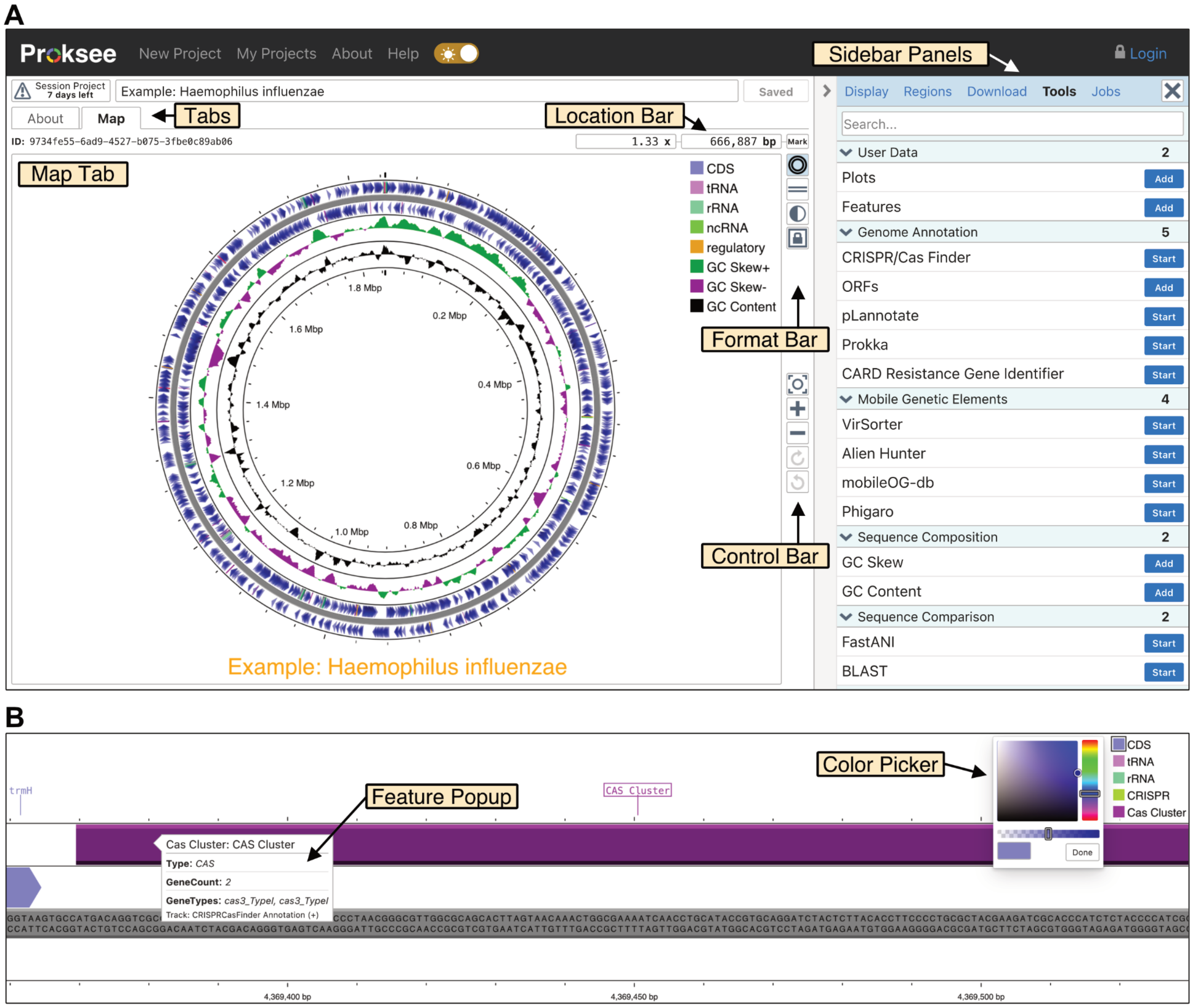
CBW IDE Module 9 Lab — Genome Annotation and Phylogenomics Using Proksee and ARETE

Proksee

Author: Gary Van Domselaar

Proksee is a bacterial genome assembly, annotation, and visualization tool that is designed to be easy to use. Due to time constraints, we will skip the assembly part and just do a couple of simple annotations and analyses of AMR and lateral gene transfer in an E. coli plasmid.

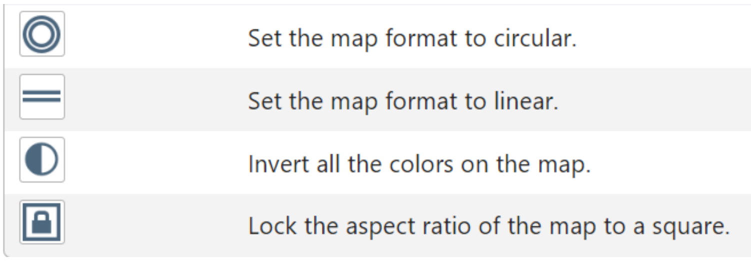
Proksee Overview



Proksee has several main components that you can use to generate and visualize annotations on a bacterial genome:

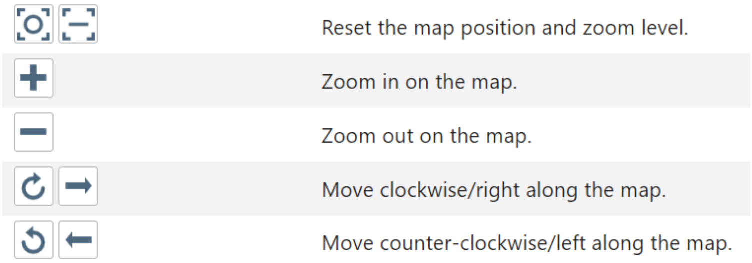
- **Tabs**
  - About - provides summary information for an annotation project

    - Map - the main tab for viewing your genome map
    - Tools - these will appear when you run annotation tools. They can be used to monitor the progress of annotation jobs, add annotations to the map, and download output files.
- **Location Bar**
    - More useful in linear layout - shows the coordinates you are currently viewing. You can ‘Mark’ a view to create a bookmark for views you like and want to return to later.
- **Sidebar Panels** - the **Sidebar** contains panels for
    - Customizing the look of the map (**Display**)
    - Investigating features, plots and contigs on the map (**Regions**)
    - Downloading images, JSON and sequences (**Download**)
    - Launching annotation and analyses (**Tools**)
    - Checking job results (**Jobs**).
- **Format Bar** - controls for altering the overall map format.



◎	Set the map format to circular.
═	Set the map format to linear.
◐	Invert all the colors on the map.
🔒	Lock the aspect ratio of the map to a square.

- Control Bar

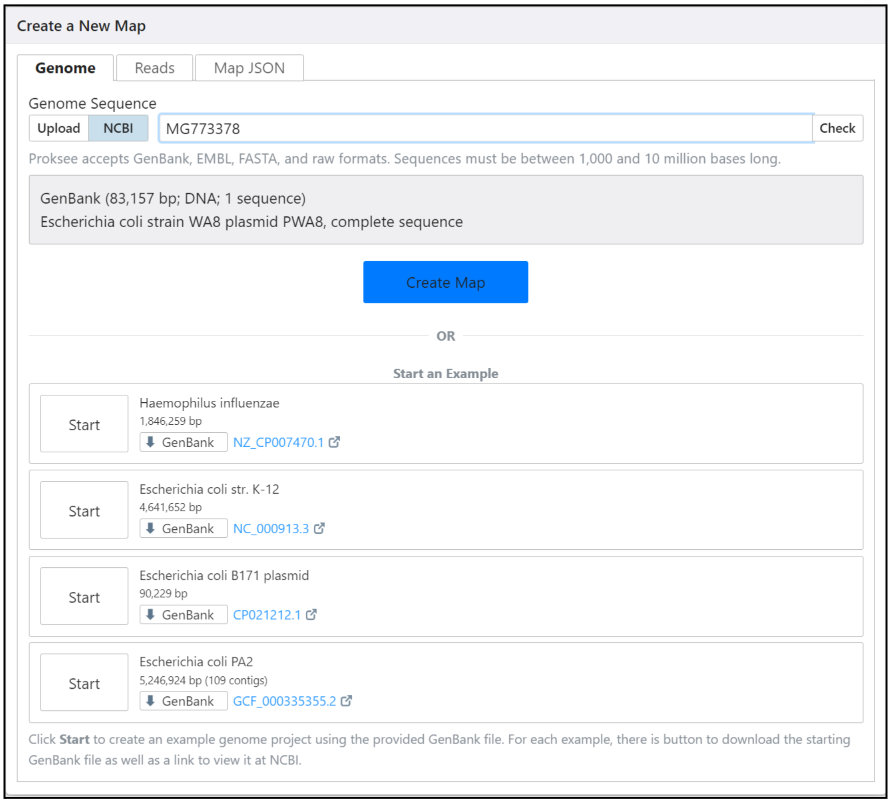


⊙ ⊟	Reset the map position and zoom level.
+	Zoom in on the map.
−	Zoom out on the map.
↻ →	Move clockwise/right along the map.
↺ ←	Move counter-clockwise/left along the map.

- **Feature Popup** - more information about a feature
- **Color Picker** - customize the color of a feature

Proksee exercise: annotate and interpret a plasmid

1. Go to the proksee homepage: https://proksee.ca
2. Create a new map:
   a. Click on the **NCBI** tab.
   b. Enter Genbank Accession **MG773378**.
   c. Click **Create Map**.



Create a New Map

Genome Sequence

MG773378

Proksee accepts GenBank, EMBL, FASTA, and raw formats. Sequences must be between 1,000 and 10 million bases long.

