



THE UNIVERSITY OF  
**SYDNEY**

—  
**Sydney  
Informatics Hub**

[informatics.sydney.edu.au](http://informatics.sydney.edu.au)  
[sih.info@sydney.edu.au](mailto:sih.info@sydney.edu.au)

The Sydney Informatics Hub provides support, training, and advice on research data, analyses and computing. Talk to us about your computing infrastructure, digital tools and data governance needs. You can also collaborate on grants and projects with our Data Science experts.

# Introduction to DNA-Sequencing

## Part 1: DNA/RNA Sequencing Course, Westmead

### Course Overview

Introduction to DNA sequencing

Group activity: Reference mapping

Genome assembly pipeline

Hands-on activity: Reference mapping on Galaxy

## Part B: Group activity – The DNA detectives in a jam

Farmer McSweetie has a problem. He grows the world's sweetest strawberries but something has gone badly wrong. His latest crop taste terrible and no-one will buy them.

*What could have happened?*

He has had his strawberries tested and discovered that they don't have enough of the right sort of sugar to make them sweet. An enzyme called Acid Invertase converts normal sugar (sucrose) to extra sweet fruit sugar (fructose) but it doesn't seem to be working properly and he doesn't know why.

It's time to call in the DNA detectives (that is you, by the way) to try to unravel the mystery of what is wrong with McSweetie's strawberries.

First, we will get the DNA sequence from the mutant strawberries. (A mutation is just a change in the DNA sequence compared to normal). Unfortunately, we can't read all of it at once so we have to read lots of smaller fragments and put them together so they overlap and we can get the whole sequence. A picture of some students doing this is shown below.



See how each bit of DNA sequence overlaps with the next one so we can read the whole thing.

When we have built our sequence we can compare it to what it should be. We will use a program to search through all the known DNA sequences for strawberry. We need to be very careful as we copy the A's, C's, G's and T's as just one letter out of place could spell disaster.

When you have got your DNA sequence assembled to make a long molecule (called a contig), carefully read off the DNA letters in order.

Here is the sequence for a juicy strawberry. Find the base in your sequence that is different and circle it. Write the letter you find underneath.

TATCATTTCCAGCCTTGCAAGAATTGGATGAACGATC  
CTAATGGGCCAATGGTTTACAAGAACGTATACCATCA  
TTTTTATCAATATAATCCCGATGGT

Let's learn more about this DNA sequence.

In a web browser go to:

### Web BLAST



## Basic Local Alignment Search Tool

**BLAST** finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

[Learn more](#)

[https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE\\_TYPE=BlastSearch&LINK\\_LOC=blasthome](https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome)

Copy your sequence (the above sweet strawberry DNA sequence, modified with the variant you identified) and paste it into the 'Enter Query Sequence' box:

Enter accession number(s), gi(s), or FASTA sequence(s) ?

**BLAST**

Then press

A comparison with DNA sequences in the NCBI BLAST database your sequence is similar to:



## Part C: A genome assembly pipeline in Galaxy

### Case study

# Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study

Simon R Harris\*, Edward J P Cartwright\*, M Estée Török, Matthew T G Holden, Nicholas M Brown, Amanda L Ogilvy-Stuart, Matthew J Ellington, Michael A Quail, Stephen D Bentley, Julian Parkhill†, Sharon J Peacock†

### Summary

**Background** The emergence of meticillin-resistant *Staphylococcus aureus* (MRSA) that can persist in the community and replace existing hospital-adapted lineages of MRSA means that it is necessary to understand transmission dynamics in terms of hospitals and the community as one entity. We assessed the use of whole-genome sequencing to enhance detection of MRSA transmission between these settings.

**Methods** We studied a putative MRSA outbreak on a special care baby unit (SCBU) at a National Health Service Foundation Trust in Cambridge, UK. We used whole-genome sequencing to validate and expand findings from an infection-control team who assessed the outbreak through conventional analysis of epidemiological data and antibiogram profiles. We sequenced isolates from all colonised patients in the SCBU, and sequenced MRSA isolates from patients in the hospital or community with the same antibiotic susceptibility profile as the outbreak strain.

**Findings** The hospital infection-control team identified 12 infants colonised with MRSA in a 6 month period in 2011, who were suspected of being linked, but a persistent outbreak could not be confirmed with conventional methods. With whole-genome sequencing, we identified 26 related cases of MRSA carriage, and showed transmission occurred within the SCBU, between mothers on a postnatal ward, and in the community. The outbreak MRSA type was a new sequence type (ST) 2371, which is closely related to ST22, but contains genes encoding Panton-Valentine leucocidin. Whole-genome sequencing data were used to propose and confirm that MRSA carriage by a staff member had allowed the outbreak to persist during periods without known infection on the SCBU and after a deep clean.

**Interpretation** Whole-genome sequencing holds great promise for rapid, accurate, and comprehensive identification of bacterial transmission pathways in hospital and community settings, with concomitant reductions in infections, morbidity, and costs.

