# EnteroBase Manual

Version 1.0.0 – May 2016

Nabil-Fareed Alikhan <n-f.alikhan@warwick.ac.uk>

Martin Sergeant<M.J.Sergeant@warwick.ac.uk>

Zhemin Zhou <Zhemin.Zhou@warwick.ac.uk>

## Contents

## About EnteroBase

EnteroBase aims to establish a world-class, one-stop, user-friendly, backwards-compatible but forward-looking genome database, EnteroBase—together with a set of web-based tools, EnteroTools—to enable bacteriologists to identify, analyse, quantify and visualize genomic variation principally within the genera:

- *Escherichia*
- *Salmonella*
- The *Yersiniae*
- *Moraxella*

EnteroBase is populated with over 100,000 of genomic assemblies derived from publicly available complete genomes, sequence read archives and user uploads.

Funded by BBSRC research grant BB/L020319/1.

### Implementation

EnteroBase is strain based. Each strain is associated with metadata and genomic assemblies, as well as with deduced genotyping data. All assemblies are performed *de novo* from Illumina reads using a standardised, versioned pipeline. Unless explicitly chosen, only assemblies that match pre-defined criteria are displayed, and where multiple assemblies are associated with a strain, only the best assembly according to assembly criteria is displayed.

Genotyping data is deduced exclusively from assemblies. MLST data is called by uBlast and amino acid BLAST+ against a defined dataset of reference allelic sequences. The full data including assemblies can be downloaded freely, in accordance with Fair Usage (see below).

### Fair Usage

All metadata, assemblies and genotyping data can be freely downloaded for academic purposes. In order to allow users who upload unpublished data sufficient time to perform their own analyses, we request that no analyses of user data be published without their explicit permission prior to the release date. Both metadata and genomic data will be clearly marked if it is downloaded prior to the release date. We would also consider it fair usage that users who wish to analyse very large amounts of the data stored in EnteroBase also contribute software tools to EnteroBase that facilitate the presentation and analysis of their results. Downloading and analyses of data by commercial enterprises can only be performed after explicit permission by the administrators, which may involve legal agreements regarding material transfer.

### Data Privacy

EnteroBase users are encouraged to upload their own reads to the website, which will be assembled and genotyped like existing public data. Submitters should note that raw data (sequence reads) will never be made public through the website to other users. The genome assembly will only be accessible to the data submitter and their buddies for 6 months after uploads. Assembly data will then be made public, longer release dates can be negotiated by contacting Martin Sergeant on M.J.Sergeant@warwick.ac.uk.

Genotyping results i.e. MLST, ribosomal MLST, core genome MLST, *in silico* serotyping, will be made public as soon as the uploaded data has been processed.

User passwords on the website are encrypted and no one, including administrators, can easily access them. However, we would advise you NOT to use the same password you would use for important accounts, such as internet banking.

## Citation

EnteroBase has not been formally published, yet. If you use data/metadata from the website, or the analysis based on these data, please cite EnteroBase website directly: http://enterobase.warwick.ac.uk

An extend citation could be:

EnteroBase. [online] Enterobase.warwick.ac.uk. Available at:
http://enterobase.warwick.ac.uk [Accessed 1 January 2016].

## How to use this manual

Chapters 1-4 provide a guided tour with worked examples of how to use the website and should be read sequentially, like a tutorial.

- Chapter 1: Basic interactions like logging in and searching the database
- Chapter 2: Uploading, editing and analysing your own data.
- Chapter 3: How to interrogate real world datasets using EnteroBase.
- Chapter 4: More complex interactions with real world datasets

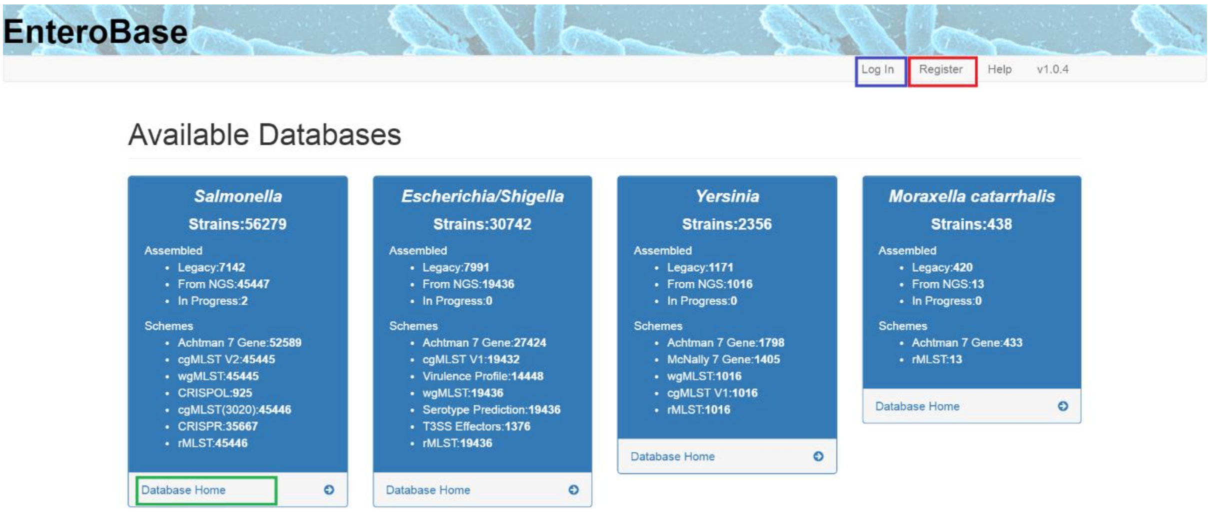
Chapters from 5 and up provide a reference for concepts and systems within EnteroBase. You may want to jump to aspects of these chapters for clarification about a specific aspect of EnteroBase.

- Chapter 5: Documentation of pipelines and analysis methods.

## Chapter 1: Getting started, searching and browsing

EnteroBase runs entirely online, all you require is an updated web browser; Ideally Google Chrome (https://www.google.com/chrome/). EnteroBase is available at http://enterobase.warwick.ac.uk/ . The main page (Figure 1) presents available databases (*Salmonella, Escherichia, Moraxella* and *Yersinia*) with an overview of the number of records. From here you can:

- **Access a database** by clicking the 'Database Home' link (green).
- **Access User account options**, such as logging in (blue), registering a new account (red) and changing email/passwords are found in the top right.



Available Databases

***Salmonella***
**Strains:56279**
Assembled
- Legacy:**7142**
- From NGS:**45447**
- In Progress:**2**

Schemes
- Achtman 7 Gene:**52589**
- cgMLST V2:**45445**
- wgMLST:**45445**
- CRISPOL:**925**
- cgMLST(3020):**45446**
- CRISPR:**35667**
- rMLST:**45446**

Database Home

***Escherichia/Shigella***
**Strains:30742**
Assembled
- Legacy:**7991**
- From NGS:**19436**
- In Progress:**0**

Schemes
- Achtman 7 Gene:**27424**
- cgMLST V1:**19432**
- Virulence Profile:**14448**
- wgMLST:**19436**
- Serotype Prediction:**19436**
- T3SS Effectors:**1376**
- rMLST:**19436**

Database Home

***Yersinia***
**Strains:2356**
Assembled
- Legacy:**1171**
- From NGS:**1016**
- In Progress:**0**

Schemes
- Achtman 7 Gene:**1798**
- McNally 7 Gene:**1405**
- wgMLST:**1016**
- cgMLST V1:**1016**
- rMLST:**1016**

Database Home

***Moraxella catarrhalis***
**Strains:438**
Assembled
- Legacy:**420**
- From NGS:**13**
- In Progress:**0**

Schemes
- Achtman 7 Gene:**433**
- rMLST:**13**

Database Home

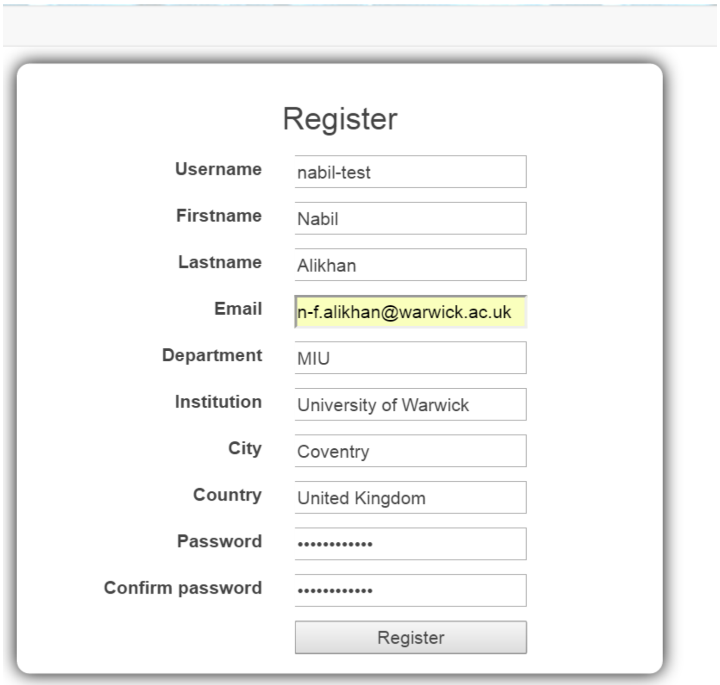
**Figure 1 Enterobase main page**

## Registering on EnteroBase

To Register on EnteroBase, visit the main webpage (http://enterobase.warwick.ac.uk) and click 'Register' in the top right (Figure 1).

This will direct you to the registration form (Figure 2) where you should fill in a username, password, email and details about yourself. Once the form is filled, click 'Register'.

This will send a verification email to your specified email address. Click the link in this email to confirm your registration and you can then log into the EnteroBase website.



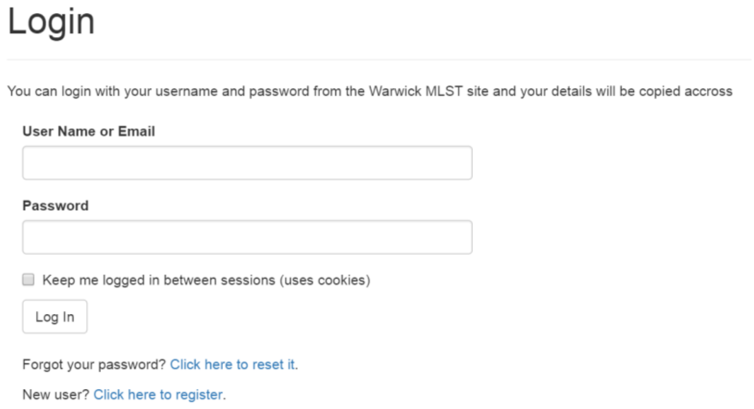
Register

| | |
|---|---|
| Username | nabil-test |
| Firstname | Nabil |
| Lastname | Alikhan |
| Email | n-f.alikhan@warwick.ac.uk |
| Department | MIU |
| Institution | University of Warwick |
| City | Coventry |
| Country | United Kingdom |
| Password | •••••••••••• |
| Confirm password | •••••••••••• |

**Figure 2 Registration page**

## Logging in

To Log into EnteroBase, visit the main webpage (http://enterobase.warwick.ac.uk) and click 'Login' in the top right (Figure 1).
This will direct you to the login page (Figure 3) where you can enter your username and password. If you have not create an account or have forgotten your password there are links below to help you resolve this.