GenBank (83,157 bp; DNA; 1 sequence)
Escherichia coli strain WA8 plasmid PWA8, complete sequence

Start an Example

Haemophilus influenzae
1,846,259 bp
GenBank NZ_CP007470.1

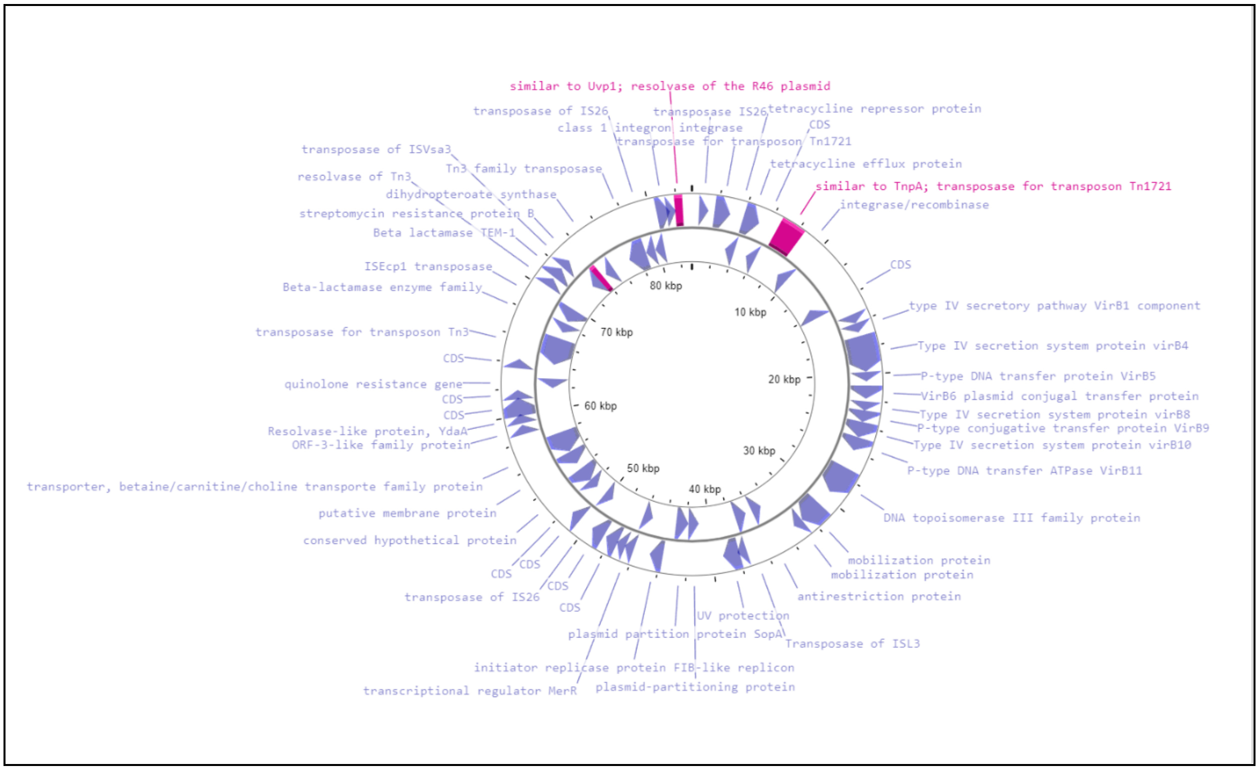
Escherichia coli str. K-12
4,641,652 bp
GenBank NC_000913.3

Escherichia coli B171 plasmid
90,229 bp
GenBank CP021212.1

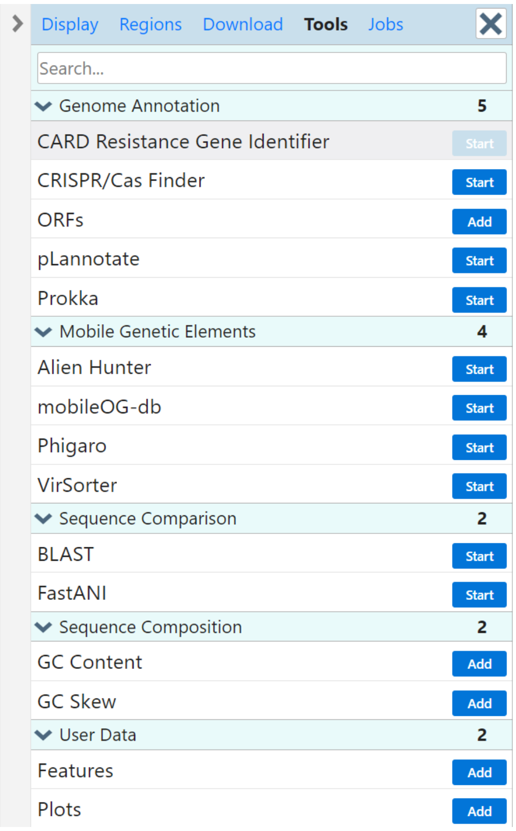
Escherichia coli PA2
5,246,924 bp (109 contigs)
GenBank GCF_000335355.2

Click **Start** to create an example genome project using the provided GenBank file. For each example, there is button to download the starting GenBank file as well as a link to view it at NCBI.

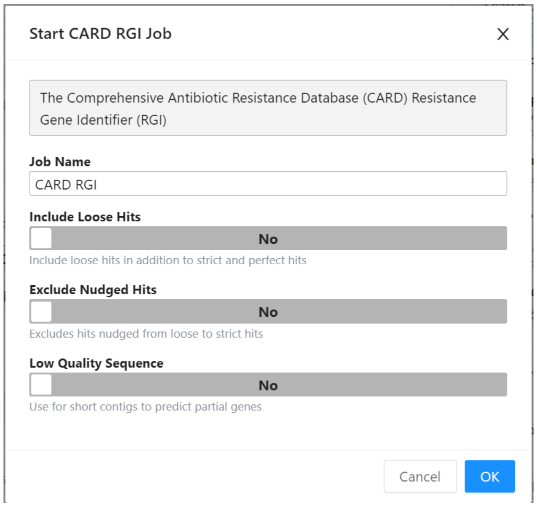
3. **View the map**. You will see a lot of annotated coding sequences (CDS). Review the **Additional Slides section** in the **Module 9 lecture**. Can you identify any characteristic plasmid genes?


similar to Uvp1; resolvase of the R46 plasmid
transposase of IS26
class 1 integron integrase
transposase IS26
tetracycline repressor protein
CDS
transposase for transposon Tn1721
transposase of ISVsa3
Tn3 family transposase
tetracycline efflux protein
resolvase of Tn3
similar to TnpA; transposase for transposon Tn1721
dihydropteroate synthase
integrase/recombinase
streptomycin resistance protein B
Beta lactamase TEM-1
ISEcp1 transposase
CDS
Beta-lactamase enzyme family
80 kbp
10 kbp
type IV secretory pathway VirB1 component
70 kbp
transposase for transposon Tn3
Type IV secretion system protein virB4
CDS
P-type DNA transfer protein VirB5
20 kbp
quinolone resistance gene
VirB6 plasmid conjugal transfer protein
CDS
60 kbp
CDS
Type IV secretion system protein virB8
P-type conjugative transfer protein VirB9
Resolvase-like protein, YdaA
ORF-3-like family protein
Type IV secretion system protein virB10
30 kbp
50 kbp
P-type DNA transfer ATPase VirB11
transporter, betaine/carnitine/choline transporte family protein
40 kbp
putative membrane protein
DNA topoisomerase III family protein
conserved hypothetical protein
CDS
CDS
mobilization protein
mobilization protein
CDS
transposase of IS26
antirestriction protein
CDS
UV protection
plasmid partition protein SopA
Transposase of ISL3
initiator replicase protein FIB-like replicon
transcriptional regulator MerR
plasmid-partitioning protein

4. **Look for AMR genes**.
   a. Click on the **Tools** panel in the **Sidebar**. Locate **CARD Resistance Gene Identifier** and click **Start**.



b. Accept the default values and click **OK**. The server will identify AMR features using RGI and pre-load the results so they can be added to the plot.



Start CARD RGI Job

The Comprehensive Antibiotic Resistance Database (CARD) Resistance Gene Identifier (RGI)

Job Name

CARD RGI

Include Loose Hits

No

Include loose hits in addition to strict and perfect hits

Exclude Nudged Hits

No

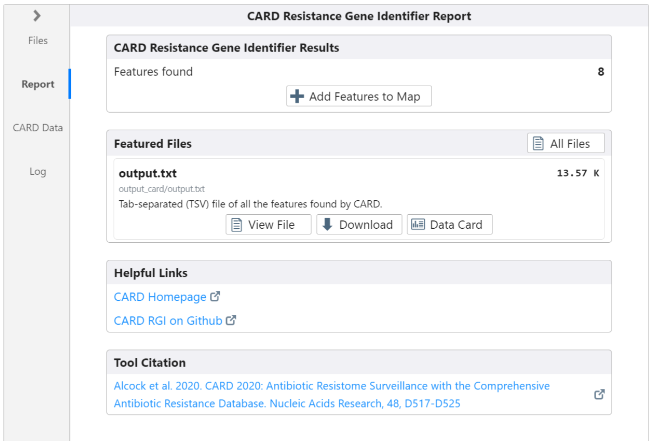
Excludes hits nudged from loose to strict hits

Low Quality Sequence

No

Use for short contigs to predict partial genes

c. When the job is done, you will get a **CARD Resistance Gene Identifier Report**.



CARD Resistance Gene Identifier Report

CARD Resistance Gene Identifier Results

Features found **8**

Featured Files

output.txt 13.57 K

output_card/output.txt

Tab-separated (TSV) file of all the features found by CARD.

Helpful Links

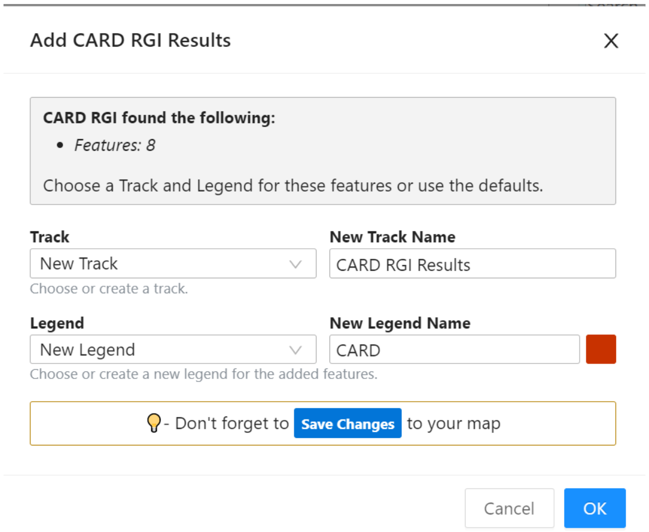
CARD Homepage

CARD RGI on Github

Tool Citation

Alcock et al. 2020. CARD 2020: Antibiotic Resistome Surveillance with the Comprehensive Antibiotic Resistance Database. Nucleic Acids Research, 48, D517-D525

d. Click **Add Features to Map**. In the popup dialogue, accept the defaults and click **OK**.



Add CARD RGI Results ✕

CARD RGI found the following:

- *Features: 8*

Choose a Track and Legend for these features or use the defaults.

Track: New Track
Choose or create a track.

New Track Name: CARD RGI Results

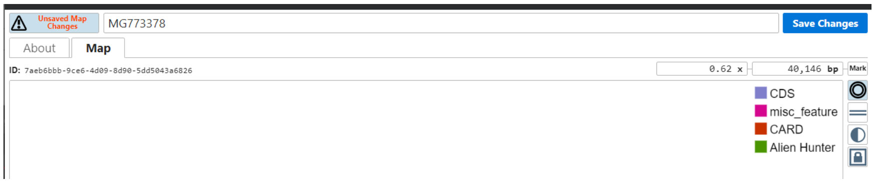
Legend: New Legend
Choose or create a new legend for the added features.

