Galaxy Project Update

ISMB / ECCB 2015
Dublin, Ireland
Dave Clements
and the Galaxy Team
Johns Hopkins University
Multi-Sample Analysis

Need to scale to support current and coming experimental designs:

500 specimens
or
20 different conditions with 4 replicates each
or
...

Multi-Sample Analysis: Workflows
Multi-Sample:
Multiple dataset selection within a tool

FastQC Read Quality reports (Galaxy Tool Version 0.63)

Short read data from your current history

File: MeOH_REP1_R1.fastq

Multiple datasets

File: No selection

Tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Submodule and Limit specifying file

File: No selection

A file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

Execute
NGS Data Quality Assessment: 11 more to go!

Short read data from your current history

12: R3G_REP3_R2.fastq
11: R3G_REP3_R1.fastq
10: R3G_REP2_R2.fastq
9: R3G_REP2_R1.fastq
8: R3G_REP1_R2.fastq

*This is a batch mode input field. A separate job will be triggered for each dataset.*

Contaminant list

No selection
tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGCGCATACGA

Submodule and Limit specifying file

No selection

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter
## Multi-Sample Analysis: Collections

### History

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cam RNA-Seq Day 2 Test</td>
<td><img src="image" alt="Actions" /></td>
</tr>
<tr>
<td>189.3 MB</td>
<td><img src="image" alt="Actions" /></td>
</tr>
<tr>
<td>8: genes_chr12.g 6: R3G_REP3 Mapped &amp; Filtered</td>
<td><img src="image" alt="Actions" /></td>
</tr>
<tr>
<td>7: chr12.fa 5: R3G_REP2 Mapped &amp; Filtered</td>
<td><img src="image" alt="Actions" /></td>
</tr>
<tr>
<td>6: R3G_REP1 Mapped &amp; Filtered</td>
<td><img src="image" alt="Actions" /></td>
</tr>
<tr>
<td>4: MeOH_REP3 Mapped &amp; Filtered</td>
<td><img src="image" alt="Actions" /></td>
</tr>
<tr>
<td>3: MeOH_REP2 Mapped &amp; Filtered</td>
<td><img src="image" alt="Actions" /></td>
</tr>
<tr>
<td>2: MeOH_REP1 Mapped &amp; Filtered</td>
<td><img src="image" alt="Actions" /></td>
</tr>
<tr>
<td>1: MeOH_REP1 Mapped &amp; Filtered</td>
<td><img src="image" alt="Actions" /></td>
</tr>
</tbody>
</table>
Multi-Sample Analysis: Collections

Collections:
1: MeOH REP1 Mapped & Filtered
2: MeOH REP2 Mapped & Filtered
3: MeOH REP3 Mapped & Filtered
4: R3G REP1 Mapped & Filtered
5: R3G REP2 Mapped & Filtered
6: R3G REP3 Mapped & Filtered
7: chr12.fa
8: genes chr12.gtf

Operations on multiple datasets:
- All
- None
- For all selected...
Multi-Sample Analysis: Collections

History

Cam RNA-Seq Day 2 Test
8 shown
189.3 MB

8: genes_chr12.gtf
7: chr12.fa
6: R3G REP3 Mapped & Filtered
5: R3G REP2 Mapped & Filtered
4: R3G REP1 Mapped & Filtered
3: MeOH REP3 Mapped & Filtered
2: MeOH REP2 Mapped & Filtered
1: MeOH REP1 Mapped & Filtered

Operations on multiple datasets

Search datasets

189.3 MB

All | None
For all selected

1: MeOH REP1 Mapped & Filtered
2: MeOH REP2 Mapped & Filtered
3: MeOH REP3 Mapped & Filtered
4: R3G REP1 Mapped & Filtered
5: R3G REP2 Mapped & Filtered
6: R3G REP3 Mapped & Filtered
7: chr12.fa
8: genes_chr12.gtf

Hide datasets
Unhide datasets
Delete datasets
Undelete datasets
Permanently delete datasets
Build Dataset List
Build Dataset Pair
Build List of Dataset Pairs
Multi-Sample Analysis: Collections