Login

You can login with your username and password from the Warwick MLST site and your details will be copied accross

User Name or Email

Password

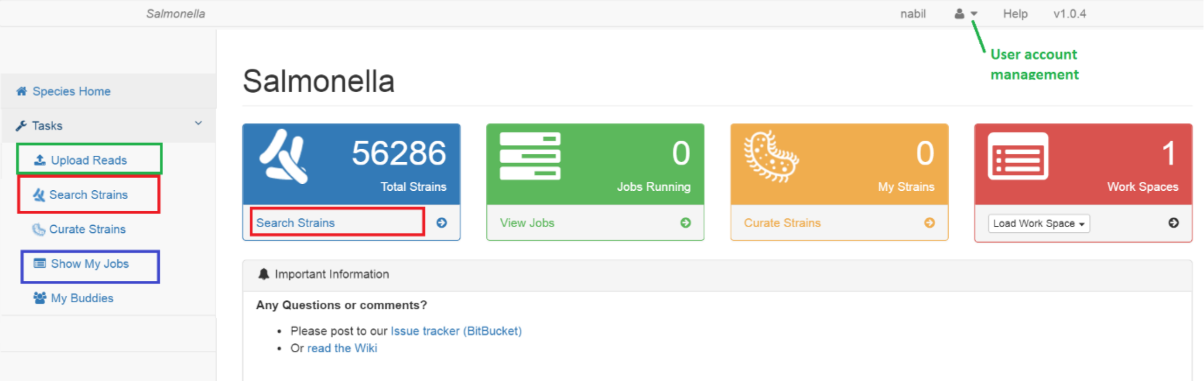
□ Keep me logged in between sessions (uses cookies)

Forgot your password? Click here to reset it.

New user? Click here to register.

**Figure 3 Login page**

## Basic navigation and searching strains



Important Information

**Any Questions or comments?**

- Please post to our Issue tracker (BitBucket)
- Or read the Wiki

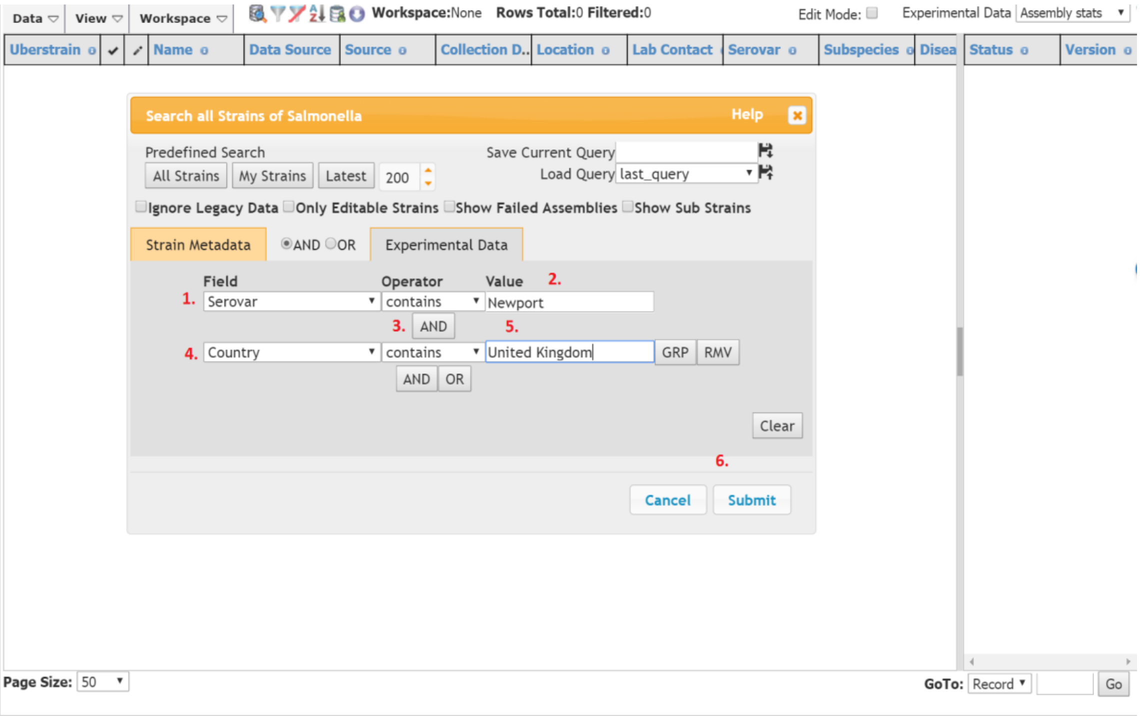
**Figure 4 Database homepage**

From the main page (Figure 1), click on database home link for *Salmonella* to be brought the dashboard for this database. This page (Figure 4) provides a detailed overview of each genera and there are links to a number of tasks you can perform in EnteroBase.

- **'Search strains'** will allow you to query the database for records of interest (red).
- **'Upload Reads'** will allow you to upload your own sequence reads for analysis (green).
- **'Show my Jobs'** will show analysis jobs related to your data (blue).
- **Manage your account** through the dropdown in the top right.

These links on the side bar and along the top (grey) will always be present as you navigate deeper into the website so you can easily jump to another task.

By clicking through 'Search Strains' you are presented with the Search menu with the data panels in the background (Figure 5). Once you submit a search, these panels will fill with your search results.



Workspace:None Rows Total:0 Filtered:0

Load Query last_query

200

1. Serovar contains Newport

4. Country contains United Kingdom

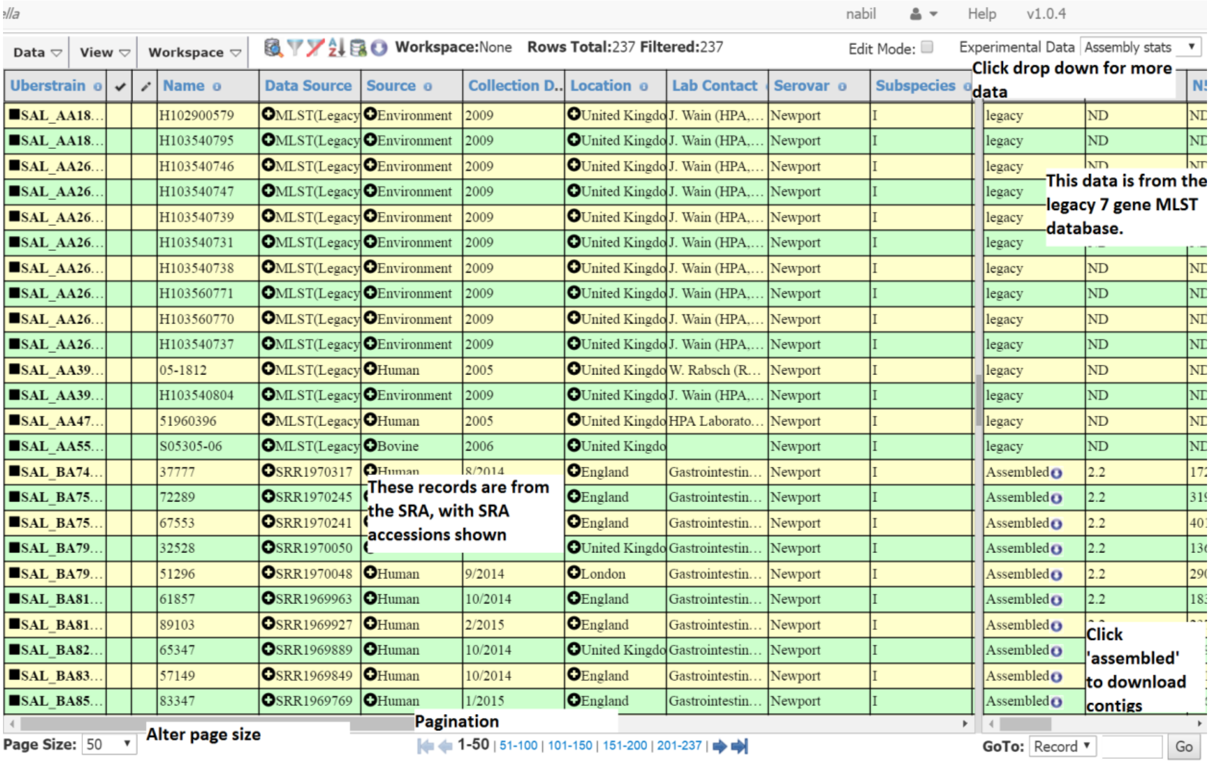
Page Size: 50

**Figure 5 Search menu and data panes**

Let's perform a search together (Figure 5) by following these steps:

1. Click the 'Field' dropdown and select 'Serovar'. *N.B. A 'Serovar' is just a sub-division of the species based on antigenicity, in* Salmonella *these are often named after geographic locations.*
2. In the 'Value' field to the right, type in 'Newport' (a common *Salmonella* serovar). You should notice that a dropdown will appear below giving suggestions. You can click on 'Newport' in this list or continue typing the full word.
3. Click the AND button. We want add an additional condition to our search.
4. Click the 2nd 'Field' dropdown and select 'County.
5. In the 'Value' field to the right, type in 'United Kingdom'. You should notice that a dropdown will appear below giving suggestions. You can click on 'United Kingdom' in this list or continue typing the full phrase.

You have now prepared a reasonably complicated query. We are searching for strains of serovar Newport that were isolated in the United Kingdom. Click 'Submit' (6).



Workspace:None Rows Total:237 Filtered:237

Edit Mode: □

Experimental Data: Assembly stats ▾

Click drop down for more data

| Uberstrain | ✓ | ✎ | Name | Data Source | Source | Collection D.. | Location | Lab Contact | Serovar | Subspecies | | | N5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ■SAL_AA18… | | | H102900579 | MLST(Legacy | Environment | 2009 | United Kingdo | J. Wain (HPA,… | Newport | I | legacy | ND | ND |
| ■SAL_AA18… | | | H103540795 | MLST(Legacy | Environment | 2009 | United Kingdo | J. Wain (HPA,… | Newport | I | legacy | ND | ND |
| ■SAL_AA26… | | | H103540746 | MLST(Legacy | Environment | 2009 | United Kingdo | J. Wain (HPA,… | Newport | I | legacy | ND | ND |
| ■SAL_AA26… | | | H103540747 | MLST(Legacy | Environment | 2009 | United Kingdo | J. Wain (HPA,… | Newport | I | legacy | | |
| ■SAL_AA26… | | | H103540739 | MLST(Legacy | Environment | 2009 | United Kingdo | J. Wain (HPA,… | Newport | I | legacy | | |
| ■SAL_AA26… | | | H103540731 | MLST(Legacy | Environment | 2009 | United Kingdo | J. Wain (HPA,… | Newport | I | legacy | | |
| ■SAL_AA26… | | | H103540738 | MLST(Legacy | Environment | 2009 | United Kingdo | J. Wain (HPA,… | Newport | I | legacy | ND | ND |
| ■SAL_AA26… | | | H103560771 | MLST(Legacy | Environment | 2009 | United Kingdo | J. Wain (HPA,… | Newport | I | legacy | ND | ND |
| ■SAL_AA26… | | | H103560770 | MLST(Legacy | Environment | 2009 | United Kingdo | J. Wain (HPA,… | Newport | I | legacy | ND | ND |
| ■SAL_AA26… | | | H103540737 | MLST(Legacy | Environment | 2009 | United Kingdo | J. Wain (HPA,… | Newport | I | legacy | ND | ND |
| ■SAL_AA39… | | | 05-1812 | MLST(Legacy | Human | 2005 | United Kingdo | W. Rabsch (R… | Newport | I | legacy | ND | ND |
| ■SAL_AA39… | | | H103540804 | MLST(Legacy | Environment | 2009 | United Kingdo | J. Wain (HPA,… | Newport | I | legacy | ND | ND |
| ■SAL_AA47… | | | 51960396 | MLST(Legacy | Human | 2005 | United Kingdo | HPA Laborato… | Newport | I | legacy | ND | ND |
| ■SAL_AA55… | | | S05305-06 | MLST(Legacy | Bovine | 2006 | United Kingdo | | Newport | I | legacy | ND | ND |
| ■SAL_BA74… | | | 37777 | SRR1970317 | Human | 8/2014 | England | Gastrointestin… | Newport | I | Assembled | 2.2 | 17 |
| ■SAL_BA75… | | | 72289 | SRR1970245 | | | England | Gastrointestin… | Newport | I | Assembled | 2.2 | 31 |
| ■SAL_BA75… | | | 67553 | SRR1970241 | | | England | Gastrointestin… | Newport | I | Assembled | 2.2 | 40 |
| ■SAL_BA79… | | | 32528 | SRR1970050 | | | United Kingdo | Gastrointestin… | Newport | I | Assembled | 2.2 | 13 |
| ■SAL_BA79… | | | 51296 | SRR1970048 | Human | 9/2014 | London | Gastrointestin… | Newport | I | Assembled | 2.2 | 29 |
| ■SAL_BA81… | | | 61857 | SRR1969963 | Human | 10/2014 | England | Gastrointestin… | Newport | I | Assembled | 2.2 | 18 |
| ■SAL_BA81… | | | 89103 | SRR1969927 | Human | 2/2015 | England | Gastrointestin… | Newport | I | Assembled | | |
| ■SAL_BA82… | | | 65347 | SRR1969889 | Human | 10/2014 | United Kingdo | Gastrointestin… | Newport | I | Assembled | | |
| ■SAL_BA83… | | | 57149 | SRR1969849 | Human | 10/2014 | England | Gastrointestin… | Newport | I | Assembled | | |
| ■SAL_BA85… | | | 83347 | SRR1969769 | Human | 1/2015 | England | Gastrointestin… | Newport | I | Assembled | | |

This data is from the legacy 7 gene MLST database.