New Legend Name: CARD 🟥

💡- Don't forget to **Save Changes** to your map

Cancel OK

Click **Save Changes** to save your new annotations to the map.



⚠ Unsaved Map Changes MG773378 **Save Changes**

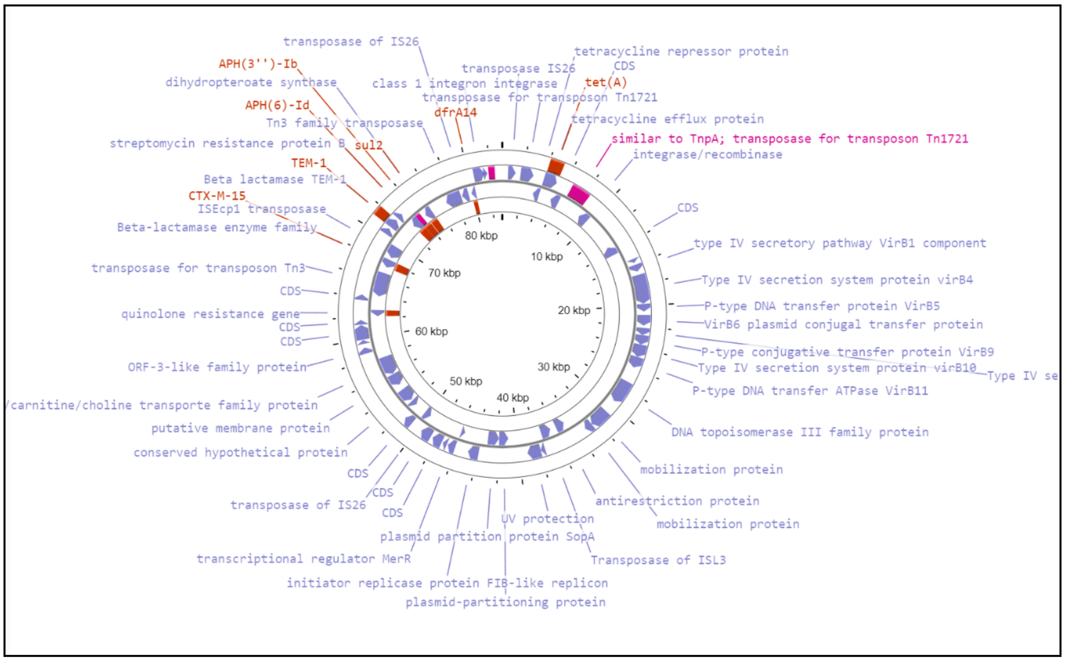
About **Map**

ID: 7aeb6bbb-9ce6-4d09-8d90-5dd5043a6826

0.62 x 40,146 bp Mark

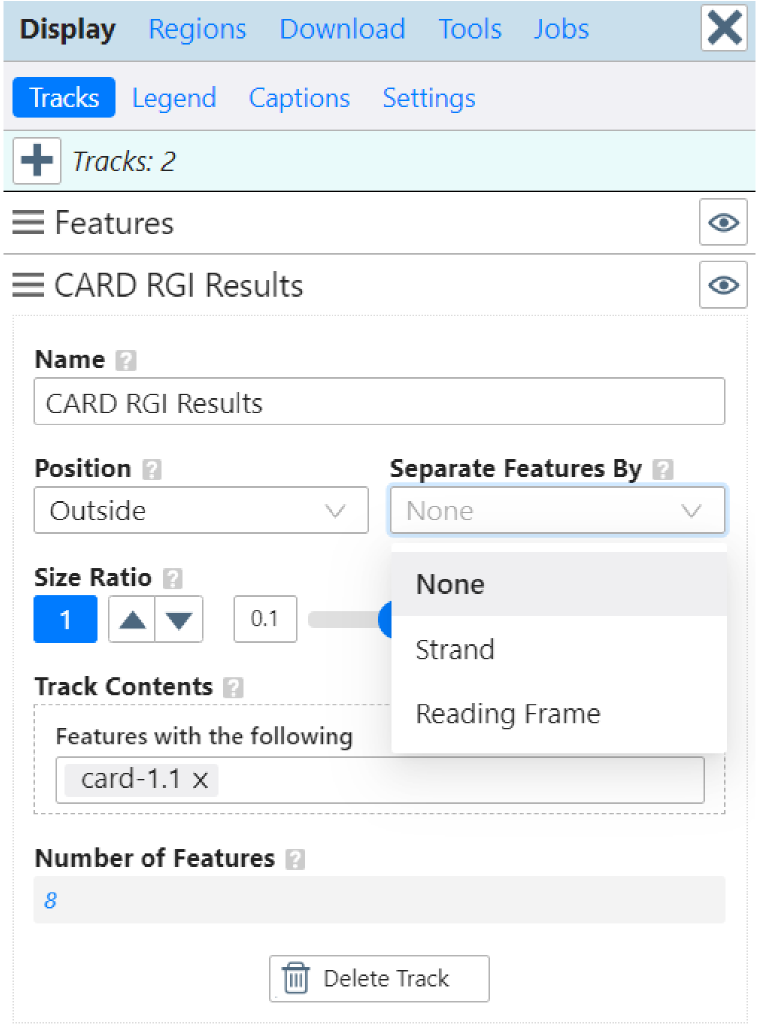
🟪 CDS
🟪 misc_feature
🟥 CARD
🟩 Alien Hunter

e. View the Map. The AMR genes predicted by CARD/RGI are plotted on the map in red.


transposase of IS26
APH(3'')-Ib
dihydropteroate synthase
APH(6)-Id
Tn3 family transposase
streptomycin resistance protein B
sul2
TEM-1
Beta lactamase TEM-1
CTX-M-15
ISEcp1 transposase
Beta-lactamase enzyme family
transposase for transposon Tn3
CDS
quinolone resistance gene
CDS
CDS
ORF-3-like family protein
/carnitine/choline transporte family protein
putative membrane protein
conserved hypothetical protein
CDS
CDS
transposase of IS26
CDS
plasmid partition protein SopA
transcriptional regulator MerR
initiator replicase protein FIB-like replicon
plasmid-partitioning protein
UV protection
Transposase of ISL3
antirestriction protein
mobilization protein
mobilization protein
DNA topoisomerase III family protein
P-type DNA transfer ATPase VirB11
Type IV secretion system protein virB10
Type IV se
P-type conjugative transfer protein VirB9
VirB6 plasmid conjugal transfer protein
P-type DNA transfer protein VirB5
Type IV secretion system protein virB4
type IV secretory pathway VirB1 component
CDS
integrase/recombinase
similar to TnpA; transposase for transposon Tn1721
tetracycline efflux protein
tet(A)
CDS
tetracycline repressor protein
transposase IS26
class 1 integron integrase
transposase for transposon Tn1721
dfrA14
80 kbp
10 kbp
20 kbp
30 kbp
40 kbp
50 kbp
60 kbp
70 kbp

f. The AMR genes harboured by each (direct and opposite) strand are plotted in separate tracks by default. Let's combine them into a single track and change their decoration to show their orientation.

i. Click on the **Display** panel in the **Sidebar**. Click the **CARD RGI Results Track** to expose the options. Open the **Position** drop-down and choose **Outside**. Open the **Separate Features By** drop-down and choose **None**



Name: CARD RGI Results

Position: Outside

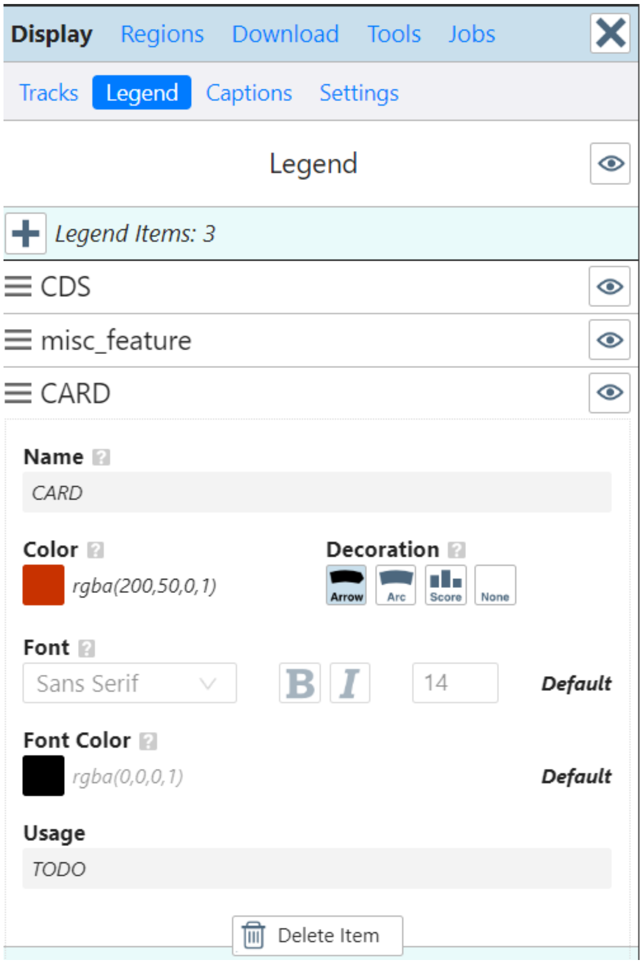
Separate Features By: None

Size Ratio: 1 0.1

Track Contents: Features with the following card-1.1 ×

Number of Features: 8

ii. Click the **Legend** tab inside the **Display** tab. Click the **CARD Legend Item** to expose the options. Change the **Decoration** from **Arc** to **Arrow**.



Legend

Legend Items: 3

CDS

misc_feature

CARD

Name

CARD

Color 🟥 *rgba(200,50,0,1)*

Decoration

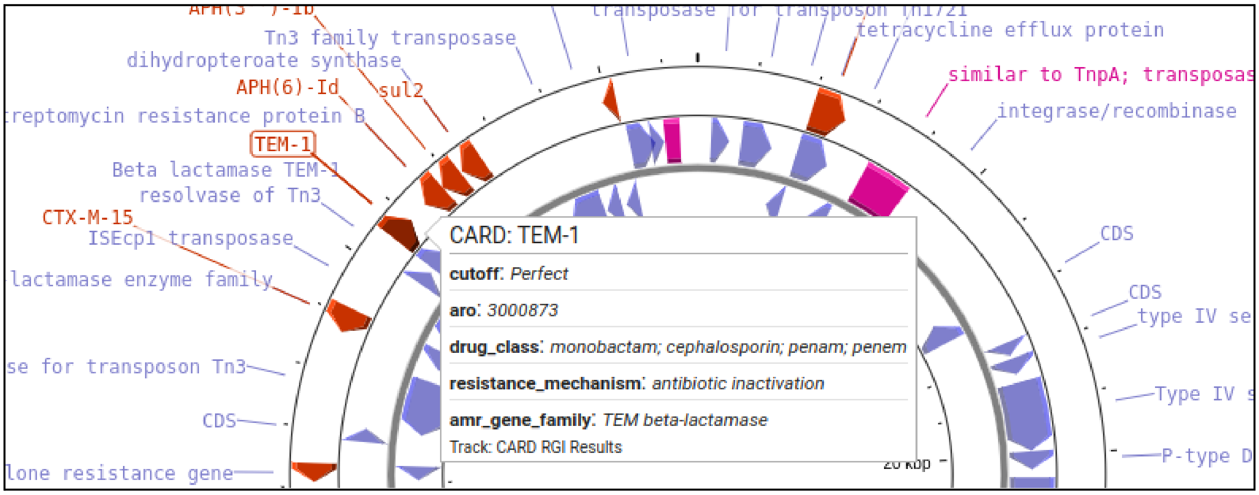
Font Sans Serif 14 ***Default***

Font Color ⬛ *rgba(0,0,0,1)* ***Default***

Usage

TODO

g. View the Map. **Hover over some genes** identified by RGI/CARD to see more info.