Lancet Infectious Diseases, 2013, 13(2):130-136  
[https://dx.doi.org/10.1016/S1473-3099\(12\)70268-2](https://dx.doi.org/10.1016/S1473-3099(12)70268-2)

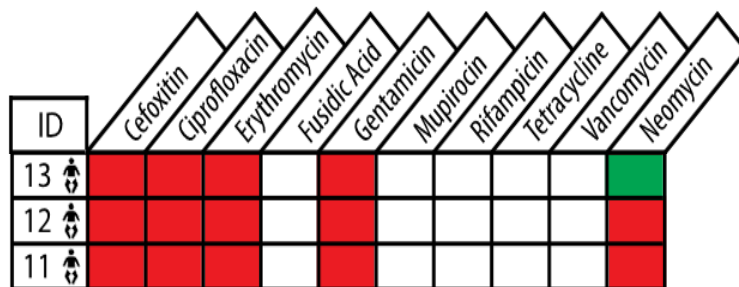




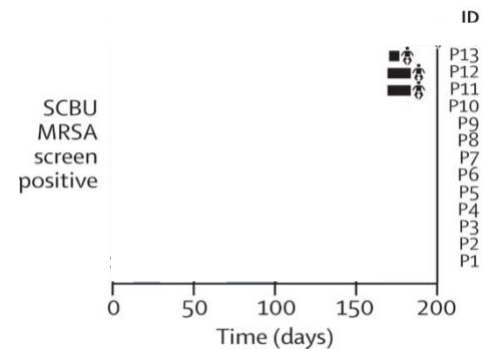
## Background

A potential outbreak of MRSA in the Special Care Babies Unit (SCBU) was indicated by three babies having similar antibiotic susceptibility profiles and who were admitted within days of each other.

**A.**

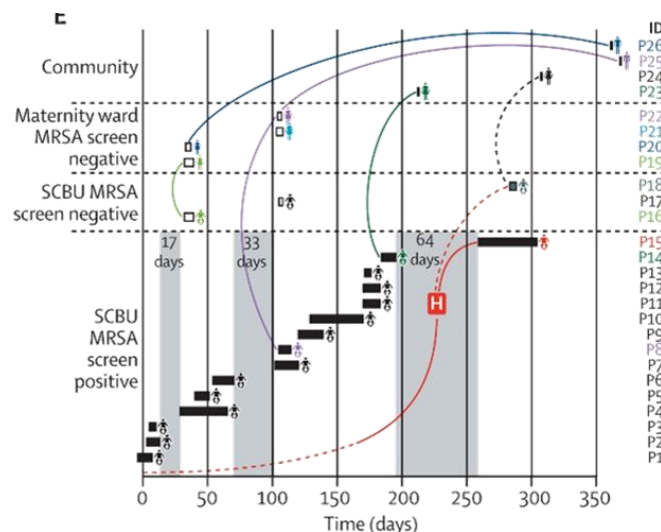


**B.**



**Figure 1. (A)** Antibiogram and **(B)** epidemiology of three possible outbreak cases in the SCBU.

Infection control responded to the outbreak by performing a deep clean, implemented weekly MRSA screening, and reinforced infection control. 13 (+3) babies with MRSA were identified within 6 months prior to the 3 outbreak cases. 1 case was identified after the deep clean of the SCBU. Infection control investigated 17 cases: 12 cases were classified as part of the outbreak and 5 cases were classified as unrelated to the outbreak.



**Figure 2.** Epidemiology of patients in the SCBU and other patients with linked MRSA infection detected in the community.

## Whole Genome Sequencing

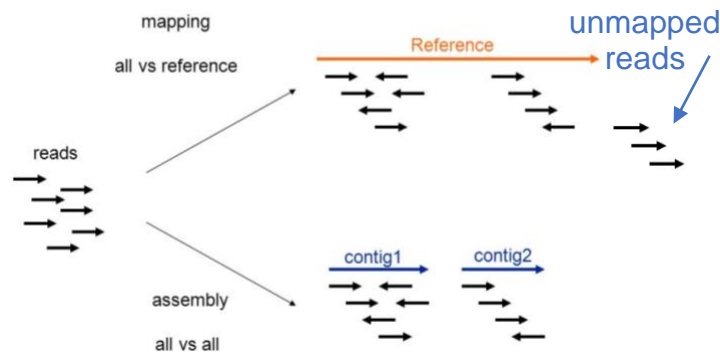
Did the outbreak extend over 6 months?

How accurate was the infection control investigation?

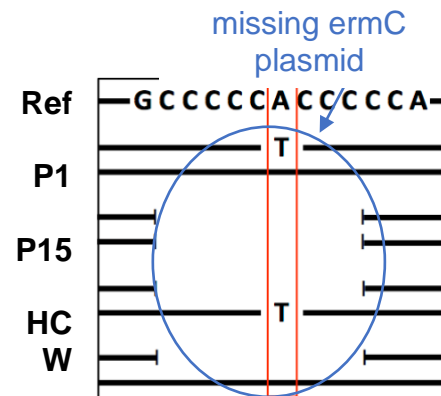
Reference mapping

(approach used by Harris et al.)

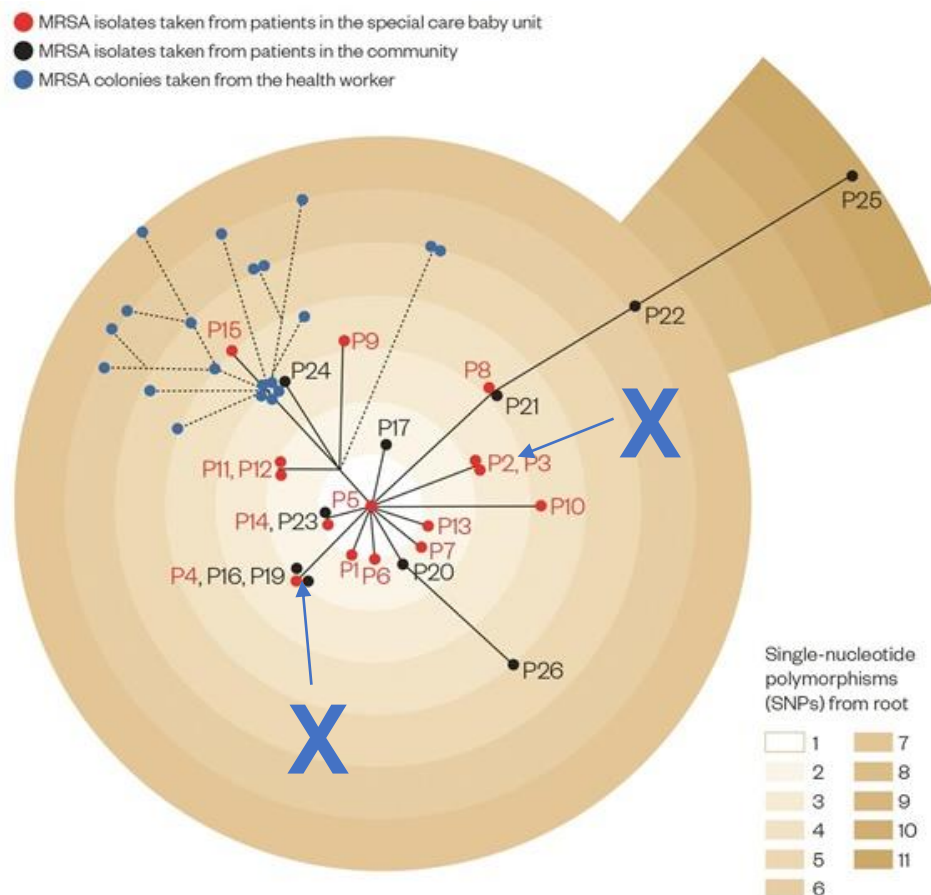
A.