These records are from the SRA, with SRA accessions shown

Click 'assembled' to download contigs

Pagination

Alter page size

Page Size: 50 ▾

1-50 | 51-100 | 101-150 | 151-200 | 201-237 |

GoTo: Record ▾

**Figure 6 Search results - Newport strains from the UK**

The exact data will change as new data would have been added since the preparation of this tutorial. However, you should see a number of strain records with metadata on the left pane and experimental data on the right pane.

The first few rows (Figure 6) show records from the legacy MLST database (http://mlst.warwick.ac.uk), and as such they have 'MLST(legacy)' as the data source. These data are derived from Sanger Traces and have no NGS data, so a number of assembly statistics and genotyping information is blank. While EnteroBase shows past MLST data, EnteroBase does not accept new data based on Sanger Traces.

The other rows (at the bottom) are derived from sequenced reads from the SRA. Their status shows they've been assembled and the data Source shows the SRA accession number.

Let's revisit the search function. It may be blank, which means repeat the steps from before to search for strains of serovar Newport from the United Kingdom.

## Saving and loading queries

As you can see, it is a little time consuming to repeatedly enter all the information for complex queries. EnteroBase has a feature where you can save an important query and load it on demand (Figure 7).

**To save the current query for later:**

1. Enter an informative query name e.g. 'Newport_UK' in the text box right of 'Save Current Query' (2).
2. Click the 'Save' button (floppy disk icon with a down arrow)

Now press the 'Clear' button near the bottom right and try loading your query

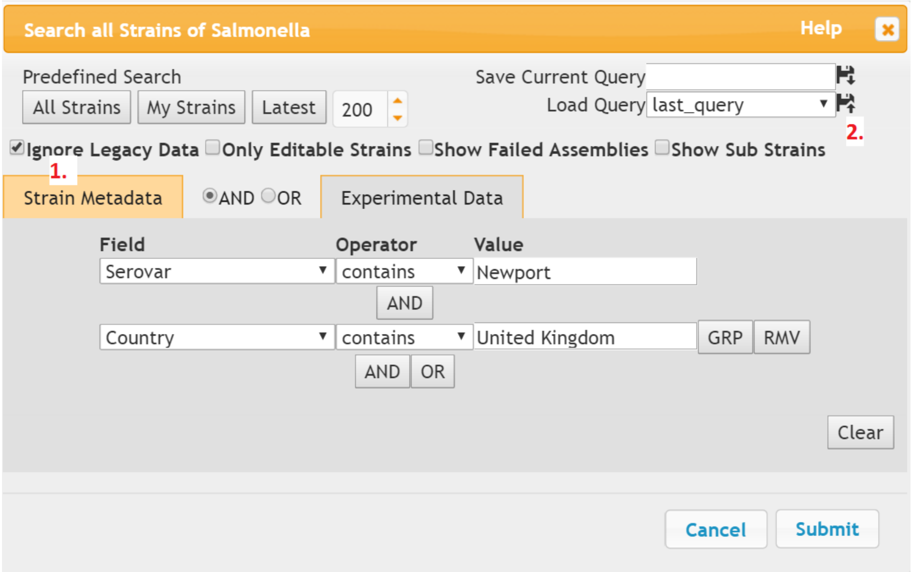
**To load a query:**

1. Click the dropdown, and you should see your query name e.g. 'Newport_UK'. Select it.
2. Click the 'Load button (floppy disk icon with an up arrow)

## Advanced query function

**There are a number of extra options to enhance your searches (Figure 7):**

1. **'Ignore Legacy Data'**: You can hide legacy data by checking this box (1)
2. **'Only Editable Strains'**: You can show strains only you own/or can edit. Usually there will be no results if you haven't uploaded anything to EnteroBase.
3. **'Show failed assemblies'**: Show assemblies that have failed the quality control. These are usually hidden. These strains will not have any genotyping or other analysis run on them, but it maybe useful to check the assembly statistics and download the contigs to see what went wrong.
4. **'Show sub strains'**: Some strains have been grouped together for various reasons (see Section Ueberstrains). These are usually hidden from search results but are shown if this is checked.



200

Save Current Query

Load Query last_query

☑ Ignore Legacy Data ☐ Only Editable Strains ☐ Show Failed Assemblies ☐ Show Sub Strains

● AND ○ OR

| Field | Operator | Value |
|---|---|---|
| Serovar | contains | Newport |
| Country | contains | United Kingdom |

**Figure 7 Advanced query functions**

There are also predefine searches that can be run with one-click, under 'Predefined Search' in the top right:

1. '**All Strains**': Fetches all strains records for the whole database. This can be slow on large databases.
2. '**My Strains**': Fetches strains that belong to you.
3. '**Latest XXX**': By setting a number in the number field, the search will fetch the last X number of strain records entered in the database.

Performing searches based on Experimental data will be discussed in '*Chapter 4: Exploring deeper lineages with MLST, rMLST and cgMLST*' on page 21.

## Workspaces

Workspaces allow you to define a set of strains in a defined group, which you can revisit at any time. Such groups should be considered as a list of strains used in a given study and can be shared with other users (this functionality is covered in the section Buddies on page 10. Workspaces are also the main method to defined input data sets for certain analysis pipelines in EnteroBase.

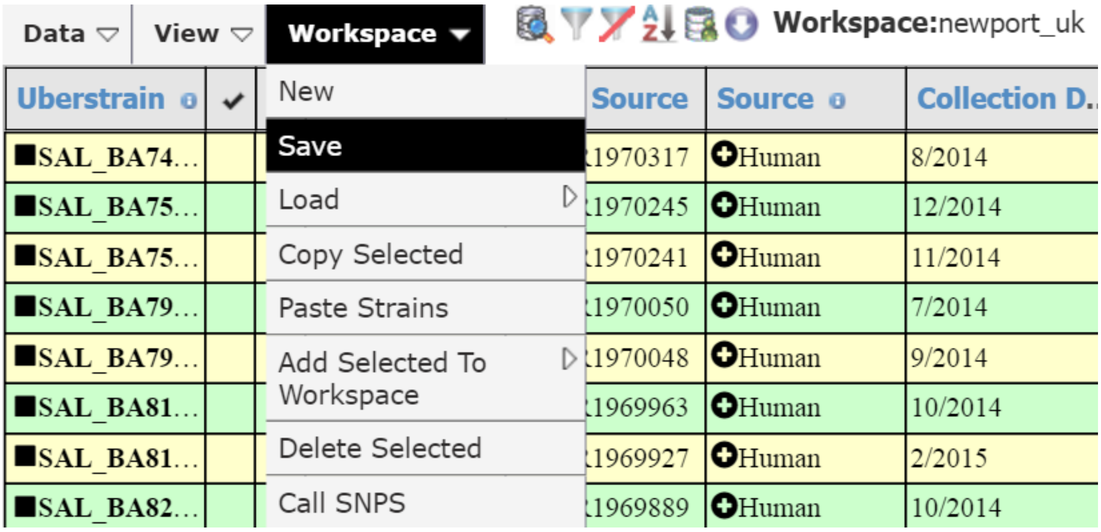
Take an existing set of search results and save a workspace via:

1. From the menu above the strain records, select 'Workspace' > 'Save'
2. Enter a memorable name for this workspace.

To retrieve a workspace:

1. From the menu above the strain records, select 'Workspace' > 'Load'
2. Choose from the names of workspaces.

You can also create a



| Uberstrain | ✓ | Source | Source | Collection D.. |
|---|---|---|---|---|
| SAL_BA74… | | 1970317 | Human | 8/2014 |
| SAL_BA75… | | 1970245 | Human | 12/2014 |
| SAL_BA75… | | 1970241 | Human | 11/2014 |
| SAL_BA79… | | 1970050 | Human | 7/2014 |
| SAL_BA79… | | 1970048 | Human | 9/2014 |
| SAL_BA81… | | 1969963 | Human | 10/2014 |
| SAL_BA81… | | 1969927 | Human | 2/2015 |
| SAL_BA82… | | 1969889 | Human | 10/2014 |

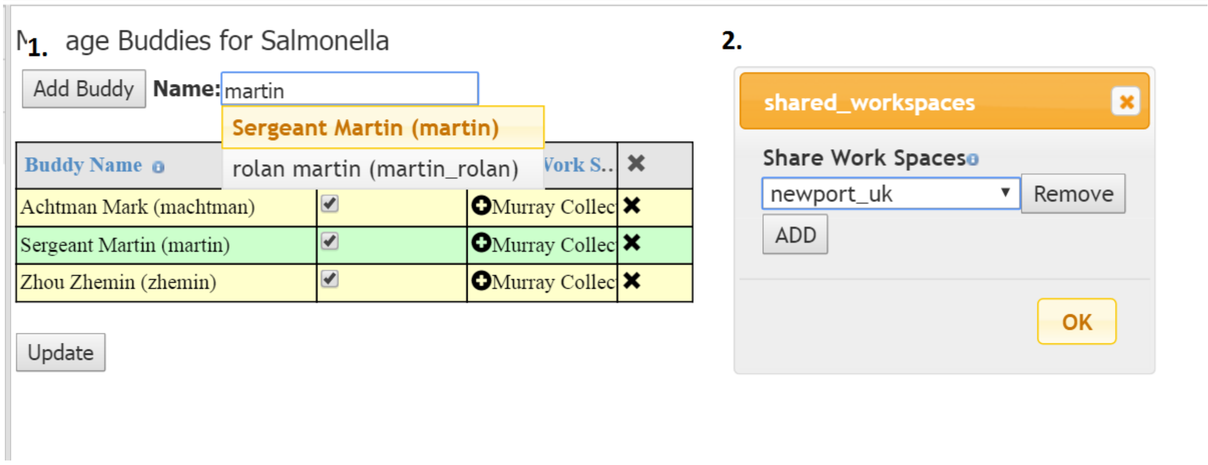
**Figure 8 Workspaces menu**

## Buddies

Editing rights to your data and access to your workspaces can be controlled through the buddies system. Buddies permission control can be accessed through the left hand sidebar: Tasks > My Buddies.

To add a buddy, fill their user name in into the Name text box, it will try to suggest possible users with a drop down (Figure 9)(1). Select the correct User and click 'Add Buddy', this will add them to the table below.

Checking the 'edit strain metadata' box will give this user global access to all your strain records in that database. Clicking on the 'shared workspaces +' will allow you to choose a previously defined workspace from the dropdown to share with that buddy (2). The 'X' deletes the buddy from your list.