CARD: TEM-1

cutoff: *Perfect*

aro: *3000873*

drug_class: *monobactam; cephalosporin; penam; penem*

resistance_mechanism: *antibiotic inactivation*

amr_gene_family: *TEM beta-lactamase*

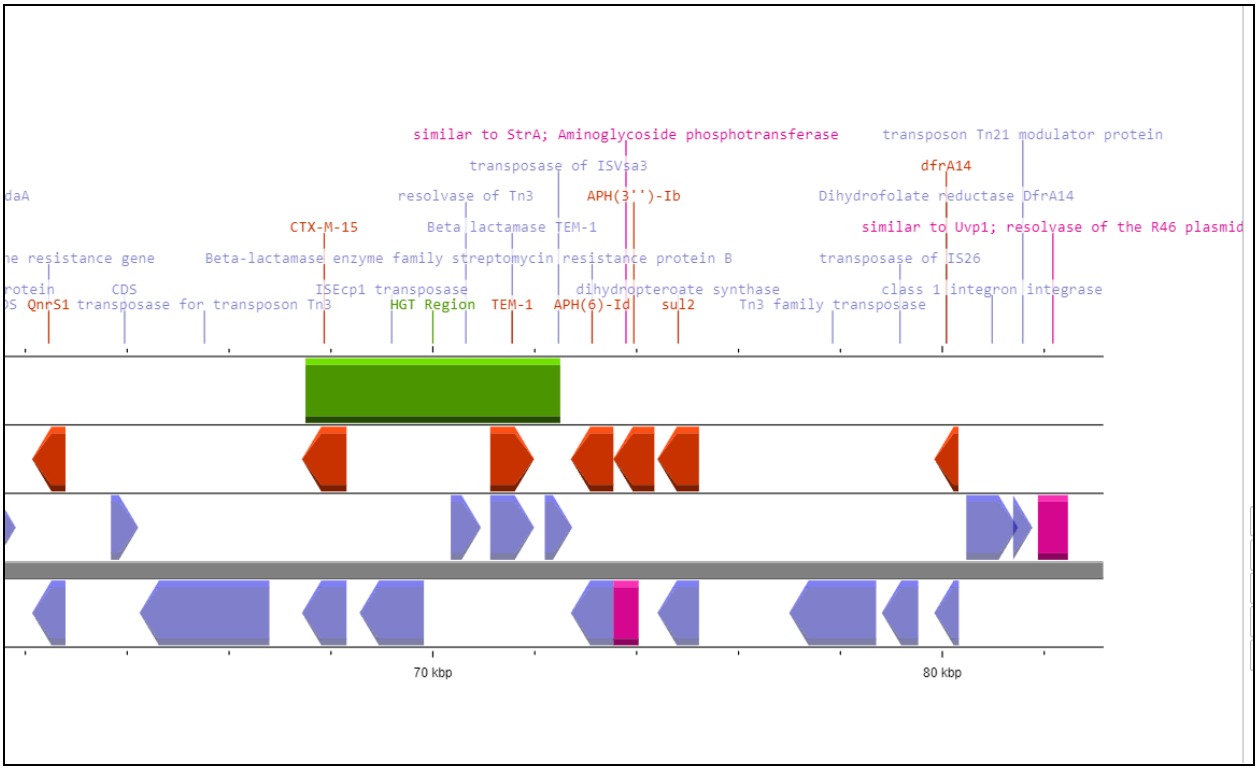
Track: CARD RGI Results

h. Can you identify any AMR genes that CARD/RGI may have missed? (Hint: the plasmid is already expertly annotated.)

5. **Look for mobile elements**. Plasmids are mobile elements, but they themselves can harbour additional mobile elements. Let's see if we can find some!
    a. Locate **Alien Hunter** in the **Tools** tab and click **Start**.
    b. Accept the defaults and click **OK**.
    c. Once the job is complete, click **Add Features to Map**. In the popup dialogue, accept the defaults and click **OK**.
    d. Click **Save Changes**.

    The Horizontal Gene Transfer (HGT) regions predicted by Alien Hunter are plotted on the map in green. Note that Alien Hunter uses interpolated variable order motifs try to identify HGT elements, but this is not hard proof that the region is truly an HGT. Alien Hunter uses hidden Markov models (HMMs) to try to define the boundaries of an HGT region, but the 'signal' for the start and end of an HGT is weak, so these predictions may not be exact.

    Take a look at the AMR cluster spanning the cluster of AMR genes from QnrS1 to dfrA14


similar to StrA; Aminoglycoside phosphotransferase
transposon Tn21 modulator protein
transposase of ISVsa3
dfrA14
daA
resolvase of Tn3
APH(3'')-Ib
Dihydrofolate reductase DfrA14
CTX-M-15
Beta lactamase TEM-1
similar to Uvp1; resolvase of the R46 plasmid
ne resistance gene
Beta-lactamase enzyme family streptomycin resistance protein B
transposase of IS26
rotein
CDS
ISEcp1 transposase
dihydropteroate synthase
class 1 integron integrase
DS QnrS1 transposase for transposon Tn3
HGT Region
TEM-1
APH(6)-Id
sul2
Tn3 family transposase
70 kbp
80 kbp

    Review the **Additional Slides** in the **Module 9 Lecture**. Are there any genes within or flanking this region that can corroborate the HGT prediction made by Alien Hunter?

ARETE

Author: Robert Beiko

This is an introductory exercise meant to demonstrate the key features of the ARETE pipeline. Given time constraints, you will not be running ARETE, rather you will be shown a demonstration of how the pipeline runs and be provided with the generated outputs that are generated by the demo. The interactive portion of the practical will be focused on visualising some of these results in microreact and exploring some of the output files in more depth.

Some of the information below is condensed from the ARETE Github page (https://github.com/beiko-lab/arete); please check there for the latest developments and updates to the software. You will also find links to the software tools mentioned below at the Github page.

ARETE Overview

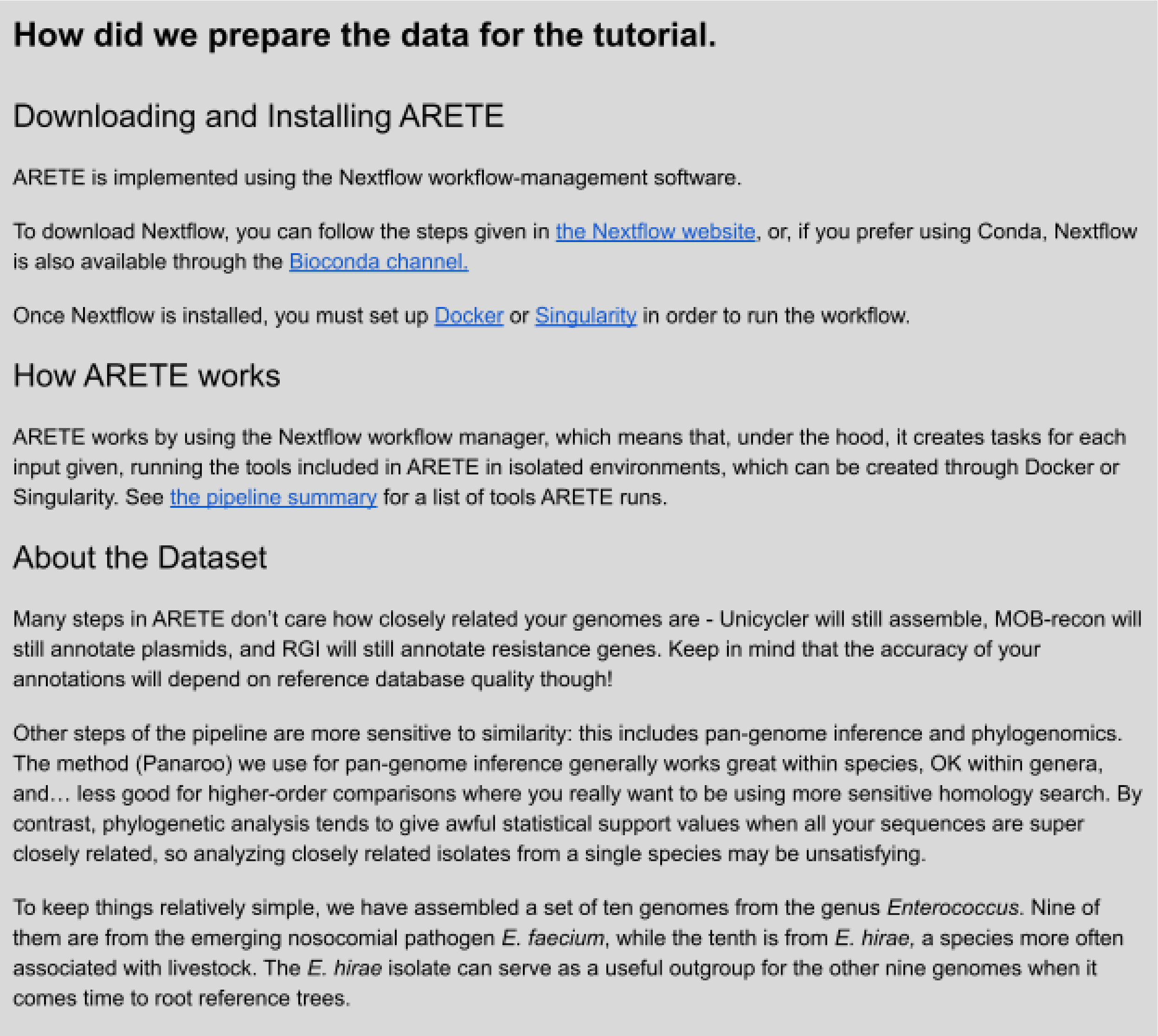
“Antimicrobial Resistance: Emergence, Transmission, and Ecology” (ARETE) is a software pipeline that takes either unassembled reads or assembled genomes as input, performs assembly and quality control if necessary, then performs a series of gene and mobile genetic element annotations and a phylogenomic analysis. The ultimate goal is to identify key determinants of resistance (and other functions) and map their pathways of transmission across a set of genomes. Doing so allows us to propose hypotheses about which genes are being mobilized and which habitats constitute “highways of gene sharing”.