B.



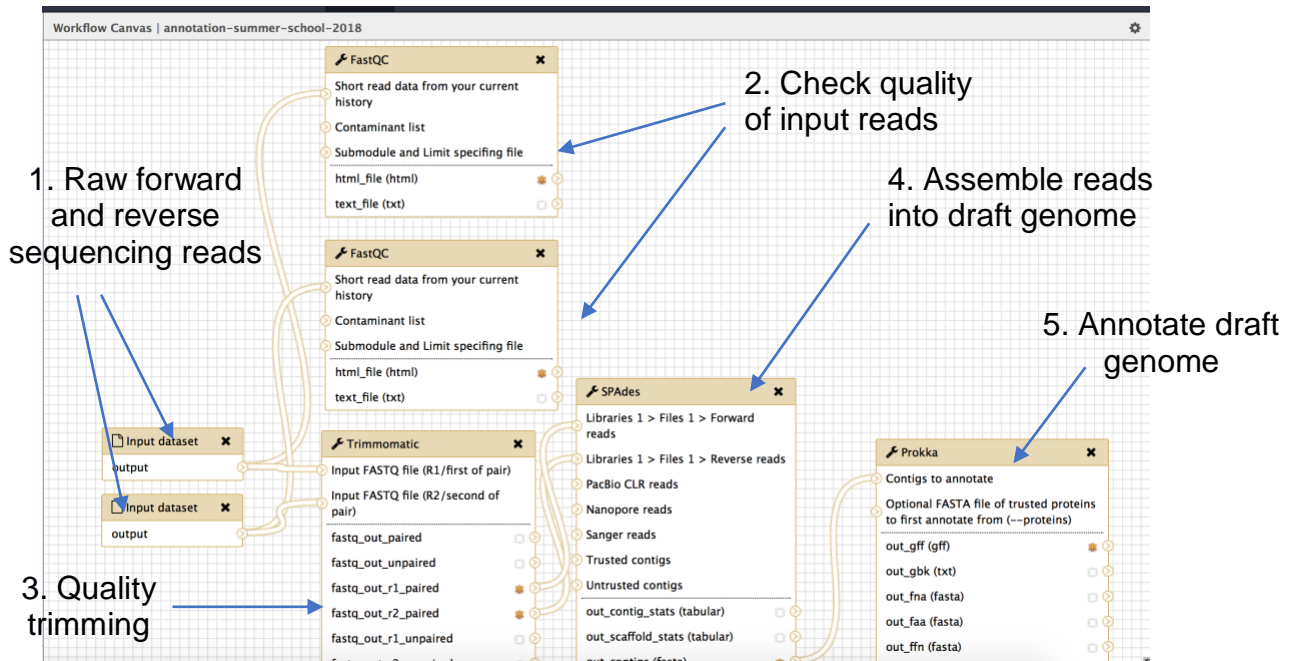
C.



**Figure 3. (A)** Creating a draft genome by reference mapping vs de novo assembly. **(B)** A missing plasmid found in samples when compared to the reference. **(C)** Phylogenetic tree based on core single nucleotide polymorphisms (SNPs), including P1 and P3 who infection control had initially excluded from the outbreak based on their antibiogram.

## Pipelines

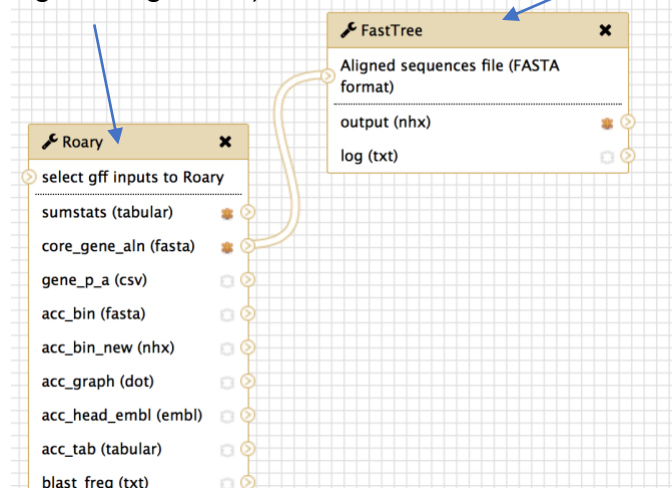
## Genome assembly and annotation



## Pangenome-based phylogeny

6. Compare genes  
(core gene alignment)

7. Make  
phylogenetic tree



These workflows can be found here:

[https://galaxy-mel.genome.edu.au/galaxy/workflow/list\\_published](https://galaxy-mel.genome.edu.au/galaxy/workflow/list_published)

