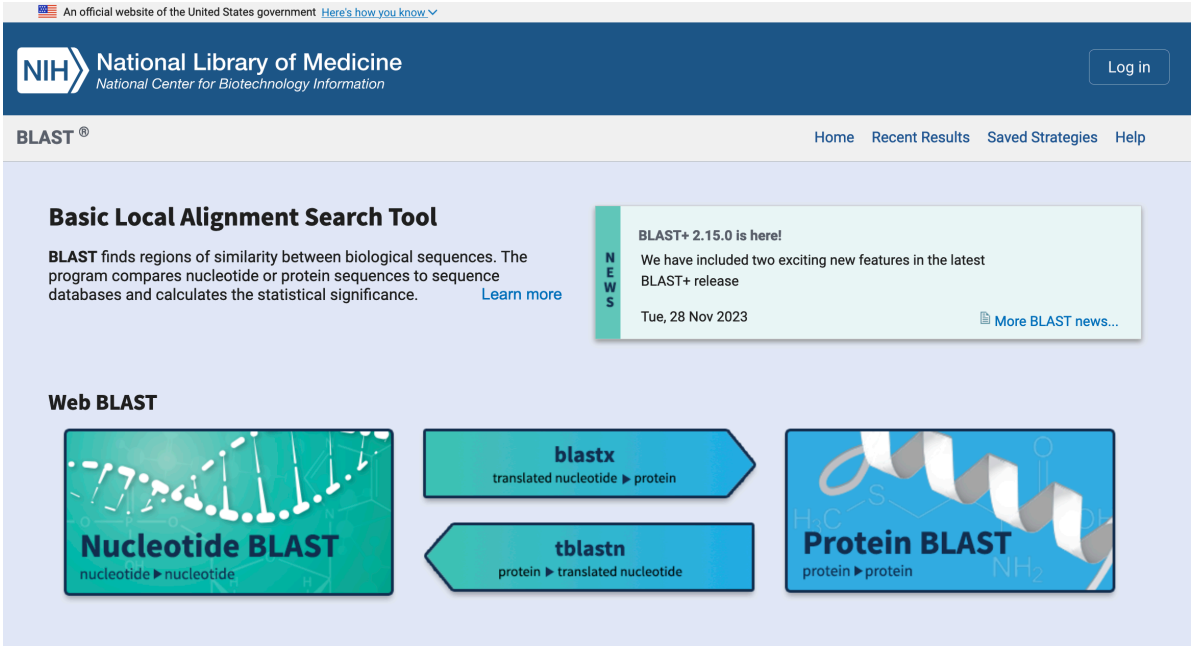
NCBI are the repository of record for published biological sequence data, and provide several BLAST databases, including one that contains all publicly-available 16S sequence data. You will use the NCBI-BLAST webservice, with your isolate's 16S sequence as the query, to search this database.

> BLAST reports a number of useful measures for each match
>
> - Percent coverage: what proportion (percentage) of the *query* is similar enough to the match that it has been aligned
> - Percentage identity: what proportion of the aligned region is identical in both the *query* and the reported match

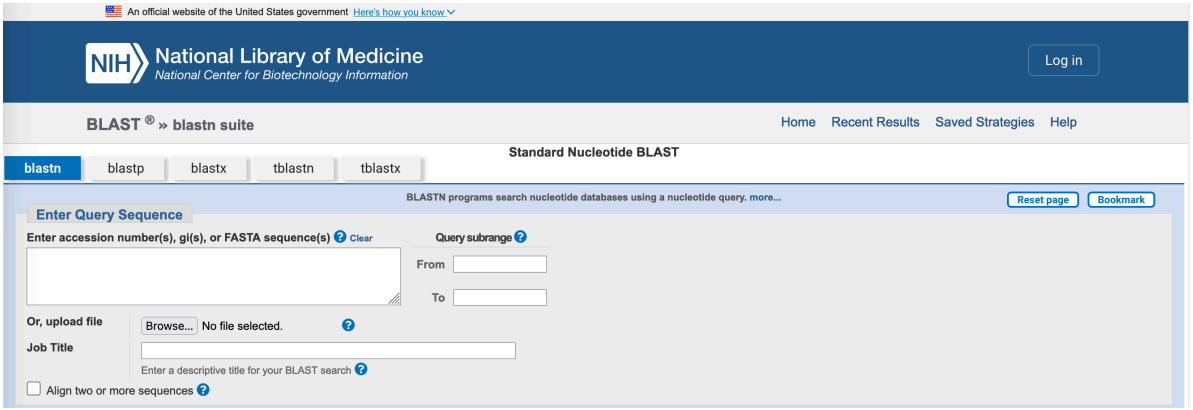
- Expectation (E-value): the number of matches you would expect by chance from this database that are *at least as similar as this one*, if you used the *query* sequence to search in a database of the same size made up of completely random sequences

3.2.1 Analysing your data

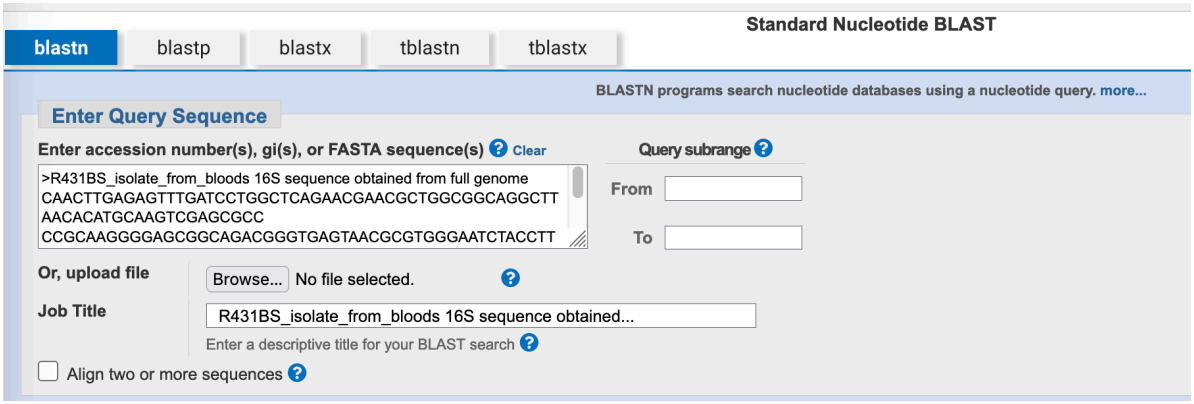
1. Go to the NCBI Blast webpage. You will see the landing page.