ARETE is under active development and we are currently adding functionality for the following:

- ***Dataset subsetting.*** We are currently running ARETE on datasets that comprise thousands of genomes and plan to scale to tens of thousands in the very near future. Annotating this many genomes is not painful if you have access to a big high-performance computing cluster, but analyses that combine all these genomes (pan-genome inference and phylogenomics, for example) scale horribly. We use PopPUNK to generate representative subsets that are tractable for these analyses.
- ***Inference of gene transfer and recombination.*** We are adding several applications to ARETE that can identify probable events of gene sharing, including phylogenetic tree comparisons using rSPR, inference of correlated gene gains and losses using EvolCCM, and recombination detection. We are currently developing methods to aggregate the full set of predictions to show the “big picture” of transmission in large genomic datasets.
- ***Gene-order analysis.*** Gene order conservation can provide important clues about evolutionary mechanisms; for example, if a set of genes are found in the same order in two distantly related genomes, we might infer that a gene-transfer event has occurred. “Guilt by association” can also be used to clarify the role of a gene: if a putative resistance gene is surrounded by other such genes or genes with related functions such

as virulence factors or transposases, we are more likely to believe that the putative gene does potentially confirm resistance..

- ***Output visualizations.*** ARETE results are meant to feed into existing visualization tools as well as new ones under development, including Indizio for visualization of co-evolving genes and Coeus for comparative gene-order visualization.



How did we prepare the data for the tutorial.

Downloading and Installing ARETE

ARETE is implemented using the Nextflow workflow-management software.

To download Nextflow, you can follow the steps given in the Nextflow website, or, if you prefer using Conda, Nextflow is also available through the Bioconda channel.

Once Nextflow is installed, you must set up Docker or Singularity in order to run the workflow.

How ARETE works

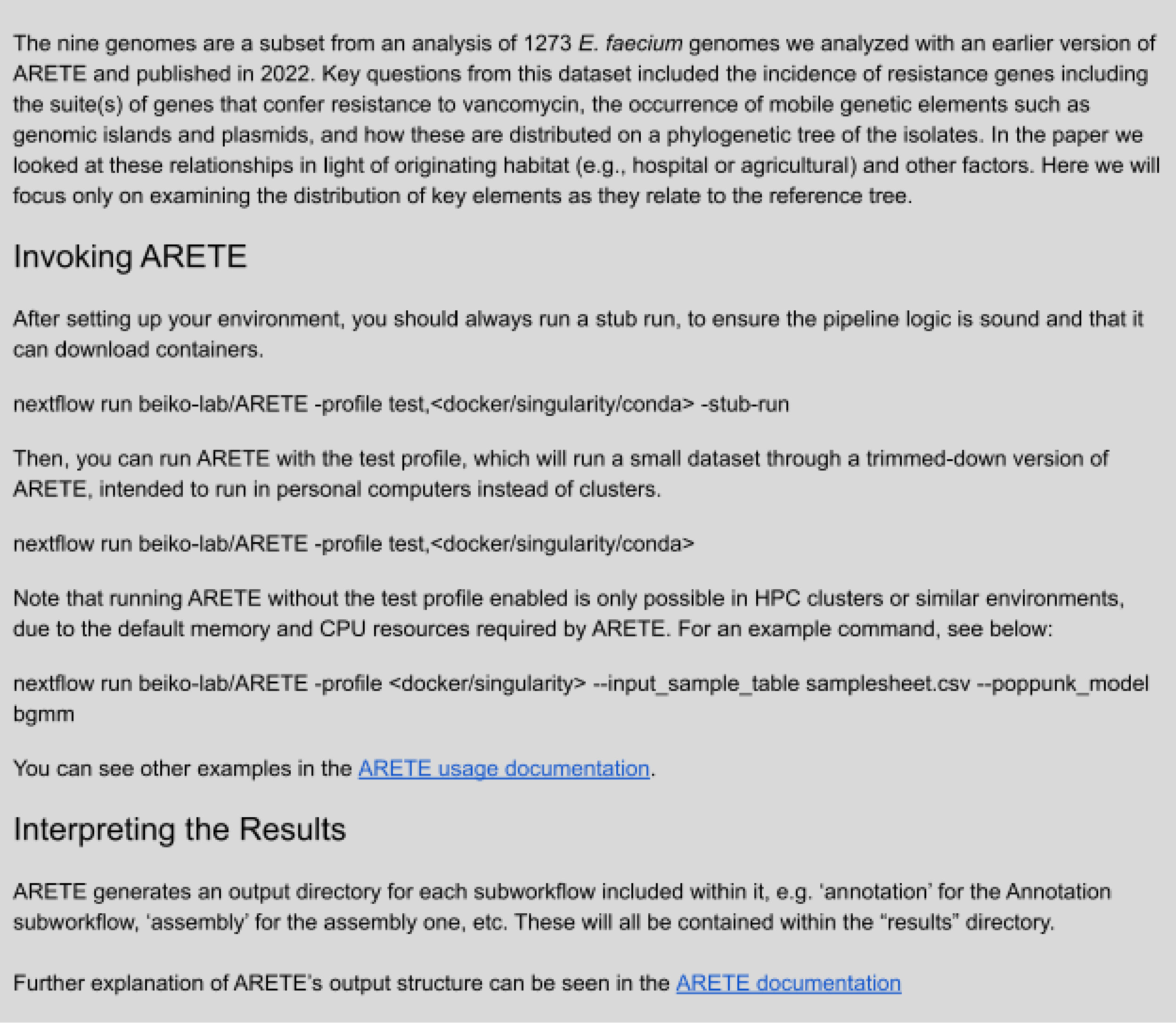
ARETE works by using the Nextflow workflow manager, which means that, under the hood, it creates tasks for each input given, running the tools included in ARETE in isolated environments, which can be created through Docker or Singularity. See the pipeline summary for a list of tools ARETE runs.

About the Dataset

Many steps in ARETE don't care how closely related your genomes are - Unicycler will still assemble, MOB-recon will still annotate plasmids, and RGI will still annotate resistance genes. Keep in mind that the accuracy of your annotations will depend on reference database quality though!

Other steps of the pipeline are more sensitive to similarity: this includes pan-genome inference and phylogenomics. The method (Panaroo) we use for pan-genome inference generally works great within species, OK within genera, and… less good for higher-order comparisons where you really want to be using more sensitive homology search. By contrast, phylogenetic analysis tends to give awful statistical support values when all your sequences are super closely related, so analyzing closely related isolates from a single species may be unsatisfying.

To keep things relatively simple, we have assembled a set of ten genomes from the genus *Enterococcus*. Nine of them are from the emerging nosocomial pathogen *E. faecium*, while the tenth is from *E. hirae,* a species more often associated with livestock. The *E. hirae* isolate can serve as a useful outgroup for the other nine genomes when it comes time to root reference trees.



The nine genomes are a subset from an analysis of 1273 *E. faecium* genomes we analyzed with an earlier version of ARETE and published in 2022. Key questions from this dataset included the incidence of resistance genes including the suite(s) of genes that confer resistance to vancomycin, the occurrence of mobile genetic elements such as genomic islands and plasmids, and how these are distributed on a phylogenetic tree of the isolates. In the paper we looked at these relationships in light of originating habitat (e.g., hospital or agricultural) and other factors. Here we will focus only on examining the distribution of key elements as they relate to the reference tree.

Invoking ARETE

After setting up your environment, you should always run a stub run, to ensure the pipeline logic is sound and that it can download containers.

```
nextflow run beiko-lab/ARETE -profile test,<docker/singularity/conda> -stub-run
```

Then, you can run ARETE with the test profile, which will run a small dataset through a trimmed-down version of ARETE, intended to run in personal computers instead of clusters.

```
nextflow run beiko-lab/ARETE -profile test,<docker/singularity/conda>
```

Note that running ARETE without the test profile enabled is only possible in HPC clusters or similar environments, due to the default memory and CPU resources required by ARETE. For an example command, see below:

```
nextflow run beiko-lab/ARETE -profile <docker/singularity> --input_sample_table samplesheet.csv --poppunk_model bgmm
```

You can see other examples in the ARETE usage documentation.

Interpreting the Results

ARETE generates an output directory for each subworkflow included within it, e.g. 'annotation' for the Annotation subworkflow, 'assembly' for the assembly one, etc. These will all be contained within the "results" directory.

Further explanation of ARETE's output structure can be seen in the ARETE documentation

Using Microreact to Visualise Results

To explore these results a bit further we are going to use microreact. This is a really handy website/service that lets you quickly visualise and explore data and metadata.

First, we need to download two key files from the results under `~/CourseData/IDE_data/module9/arete_results/` to our workspace. Create a `module9` directory within your workspace and navigate into it:

- `mkdir module9`
- `cd module9`

Copy over the directory of files needed to complete this portion of the lab:

- `cp -r ~/CourseData/IDE_data/module9/arete_results/ .`

Once the directory is copied over, following the same approach as other modules, identify your IP address on AWS.

- `curl http://checkip.amazonaws.com`

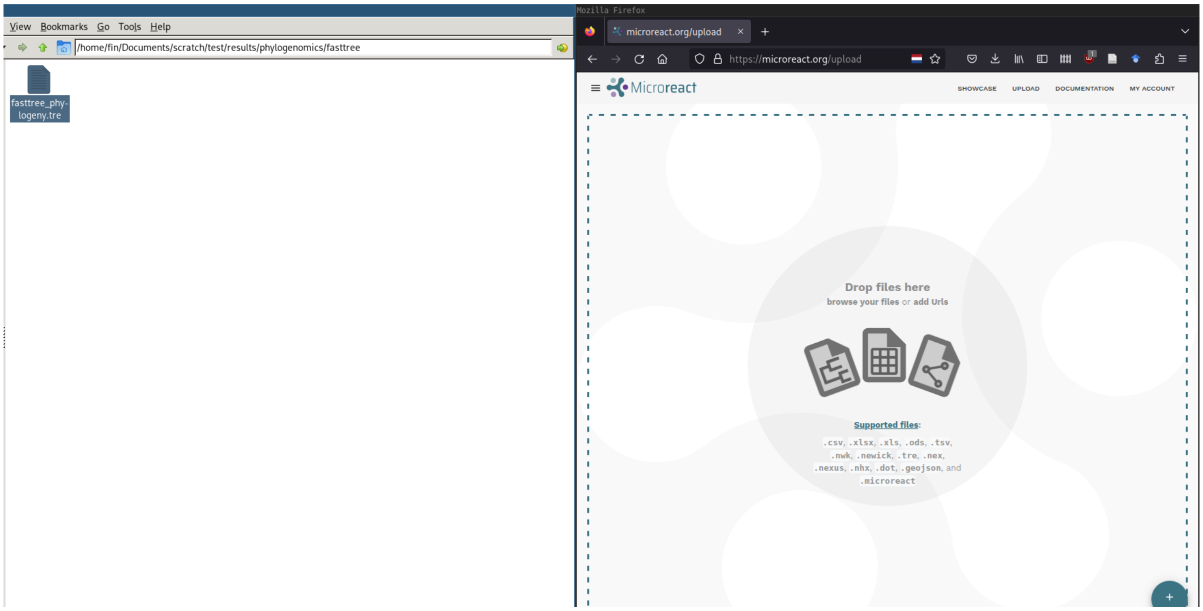
Paste the IP address (formatted as XX.XXX.XX.XX) into your web browser in order to access your new module9 directory and download the following two files (you can `right click > save link as…` to download these files).