M age Buddies for Salmonella

Name: martin

Sergeant Martin (martin)
rolan martin (martin_rolan)

| Buddy Name | | Vork S.. | ✖ |
|---|---|---|---|
| Achtman Mark (machtman) | ☑ | ⊕Murray Collec | ✖ |
| Sergeant Martin (martin) | ☑ | ⊕Murray Collec | ✖ |
| Zhou Zhemin (zhemin) | ☑ | ⊕Murray Collec | ✖ |

shared_workspaces ✖

Share Work Spaces

newport_uk ▾

**Figure 9 Buddy management**

# Chapter 2: Uploading data, correcting, visualising and downloading data.

## Uploading reads

### 1. Adding Reads Manually

Before any reads can be uploaded, metadata concerning the reads needs to be entered into Enterobase

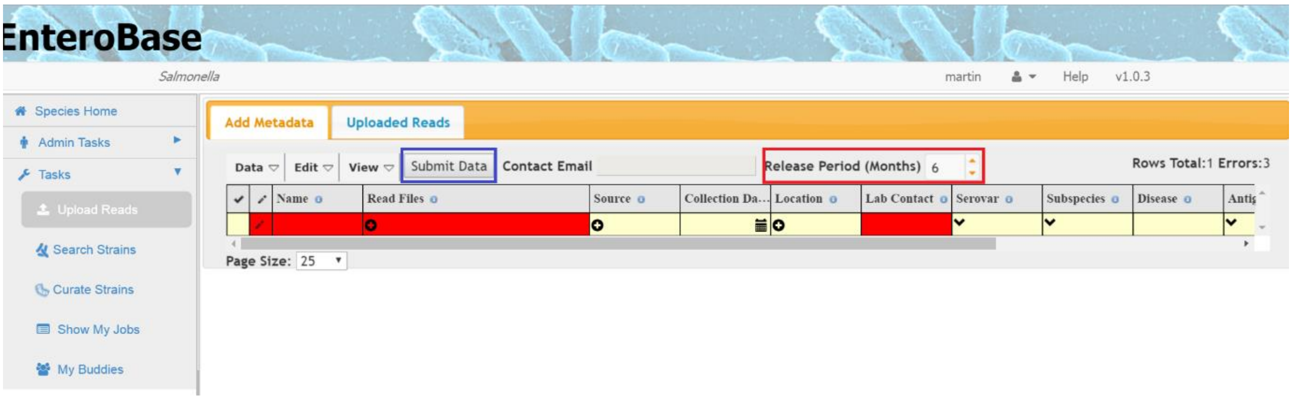
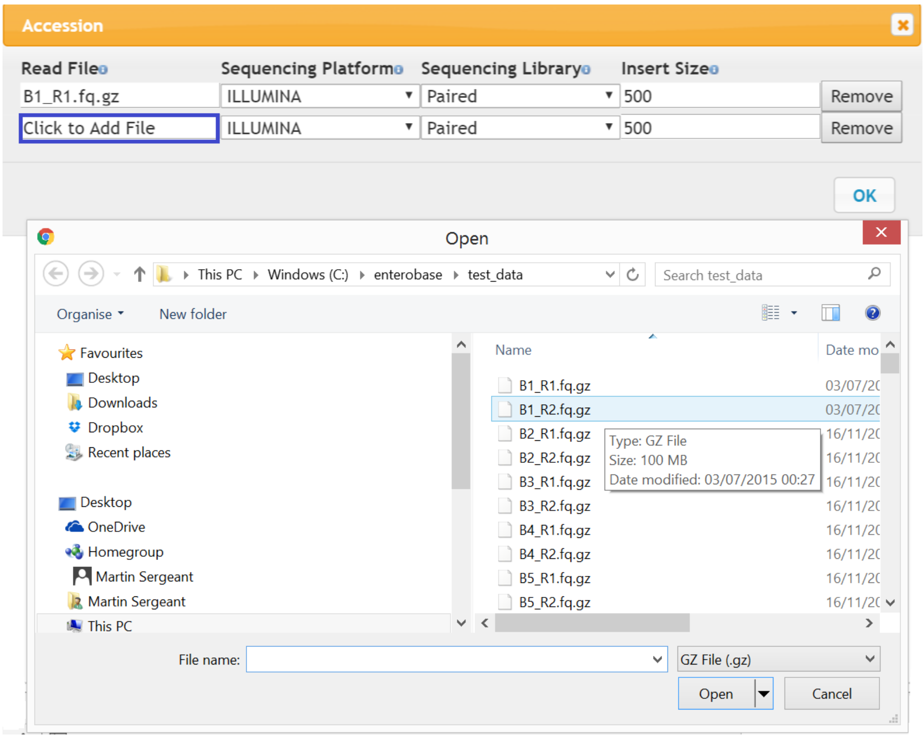


**Figure 10 Initial View of the Upload Read Page**

1. Go the upload reads page (Tasks > Upload Reads in the left hand menu) and you will confronted with a blank row of a table (Figure 10) for you to enter metadata concerning the reads you are going to upload. Any compulsory fields are shown in red. Clicking on individual cells allows you to enter data either by typing directly or selecting from a drop down box, depending on the field. Extra rows can be added by Edit > Add Blank Row or by right clicking on the table and selecting Insert Row.
2. Type in strain name and lab contact (make up suitably witty names)
3. Click on the Read Files cell and a dialog will appear (Figure 11). By default the reads are Illumina, paired with an insert size of 500. Click on Click to Add File (blue box in Figure 11) and a file dialog will appear, select a read from a local directory (In this case sal1_R1.fastq.gz and sal1_R2.fastq.gz) and press OK



Accession

| Read File | Sequencing Platform | Sequencing Library | Insert Size | |
|---|---|---|---|---|
| B1_R1.fq.gz | ILLUMINA | Paired | 500 | Remove |
| Click to Add File | ILLUMINA | Paired | 500 | Remove |

OK

Open

This PC ▸ Windows (C:) ▸ enterobase ▸ test_data

Search test_data

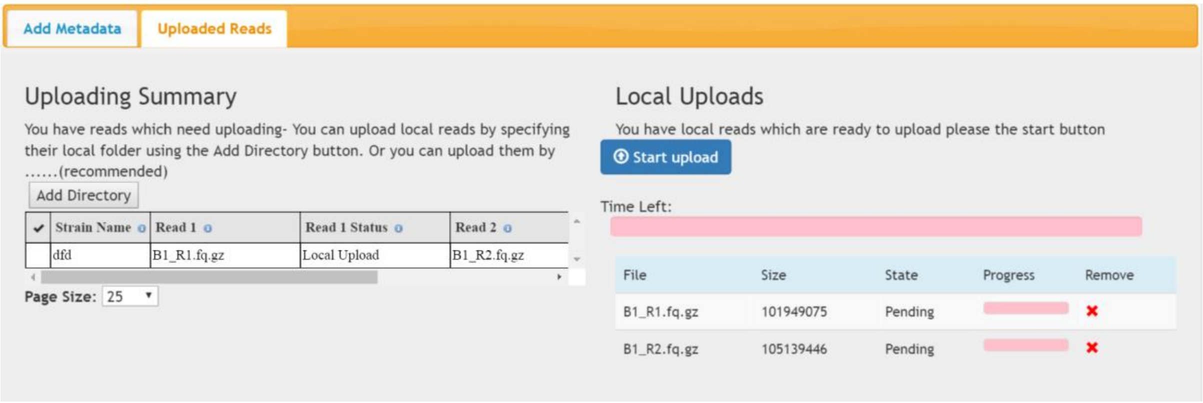
| Name | Date mo |
|---|---|
| B1_R1.fq.gz | 03/07/2( |
| B1_R2.fq.gz | 03/07/2( |
| B2_R1.fq.gz | 16/11/2( |
| B2_R2.fq.gz | 16/11/2( |
| B3_R1.fq.gz | 16/11/2( |
| B3_R2.fq.gz | 16/11/2( |
| B4_R1.fq.gz | 16/11/2( |
| B4_R2.fq.gz | 16/11/2( |
| B5_R1.fq.gz | 16/11/2( |
| B5_R2.fq.gz | 16/11/2( |

Type: GZ File
Size: 100 MB
Date modified: 03/07/2015 00:27

File name:
GZ File (.gz)

**Figure 11 Specifying Read Files**

4. When all the data has been filled in correctly, the Submit Data button will become active and can be pressed (blue box in Figure 10)
5. Once the metadata has been submitted, you will be taken to the Uploaded Reads tab and information about tour reads will be in the left hand table (Figure 12) with the status 'Local Upload'. The reads should also appear on the right hand side with the status pending. The Start Upload button can be pressed and the progress bars will show you the state of the upload. Once all files have been uploaded you will be informed by a dialog box



Uploading Summary

You have reads which need uploading- You can upload local reads by specifying their local folder using the Add Directory button. Or you can upload them by ......(recommended)

| ✓ | Strain Name | Read 1 | Read 1 Status | Read 2 |
|---|---|---|---|---|
| | dfd | B1_R1.fq.gz | Local Upload | B1_R2.fq.gz |

Page Size: 25

Local Uploads

You have local reads which are ready to upload please the start button

Time Left:

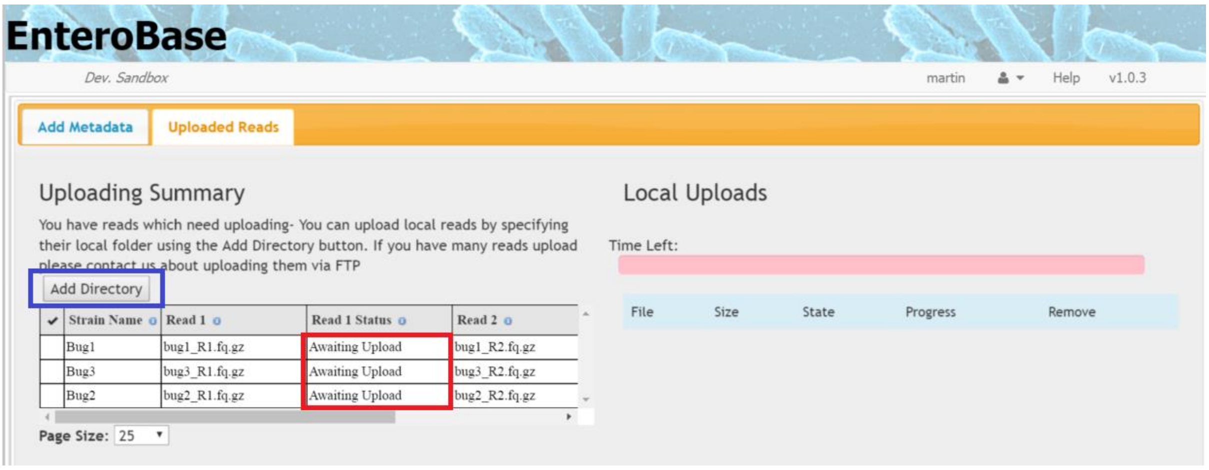
| File | Size | State | Progress | Remove |
|---|---|---|---|---|
| B1_R1.fq.gz | 101949075 | Pending | | ✗ |
| B1_R2.fq.gz | 105139446 | Pending | | ✗ |

**Figure 12 The Uploaded Reads Tab of the Uploads Page**

### 2. Adding Reads from a File

If you have many reads to upload it may be easier to add the Metadata from a file generated in excel or using a script. The file needs to be text delimited and a template is available at Enterobase.warwick.ac.uk/static/example/upload_template.txt

1. Open the file upload_example.txt in excel and add (or change) data in some of the columns, then save it (being sure to save it as tab delimited text)
2. Go to data > Load Local File and open the file. The data should appear in the table. If there are errors, mouse over the red cells to see what the problem is. When all errors have been corrected, press Submit Data.
3. This time, read status in the left hand table will be 'Awaiting Upload' (Figure 13 red box) and the files will not have appeared in the right hand panel, because you have not specified their location in the spread sheet. Even if you had, for security reasons browsers insist that you physically click on the file or a directory before it can take control of them. Therefore, you must either press the Add Files button (blue box Figure 13 to specify the directory where the files reside or drag and drop the files onto the right hand area of the screen.