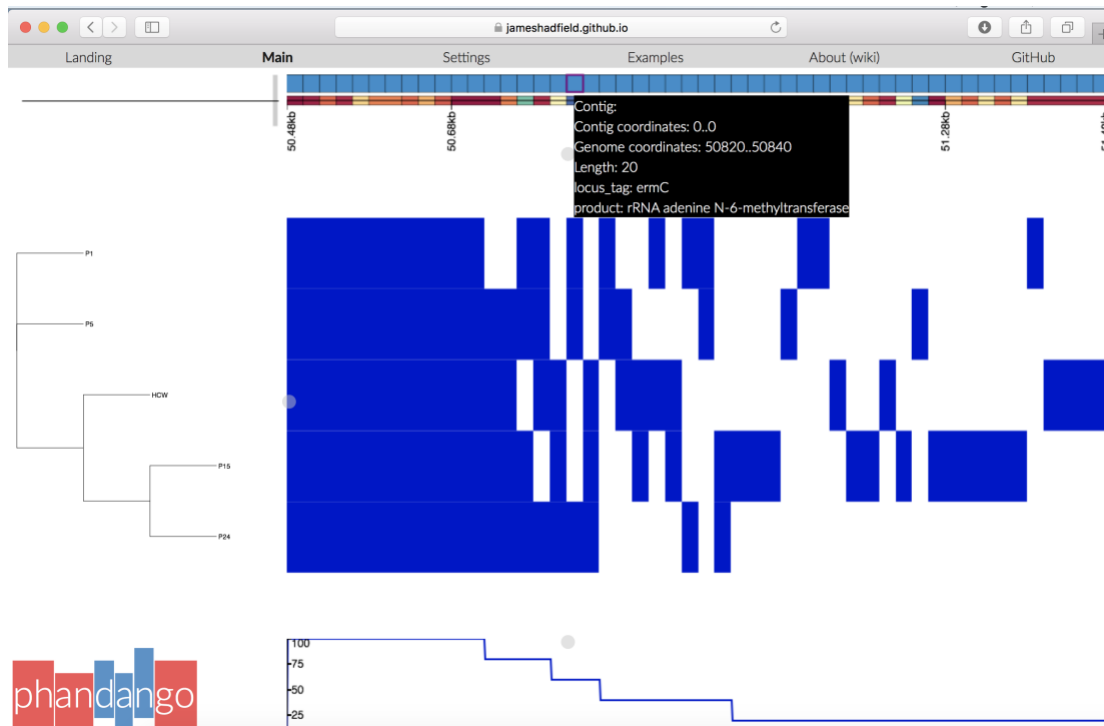


## Results

This is the phylogenetic tree produced by the pipeline and viewed in *Phandango* (Hadfield, et al. *Bioinformatics*, 2017).






<https://jameshadfield.github.io/phandango>

Here we are focusing on the *ermC* gene (rRNA adenine N-6-methyltransferase, confers Erythromycin resistance).





Notice that 3 samples, P1, P5, and P24 have the ermC gene. However, the paper reports that P24 was susceptible to Erythromycin and was missing the plasmid carrying ermC.

Table S1. Antimicrobial susceptibility pattern of MRSA isolates.

ID	Cefoxitin	Ciprofloxacin	Erythromycin	Fusidic Acid	Gentamicin	Mupirocin	Rifampicin	Tetracycline	Vancomycin	Neomycin
 H	Red	Red	Blue	Red	White	White	White	White	White	Red
24 	Red	Red	White	Red	White	White	White	White	White	Red
15 	Red	Red	White	Red	White	White	White	White	White	Red
5 	Red	Red	Red	Red	White	White	White	White	White	Red
1 	Red	Red	Red	Green	White	White	White	White	White	Green

Infant  
Healthcare  
worker

Red means resistant and white means susceptible to a given antibiotic. Green, initial result proved incorrect on repeat testing and changed designation from susceptible to resistant. A healthcare worker had twenty MRSA colonies taken from the primary plate; the blue box for erythromycin signifies that some colonies were susceptible to this antibiotic (n=18) and some were resistant (n=2).

Let’s explore this.