2. Select `Nucleotide BLAST`. This will take you to a new BLASTN search.

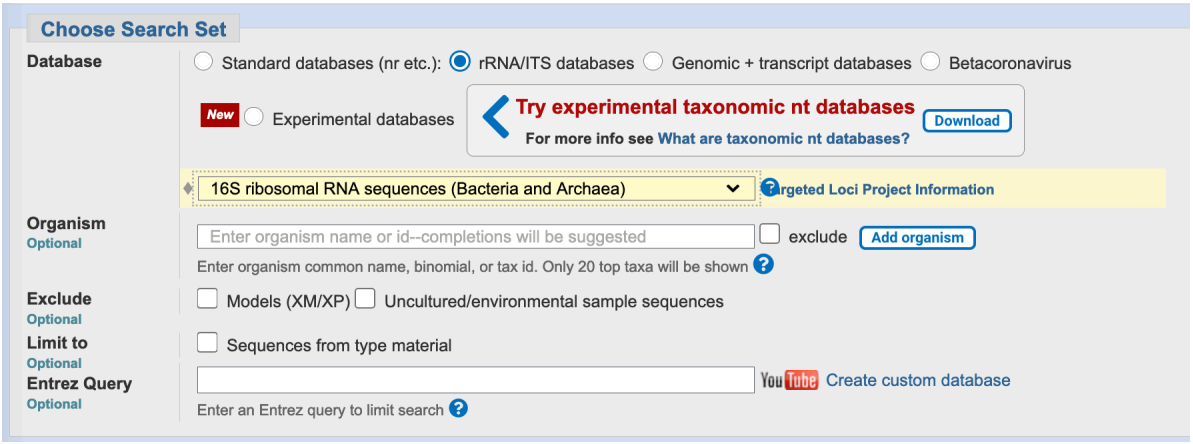


3. Upload your `isolate_16S.fasta` file, or copy and paste the contents of the file into the box marked `Enter accession number(s), gi(s) or FASTA sequence(s)`.



4. In the section headed `Choose Search Set`, select rRNA/ITS databases. This will autopopulate the database with `16S Ribosomal RNA sequences (Bacteria and Archaea)`, NCBI's complete database of 16S rRNA sequences.

> **Caution**
>
> NCBI do not manually curate (i.e. confirm sequence quality or correct identity) the data in their databases. Whole genome sequences are checked by classification algorithms for taxonomic placement, but 16S sequences are not. Also, NCBI contains a large amount of historical data which may have been annotated using outdated or superseded taxonomies.



5. Click on the `BLAST` button (towards the bottom of the page) and wait for the result to appear.

! Questions

1. What organism is the top hit to your query?
2. How similar is the top hit to your query (`Per. ident`/percentage identity column)? What is the alignment coverage?
3. What are the remaining top five matches to your query?
4. How similar are the top five matches to your query sequence?
5. What conclusion would you draw about the identity of your isolate from this BLAST search?

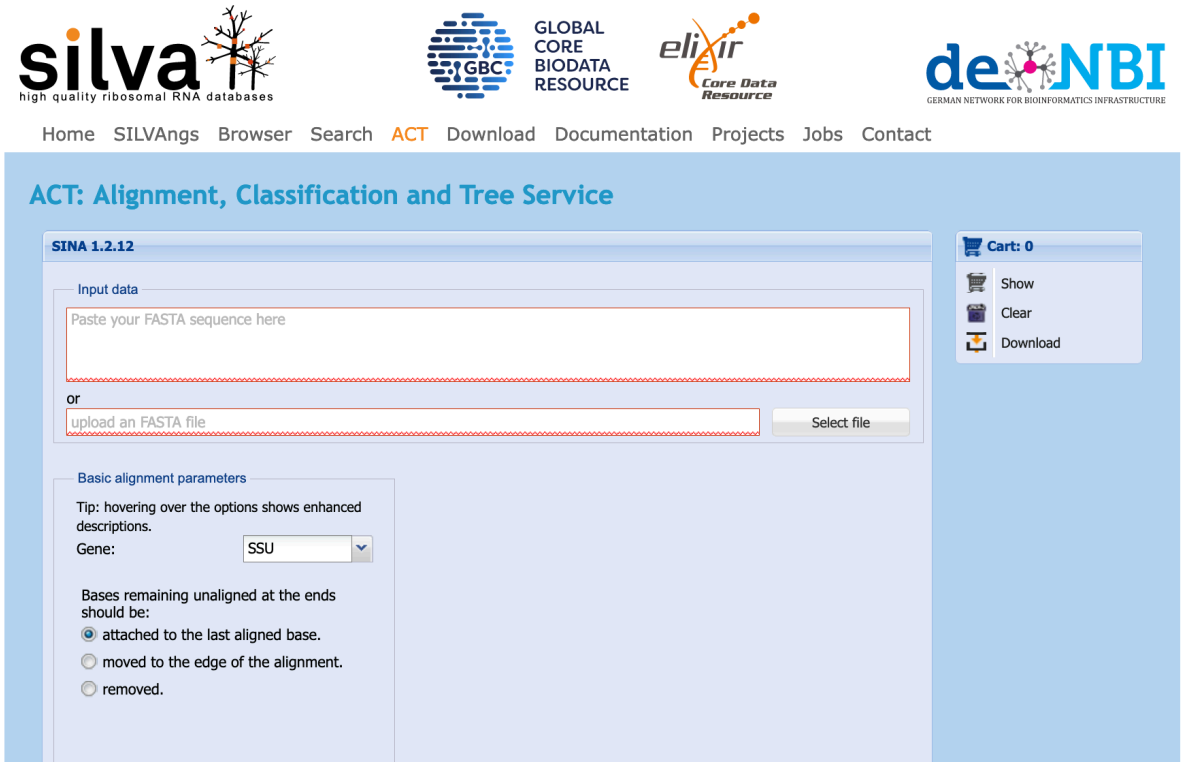
3.3 SILVA

SILVA is a curated, quality-checked database of rRNA sequence data that has been run by the Liebniz Institute DSMZ German Collection of Microorganisms and Cell Cultures for about two decades. The SILVA site provides a number of online tools and services, including the ACT service that enables users to search the database with their own rRNA sequences to find the best matches and identify their organism.