Uploading Summary

You have reads which need uploading- You can upload local reads by specifying their local folder using the Add Directory button. If you have many reads upload please contact us about uploading them via FTP

| ✔ | Strain Name | Read 1 | Read 1 Status | Read 2 |
|---|---|---|---|---|
| | Bug1 | bug1_R1.fq.gz | Awaiting Upload | bug1_R2.fq.gz |
| | Bug3 | bug3_R1.fq.gz | Awaiting Upload | bug3_R2.fq.gz |
| | Bug2 | bug2_R1.fq.gz | Awaiting Upload | bug2_R2.fq.gz |

Local Uploads

Time Left:

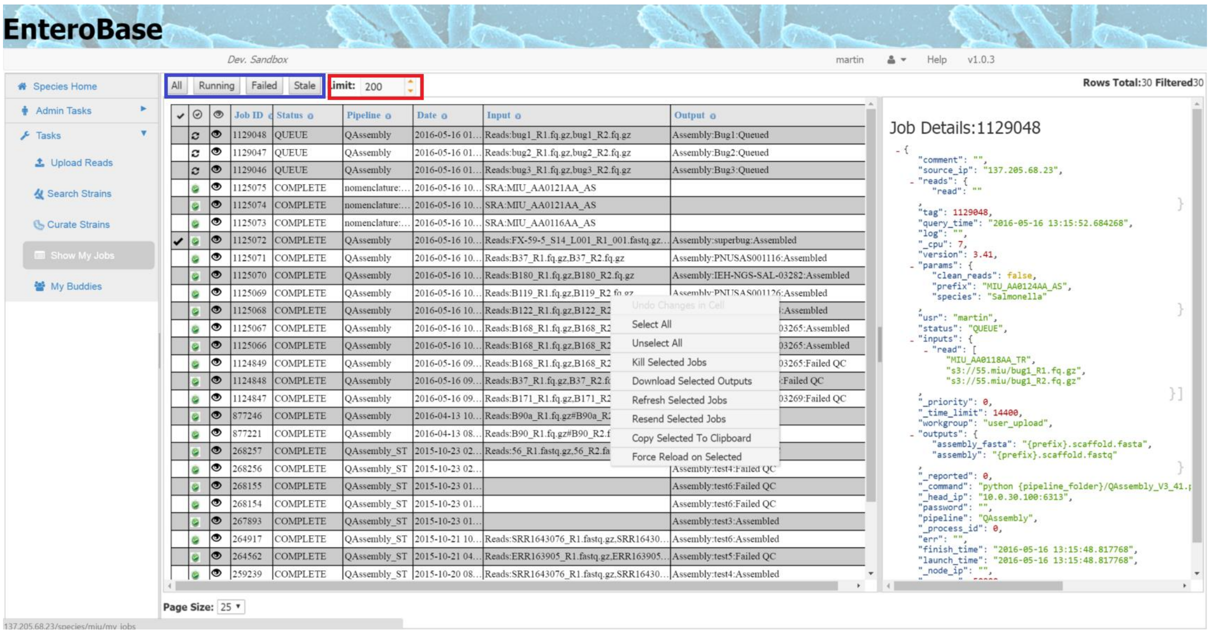
**Figure 10 Reads Awaiting Upload**

4. Once the files have been identified they will appear on the right hand screen and they can uploaded as in the previous example

> *Notes*
> The Release Period input (Figure 10 red box) specifies the amount of time in months that your reads will remain private. This means that assemblies (and annotation files) cannot be downloaded by others unless you give them permission.

## Viewing job progress

The progress of your jobs can be views by Tasks>Show My Jobs (see Figure 14)



Limit: 200

Rows Total:30 Filtered30

| ✓ | ⊙ | 👁 | Job ID | Status | Pipeline | Date | Input | Output |
|---|---|---|---|---|---|---|---|---|
|  | ↻ | 👁 | 1129048 | QUEUE | QAssembly | 2016-05-16 01… | Reads:bug1_R1.fq.gz,bug1_R2.fq.gz | Assembly:Bug1:Queued |
|  | ↻ | 👁 | 1129047 | QUEUE | QAssembly | 2016-05-16 01… | Reads:bug2_R1.fq.gz,bug2_R2.fq.gz | Assembly:Bug2:Queued |
|  | ↻ | 👁 | 1129046 | QUEUE | QAssembly | 2016-05-16 01… | Reads:bug3_R1.fq.gz,bug3_R2.fq.gz | Assembly:Bug3:Queued |
|  | ✔ | 👁 | 1125075 | COMPLETE | nomenclature:… | 2016-05-16 10… | SRA:MIU_AA0121AA_AS |  |
|  | ✔ | 👁 | 1125074 | COMPLETE | nomenclature:… | 2016-05-16 10… | SRA:MIU_AA0121AA_AS |  |
|  | ✔ | 👁 | 1125073 | COMPLETE | nomenclature:… | 2016-05-16 10… | SRA:MIU_AA0116AA_AS |  |
| ✓ | ✔ | 👁 | 1125072 | COMPLETE | QAssembly | 2016-05-16 10… | Reads:FX-59-5_S14_L001_R1_001.fastq.gz… | Assembly:superbug:Assembled |
|  | ✔ | 👁 | 1125071 | COMPLETE | QAssembly | 2016-05-16 10… | Reads:B37_R1.fq.gz,B37_R2.fq.gz | Assembly:PNUSAS001116:Assembled |
|  | ✔ | 👁 | 1125070 | COMPLETE | QAssembly | 2016-05-16 10… | Reads:B180_R1.fq.gz,B180_R2.fq.gz | Assembly:IEH-NGS-SAL-03282:Assembled |
|  | ✔ | 👁 | 1125069 | COMPLETE | QAssembly | 2016-05-16 10… | Reads:B119_R1.fq.gz,B119_R2.fq.gz | Assembly:PNUSAS001126:Assembled |
|  | ✔ | 👁 | 1125068 | COMPLETE | QAssembly | 2016-05-16 10… | Reads:B122_R1.fq.gz,B122_R2 | :Assembled |
|  | ✔ | 👁 | 1125067 | COMPLETE | QAssembly | 2016-05-16 10… | Reads:B168_R1.fq.gz,B168_R2 | 03265:Assembled |
|  | ✔ | 👁 | 1125066 | COMPLETE | QAssembly | 2016-05-16 10… | Reads:B168_R1.fq.gz,B168_R2 | 03265:Assembled |
|  | ✔ | 👁 | 1124849 | COMPLETE | QAssembly | 2016-05-16 09… | Reads:B168_R1.fq.gz,B168_R2 | 03265:Failed QC |
|  | ✔ | 👁 | 1124848 | COMPLETE | QAssembly | 2016-05-16 09… | Reads:B37_R1.fq.gz,B37_R2.fq | :Failed QC |
|  | ✔ | 👁 | 1124847 | COMPLETE | QAssembly | 2016-05-16 09… | Reads:B171_R1.fq.gz,B171_R2 | 03269:Failed QC |
|  | ✔ | 👁 | 877246 | COMPLETE | QAssembly | 2016-04-13 10… | Reads:B90a_R1.fq.gz#B90a_R2 |  |
|  | ✔ | 👁 | 877221 | COMPLETE | QAssembly | 2016-04-13 08… | Reads:B90_R1.fq.gz#B90_R2.f |  |
|  | ✔ | 👁 | 268257 | COMPLETE | QAssembly_ST | 2015-10-23 02… | Reads:56_R1.fastq.gz,56_R2.fa |  |
|  | ✔ | 👁 | 268256 | COMPLETE | QAssembly_ST | 2015-10-23 02… |  | Assembly:test4:Failed QC |
|  | ✔ | 👁 | 268155 | COMPLETE | QAssembly_ST | 2015-10-23 01… |  | Assembly:test6:Failed QC |
|  | ✔ | 👁 | 268154 | COMPLETE | QAssembly_ST | 2015-10-23 01… |  | Assembly:test6:Failed QC |
|  | ✔ | 👁 | 267893 | COMPLETE | QAssembly_ST | 2015-10-23 01… |  | Assembly:test3:Assembled |
|  | ✔ | 👁 | 264917 | COMPLETE | QAssembly_ST | 2015-10-21 10… | Reads:SRR1643076_R1.fastq.gz,SRR16430… | Assembly:test6:Assembled |
|  | ✔ | 👁 | 264562 | COMPLETE | QAssembly_ST | 2015-10-21 04… | Reads:ERR163905_R1.fastq.gz,ERR163905… | Assembly:test5:Failed QC |
|  | ✔ | 👁 | 259239 | COMPLETE | QAssembly_ST | 2015-10-20 08… | Reads:SRR1643076_R1.fastq.gz,SRR16430… | Assembly:test4:Assembled |

Page Size: 25

Job Details:1129048

```
- {
    "comment": "",
    "source_ip": "137.205.68.23",
  - "reads": {
      "read": ""
                                                    }
    ,
    "tag": 1129048,
    "query_time": "2016-05-16 13:15:52.684268",
    "log": "",
    "_cpu": 7,
    "version": 3.41,
  - "params": {
      "clean_reads": false,
      "prefix": "MIU_AA0124AA_AS",
      "species": "Salmonella"
                                                    }
    ,
    "usr": "martin",
    "status": "QUEUE",
  - "inputs": {
    - "read": [
        "MIU_AA0118AA_TR",
        "s3://55.miu/bug1_R1.fq.gz",
        "s3://55.miu/bug1_R2.fq.gz"
                                                   }]
    ,
    "_priority": 0,
    "_time_limit": 14400,
    "workgroup": "user_upload",
  - "outputs": {
      "assembly_fasta": "{prefix}.scaffold.fasta",
      "assembly": "{prefix}.scaffold.fastq"
                                                    }
    ,
    "_reported": 0,
    "_command": "python {pipeline_folder}/QAssembly_V3_41.p
    "_head_ip": "10.0.30.100:6313",
    "password": "",
    "pipeline": "QAssembly",
    "_process_id": 0,
    "err": "",
    "finish_time": "2016-05-16 13:15:48.817768",
    "launch_time": "2016-05-16 13:15:48.817768",
    "_node_ip": "",
```

**Figure 14 Monitoring Jobs**

By default, the last 200 jobs are shown in all states. To show more jobs, increase the value in the Limit input (Figure 14 red box) and press one of the buttons in the blue box:

All - Shows all jobs

Running - Shows jobs in wait resource, queue or running state

Failed – Shows jobs which are failed or have been killed

Stale - Shows jobs which have been running longer than a day

By clicking on the eye, information about the job is displayed in the right hand pane (although this is not very readable). When the page first loads information about the top job in the table will be displayed in this pane.

Usually all jobs are run automatically, once your assembly job has finished, nomenclature and any other jobs that act upon the assembly will be initiated. In addition every hour all databases are automatically checked and any outstanding or failed jobs are resent. However, in certain circumstance jobs can be manipulated by right clicking on the grid which brings up the context menu shown in Figure 14.

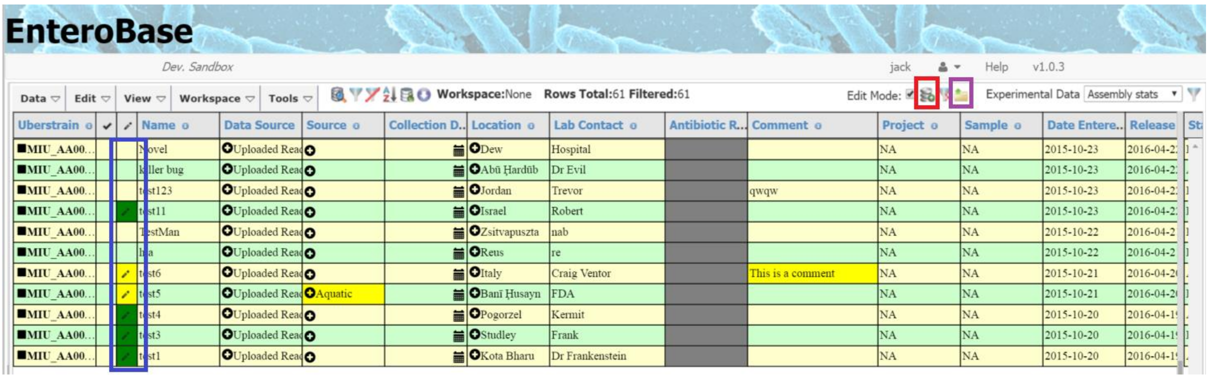
Refresh Selected Jobs – Jobs go through four states Wait Resource > Queued > Running >Complete/Failed. The server is informed only when a job has completed or fails. Therefore refreshing a job will update its intermediate status in the grid