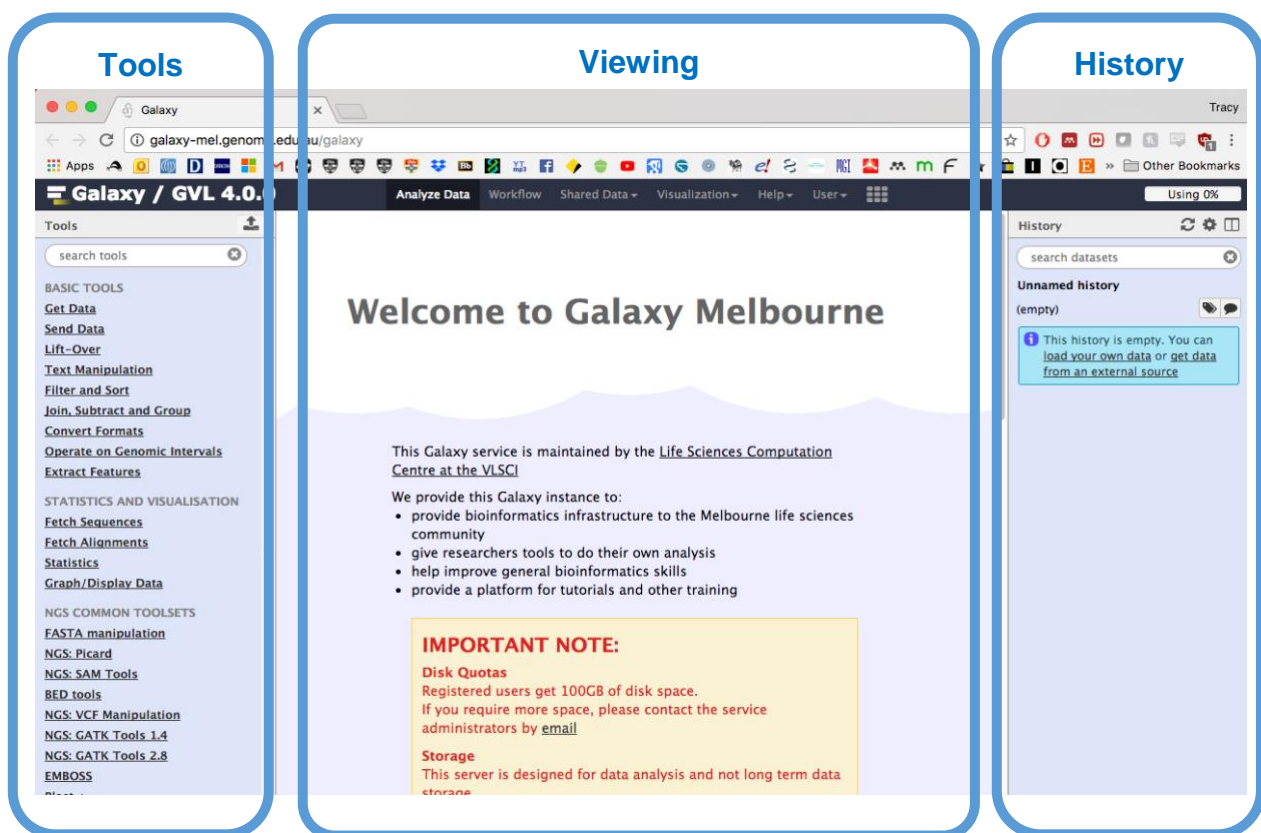
## Part C. Reference mapping in Galaxy

### Galaxy

This part will be based on a web-based platform called **Galaxy**. It is a useful tool for biologists to perform analysis workflows and contains a lot of common bioinformatics tools. Registered users can use 100Gb of space for free.

We will use Galaxy (Melbourne instance) for aligning reads to the plasmid carrying the ermC gene ([CP002148](http://CP002148)).

- Go to the Melbourne Galaxy Server (in Firefox, Chrome or Safari): <http://galaxy-mel.genome.edu.au/galaxy>
- If you haven't already, register (click User > Register), then log in




## Part C1. Import your data

1. Open the public histories:

[https://galaxy-mel.genome.edu.au/galaxy/history/list\\_published](https://galaxy-mel.genome.edu.au/galaxy/history/list_published)

2. Click on SIH\_DNASeq\_Mapping

### Published Histories



Advanced Search

Name	Annotation	Owner
SIH_DNASeq_Mapping		sih.training



3. Import the dataset into your 'history'

Galaxy / GVL 4.0.0

Analyze Data Workflow Shared Data Visualization Help User

Published Histories | sih.training | SIH\_DNASeq\_Mapping

SIH\_DNASeq\_Mapping

1.02 GB

search datasets

Dataset

Dataset	Annotation
8: P24 reads mapped to plasmid-BAM	
7: P1 reads mapped to plasmid-BAM	
6: Plasmid-CP002148-1.gb	

Import history

Make a copy of this history and switch to it

sih.trai

Relatu

All pub

Publsh

Ratin

Commu

(0 rating

Yours

Taos

### Importing history "SIH\_DNASeq\_Mapping"

Enter a title for the new history:

imported: SIH\_DNASeq\_Mapping

Cancel Import

Galaxy / GVL 4.0.0

Analyze Data Workflow Shared Data Visualization Help User

Using 5%

Tools

search tools

BASIC TOOLS

- Get Data
- Send Data
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Operate on Genomic Intervals
- Extract Features

STATISTICS AND VISUALISATION

- Fetch Sequences
- Fetch Alignments
- Statistics
- Graph/Display Data

NGS COMMON TOOLSETS

- FASTA manipulation
- NGS Pipel

Welcome to Galaxy Melbourne

This Galaxy service is maintained by the Life Sciences Computation Centre at the VLSCI

We provide this Galaxy instance to:

- provide bioinformatics infrastructure to the Melbourne life sciences community
- give researchers tools to do their own analysis
- help improve general bioinformatics skills
- provide a platform for tutorials and other training

**IMPORTANT NOTE:**

Disk Quotas

Registered users get 100GB of disk space

History

search datasets

imported: SIH\_DNASeq\_Mapping

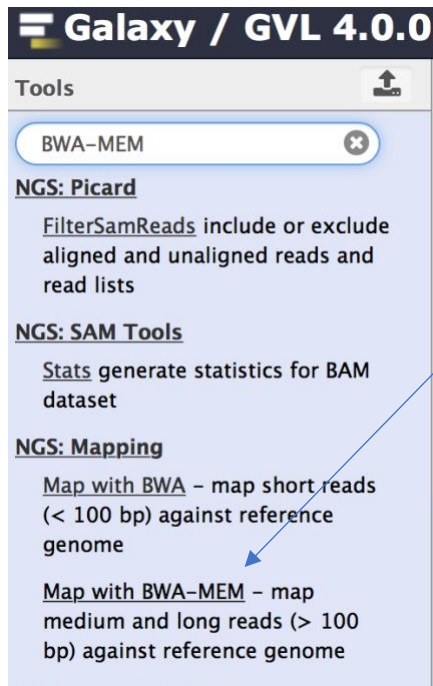
8 shown

1.02 GB

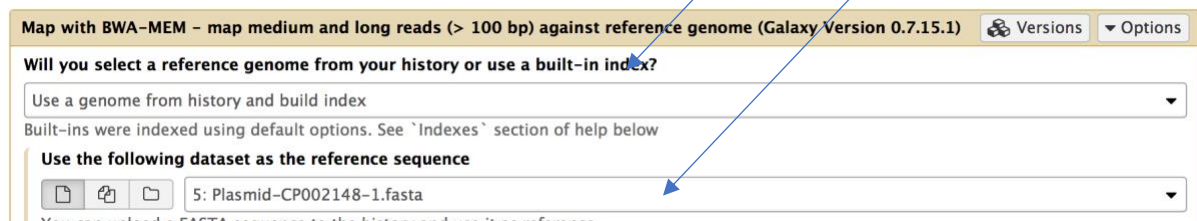
- 8: P24 reads mapped to plasmid-BAM
- 7: P1 reads mapped to plasmid-BAM
- 6: Plasmid-CP002148-1.gb
- 5: Plasmid-CP002148-1.fasta
- 4: P24\_2.fastq
- 3: P24\_1.fastq
- 2: P1\_2.fastq
- 1: P1\_1.fastq

