

A Galaxy of learning:

Bioinformatics tutorials based on Galaxy

Simon Gladman



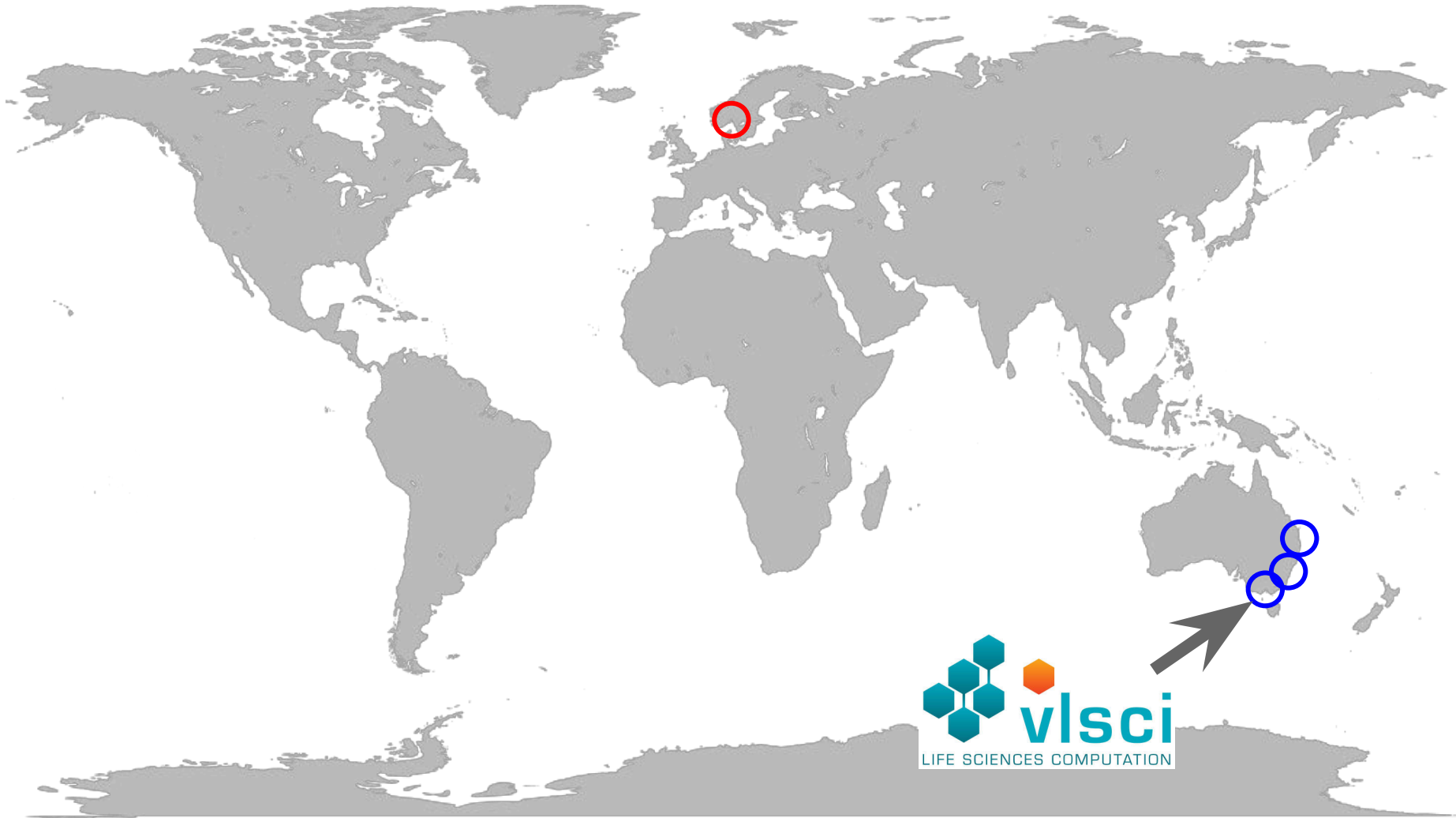
MONASH
University



*Victorian
Bioinformatics
Consortium*

Background and the Genomics Virtual Lab

We come from a land down-under



The Genomics Virtual Lab

Nationally funded and distributed platform for genomic analyses

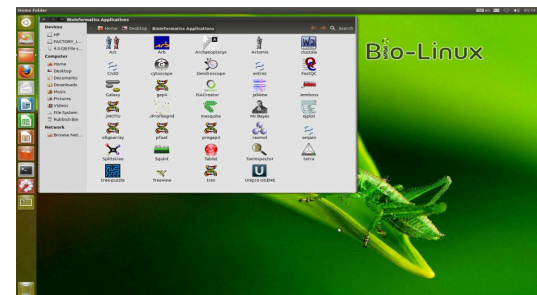
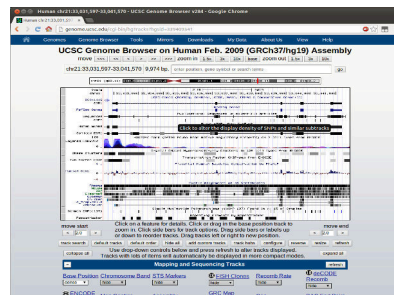
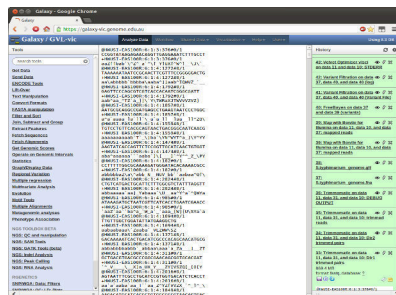
Genomics Virtual Lab

Galaxy

UCSC Browser

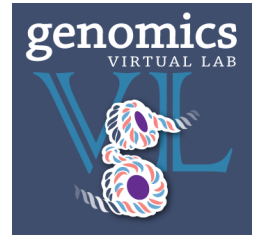
Bio Linux

- Tools
- Workflows
- Visualisation
- Tutorials



Australian Research
Cloud Infrastructure

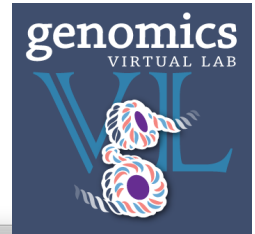
The Genomics Virtual Lab



- Sets of scripts and machine images for building GVL instances
 - Including Galaxy and CloudBioLinux
 - Reference datasets
- Can run on:
 - Stand alone systems
 - Cloud via Cloudman
- Public Galaxy service for training
 - galaxy-tut.genome.edu.au
- A suite of tutorials and protocols for common genomics tasks

The Genomics Virtual Lab

www.genome.edu.au



Screenshot of the Genomics Virtual Laboratory website interface.

The browser address bar shows <https://genome.edu.au/wiki/GVL>.

The page title is "Welcome to the Genomics Virtual Laboratory".

The page content is organized into a grid of tiles:

- USE** (Genomics Virtual Lab logo)
- LEARN** (Genomics Virtual Lab logo) - This tile is circled in red.
- GET** (Genomics Virtual Lab logo)
- BROWSE** (UCSC logo)
- DO** (Genomics Virtual Lab logo)
- Get GVL Data** (Genomics Virtual Lab logo)
- HELP** (Genomics Virtual Lab logo)
- Publications** (Nature and Science logos)
- ABOUT** (Genomics Virtual Lab logo)
- PROJECT UPDATES** (Genomics Virtual Lab logo)

The left sidebar contains a navigation menu:

- GVL Home
- About GVL
- USE Galaxy
- LEARN Galaxy
- GET Galaxy
- BROWSE UCSC
- DO Genomics
- Get GVL Data
- GVL Help
- Publications
- ENCODE Data
- Help Centre

The right sidebar contains a search bar and a "Go" button.