Resend Selected Jobs – This will resend any selected jobs which have failed

Force Reload on Selected - If there has been major network issues or the server was temporarily down, the server may be unaware that a job has. In this case the job status in the grid will be Running/Queued but the information on the right hand pane will show the job has been completed. Under these circumstances, forcing a reload will ensure that the job is processed

## Editing metadata

1. Load all the records in the test database and click on the Edit mode check box. You will get a dialog with some information – click OK. Records with a pencil icon (Figure 15 blue box) show that you have permission to edit the metadata (In this case, it will only be the four strain you have uploaded). Click on any cell and alter its content and the cell should turn yellow. Once you have made all the changes you require you need to upload these changes to the database (press the upload changes icon – red box in Figure 15) or right click on  row containing an edited cell and select Upload Changes in Row.



| Uberstrain | ✓ | ✎ | Name | Data Source | Source | Collection D.. | Location | Lab Contact | Antibiotic R... | Comment | Project | Sample | Date Entere.. | Release |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MIU_AA00... | | | Novel | Uploaded Read | | | Dew | Hospital | | | NA | NA | 2015-10-23 | 2016-04-2 |
| MIU_AA00... | | | killer bug | Uploaded Read | | | Abū Ḩardūb | Dr Evil | | | NA | NA | 2015-10-23 | 2016-04-2 |
| MIU_AA00... | | | test123 | Uploaded Read | | | Jordan | Trevor | | qwqw | NA | NA | 2015-10-23 | 2016-04-2 |
| MIU_AA00... | | ✎ | test11 | Uploaded Read | | | Israel | Robert | | | NA | NA | 2015-10-23 | 2016-04-2 |
| MIU_AA00... | | | TestMan | Uploaded Read | | | Zsitvapuszta | nab | | | NA | NA | 2015-10-22 | 2016-04-2 |
| MIU_AA00... | | | ha | Uploaded Read | | | Reus | re | | | NA | NA | 2015-10-22 | 2016-04-2 |
| MIU_AA00... | | ✎ | test6 | Uploaded Read | | | Italy | Craig Ventor | | This is a comment | NA | NA | 2015-10-21 | 2016-04-2 |
| MIU_AA00... | | ✎ | test5 | Uploaded Read | Aquatic | | Banī Ḩusayn | FDA | | | NA | NA | 2015-10-21 | 2016-04-2 |
| MIU_AA00... | | ✎ | test4 | Uploaded Read | | | Pogorzel | Kermit | | | NA | NA | 2015-10-20 | 2016-04-1 |
| MIU_AA00... | | ✎ | test3 | Uploaded Read | | | Studley | Frank | | | NA | NA | 2015-10-20 | 2016-04-1 |
| MIU_AA00... | | ✎ | test1 | Uploaded Read | | | Kota Bharu | Dr Frankenstein | | | NA | NA | 2015-10-20 | 2016-04-1 |

**Figure 15 Editing Metadata**

After updating changes the cells should turn back to their normal colour and a dialog will inform you whether the update has been successful.

A Search and Replace (Edit >Search and Replace) and undo function (Edit > Undo or ctrl+Z) are available but for large scale editing it may be easier to load the data into excel.

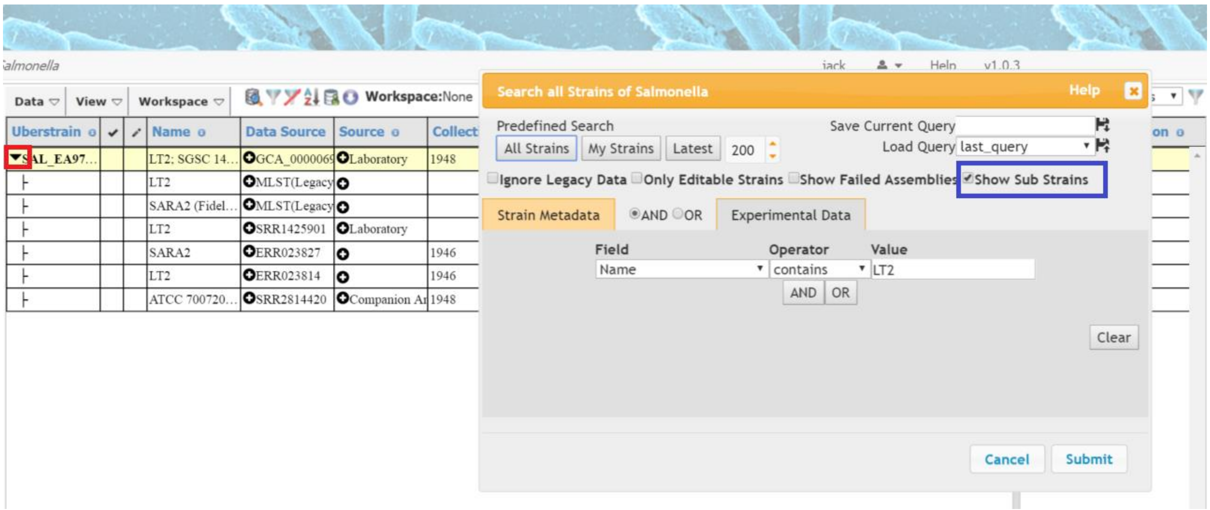
2. Click on the 'My strains Icon' or Data > My strains or and the four strains you uploaded previously should be present – all of which you have editing permission. Then save them to a local, file Data > Save to Local File. Open this file in excel and change some values. Then re-save the file (making sure it is in tab delimited text). Reload the file into Enterobase, by first clicking on the edit mode check box and then clicking The Load Modified File icon (purple box in Figure 15). Any changes or errors in your modified file will be shown as yellow or red cells.

In edit mode you can send jobs / assemblies on any strains that you have editing rights to by using Tools > Assemble Selected or Tools>Call Scheme for Selected

## Ueberstrains

You may have noticed the Ueberstrain column at the very left of the strains table and wondered what it is all about. Certain records are duplicated in that there are many entries for what is essentially the same strain. This can skew analysis because your analysis may produce false clusters, which in reality are just the same strain. Thus in Enterobase such entries are 'merged' and a single Ueberstrain is created. Normally only the Ueberstrain is shown and so you do not need to worry about this de-replication.

However you can still examine the sub-strains associated with an Ueberstrain (see below)



| Uberstrain | ✓ | | Name | Data Source | Source | Collect |
|---|---|---|---|---|---|---|
| ▼SAL_EA97… | | | LT2; SGSC 14… | GCA_0000069 | Laboratory | 1948 |
| ├ | | | LT2 | MLST(Legacy | | |
| ├ | | | SARA2 (Fidel… | MLST(Legacy | | |
| ├ | | | LT2 | SRR1425901 | Laboratory | |
| ├ | | | SARA2 | ERR023827 | | 1946 |
| ├ | | | LT2 | ERR023814 | | 1946 |
| ├ | | | ATCC 700720… | SRR2814420 | Companion An | 1948 |

Predefined Search: 200

Save Current Query

Load Query: last_query

□ Ignore Legacy Data □ Only Editable Strains □ Show Failed Assemblies ☑ Show Sub Strains

● AND ○ OR

| Field | Operator | Value |
|---|---|---|
| Name | contains | LT2 |

**Figure 16 Uberstrains**

1. Go the Salmonella database and search for LT2, make sure that the Show Sub Strains box in checked in the query dialog (Figure 16 blue box). A single record should load but it will have an expand icon in the Ueberstrain column (Figure 16 red box). Clicking on this icon will show all the sub strains associated with this master strain. The master strain is usually the most complete (in this case it is the complete closed genome)

## Downloading data

The current data can always be saved to file by Data > Save To Local File. Because of browser restrictions, the data is actually treated as a download so it may probably end up in the downloads directory that your browser uses. Some browsers however let you choose the location. The file is just a tab delimited text file, which can be opened in any spread sheet. The file will contain all the strain metadata and any associated experimental data in the right hand pane. For large schemes, this data is not very useful therefore Enterobase enables you to download all the allele information separately

(1)In the Salmonella database query on serovar equals Dublin. By default the experimental data will be assembly stats. In the Experimental Data dropdown choose cgMLST (3020). The data in the right hand panel with then show the cgMLST ST for each record, which is not very useful (but then would you want to look a 3020 columns of allele numbers). Right Click on the right hand panel and select 'Save all'. A dialog will appear showing the progress of retrieving the information (100 records are obtained at one time). Once the data has been retrieved, it can be saved as with any other file (type a file name in the text box and press save)

# Chapter 3: Outbreak! Applying EnteroBase to real data

## Search an outbreak strain

In the Salmonella database, search for name equals CFSAN037834. Let's pretend we have just uploaded this strain and want to find out if it part of an outbreak

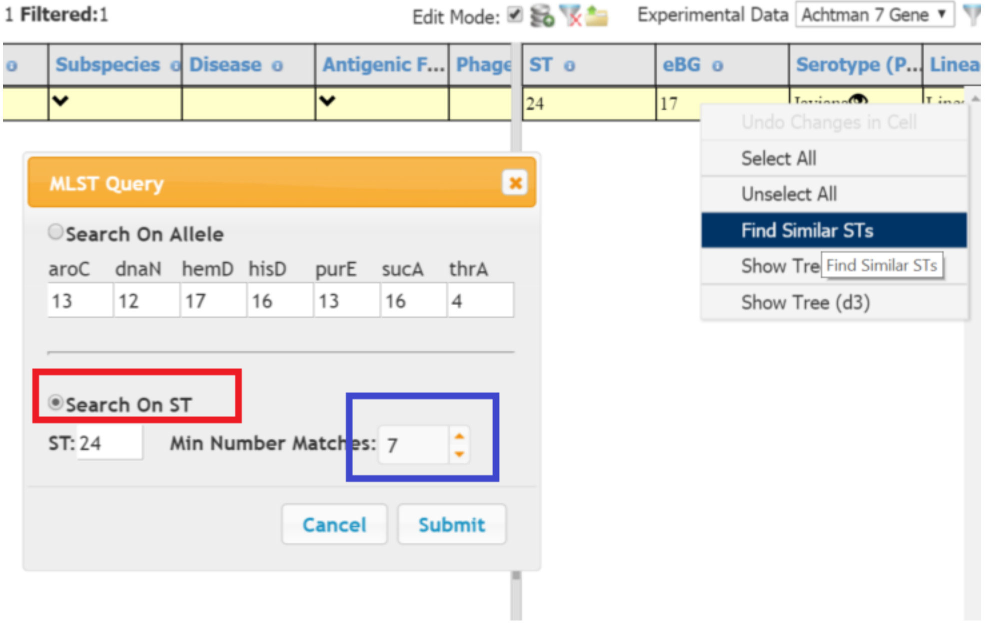


**Figure 17 Searching For Similar STs**

## Finding NearestNeighbours

(1)In Experimental Data dropdown choose Achtman 7 Gene. Right click on the row in the experimental pane and select Find Similar STs in the context menu

(2)In the dialog be sure that search on ST is checked (Figure 17 red box). In this case we are finding complete matches i.e. 7 out of 7 alleles (blue box) must match, however we can be less stringent and reduce this number , which we will want to do in larger schemes where there are more alleles (e.g. cgMLST and rMLST).Press subit

(3) Over 400 strains will be loaded into the browser. Now save this Workspace > save. Call it MLST_Achtman_ST24 or something else equally descriptive.