## Part C2. Perform the reference mapping task


1. In the left 'Tools' pane, type in 'BWA-MEM', and click on 'Map with BWA-MEM'.

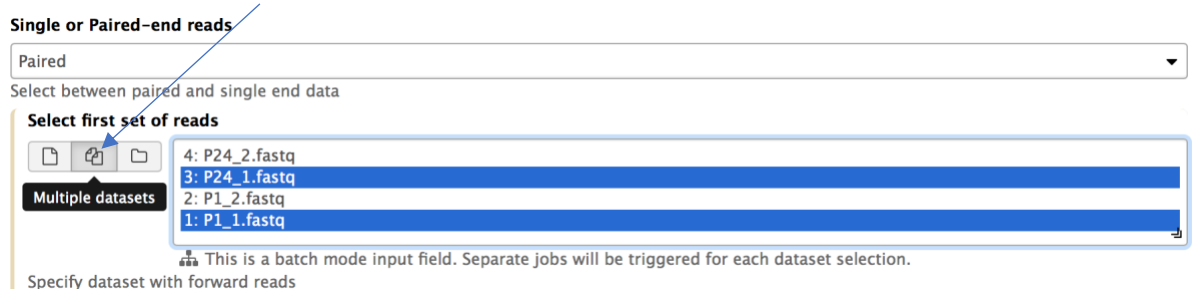



2. Select 'Use a genome from history and build index', then select 'Plasmid-CP002148-1.fasta'



3. Select 'Paired-Ends'

4. For forward read selection, click the multiple datasets button  and using (ctrl on Windows, command on Mac) select both P1\_1.fastq and P24\_1.fastq.



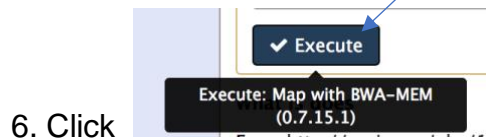
5. For reverse read selection, click the multiple datasets button  and select both P1\_2.fastq and P24\_2.fastq.



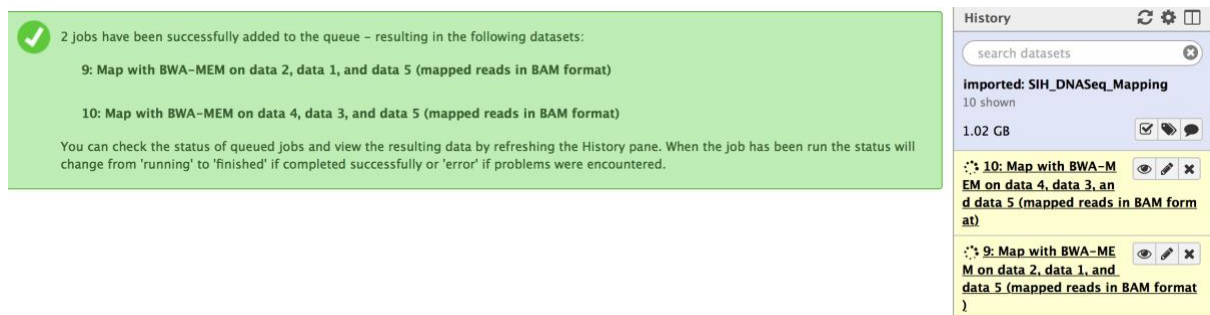
Select second set of reads



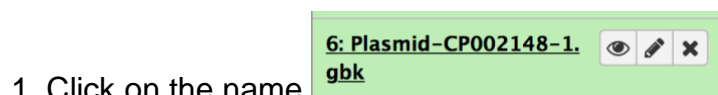
Specify dataset with reverse reads




10. The mapping jobs will be submitted to the Galaxy cluster for processing – **Be Patient!** – a key lesson in Genomics.