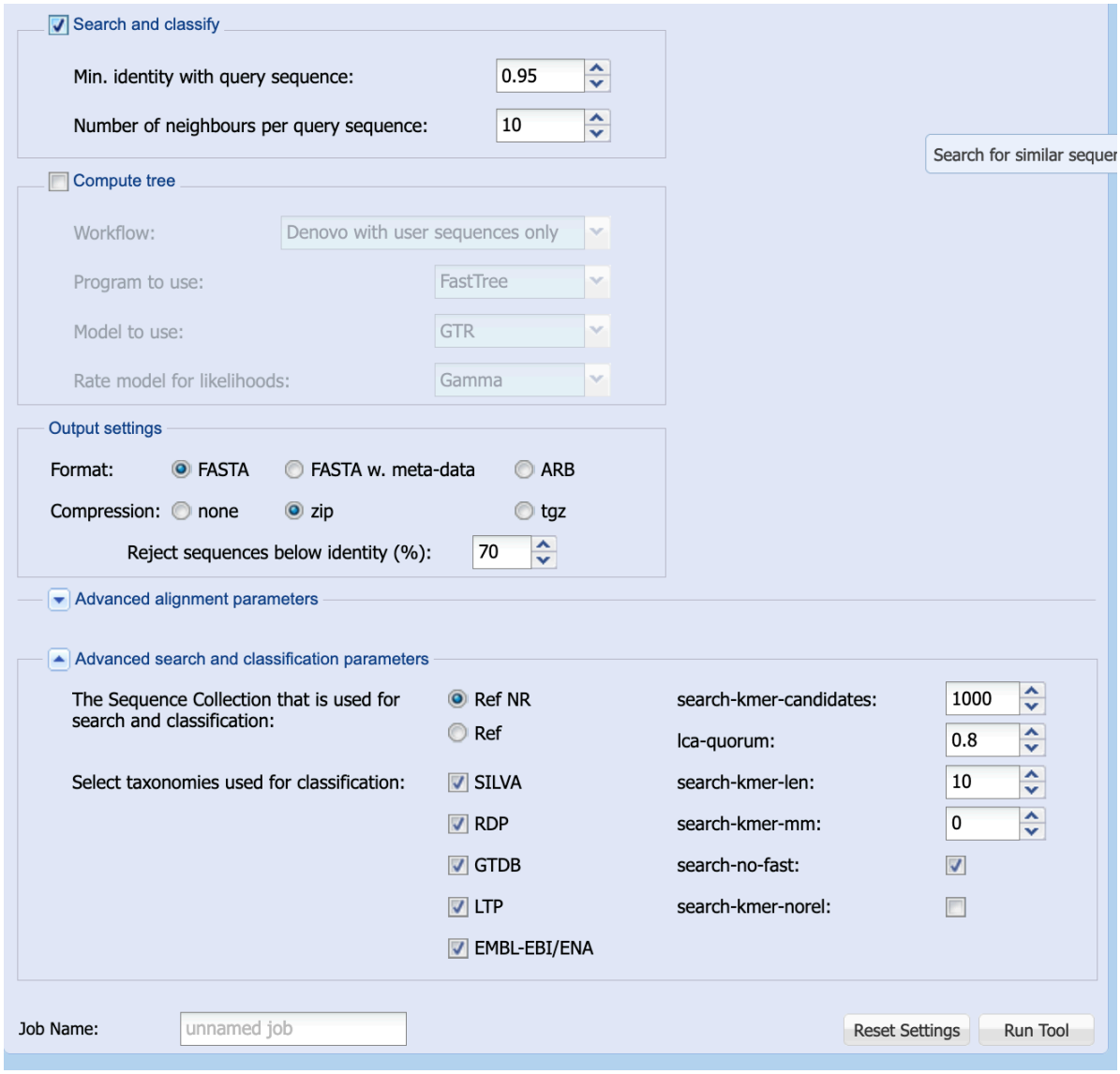
- SILVA rRNA database project
- SILVA Alignment, Classification, and Tree (ACT) webservice
- ACT online tutorial

3.3.1 Analysing your data

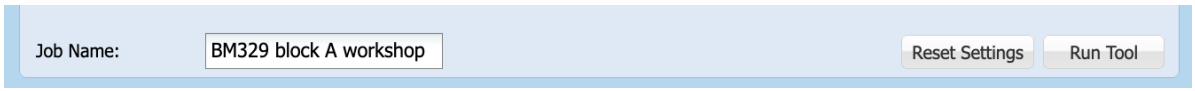
1. Go to the SILVA ACT (Alignment, Classification, and Tree) service



2. Upload your `isolate_16S.fasta` file, or copy and paste the contents of the file into the box marked `Input data`.
3. Check the box marked `Search and classify`, but leave all the other options as they are.

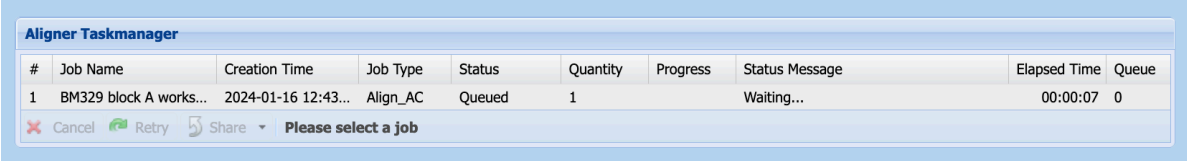


4. Enter a name for your job in the field marked `Job Name:` (e.g. "`BM432 Block A workshop`" or "`blood isolate`")



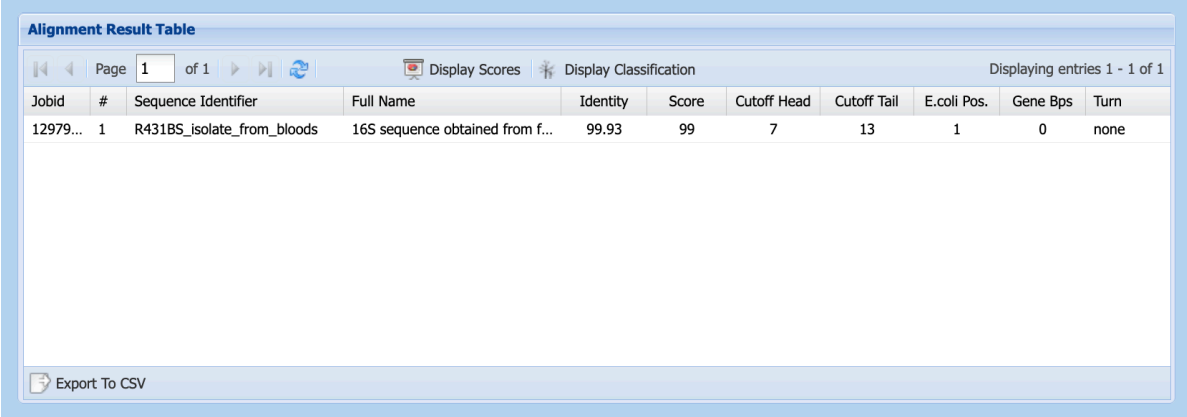
5. Click on the `Run Tool` button and wait for the result to appear.

> **Note**
>
> While the job is waiting it will appear in the `Aligner Taskmanager` as *Queued*, then *Starting*, then *Processing*. It may take a couple of minutes to produce a result.



6. When the job is complete, select your result in the `Alignment Result Table`, and click on `Display Classification`

> **Tip**
>
> You may need to adjust the widths of the headers in the result table to see the full classification



7. Click on `Export to CSV` (lower left of `Alignment Result Table`) to download the classification result. You can view this file in `Excel` or in a plain text editor.

> **Questions**
>
> 1. What is the taxonomic identity of the last common ancestor (LCA) that SILVA assigns to your isolate (`lca_tax_slv` or `LCA tax. SILVA` column)?
> 2. How similar (`Identity` column) does SILVA say your isolate is to that last common ancestor sequence?

3.4 Summary

You have used two different databases, and two different techniques, to assign taxonomic identity to a 16S sequence. Are the results consistent with each other?

Questions

1. Do the BLAST and SILVA analyses give the same taxonomic identity for your 16S sequence?
2. If the two results differ, how do they differ?
3. If the results differ, how important do you think is the difference between the results?
4. What is your current opinion about the identity of your isolate? Have you revised it since the BLAST search? How confident are you in the identification?

Continue your identification

Now you have a possible identification from 16S, you should continue the classification of your isolate by using Multi-Locus Sequence Typing (MLST) to try to gain a more precise taxonomic placement. Click on the link to MLST (here, or below), to get started.

4 Multi-Locus Sequence Typing (MLST)

4.1 MLST identification of your isolate

In this part of the workshop, you will use a public Multi-Locus Sequence Typing (MLST) database and resources to help identify your isolate.

! Important

Please ensure that you have downloaded the genome file for your isolate (`isolate_genome.fasta`) to a suitable location on your computer.