- `module9/arete_results/phylogenomics/fasttree/fasttree_phylogeny.tre`
- `module9/arete_results/esbl_carb_plasmid_tally.tsv`

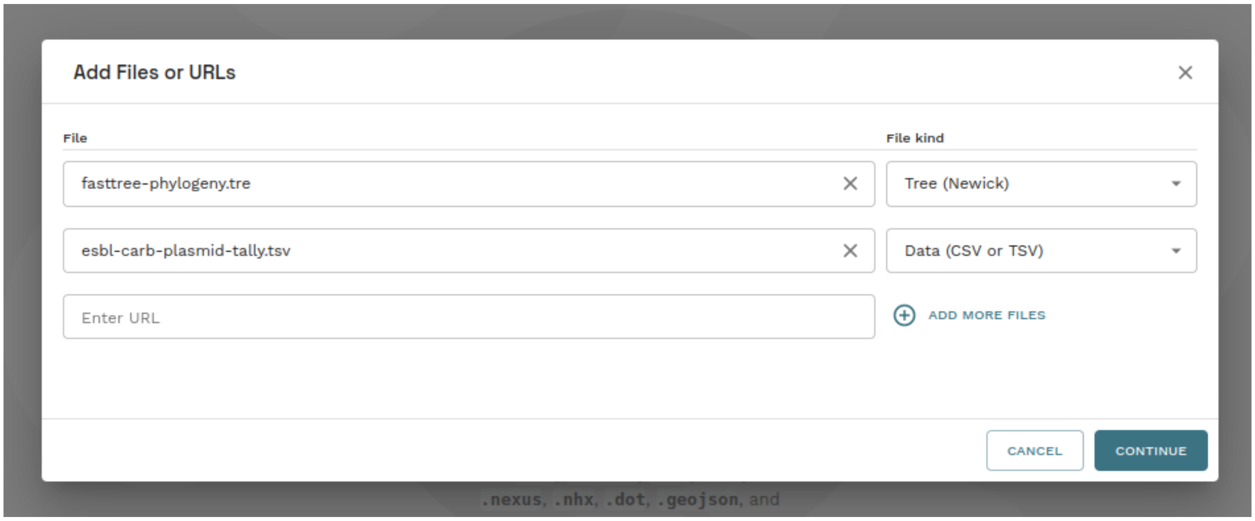
The first of these is the maximum-likelihood phylogeny created from the core-genome alignment generated by **panaroo** and **fasttree2**. This will be used to provide a structure on which we visualise data generated from other stages of the workflow.

The second file is a pre-generated merged data file with (fictional) metadata of where and when each genome was sampled, the real source material, and the presence or absence of different plasmid clusters as well as the extended-spectrum beta-lactamases and carbapenemase gene families detected on these plasmids.

Then navigate your browser to https://microreact.org/upload and drag and drop `fasttree_phylogeny.tre` phylogeny file into the window.


View Bookmarks Go Tools Help
/home/fin/Documents/scratch/test/results/phylogenomics/fasttree
fasttree_phylogeny.tre
Mozilla Firefox
microreact.org/upload
https://microreact.org/upload
Microreact
SHOWCASE
UPLOAD
DOCUMENTATION
MY ACCOUNT
Drop files here
browse your files or add Urls
Supported files:
.csv, .xlsx, .xls, .ods, .tsv, .nwk, .newick, .tre, .nex, .nexus, .nhx, .dot, .geojson, and .microreact

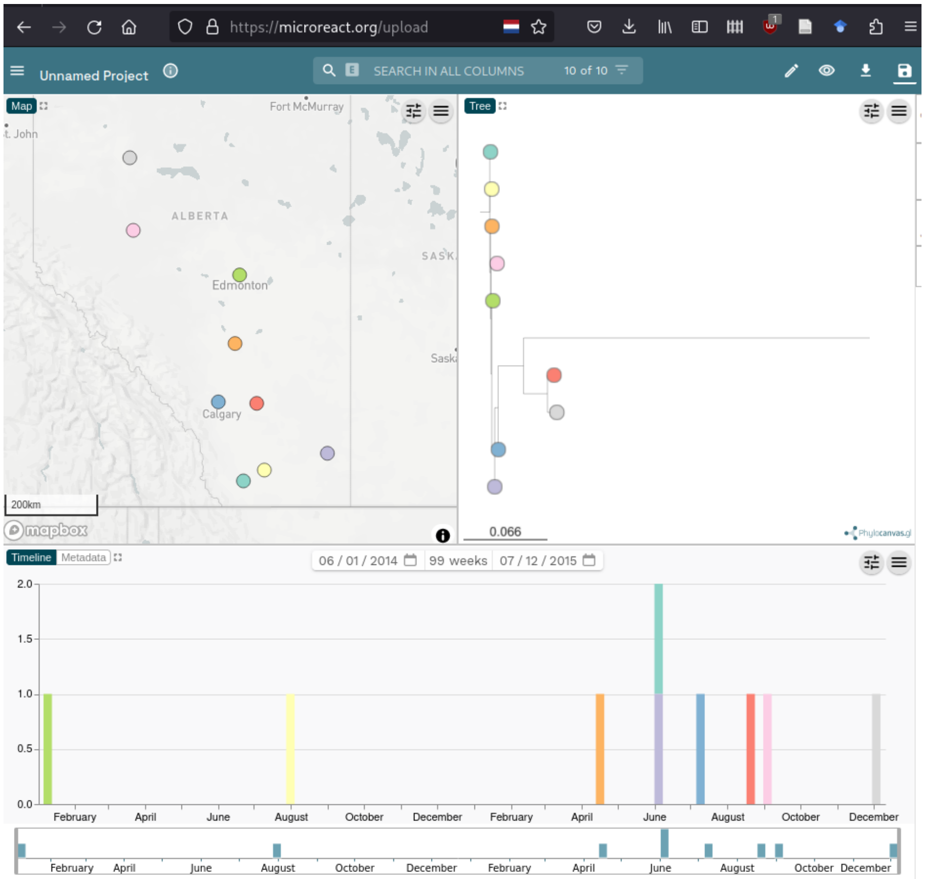
You will then get the following prompt to add more files, click the plus symbol, and add the metadata and parsed results file `esbl_carb_plasmid_tally.tsv` and hit **continue twice**.



Add Files or URLs ×

File	File kind
fasttree-phylogeny.tre ×	Tree (Newick) ▾
esbl-carb-plasmid-tally.tsv ×	Data (CSV or TSV) ▾
Enter URL	⊕ ADD MORE FILES

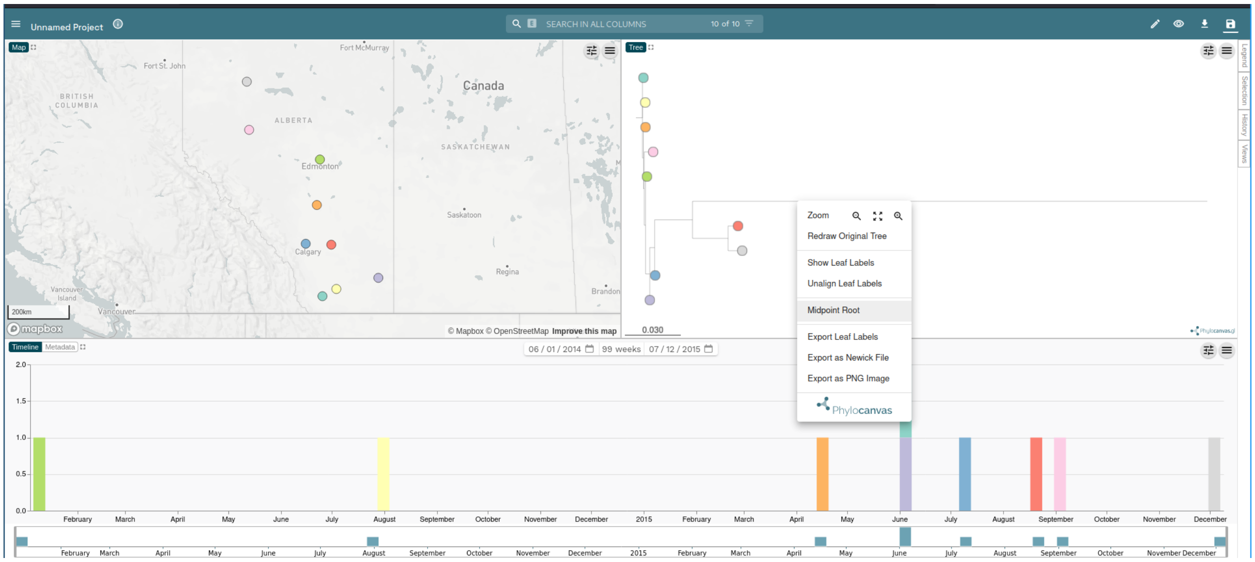
Microreact will automatically parse these files and generate a few visualisations of the data. This includes a map of sample locations, the phylogeny, a timeline of sampling dates, and a metadata sheet view. **(if having issues try to use a different browser: Mozilla, Edge, or Safari**)