## Create a Minimum Spanning Tree

(4)Once the workspace has been saved an Analysis tab should appear on the top menu. Before we create the MS tree we need to change the experimental data to cgMLST (3020)

(5) Next Analysis > Create MS tree. Make sure the scheme cgMLST is specified else your tree will be built on 7 gene MLST data and contain just one node. Call the tree ST_24 and press submit, a new window should open with a Waiting dialog. If this is not the case then check that your browser allows pop ups from this site. The tree can be reloaded by Analysis> Load MS Tree > *Tree_name*

(6) Once the tree has loaded, the first thing we are going to do is find our original strain. In the main window we want to filter our data wither the filter icon (funnel) or view > Filter Data and select name and type (or paste) CFSAN037834 in the text box

(7) A single row should be visible. Select this row and then return to the Tree window and press highlight selected (Figure 18 – 8), the largest node should turn yellow

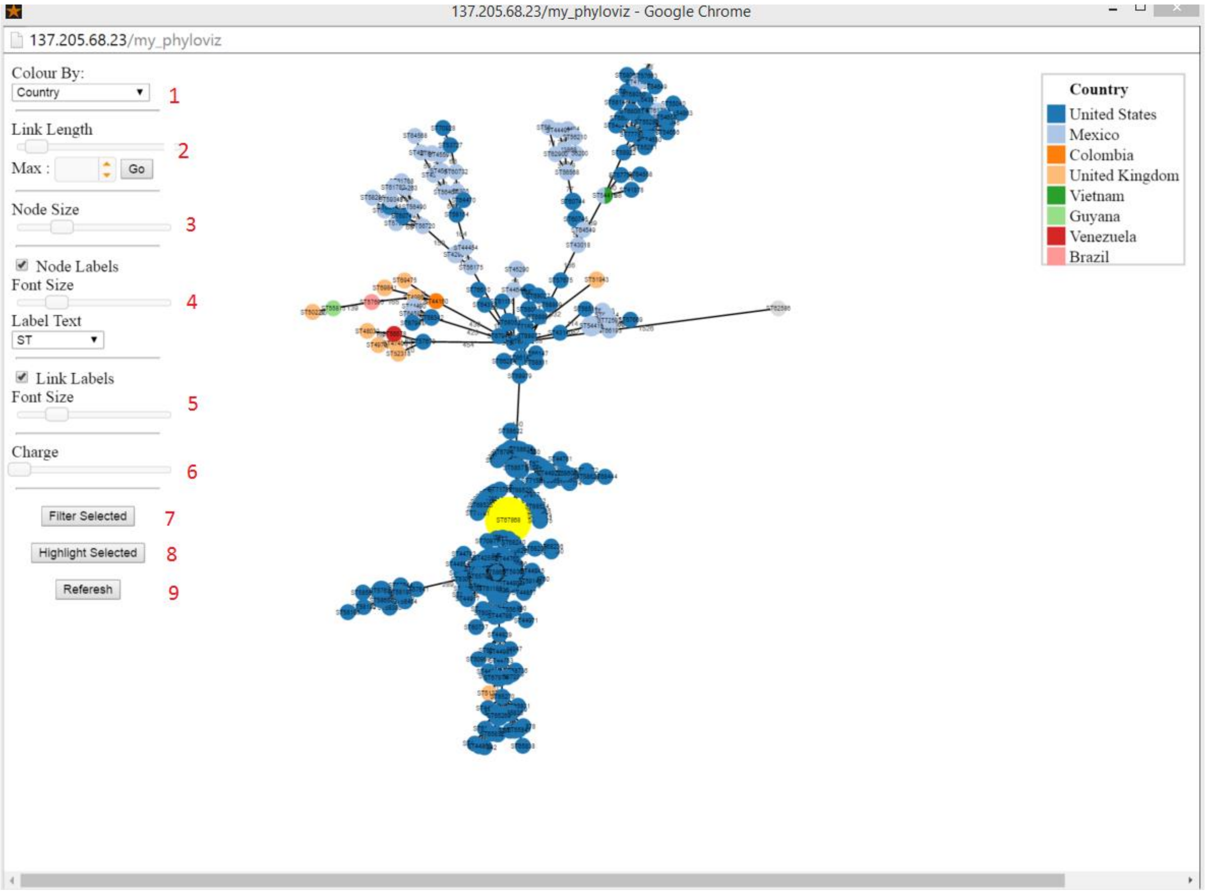
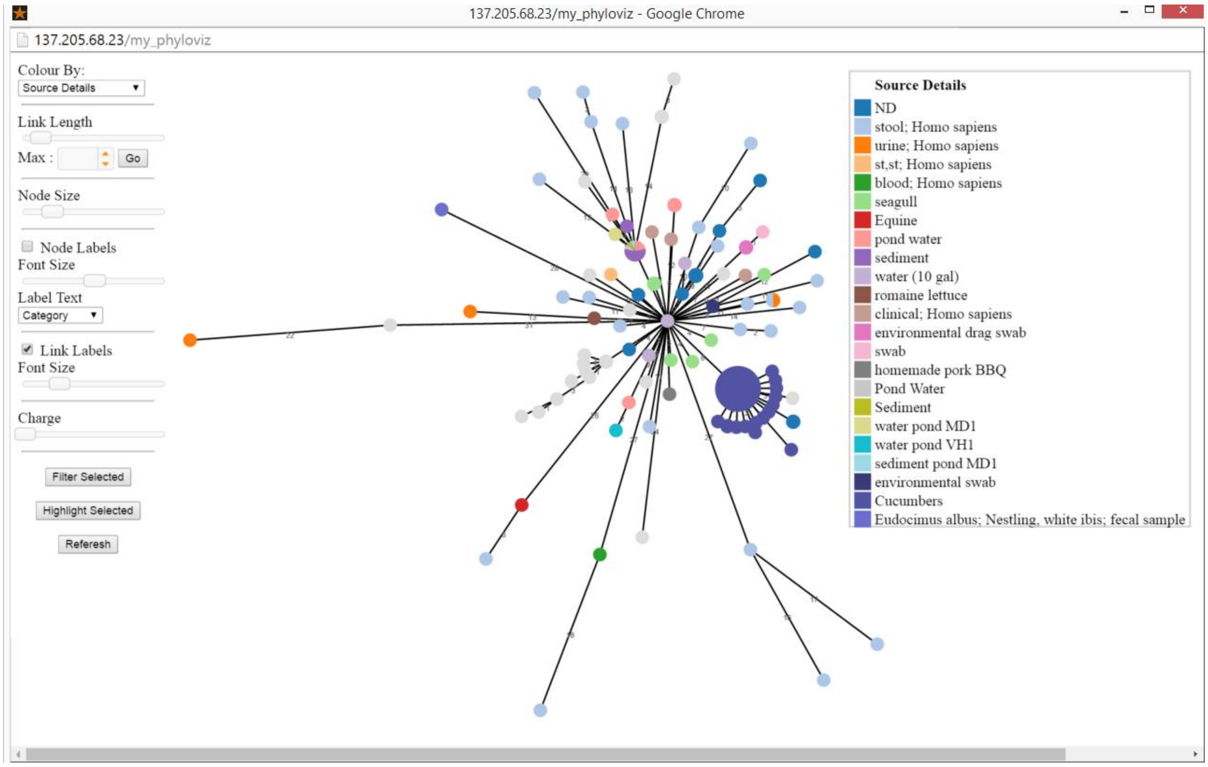


**Figure 18 Minimal Spanning Tree**

(8)In this case we are going to create a sub tree based on the closest neighbours of our original strain. Press shift and drag the mouse to select those nodes nearest the one of interest. These nodes will turn red. Then press Filter Selected

(9)You will see that on the main page the table has been filtered to contain only those strains you have selected. Create another tree Analysis > Create Tree and give it an appropriate name

(10)Analyse the data further by changing the Colour By dropdown (Figure 18 -1) to different categories (Figure 19)



137.205.68.23/my_phyloviz - Google Chrome

137.205.68.23/my_phyloviz

Colour By: Source Details

Link Length

Max : Go

Node Size

□ Node Labels

Font Size

Label Text: Category

☑ Link Labels

Font Size

Charge

**Figure 19 A More Detailed sub Tree**

By using the mouse wheel, you can zoom in and out of the tree and by dragging outside of any nodes, the position of the tree can be altered. Nodes can be coloured by metadata (Figure 18 -1) and labels shown hidden etc. as well as increasing/decreasing the length of the links or size of the nodes

Link length is proportional to the number of differences between nodes. You can set a maximum using the max input underneath the Link length slider (Figure 18-2) such that all links which have a greater value than that specified will be the same length (but the link will be a thin grey line to show that it is not proportional). This is helpful if you want to look at two or more clusters that are separated by long linkages.

The charge slider (Figure 18 -9) controls the repulsion between nodes, the default value is around 3. It can be decreased to 0, which will make the tree more stable, but nodes will tend to overlap more. Conversely it can be increased, which will separate clusters of nodes and make the tree easier to visualize, but be aware that the length of links may no longer be proportional to number of allele differences.

## Chapter 4: Exploring deeper lineages with MLST, rMLST and cgMLST

In this chapter we explore some of the broader concepts behind EnteroBase and how these can be used to get the most out of EnteroBase. One of the unique selling points of EnteroBase is that it provides a global overview of an entire genus. Allowing you to see where you strain sits within the entire population. To effectively deal with such large datasets, however, require some degree of abstraction which we will introduce here.

### Thinking about classifying a bacterial population

Typing methods based around antigenicity, pathotyping and other typing methods, some of which are the de jure standard in many reference labs, do not always correlate with the relativity of individual strains. Consider the presence of the Shiga toxin genes in Enterohaemorhaggic *E. coli,* where Shiga toxin positive *E. coli* is found in all phylogroups across the population. The designation of Enterohaemorhaggic is ultimately one of clinical manifestation rather than suggesting any shared ancestry between such strains. Likewise *Salmonella enterica* serovar Newport is made of multiple discrete lineages and to treat it as uniform is misleading.

In analyses attempting to place strains within a population, it makes sense to use a neutral set of markers from across the genome. This is the motivation behind MLST. However, classical MLST is limited in its discriminant power, as it only focuses on a handful of genes. The solution in this case is to increase the number of genes, or use SNPs, as the informative sites.

| | |
|---|---|
| ↓<br>Increasing discrimination | eBURST Group (eBG) |
| | Sequence type (MLST) |
| | Ribosomal MLST eBG |
| | Ribosomal MLST ST |
| | Core genome MLST |

Within EnteroBase we extend each species from classical MLST, rMLST, to core genome MLST.

| **MLST – Classic** | **Ribosomal MLST** | **Core Genome MLST** |
|---|---|---|
| ~5-15 Loci | 52 Loci | ~ 3000-4000 for Salmonella |
| Conserved Housekeeping genes | Ribosomal proteins | Any conserved coding sequence |
| Highly conserved;<br>Low resolution | Highly conserved;<br>Medium resolution | Variable;<br>High resolution |
| Different scheme for each<br>Species/genus | Single scheme across tree of life* | Different scheme for each<br>Species/genus |