⚠ Multi-Locus Sequence Typing (MLST)

Multi-Locus Sequence Typing (MLST) is a widely-used method for bacterial identification. It is typically more precise and has more resolving power than 16S sequence analysis, but less precise than whole-genome sequence analysis (Maiden et al. (2013)).
MLST works by defining *marker sequences* for a taxon. These are typically well-conserved ("housekeeping") genes which very relatively little between organisms in the taxon, but enough to allow discrimination between them. The number of markers varies, but is usually somewhere around seven.
Each marker sequence has many *variants* (different sequences) within the taxon, and these are known as *alleles*. Each marker allele is given a unique number (starting at 1 and counting upwards) - its *allele number*. A single organism's *sequence type (ST)* is determined by the list of allele numbers that it contains. Organisms with the same *sequence type* are considered to be part of the same group.

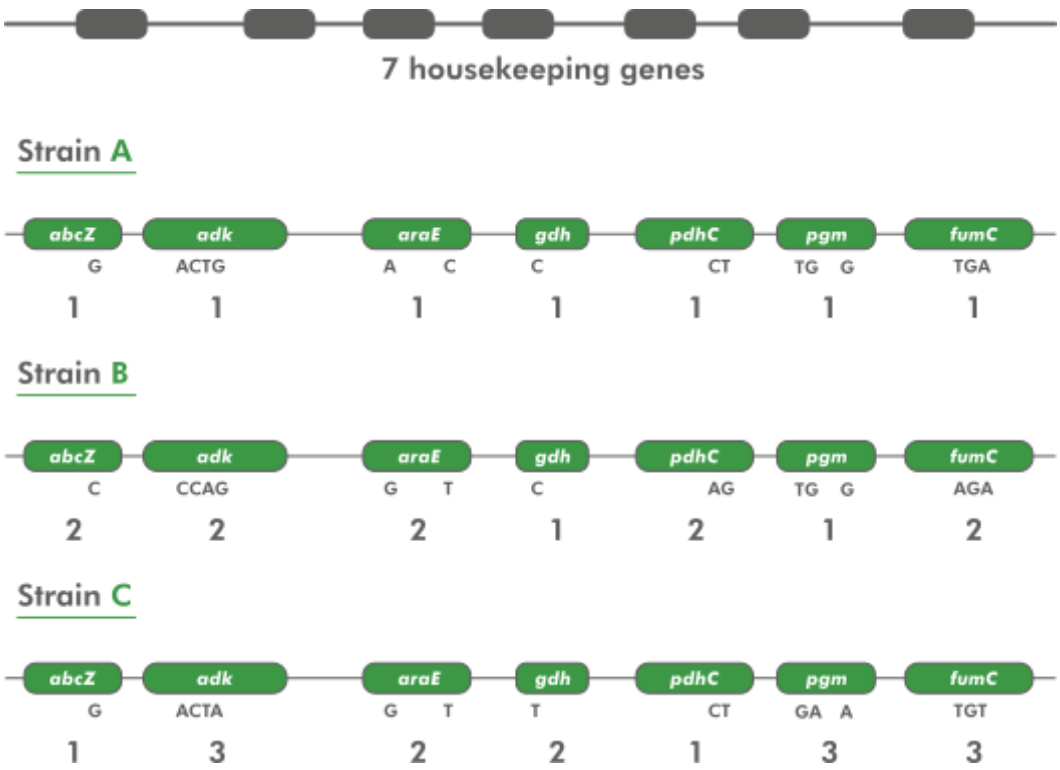
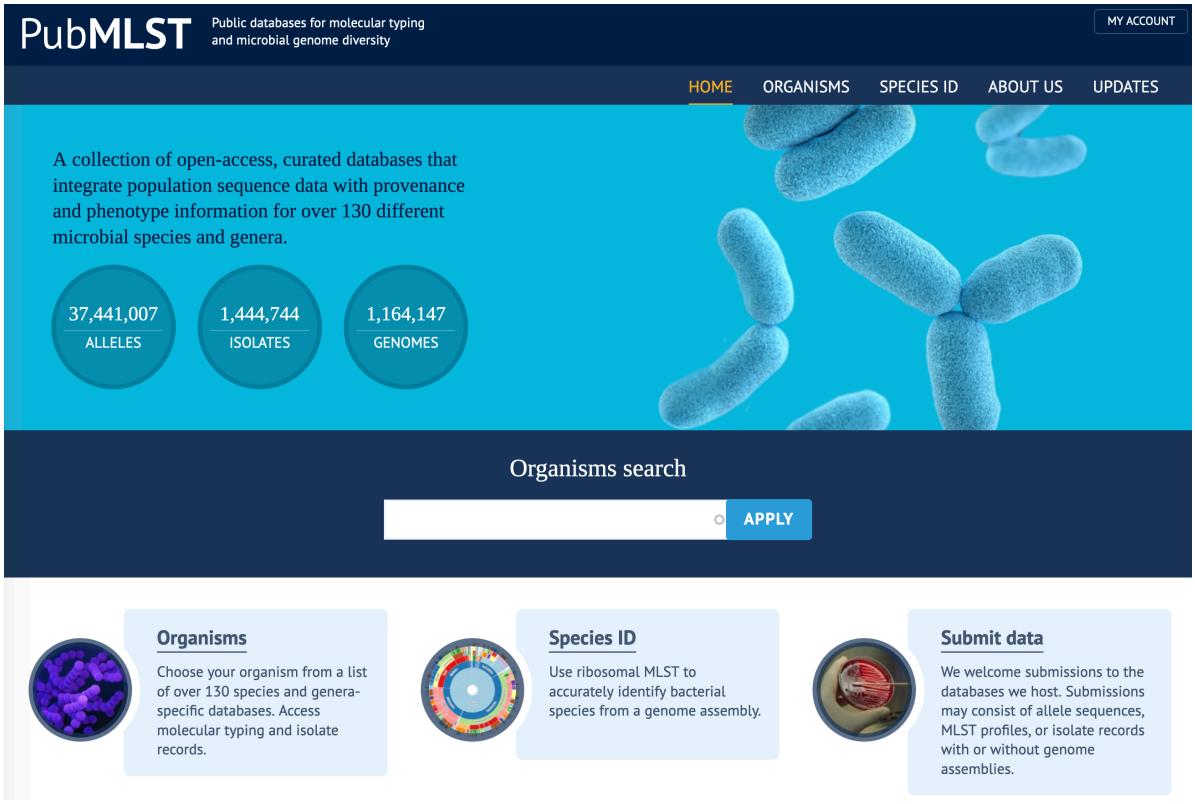