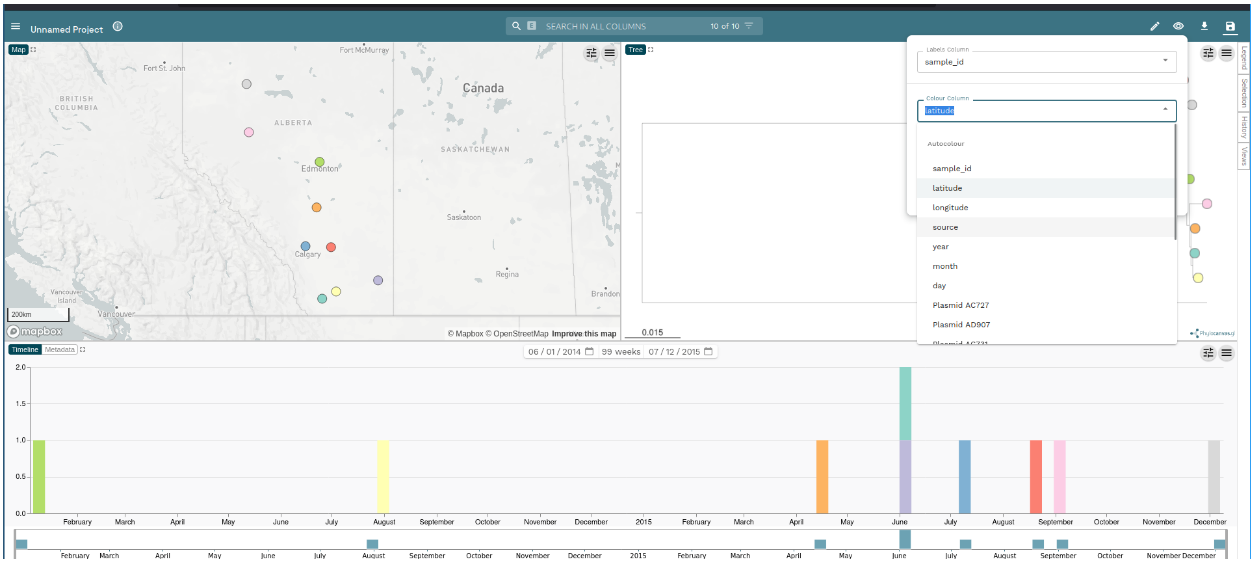
However, currently these visualisations aren't very helpful for exploring the dataset. To fix this we are going to go into the settings and make some changes. First we want to properly root our tree on the *E. hirae* outgroup.

1. **Based on the phylogenetics modules, which branches do you think represents our intended outgroup?**

Fortunately, we don't even have to directly select the correct branch in this case because a midpoint rooting will achieve our goal. To do this, right click on the phylogeny pane and select **"Midpoint Root"**

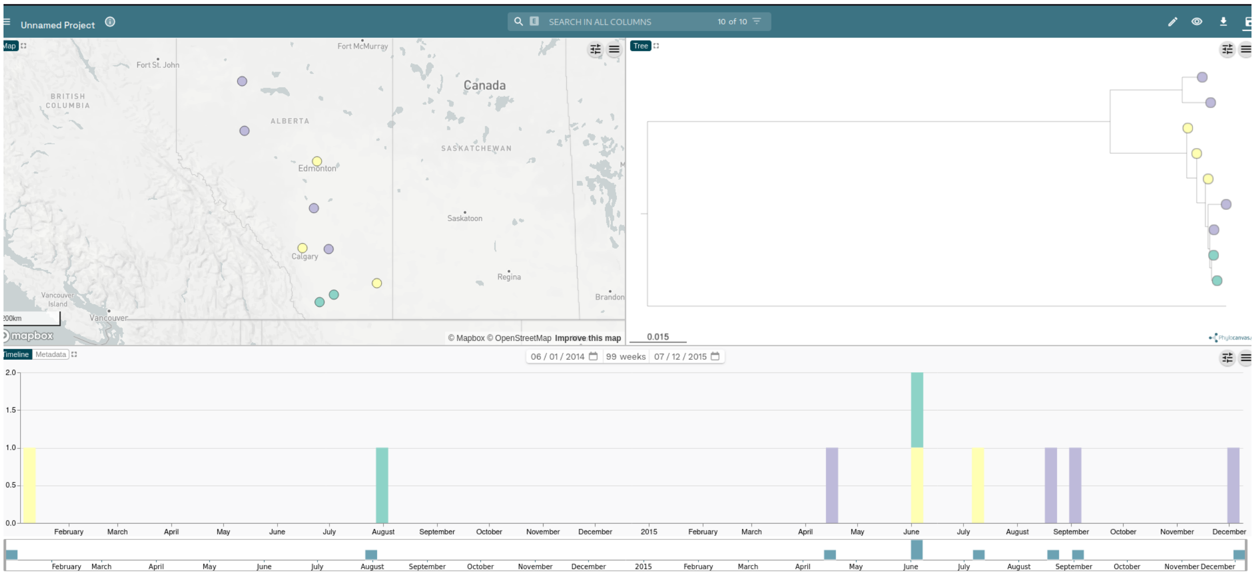


Currently, the samples are just being coloured by their sampling latitude. This doesn't really provide any useful information not present on the map so we are instead going to colour them by where that genome came from (e.g., wastewater, human clinical samples, livestock). To do click on the "eye" icon at the top right, then click on the "color column" pull-down menu and select "**source**".


Labels Column: sample_id

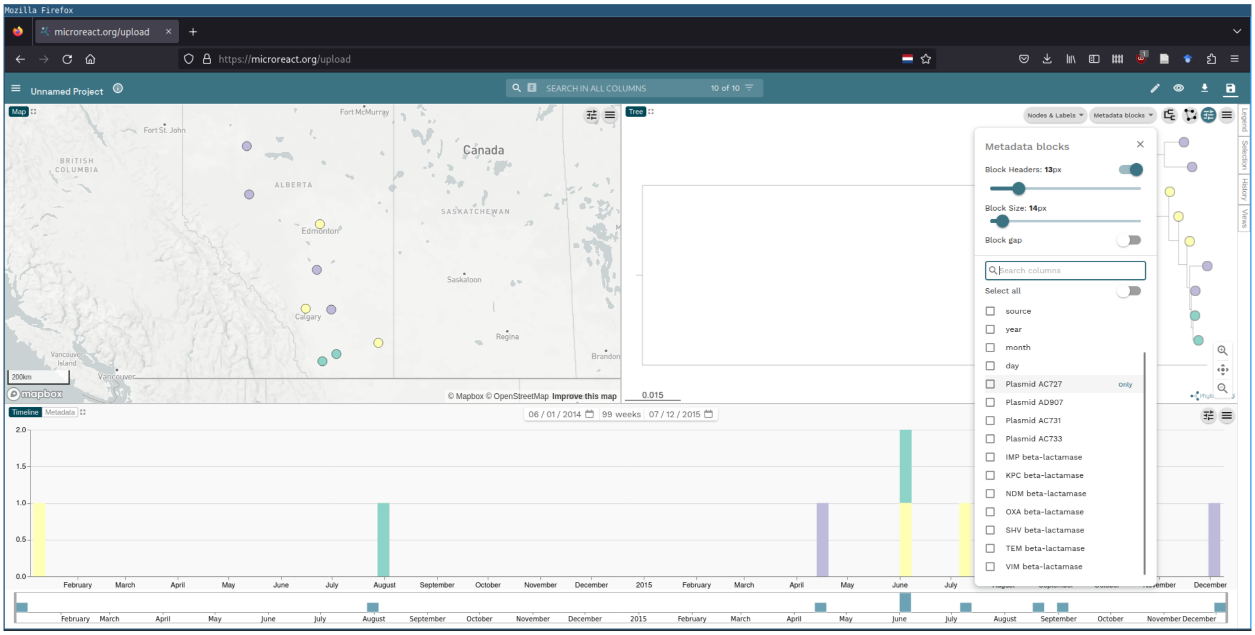
Colour Column: latitude

You can now put your cursor over individual nodes (on any panel) and get the source information (or hit the legend tab just right of the phylogeny).



2. **Now we have a phylogeny coloured by source, what relationship does the core genome phylogeny suggest there is between the clinical samples (yellow above) and the other samples?**

Now we are going to look at mobile genetic elements detected in these genomes by annotating the parsed output of **mob-suite** onto the phylogeny. This has been filtered to just plasmids inferred to be carrying worrying AMR genes (discussed below). Click on the button with horizontal lines and dots on the top right, then click on **metadata blocks** and select the 4 checkboxes that start with “**Plasmid AXXX**”