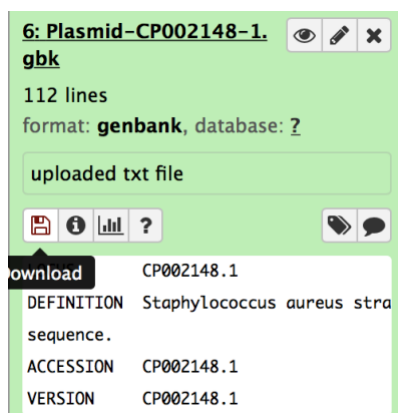


### Part C3. Visualise the results the Integrated Genomics Viewer (IGV)

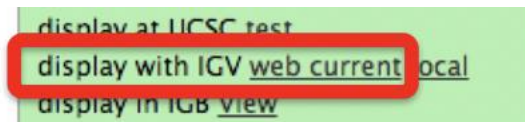


1. Click on the name

2. Click the save icon  to download the Plasmid genbank file to your desktop.

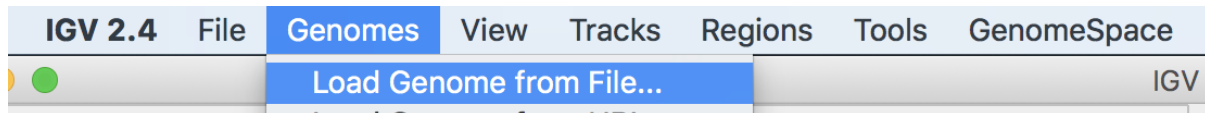


3. Click on '9 Map with BWA-MEM on data 2, data 1, and data 5 (mapped reads in BAM format)'

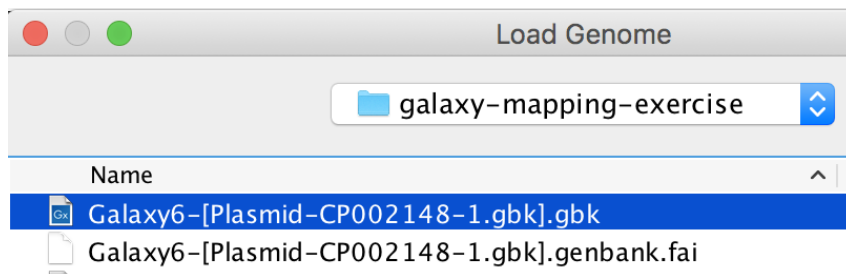


4. Click on “web\_current” next to “display with IGV”. Save the ‘igv.jnlp’ file and double click the file to open it.

5. IGV will open, defaulting to the Human (hp19) genome. Go to Genome -> Load Genome from File.



And select the Plasmid.gbk (genbank) file

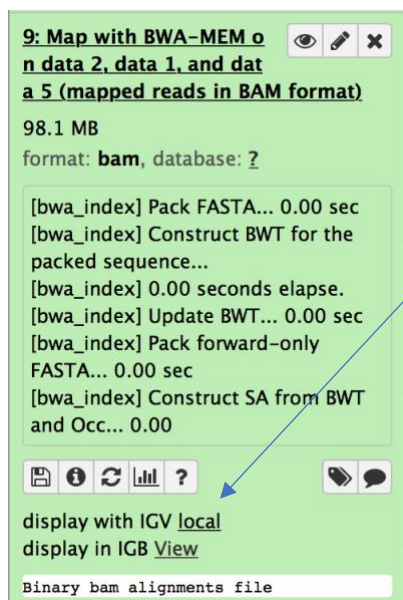


6. In Galaxy, select ‘display with IGV local’ for both:

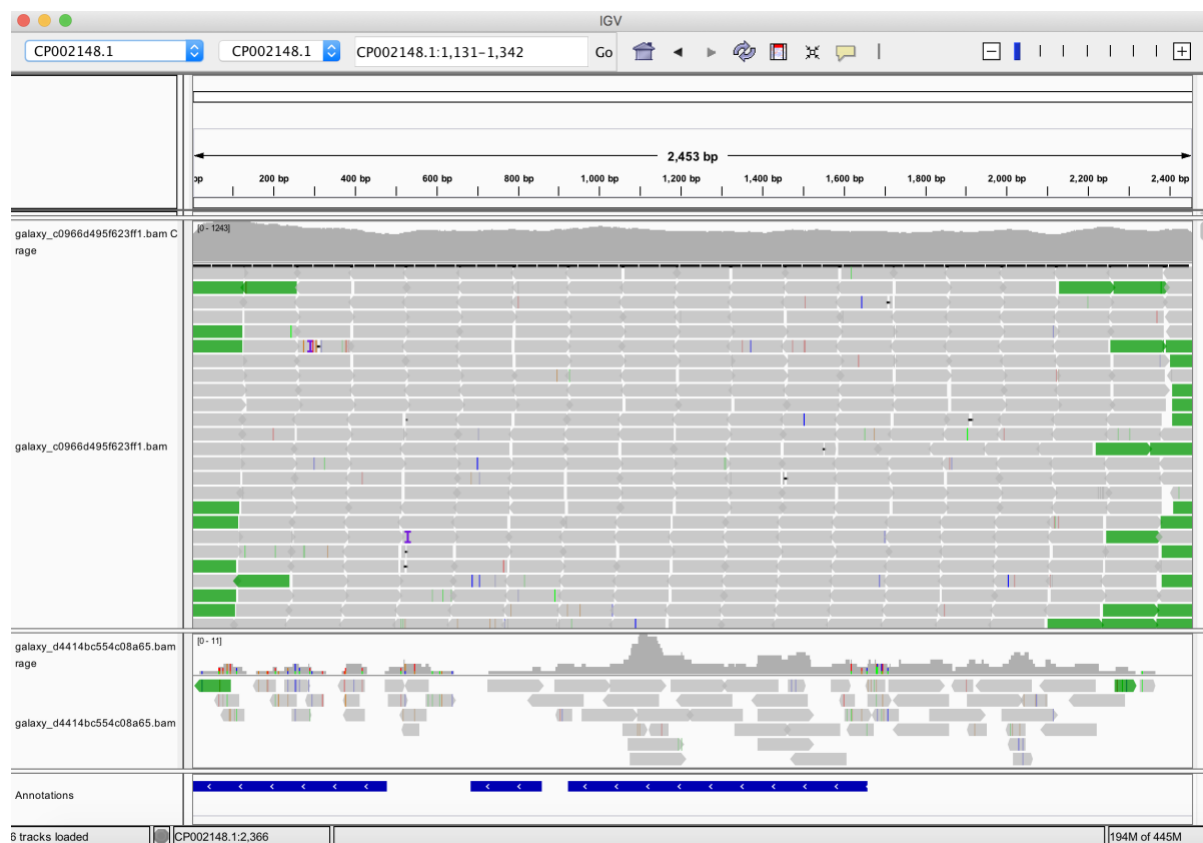
**9 Map with BWA-MEM on data 2, data 1, and data 5 (mapped reads in BAM format)**

And

**10 Map with BWA-MEM on data 4, data 3, and data 5 (mapped reads in BAM format)**



7. Navigate back to IGV, you will see the reads that mapped to this plasmid containing the Erythromycin gene. P1 has a lot of coverage, whereas P24 does not have enough coverage to support it containing this plasmid.