The bottom of the page features a footer with the text: "The Genomics Virtual Laboratory takes the IT out of Bioinformatics. It lets Biologists use a suite of genomics analysis tools that currently often require specialist assistance. GVL"

Training Using Galaxy

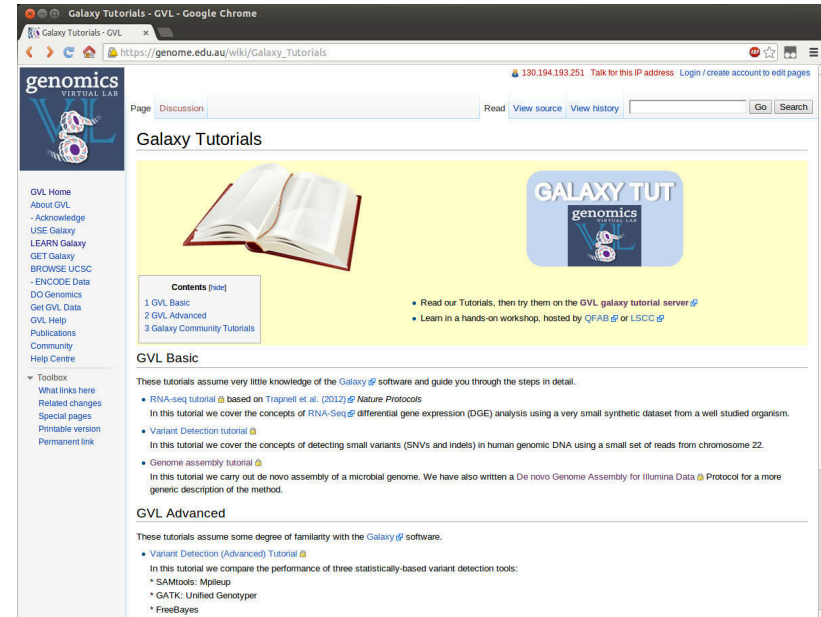
Galaxy Tutorials

Current:

- *de novo* Assembly
- RNA-Seq DGE (Basic)
- Variant detection (Basic)
- Variant detection (Advanced)

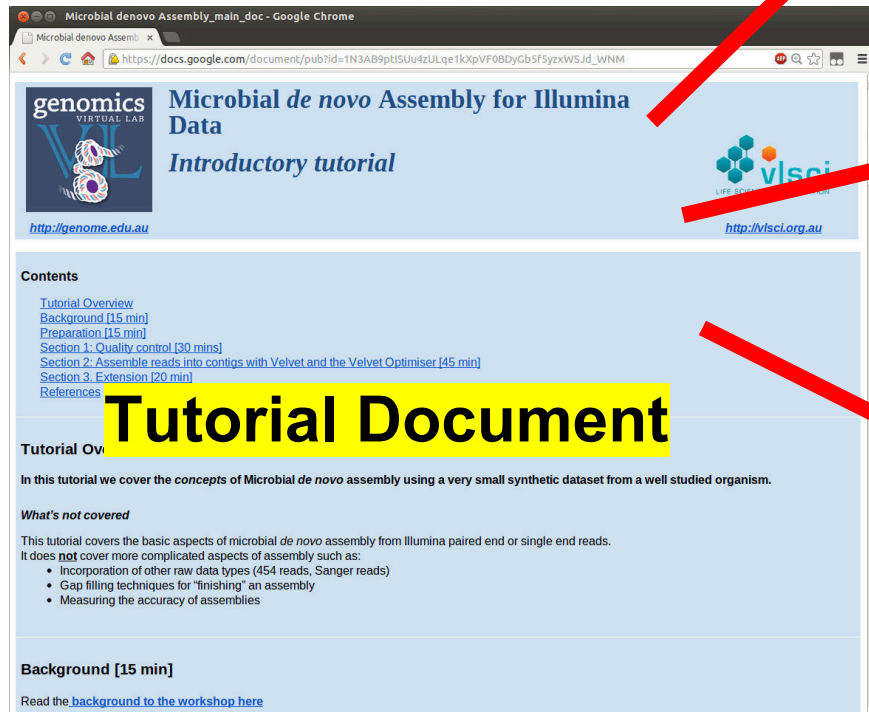
Forthcoming:

- Variant detection (Somatic)
- RNA-Seq DGE (Advanced)
- Transcriptome assembly
- Metagenomics
- ChIP-Seq
- Bacterial Genome Annotation