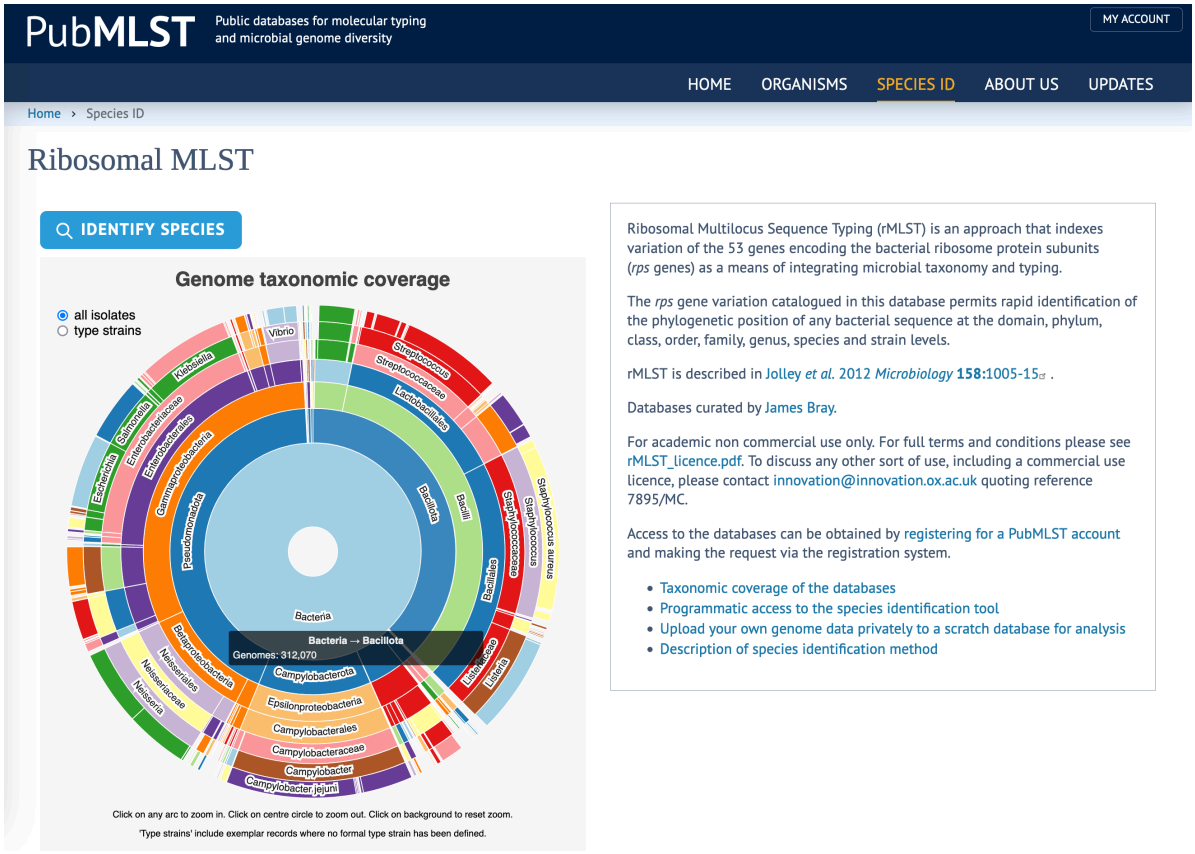
Figure 4.1: Schematic representation of MLST. In this typing scheme there are seven markers ("housekeeping genes"). One of these is the gene *adk*, which has a different sequence in each of the strains A, B, and C - so these have different allele numbers (1, 2, and 3). Another marker, *pdhC* has the same sequence (allele number) in strains A and C, but a different sequence (allele number) in strain B. The strains have allele numbers: 1,1,1,1,1,1,1 (A); 2,2,2,1,2,1,2 (B); 1,3,2,2,1,3,2 (C). These lists are different, so the strains have different sequence types.

4.2 pubMLST

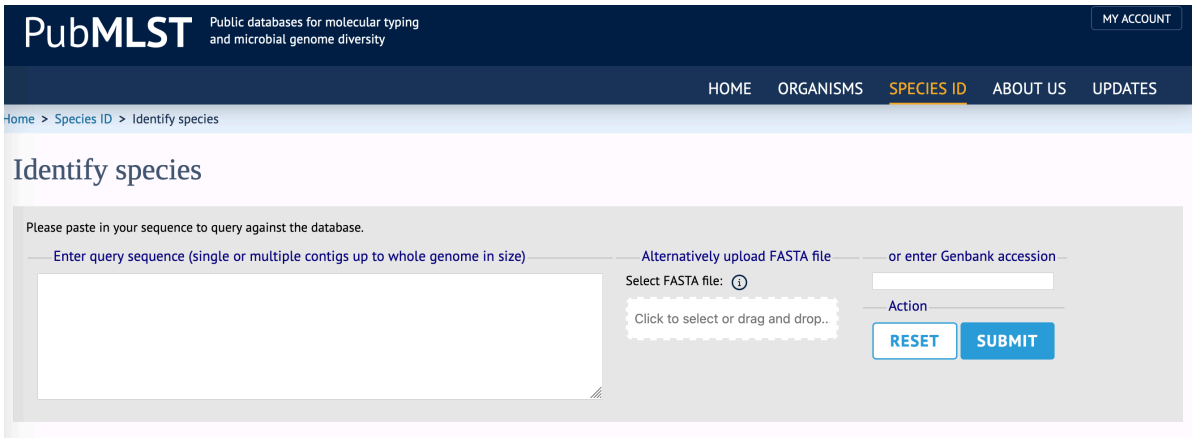
- Go to the pubMLST website



- Click on `Species ID` to open the bacterial species identifier



- Click on `Identify Species` to start the identification process



- Upload your `isolate_genome.fasta` file to the server either by clicking on `Click to select or drag and drop` and using the dialogue box to find your file, or by dragging the file onto the `Select FASTA file` field.

  - Alternatively, copy and paste the file contents - the *entire genome sequence* - into the `Enter query sequence` box.

- Click on the `Submit` button and wait for the results

Questions

1. What is the predicted taxon of your isolate's genome? What level of confidence does the pubMLST tool assign to this prediction (`Support` column)?
2. Of the 53 ribosomal genes used to classify all bacteria, how many of the genes in your isolate had an *exact* match in the database? Do you think this is a sufficient number for accurate taxonomic placement?
3. How many different species in the database had an *exact* match to markers in your isolate's genome?
4. *Which* species in the database had an *exact* match to markers in your isolate's genome? What effect does this have on your confidence in your classification?
5. What is your current opinion about the identity of your isolate? Have you revised your opinion due to the MLST analysis? How confident are you in the identification?

Continue your identification

Now you have further supporting information about the identity of your isolate, you should continue the classification of your isolate by using whole-genome comparison methods to try to pin down a more definitive taxonomic placement Click on the link to `Whole-genome Comparison` (here, or below), to get started.