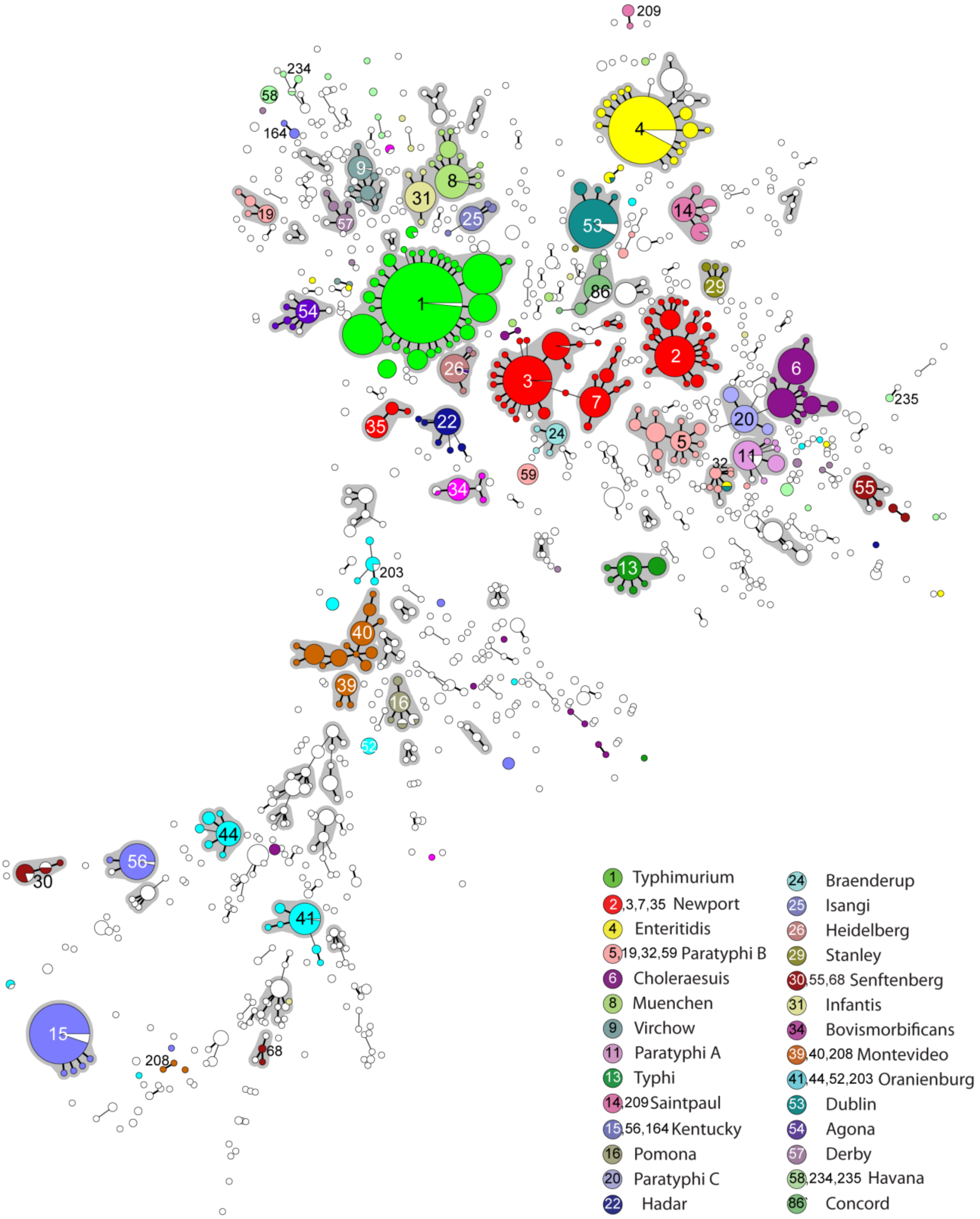


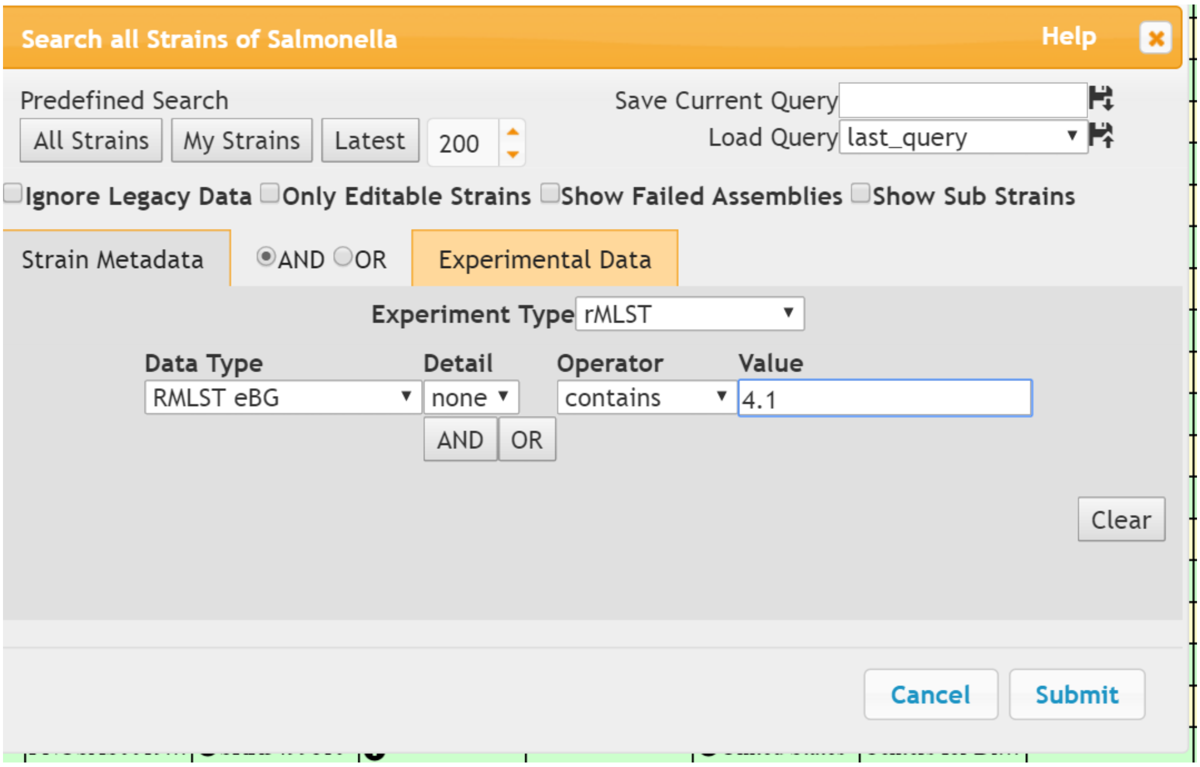
**Figure 11 Minimal spanning tree (MSTree) of MLST data on 4257 isolates of *S. enterica* subspecies enterica. From Achtman et al. (2012) PLoS Pathog 8(6): e1002776.**

## Searching deeper within clonal complexes

EnteroBase currently supports a number of population clustering approaches:

| | |
|---|---|
| MLST | eBG |
| rST | rEBG |
| cgMLST | Serovar predictions |

These methods can be searched through the Experimental Data tab on the search. The example below shows how to search rMLST eBG '4.1' which corresponds to a sub-lineage within *Salmonella* serovar Enteritidis.



Search all Strains of Salmonella

Predefined Search: 200

Save Current Query:

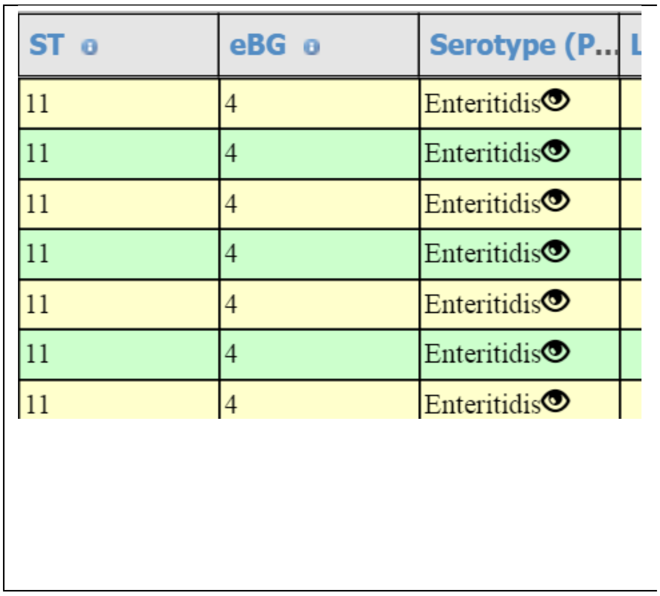
Load Query: last_query

□ Ignore Legacy Data □ Only Editable Strains □ Show Failed Assemblies □ Show Sub Strains

● AND ○ OR

Experiment Type: rMLST

| Data Type | Detail | Operator | Value |
|---|---|---|---|
| RMLST eBG | none | contains | 4.1 |



| ST | eBG | Serotype (P... |
|---|---|---|
| 11 | 4 | Enteritidis👁 |
| 11 | 4 | Enteritidis👁 |
| 11 | 4 | Enteritidis👁 |
| 11 | 4 | Enteritidis👁 |
| 11 | 4 | Enteritidis👁 |
| 11 | 4 | Enteritidis👁 |
| 11 | 4 | Enteritidis👁 |

The values can be browsed through the experimental data for each genotyping methods.
From the top right hand dropbox, you can select available genotyping schemes.

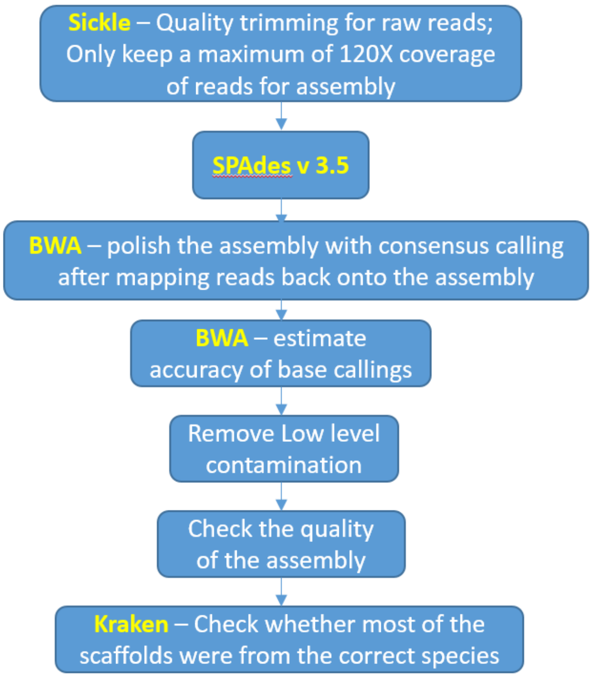
Serovar prediction (in *Salmonella)* is based on the consensus of metadata serovar designation to the strain's eBG (either rMLST or MLST). Click the eye to see an extended breakdown.

7 Gene MLST shows all allele profile in the right hand pane, if you scroll right. Larger genotyping schemes show the allele profile through the eye on the left.

# Chapter 5: EnteroTools - EnteroBase under the hood

## Workflow after receiving reads:

### 1. Automatic assembly



Sickle – Quality trimming for raw reads; Only keep a maximum of 120X coverage of reads for assembly

SPAdes v 3.5

BWA – polish the assembly with consensus calling after mapping reads back onto the assembly

BWA – estimate accuracy of base callings

Remove Low level contamination

Check the quality of the assembly

Kraken – Check whether most of the scaffolds were from the correct species

**Table 1 Assembly criteria for *Salmonella***

| Metrics | Criteria |
|---|---|
| Number of bases | 4 Mbp – 5.8 Mbp |
| N50 value | >20kb |
| Number of contigs | <600 |
| Proportion of scaffolding placeholders (N's) | <3% |
| Species assignment using Kraken | >70% contigs |

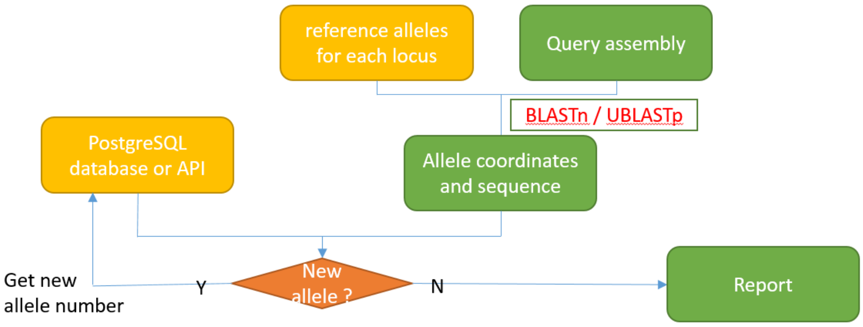
### 2. Annotation (Prokka)

**Table 3 Pan-gene sets:**

| *Species* | *No. of genes* |
|---|---|
| *Salmonella* | 21,065 |
| *Escherichia* | 25,002 |
| *Yerisnia* | 19,591 |

## 3. MLST, rMLST, wgMLST, cgMLST

### Working pipeline for ST nomenclatures :


reference alleles for each locus
Query assembly
BLASTn / UBLASTp
PostgreSQL database or API
Allele coordinates and sequence
Get new allele number
Y
New allele ?
N
Report

### cgMLST validation (Using *Salmonella* as an example):

wgMLST scheme consists of 21,065 genes and was built from 237 complete genomes from NCBI (http://www.ncbi.nlm.nih.gov/genome/), 82 PacBio assemblies from NCTC collection (http://www.sanger.ac.uk/resources/downloads/bacteria/nctc/) and 288 selected assemblies from EnteroBase.

cgMLST scheme consists of 3,002 genes and was set up as a sub-scheme from wgMLST, using only genes that :

1. present in over 98% of genomes
2. Disrupted in less than 6% of genomes
3. Not too variable