Galaxy Tutorials

- Layered Google documents
- Use real datasets
- galaxy-tut.genome.edu.au
- Histories at each step
- Galaxy server image forthcoming



Microbial de novo Assembly for Illumina Data
Introductory tutorial

<http://genome.edu.au>

Contents

- [Tutorial Overview](#)
- [Background \[15 min\]](#)
- [Preparation \[15 min\]](#)
- [Section 1: Quality control \[30 mins\]](#)
- [Section 2: Assemble reads into contigs with Velvet and the Velvet Optimiser \[45 min\]](#)
- [Section 3: Extension \[20 min\]](#)
- [References](#)

Tutorial Overview

In this tutorial we cover the concepts of Microbial *de novo* assembly using a very small synthetic dataset from a well studied organism.

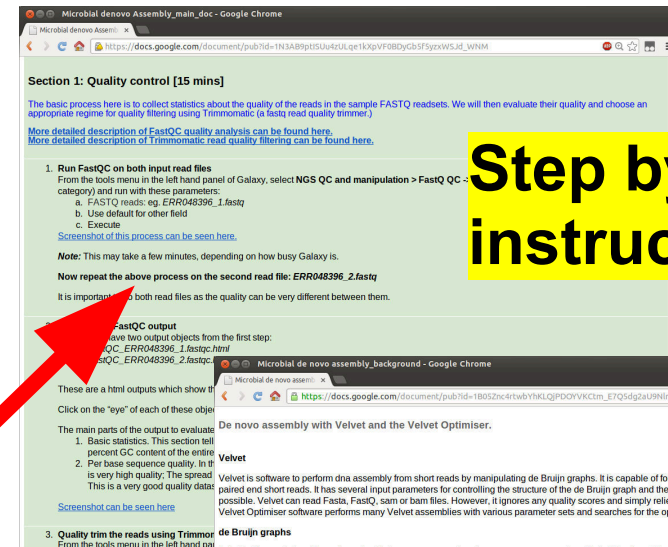
What's not covered

This tutorial covers the basic aspects of microbial *de novo* assembly from Illumina paired end or single end reads. It does **not** cover more complicated aspects of assembly such as:

- Incorporation of other raw data types (454 reads, Sanger reads)
- Gap filling techniques for "finishing" an assembly
- Measuring the accuracy of assemblies

Background [15 min]

[Read the background to the workshop here](#)



Section 1: Quality control [15 mins]

The basic process here is to collect statistics about the quality of the reads in the sample FASTQ readsets. We will then evaluate their quality and choose an appropriate regime for quality filtering using Trimmomatic (a fastq read quality trimmer.)

More detailed description of FastQC quality analysis can be found [here](#).
More detailed description of Trimmomatic read quality filtering can be found [here](#).

1. Run FastQC on both input read files

From the tools menu in the left hand panel of Galaxy, select NGS QC and manipulation > FastQC (category) and run with these parameters:

- FASTQ reads eg. ERR048396_1.fastq
- Use default for other field
- Execute

Screenshot of this process can be seen [here](#).

Note: This may take a few minutes, depending on how busy Galaxy is.

Now repeat the above process on the second read file: ERR048396_2.fastq

It is important to run both read files as the quality can be very different between them.

FastQC output

FastQC will have two output objects from the first step:

- 1. Basic statistics. This section tells you the overall quality of the reads.
- 2. Per base sequence quality. In this section you can see the quality of each base in the reads. This is a very good quality data.

These are a HTML outputs which show the results of the quality control.

Click on the "type" of each of these objects.

The main parts of the output to evaluate:

1. Basic statistics. This section tells you the overall quality of the reads.
2. Per base sequence quality. In this section you can see the quality of each base in the reads. This is a very good quality data.

Screenshot can be seen [here](#).