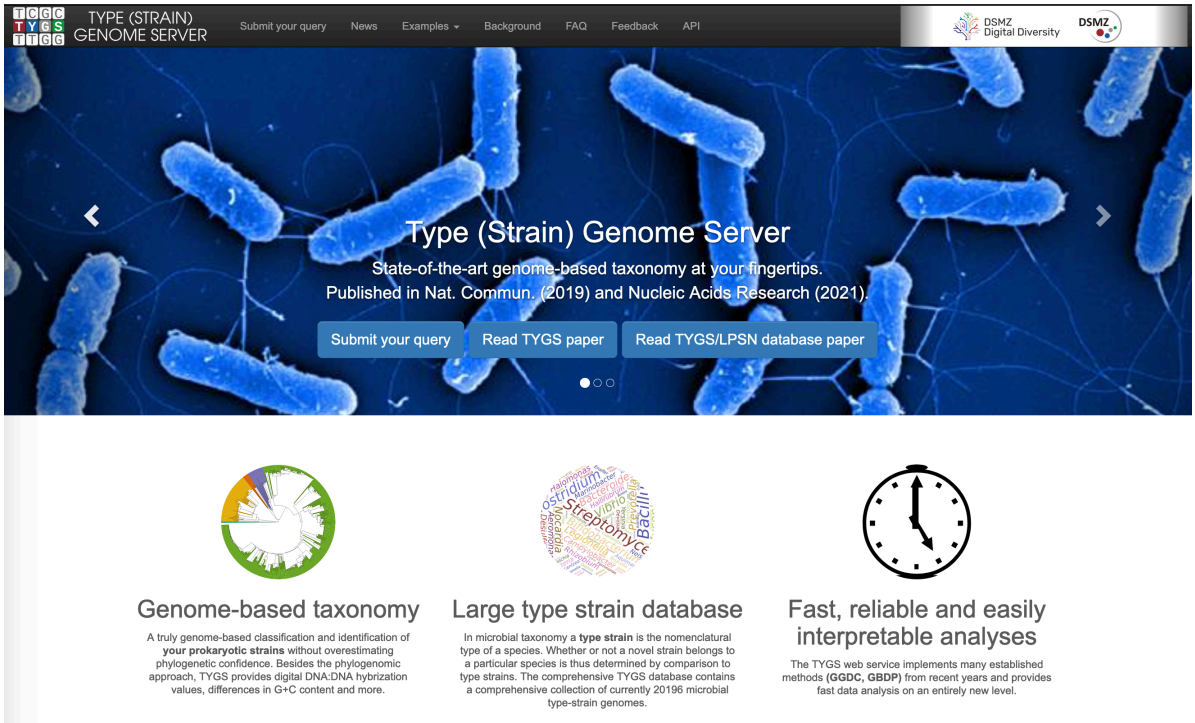
5 Whole-Genome Comparison

In this part of the workshop, you will use public resources for whole-genome identification and classification of prokaryotes to help identify your isolate.

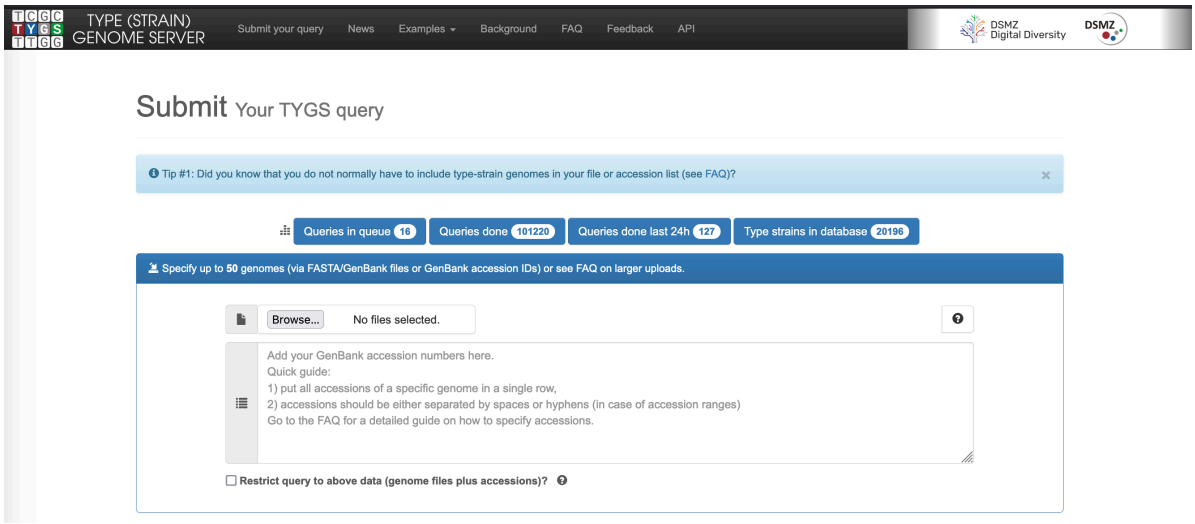
> **!** Important
>
> Please ensure that you have downloaded the genome file for your isolate (`isolate_genome.fasta`) to a suitable location on your computer.

5.1 TYGS

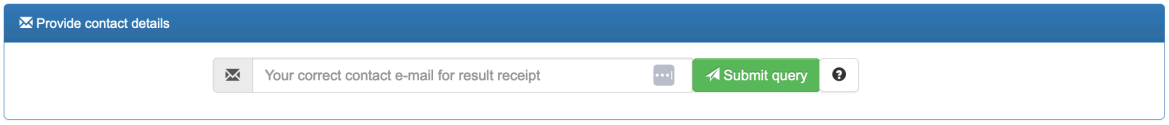
- Go to the TYGS server



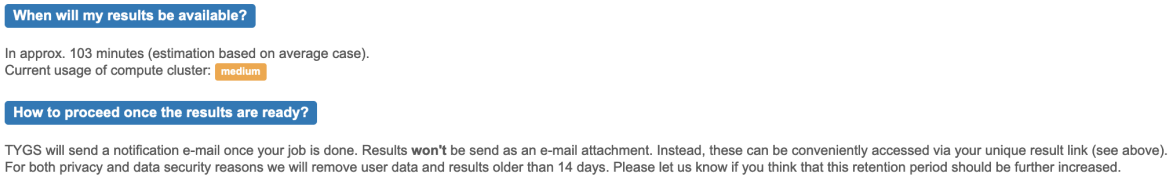
- Click on `Submit your query` (top menu), or `Submit your job` (button) to reach the query page



- Click on the `Browse…` button and navigate to your `isolate_genome.fasta` file to select it for classification.
- Enter your email address in the `Provide contact details` field



- Click on `Submit query` and wait for the results



> **Warning**
>
> This may take some time (maybe over an hour, depending on server load!), and may not be complete before the end of the workshop.
> **Please move on to the next section while you are waiting.**

- **When you get the result confirmation email, inspect the PDF**

! Questions

1. What is the predicted taxonomic classification of your isolate's genome?
2. How similar is your isolate's genome to the closest match? (the d_4 result is the dDDH (digital DNA-DNA hybridisation) score)
3. What is your current opinion about the identity of your isolate? How confident are you in the identification?

5.2 genomeRxiv

genomeRxiv is a new appraoch that promises to identify and classify prokaryotic genomes quickly and accurately into categories called LINgroups (LIN: Life Identification Number). LINgroups are a taxonomy-independent, quantitative categorisation scheme that organises genome sequences by similarity in multidimensional "space". These categorisations can then be used to relate alternative taxonomic assignments, and other annotations, to each other (Mazloom et al. (2022)).

i How `genomeRxiv` works

`genomeRxiv` works in a similar way to map grid references.