Metadata blocks

Block Headers: 13px

Block Size: 14px

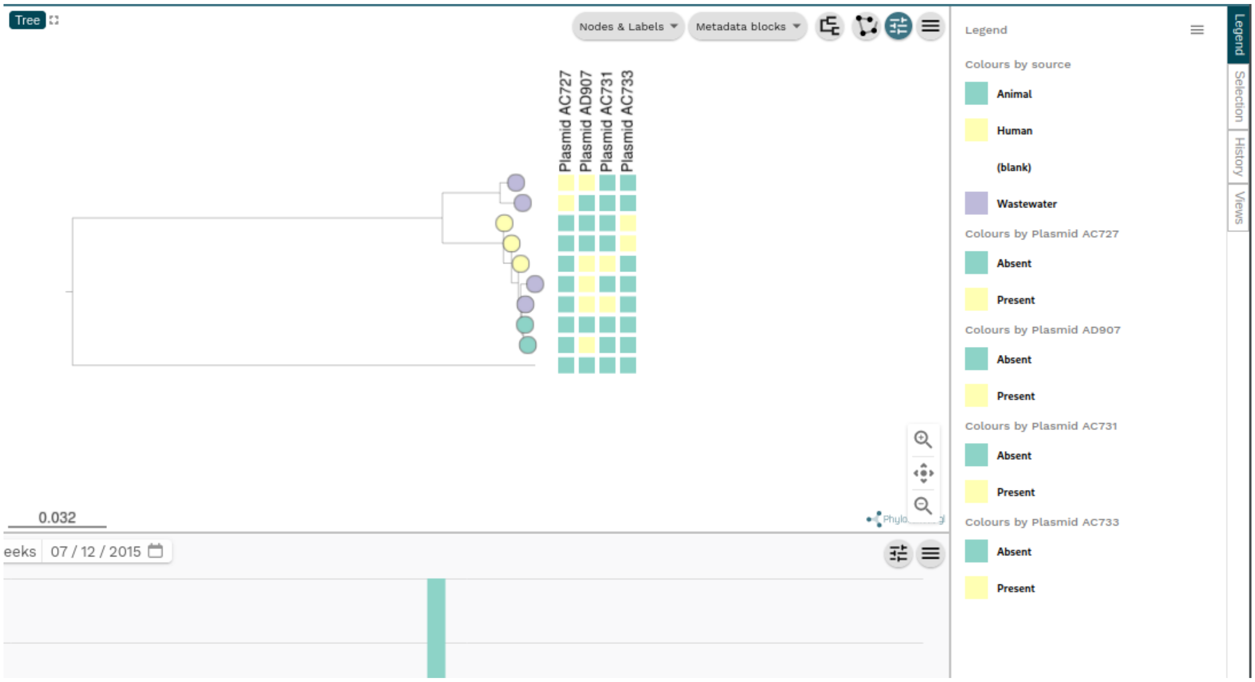
Block gap

Search columns

Select all

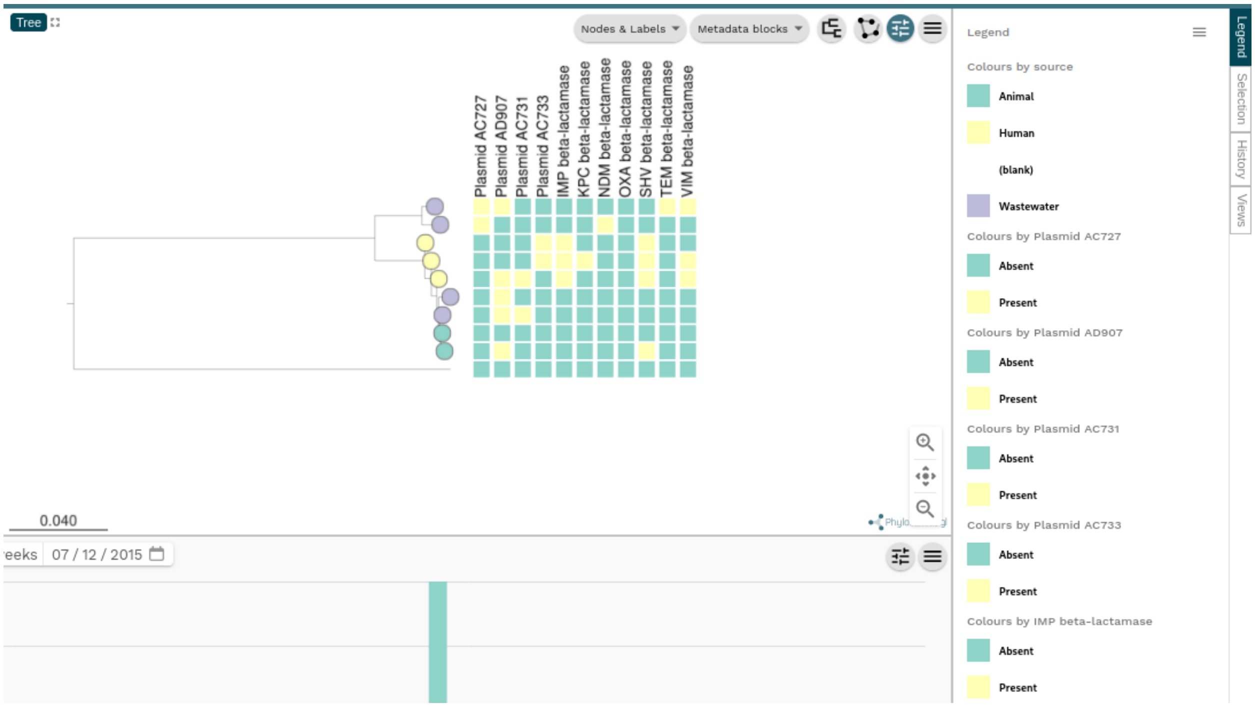
□ source
□ year
□ month
□ day
□ Plasmid AC727 Only
□ Plasmid AD907
□ Plasmid AC731
□ Plasmid AC733
□ IMP beta-lactamase
□ KPC beta-lactamase
□ NDM beta-lactamase
□ OXA beta-lactamase
□ SHV beta-lactamase
□ TEM beta-lactamase
□ VIM beta-lactamase

This will create a presence absence series of plots next to the phylogeny (yellow indicating “presence of a plasmid” in that genome and teal “absence”).



3. **Which plasmid cluster is only found in the clinical "human" samples?**

We can also annotate the extended-spectrum beta-lactamases and carbapenemases in these plasmids in the same way as we did for the plasmids.



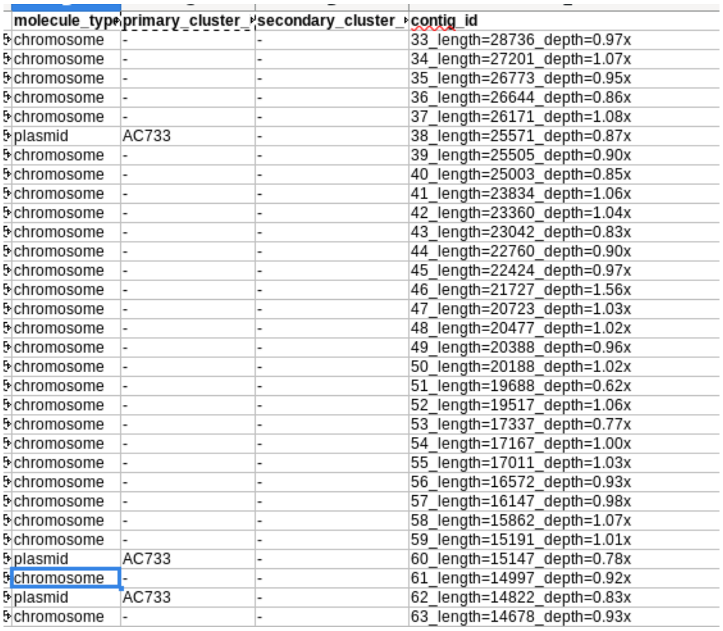
4. **Which sample has the most different classes of ESBLs/Carbapenemases associated with plasmids?**

That seems like a lot of AMR genes for one plasmid! Let's do a deeper dive into this plasmid by looking directly at the **mob-suite** and **RGI** results.

Download the contig_report output file for the mobile-AMR rich clinical sample (SRR14026555):
`module9/arete_results/annotation/mob_recon/SRR14026555_mob_recon/contig_report.txt`

This file contains the detailed plasmid prediction and classification data for this particular genome assembly. Open this file in whatever way you are most comfortable to quickly view it (libreoffice/excel, dragging and dropping into microreact, rstudio, jupyter etc).

The **molecule_type** column tells us what the corresponding contig in **contig_id** was classified as by mob-suite. As you'd expect from an assembly, most contigs are predicted to be chromosomal in origin.



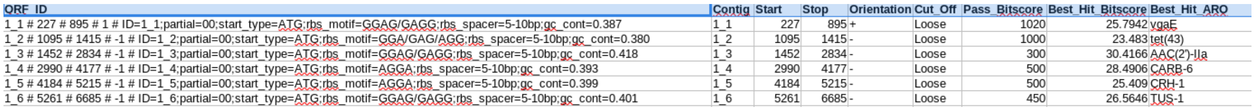
molecule_type	primary_cluster_	secondary_cluster_	contig_id
chromosome	-	-	33_length=28736_depth=0.97x
chromosome	-	-	34_length=27201_depth=1.07x
chromosome	-	-	35_length=26773_depth=0.95x
chromosome	-	-	36_length=26644_depth=0.86x
chromosome	-	-	37_length=26171_depth=1.08x
plasmid	AC733	-	38_length=25571_depth=0.87x
chromosome	-	-	39_length=25505_depth=0.90x
chromosome	-	-	40_length=25003_depth=0.85x
chromosome	-	-	41_length=23834_depth=1.06x
chromosome	-	-	42_length=23360_depth=1.04x
chromosome	-	-	43_length=23042_depth=0.83x
chromosome	-	-	44_length=22760_depth=0.90x
chromosome	-	-	45_length=22424_depth=0.97x
chromosome	-	-	46_length=21727_depth=1.56x
chromosome	-	-	47_length=20723_depth=1.03x
chromosome	-	-	48_length=20477_depth=1.02x
chromosome	-	-	49_length=20388_depth=0.96x
chromosome	-	-	50_length=20188_depth=1.02x
chromosome	-	-	51_length=19688_depth=0.62x
chromosome	-	-	52_length=19517_depth=1.06x
chromosome	-	-	53_length=17337_depth=0.77x
chromosome	-	-	54_length=17167_depth=1.00x
chromosome	-	-	55_length=17011_depth=1.03x
chromosome	-	-	56_length=16572_depth=0.93x
chromosome	-	-	57_length=16147_depth=0.98x
chromosome	-	-	58_length=15862_depth=1.07x
chromosome	-	-	59_length=15191_depth=1.01x
plasmid	AC733	-	60_length=15147_depth=0.78x
chromosome	-	-	61_length=14997_depth=0.92x
plasmid	AC733	-	62_length=14822_depth=0.83x
chromosome	-	-	63_length=14678_depth=0.93x

5. **How many different plasmids are identified by mob-suite for this genome? Note: there are more than was visualised in microreact as that data only included plasmids with AMR genes.**
6. **What do you think it means to see many different contigs being assigned as AC733?**
7. **How big is the largest AC733 contig?**

Now let's take the contig of the largest AC733 fragment (contig 38) and cross-reference this data to the RGI data.

Download the rgi report for SRR14026555 and open it in your CSV viewer of choice as above:
`module9/arete_results/annotation/rgi/SRR14026555_rgi.txt`

Now scroll down the contig column until you get to contig 38 (i.e., the values that start 38_*)



ORF_ID	Contig	Start	Stop	Orientation	Cut_Off	Pass_Bitscore	Best_Hit_Bitscore	Best_Hit_ARO
1_1 # 227 # 895 # 1 # ID=1_1;partial=00;start_type=ATG;rbs_motif=GGAG/GAGG;rbs_spacer=5-10bp;gc_cont=0.387	1_1	227	895	+	Loose	1020	25.7942	vgaE
1_2 # 1095 # 1415 # -1 # ID=1_2;partial=00;start_type=ATG;rbs_motif=GGA/GAG/AGG;rbs_spacer=5-10bp;gc_cont=0.380	1_2	1095	1415	-	Loose	1000	23.483	tet(43)
1_3 # 1452 # 2834 # -1 # ID=1_3;partial=00;start_type=ATG;rbs_motif=GGAG/GAGG;rbs_spacer=5-10bp;gc_cont=0.418	1_3	1452	2834	-	Loose	300	30.4166	AAC(2')-IIa
1_4 # 2990 # 4177 # -1 # ID=1_4;partial=00;start_type=ATG;rbs_motif=AGGA;rbs_spacer=5-10bp;gc_cont=0.393	1_4	2990	4177	-	Loose	500	28.4906	CARB-6
1_5 # 4184 # 5215 # -1 # ID=1_5;partial=00;start_type=ATG;rbs_motif=AGGA;rbs_spacer=5-10bp;gc_cont=0.399	1_5	4184	5215	-	Loose	500	25.409	CRH-1
1_6 # 5261 # 6685 # -1 # ID=1_6;partial=00;start_type=ATG;rbs_motif=GGAG/GAGG;rbs_spacer=5-10bp;gc_cont=0.401	1_6	5261	6685	-	Loose	450	26.5646	TUS-1

8. **Based on the AMR module and the explanation of cut-offs used by RGI, do you notice anything unusual about all of the AMR genes listed on contig 38?**

9. **Building off your answers to 6 and 8, do you actually think this clinical genome really contains a scary plasmid stuffed full of worrying AMR genes? Why or why not?**

This highlights a key lesson: data is great, workflows like ARETE allow you to generate it very quickly, and tools like microreact let you visualise it very quickly. However, if you aren't careful and don't actually dive into the disaggregated outputs it is very easy to be misled!