3. Quality trim the reads using Trimmomatic

From the tools menu in the left hand panel

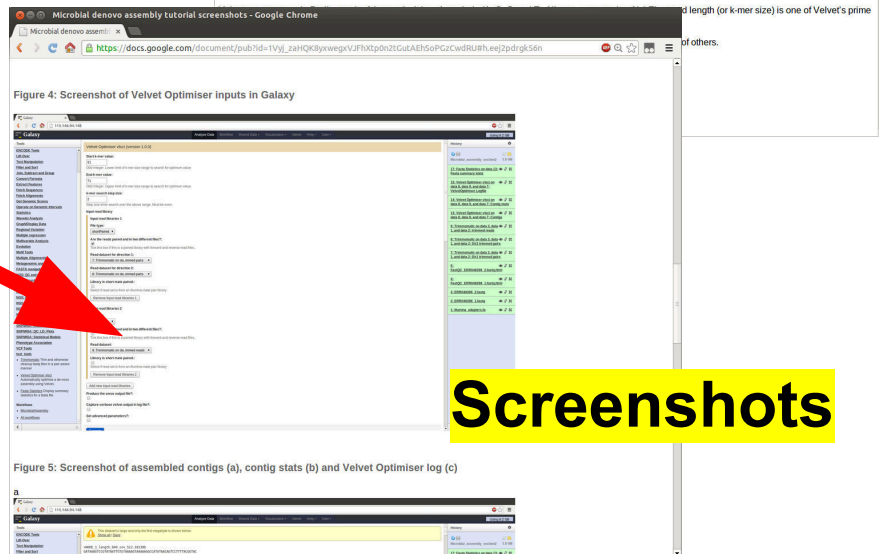
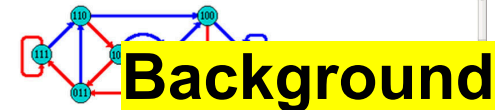
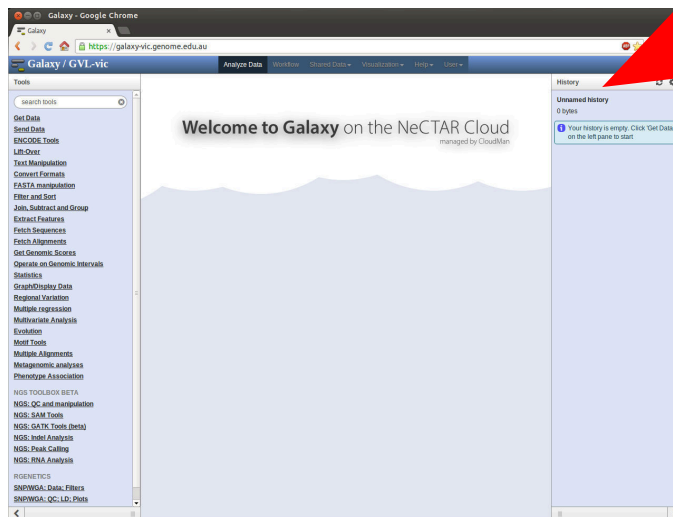
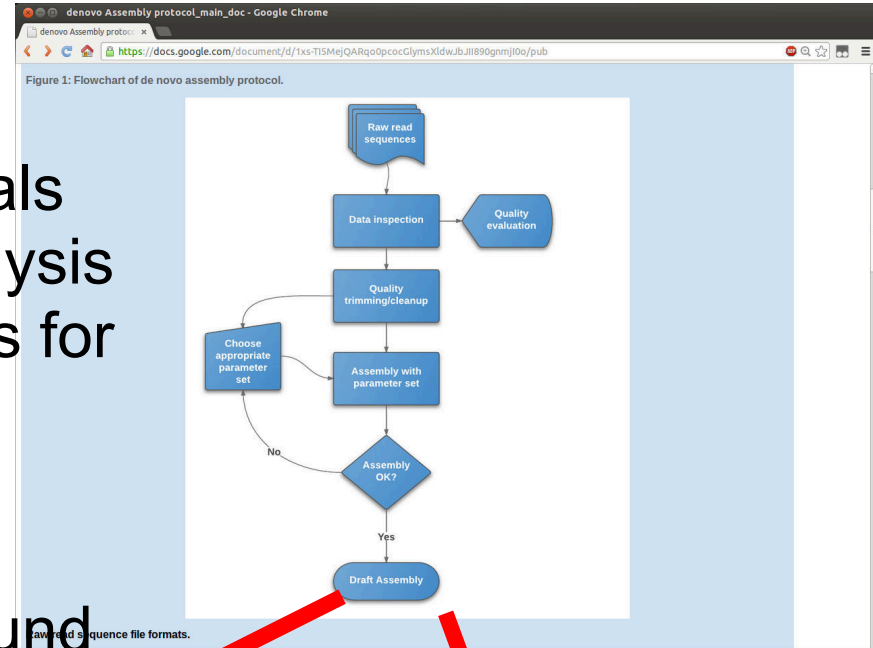