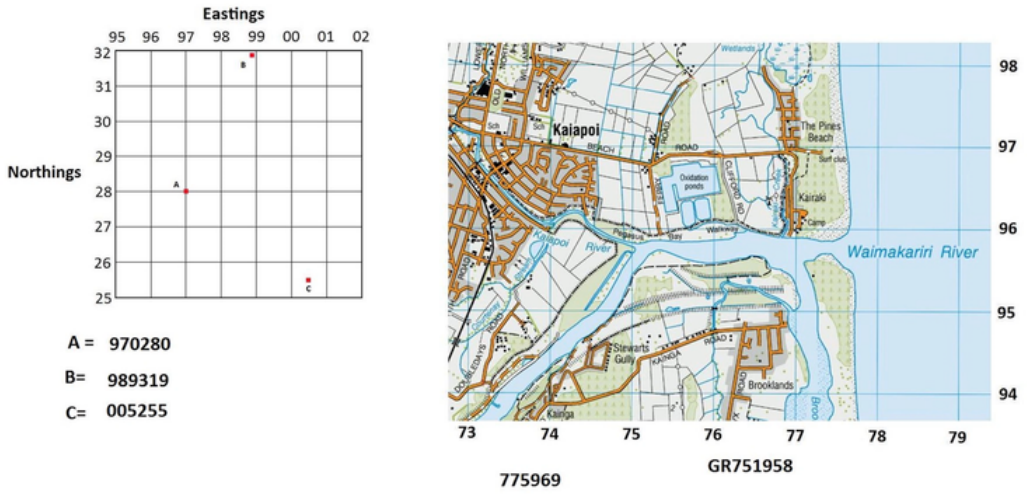


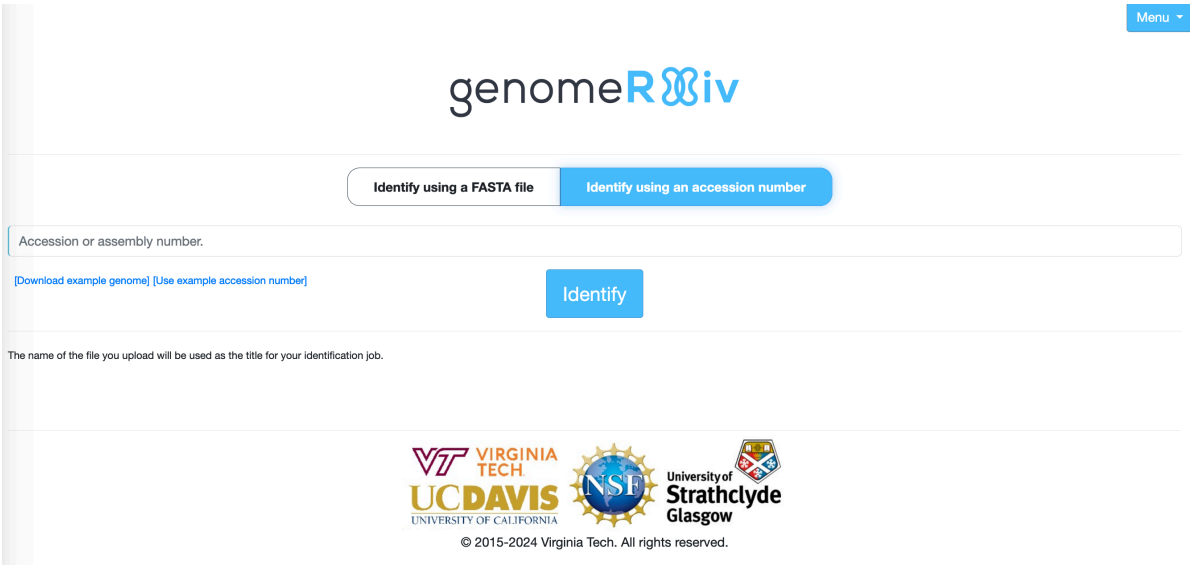
Figure 5.1: Example map grid references: point A has grid reference 970 (Easting) and 280 (Northing) as x-y co-ordinates, so the total reference is 970280. Similarly, point B is at 989319, and point C at 005255. All points in the map, however large or small, are uniquely assigned to a discrete location and, by subdividing the map into smaller and smaller squares (longer and longer numbers) an arbitrary level of precision can be reached

`genomeRxiv` compares input sequences to a reference database with a very fast bioinformatics algorithm (`sourmash`) to get a set of good matches, and then refines the match with a more precise but slower algorithm (`ANI`). `genomeRxiv` then assigns a LINgroup to the genome. The LINgroup is a string of numbers, analogous to a map co-ordinate. The shorter the LINgroup, the lower the resolution of identification (Phylum, Family, etc.), and the longer the LINgroup, the finer the resolution (species, subspecies, strain, etc.).
The first key difference between LINgroups and map co-ordinates is that LINgroups are not co-ordinates on a two-dimensional surface, but in many-dimensional space. The second is that a single number describes the location, rather than two numbers (the "Easting" and "Northing" of map co-ordinates. Finally, grid references represent a physical space (such as the surface of the Earth), and LINgroups represent "sequence space" - which is not physical (Mazloom et al. (2022)).
Taxa are then defined by the volumes in space circumscribed by genomes which are examples of each taxon. When a new unknown genome is added, a LINgroup is assigned and - if it lies within a volume of space contained only by members of a single taxon (e.g. *E. coli*), the genome is assigned that taxon.

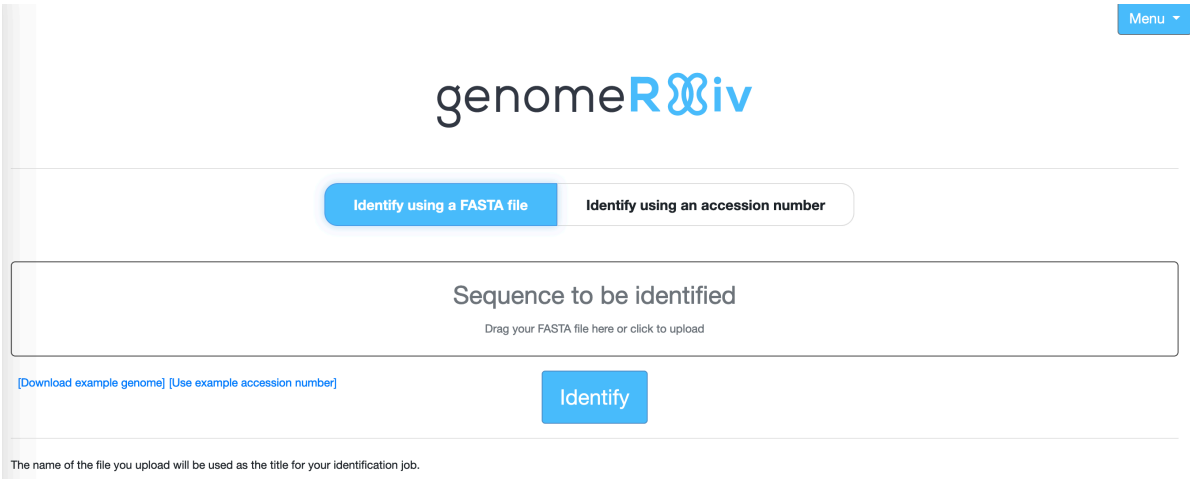
The `genomeRxiv` webservice allows users to input their bacterial genomes and rapidly obtain, or predict, taxonomic assignments on the basis of genome sequence.

> **Warning**
>
> `genomeRxiv` remains under active development and is not yet fully-released, although it is public and usable.

- Go to the genomeRxiv server



- Click on `Identify using a FASTA file`



- Either click on the `Sequence to be identified` link to bring up a dialogue box through which you can upload the `isolate_genome.fasta` file, or drag the

`isolate_genome.fasta` file onto the text.
- Click on `Identify` and wait for the results

> **Note**
>
> The genomeRxiv site may ask you to allow pop-ups. You should allow the pop-ups, as these contain the identification output you need.

> **Questions**
>
> 1. What is the predicted taxonomic classification of your isolate's genome? (Check the `Member LINgroups` result)
> 2. What is the taxonomic identity assigned to the most similar genome in the database?
> 3. How similar is your isolate's genome to the closest match? (Look at the `ANI to Target` value)
> 4. What is your current opinion about the identity of your isolate? Have you modified your classification? How confident are you in the identification?

