NGS Data Analysis and Galaxy

University of Pretoria
Pretoria, South Africa
14-18 October 2013

Dave Clements, Emory University
http://galaxyproject.org/

Fourie Joubert, Burger van Jaarsveld
Bioinformatics & Computational Biology Unit
University of Pretoria
http://science.up.ac.za/

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

AFRICAN CENTRE FOR GENE TECHNOLOGIES

Galaxy

amazon web services™

South Africa Galaxy Workshop Tour

This Week

Monday	Welcome, Project Intro, Basic Galaxy Usage NGS QualityControl
Tuesday	RNA-Seq - Mapping and Transcript Prediction RNA-Seq: Differential expression and Alternative Pipelines; SNP & Variant Analysis
Wednesday	SNP & Variant Analysis Chip-Seq Analysis
Thursday	Genome Assembly Install your own Galaxy on Amazon Cloud
Friday	Customizing Galaxy, Galaxy Tool Shed, and Wrapping Tools for Galaxy

Thursday Agenda

8:30 Welcome and Questions

8:45 *de novo* Genome Assembly, Part I

10:30 Break

11:00 *de novo* Genome Assembly, Part II

12:30 Lunch

13:30 Galaxy CloudMan on Amazon, Part I

15:00 Break

15:30 Galaxy CloudMan on Amazon, Part I

16:30 Done, Feedback

Beginner's guide to comparative bacterial genome analysis using next-generation sequence data

By David J Edwards and Kathryn E Holt

Microbial Informatics and Experimentation 2013, **3**:2

and the accompanying

Bacterial Comparative Genomics Tutorial

Create a new history

Shared Data → Data Libraries → **Assembly**

Select both FASTQ files

Illumina HiSeq paired-end reads
from *E. coli* O104:H4 strain TY-2482
(ENA accession SRR292770)

http://www.ebi.ac.uk/ena/data/view/SRR292770&display=html

http://www.ncbi.nlm.nih.gov/sra/SRX079805

NGS Assembly: Quality Control

FastQC Reports for both input datasets are in

Shared Data → Assembly

Note the very different results from RNA-Seq

Only issue appears to be duplication

(How is it possible to *have* > 25% sequence duplication and then *not have any* overrepresented sequences?)

NGS Assembly: Quality Control

The duplication will affect the assembly.

The tutorial says you can use the FASTX Toolkit for this.

NGS: QC and Manipulation → Collapse

Hmm, but

that will destroy our mate pairs

and

a mate pair when one end is not a duplicate is not a duplicate

NGS Assembly: Quality Control

NGS: QC and Manipulation → FASTQ Joiner

NGS: QC and Manipulation → Collapse

NGS: QC and Manipulation → FASTQ Splitter

But don't do this now. It is slow.

Just get the results from the Assembly Data Library

Shared Data → ...

But don't do that either.

Collapse does not find any duplicates.

(Why? And why didn't we do this with the RNA-Seq data?)

NGS Assembly: Velveth

NGS: Assembly → Velveth

Hash length?

Gives us choices from 11 to 29. But the tutorial says use 35 (because they have determined that to be optimal).

The maximum kmer-length Velvet can use is set at install/ compile time

Use 29. We will revisit this.

NGS Assembly: Velveth

Click on Add new Input Files

File format → FASTQ

Read type → shortPaired reads

Dataset → 1 (forward reads)

Repeat for Dataset 2 (reverse reads)

Produces an index of the reads using the kmer length.

Index is used by Velvetg to do actual mapping.

NGS Assembly: Velvetg

Velvetg does the actual assembly

Velvet Dataset → *Output dataset from velveth*

Check Generate unusedReads fasta file

The tutorial provides us with several "optimal" values to use.

Let's use them and then revisit them.

Coverage cutoff → Specify cutoff value → 2.81

Expected coverage of unique regions → Specify expected value → 21.0

Set minimum contig length → Yes → 200

Using paired end reads → Yes

NGS Assembly: Velvetg

Several output files

Unmapped Reads

Stats

Statistics about the graph nodes constructed during assembly.

Information about the internals of Velvetg.

Contigs

The list of contigs produced by this assembly run.

Let's take a look at the contigs

NGS Assembly: Velvetg

Contigs

FASTA Manipulation → Compute Sequence Lengths

Give it the contigs file

Statistics → Summary Statistics

Contgs file, column 2

NGS Assembly: Parameters

Remember these?

Hash size → 29

Coverage cutoff → Specify cutoff value → 2.81

Expected coverage of unique regions → Specify expected value → 21.0

Not very often will someone tell you the optimal values.

NGS Assembly: Parameters

Velvet Optimiser

Explore a range of parameter values and combinations

kmer range → 11-29

step size → 2

Click Add new input read library

File Type → shortPaired

Check Are the reads paired ...

Select read files

and ...

NGS Assembly: Velvet Optimiser

... and

Click Execute *and then wait several hours*, or

Get the results from the data library

Generate some basic statistics on this.
Assembly appears much better.

NGS Assembly: Resources and Reading

Beginner's guide to comparative bacterial genome analysis using next-generation sequence data

Bacterial Comparative Genomics Tutorial

By David J Edwards and Kathryn E Holt

A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies

Zhang, *et al.*

Whole Genome Assembly and Alignment

Michael Schatz

27 Repositories in the Assembly Category of the Galaxy Tool Shed

Thursday Agenda

8:30 Welcome and Questions

8:45 *de novo* Genome Assembly, Part I

10:30 Break

11:00 *de novo* Genome Assembly, Part II

12:30 Lunch

13:30 Galaxy CloudMan on Amazon, Part I

15:00 Break

15:30 Galaxy CloudMan on Amazon, Part I

16:30 Done, Feedback

Thanks

Dave Clements

Galaxy Project
Emory University

clements@galaxyproject.org