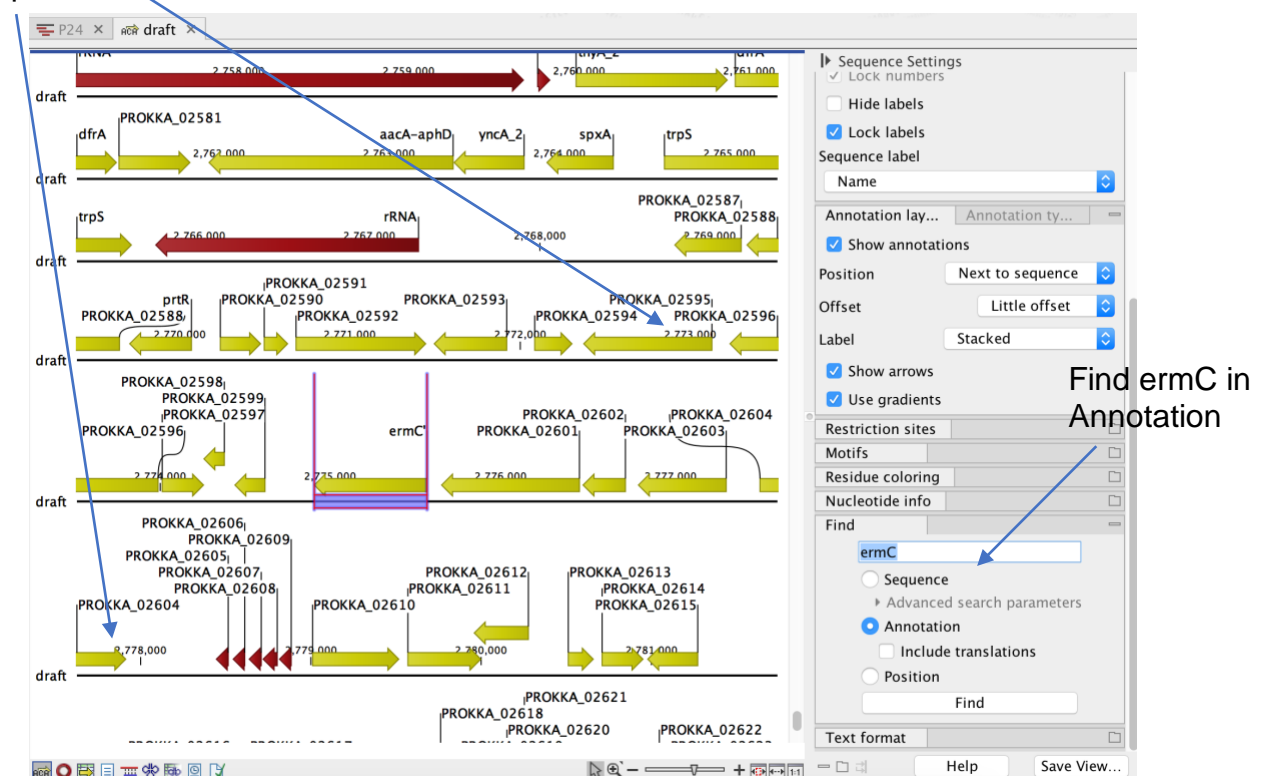


### Part C4. Further investigation

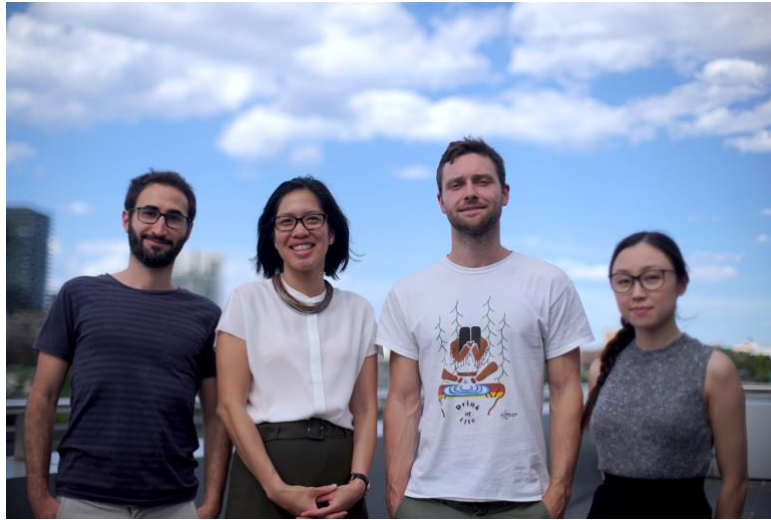
The assembly and annotation pipeline created a draft genome for P24 by denovo assembly. When looking for the *ermC* gene in the chromosome (as we did not find it on the plasmid), we see a transposase and IS protein around the *ermC* gene.

**Discussion: Did the *ermC* gene move from the plasmid to the chromosome through a transposase?**

Transposase/IS  
protein



# Thank you



Sydney Informatics Hub – Research Computing team  
bioinformatics, modelling, simulation

The Sydney Informatics Hub is a Core Research Facility. We provide support, training, and advice on research data, analyses and computing. We also collaborate on grants and projects. Talk to us about your computing infrastructure, digital tools and data governance needs. Contact [sih.info@sydney.edu.au](mailto:sih.info@sydney.edu.au) for more information.  
<https://informatics.sydney.edu.au>



THE UNIVERSITY OF  
**SYDNEY**  
—  
**Sydney  
Informatics Hub**