Figure 4: Screenshot of Velvet Optimiser inputs in Galaxy

Figure 5: Screenshot of assembled contigs (a), contig stats (b) and Velvet Optimiser log (c)

Protocols

- Less prescriptive than Tutorials
- Presents concepts of an analysis
- Suggests tools and alternates for each step:
 - Galaxy
 - Command-line
- Includes substantial background



```
simon@hyperion:~$
File Edit View Search Terminal Help

Read type options:
-short1 -shortPaired
-short2 -shortPaired2
-short3 -shortPaired3
-short4 -shortPaired4
-long -longPaired
-reference

Options:
-strand_specific : for strand specific transcriptome sequencing data (default: off)
-reuse_sequences : reuse sequences file (or link) already in directory (no need to provide original filename in this case (default: off))
-nolink : simply prepare sequences file, do not hash reads or prepare Roadmaps file (default: off)
-create_binary : create binary crynifiedseq file (default: off)

Synopsis:
- Short single end reads:
  velveth Assem 29 -short -fastq s_1_sequence.txt

- Paired-end short reads (remember to interleave paired reads):
  velveth Assem 31 -shortPaired -fasta interleaved.fna

- Paired-end short reads (using separate files for the paired reads)
  velveth Assem 31 -shortPaired -fasta -separate left.fa right.fa

- Two channels and some long reads:
  velveth Assem 43 -short -fastq unmapped.fna -longPaired -fasta SangerReads.fasta

- Three channels:
  velveth Assem 35 -shortPaired -fasta pe_lib1.fasta -shortPaired2 pe_lib2.fasta -short3 se_lib1.fasta

Output:
directory/Roadmaps
directory/Sequences
[Both files are picked up by graph, so please leave them there]

simon@hyperion:~$
```

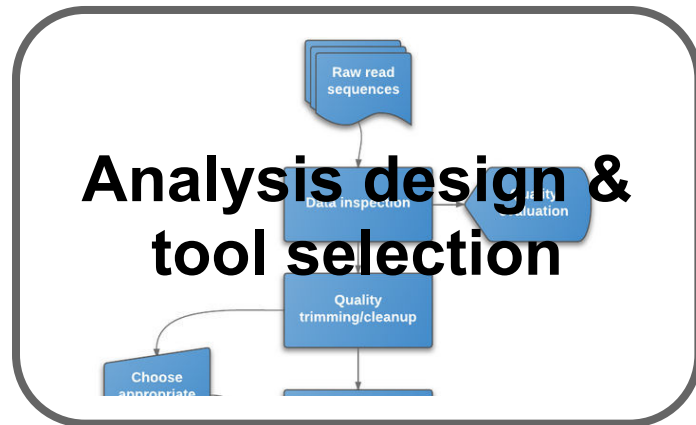
Workshops

- One day workshop per topic
- Seminar/lectures in Morning
- Afternoon
 - Hands on with Galaxy
 - Tutorial walkthrough
 - Q&A on analysis with Galaxy
- Feedback has been extremely positive
- Community of Galaxy users is growing