> **Consider your final evaluation**
>
> You have now used a number of different online bioinformatics tools to obtain a possible taxonomic classification for the isolate in your blood sample. You should, by this point, have some idea of what you think the organism is, and how confident you are. Now it's time to take a look at the official prokaryotic nomenclature database, to get a little more context around the candidate taxonomic name. Click on the link to `LPSN` (here, or below), to keep going.

6 LPSN: the List of Prokaryotic names with Standing in Nomenclature

6.1 Investigating your classification at LPSN

- Go to the LPSN server
- Enter the taxonomic classification you think should be assigned to your organism in the `Search taxonomy` field, and hit `Return/Enter`. The result should appear immediately.

! Questions

1. What kind of taxonomic category does the name of your organism represent?
2. What does the name mean, and what does it imply about the bacteria?
3. Is the name you entered a validly published taxonomic name?
4. What is the parent taxon for this name?

- Click on the link by `Parent taxon:` (the value here is also the answer to question 4. above).

! Questions

1. What kind of taxonomic category does this name represent?
2. What does the name mean, and what does it imply about the bacteria?
3. Is the name you entered a validly published taxonomic name?
4. What is the `Taxonomic status` of this name?
5. Is this name the "correct name" for the organism?

⚠ What does "correct name" mean in a taxonomic context?

The precise meaning of "correct name" in this context is provided in a glossary at LPSN. The key points in that text are:

- Any taxon may have only one *correct* name. This is usually the earliest validly published, legitimate name.
- The LPSN selects certain taxon names as *correct* names. If more than one option

is available, then the LPSN's choice reflects only *one* of the taxonomic opinions expressed in the literature.
- Valid names representing other taxonomic opinions are given *synonym* status.
- Other scientists may express distinct taxonomic opinions and are permitted to do so, by using a *synonym*.

The term *correct* is therefore used in a technical sense, in a taxonomic context. The existence of a *correct* name for a taxon does not imply that other names (*synonyms*) are invalid or cannot be used.

- Read the `Notes` section on the LPSN page.

! Questions

1. How do the notes on this page reflect the controversy over naming of *Ochrobactrum*/*Brucella* described in the Introduction?
2. Does the LPSN page affect your opinion about how you have classified your isolate?
3. Finally, what is the taxonomic name you would use for your isolate?

7 Summary

In this workshop, you have:

1. gained experience using 16S and whole-genome sequence data to identify, bioinformatically, an unknown organism, using:

- two 16S databases and associated search services to identify matching sequences from other organisms
- the main reference MLST database (pubMLST) and its search tool to obtain a classification for your organism
- two whole-genome classification approaches to obtain taxonomic classifications for your organism

2. considered and weighed the evidence from these approaches to draw an informed conclusion about the most likely taxonomic assignment, considering:

- the relative information used by each method
- the quality of the database(s) used in each approach
- the bioinformatic approach taken

3. interpreted your classification according to the rules laid down in the ICNP (International Code of Nomenclature of Prokaryotes), a.k.a. the Prokaryotic Code, and at LPSN (the List of Prokaryotic names with Standing in Nomenclature).

Thank you

That's almost the end of the workshop. Thank you for participating. We hope you found it enjoyable and interesting, and that you now understand more about how bacterial classification works, in practice.
This year is the first presentation of this particular material, and we would be very grateful to hear feedback by email or through the GitHub repository Issues page.

Completing the workshop

Please now proceed to the BM329 MyPlace page and complete the formative questions, to finish the workshop.

References

Barth, Patricia Orlandi, Eliane Wurdig Roesch, Larissa Lutz, Ândrea Celestino de Souza, Luciano Zubaran Goldani, and Dariane Castro Pereira. 2023. "Rapid Bacterial Identification by MALDI-TOF MS Directly from Blood Cultures and Rapid Susceptibility Testing: A Simple Approach to Reduce the Turnaround Time of Blood Cultures." *Braz. J. Infect. Dis.* 27 (1): 102721.

Bosilkovski, Mile, Fariba Keramat, and Jurica Arapović. 2021. "The Current Therapeutical Strategies in Human Brucellosis." *Infection* 49 (5): 823–32.

Edgar, Robert. 2018. "Taxonomy Annotation and Guide Tree Errors in 16S rRNA Databases." *PeerJ* 6 (e5030): e5030.

Edgar, Robert C. 2018a. "Accuracy of Taxonomy Prediction for 16S rRNA and Fungal ITS Sequences." *PeerJ* 6 (e4652): e4652.

———. 2018b. "Updating the 97% Identity Threshold for 16S Ribosomal RNA OTUs." *Bioinformatics* 34 (14): 2371–75.

Franco, María Pía, Maximilian Mulder, Robert H Gilman, and Henk L Smits. 2007. "Human Brucellosis." *Lancet Infect. Dis.* 7 (12): 775–86.

Hagiya, Hideharu, Kouhei Ohnishi, Miyako Maki, Naoto Watanabe, and Tomoko Murase. 2013. "Clinical Characteristics of Ochrobactrum Anthropi Bacteremia." *J. Clin. Microbiol.* 51 (4): 1330–33.

Hördt, Anton, Marina García López, Jan P Meier-Kolthoff, Marcel Schleuning, Lisa-Maria Weinhold, Brian J Tindall, Sabine Gronow, Nikos C Kyrpides, Tanja Woyke, and Markus Göker. 2020. "Analysis of 1,000+ Type-Strain Genomes Substantially Improves Taxonomic Classification of Alphaproteobacteria." *Front. Microbiol.* 11 (April): 468.

Johnson, Jethro S, Daniel J Spakowicz, Bo-Young Hong, Lauren M Petersen, Patrick Demkowicz, Lei Chen, Shana R Leopold, et al. 2019. "Evaluation of 16S rRNA Gene Sequencing for Species and Strain-Level Microbiome Analysis." *Nat. Commun.* 10 (1): 5029.

Laine, Christopher G, Valen E Johnson, H Morgan Scott, and Angela M Arenas-Gamboa. 2023. "Global Estimate of Human Brucellosis Incidence." *Emerg. Infect. Dis.* 29 (9): 1789–97.

Maiden, Martin C J, Melissa J Jansen van Rensburg, James E Bray, Sarah G Earle, Suzanne A Ford, Keith A Jolley, and Noel D McCarthy. 2013. "MLST Revisited: The Gene-by-Gene Approach to Bacterial Genomics." *Nat. Rev. Microbiol.* 11 (10): 728–36.

Mazloom, Reza, Leighton Pritchard, C Titus Brown, Boris A Vinatzer, and Lenwood S Heath. 2022. "LINgroups as a Principled Approach to Compare and Integrate Multiple Bacterial Taxonomies." In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. New York, NY, USA: ACM.

Moreno, Edgardo, José María Blasco, Jean Jacques Letesson, Jean Pierre Gorvel, and Ignacio Moriyón. 2022. "Pathogenicity and Its Implications in Taxonomy: The Brucella and Ochrobactrum Case." *Pathogens* 11 (3).

Moreno, Edgardo, Earl A Middlebrook, Pamela Altamirano-Silva, Sascha Al Dahouk, George F Araj, Vilma Arce-Gorvel, Ángela Arenas-Gamboa, et al. 2023. "If You're Not Confused, You're Not Paying Attention: Ochrobactrum Is Not Brucella." *J. Clin. Microbiol.* 61 (8): e0043823.

Woese, C R, O Kandler, and M L Wheelis. 1990. "Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya." *Proc. Natl. Acad. Sci. U. S. A.* 87 (12): 4576–79.