Tutorial Development

Tutorial development

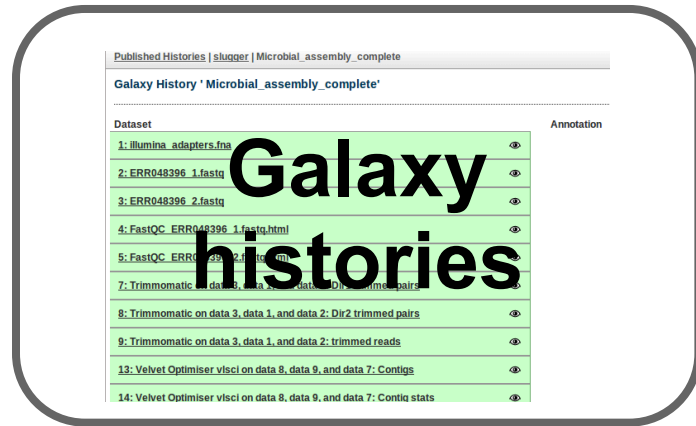


- Domain experts:
 - Design of analysis
 - Tool selection
 - Parameter selection
 - Sample data selection



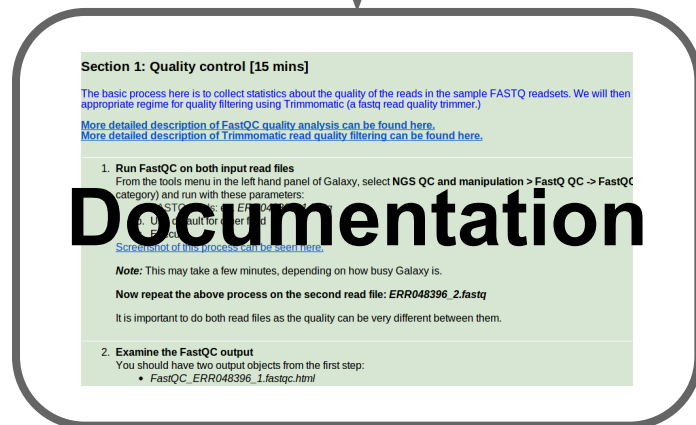
- Bioinformatician/Galaxy dev:
 - Wrapper design
 - Implementation
 - Installation on GVL image

Tutorial development



Galaxy histories

- Domain experts:
 - Run analysis on data
 - Snapshot histories at each step
 - Tweak tool interfaces



Documentation

- Team:
 - Write tutorial document
 - Collate and link background
 - Review and publish on web

Development issues



- Tool wrapping
 - Level of parameter exposure in wrapped tools
 - Test suite versus tool wrapper capabilities
- Tool dependencies and binary installations: "Shrink Wrapping"
 - Dependency handling is confusing
 - Would be good to see some more complex examples
 - Not suitable for all applications - e.g. Qiime
 - Low level dependencies - e.g. Kernel headers, BioPerl etc
 - Where to draw the line between machine/toolshed
 - Galaxy-dev mailing list is really helpful!

Updates/future work



- Keeping the tutorials up to date
 - Adding new tools as they become standard
 - Updating documentation to match
- Future/continuing work
 - New toolshed mechanisms and the toolshed police
 - Adapt to changing Galaxy landscape
 - Automated testing for all tools
 - "Shrink wrap" tool dependencies into tool repositories

Conclusions

Has anyone learned anything?

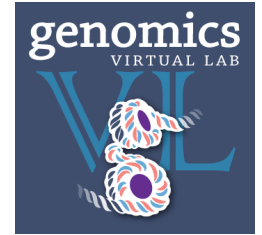


- We've learned how to set up Galaxy and wrap tools
- We are working on "Shrink Wrapping"

Most importantly:

- People who have used the tutorials / protocols / workshops have learned some useful Bioinformatics analysis techniques..
- And their level of fear regarding these analyses has been reduced!
- Lots of positive feedback.

The team



- VLSCI
 - **Andrew Lonie**
 - **Clare Sloggett***
 - Enis Afgan*
 - Nuwan Goonasekera
 - **Simon Gladman***
 - **Mahtab Mirmomeni**
 - **Franco Caramia**
- University of Queensland
 - Ron Horst*
 - **Igor Makunin**
- Garvan Institute
 - Warren Kaplan
 - **Kevin Ying***
 - Derrick Lin
- CSIRO
 - Phillippe Moncuquet*
- QFAB
 - **Mark Crowe**
 - **Pierre-Alain Chaumeil***
 - Xinyi Chua
 - Anne Kunert

Thank you.