Operations on multi-datasets:
- 8: genes_chr12.gtf
- 7: chr12.fa
- 6: R3G_REP3 Mapped & Filtered
- 5: R3G_REP2 Mapped & Filtered
- 4: R3G_REP1 Mapped & Filtered
- 3: MeOH_REP3 Mapped & Filtered
- 2: MeOH_REP2 Mapped & Filtered
- 1: MeOH_REP1 Mapped & Filtered

Hide datasets
Unhide datasets
Delete datasets
Undelete datasets
Permanently delete dataset
Build Dataset List
Build Dataset Pair
Build List of Dataset Pair
Multi-Sample Analysis: Collections

StringTie transcript assembly and quantification (Galaxy Tool, Version 1.0.3)

Mapped reads to assemble transcripts from
- 3: MeOH_REP3 Mapped & Filtered

Use dataset collection

Reference annotation to use for guiding the assembly
- G

Perform abundance estimation only of input transcripts
- e

Output additional files for use in Ballgown
- b

Options
- Use defaults

Job Resource Parameters
- Use default job resource parameters

StringTie transcript assembly and quantification (Galaxy Tool, Version 1.0.3)

Mapped reads to assemble transcripts from
- 10: R3G

Reference annotation to use for guiding the assembly process
- G

Perform abundance estimation only of input transcripts
- e

Output additional files for use in Ballgown
- b

Options
- Use defaults
User Interface: Scratchbook

ESHG 1: Through BWA Mapping & Merge w/ RG
13 shown, 2 deleted
3.4 GB

15: M512 Fully Mapped, pre manip

12: Map with BWA-ME M on data 8 and data 7 (mapped reads in BAM format)
### User Interface: Scratchbook

**ESHG 1: Through BWA Mapping & Merge w/ RG**

![Scratchbook user interface](https://test.galaxyproject.org)

#### Significant

<table>
<thead>
<tr>
<th>test_id</th>
<th>gene_id</th>
<th>gene</th>
<th>locus</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>XLOC_00005</td>
<td>XLOC_00005</td>
<td>WNK1</td>
<td>chr12:862088-1020664</td>
<td>sam</td>
</tr>
<tr>
<td>XLOC_00013</td>
<td>XLOC_00013</td>
<td>FKBP4</td>
<td>chr12:2904107-2914587</td>
<td>MeO</td>
</tr>
<tr>
<td>XLOC_00021</td>
<td>XLOC_00021</td>
<td>CCND2</td>
<td>chr12:4382901-4414522</td>
<td>MeO</td>
</tr>
<tr>
<td>XLOC_00023</td>
<td>XLOC_00023</td>
<td>RAD51AP1</td>
<td>chr12:4647949-4669213</td>
<td>MeO</td>
</tr>
<tr>
<td>XLOC_00035</td>
<td>XLOC_00035</td>
<td>TAPBPL</td>
<td>chr12:6561176-6579843</td>
<td>MeO</td>
</tr>
<tr>
<td>XLOC_00037</td>
<td>XLOC_00037</td>
<td>SCARNA10</td>
<td>chr12:6603297-6641132</td>
<td>MeO</td>
</tr>
<tr>
<td>XLOC_00038</td>
<td>XLOC_00038</td>
<td>GAPDH</td>
<td>chr12:6643584-6647537</td>
<td>MeO</td>
</tr>
<tr>
<td>XLOC_00039</td>
<td>XLOC_00039</td>
<td>COP57A</td>
<td>chr12:6833149-6841041</td>
<td>MeO</td>
</tr>
<tr>
<td>XLOC_00042</td>
<td>XLOC_00042</td>
<td>CD4</td>
<td>chr12:6898637-6929976</td>
<td>MeO</td>
</tr>
</tbody>
</table>

#### Enough data to make a call

<table>
<thead>
<tr>
<th>test_id</th>
<th>gene_id</th>
<th>gene</th>
<th>locus</th>
</tr>
</thead>
<tbody>
<tr>
<td>XLOC_00001</td>
<td>XLOC_00001</td>
<td>LOC100288778</td>
<td>chr12:87983-91263</td>
</tr>
<tr>
<td>XLOC_00003</td>
<td>XLOC_00003</td>
<td>CCDC77</td>
<td>chr12:389222-551806</td>
</tr>
<tr>
<td>XLOC_00005</td>
<td>XLOC_00005</td>
<td>WNK1</td>
<td>chr12:862088-1020664</td>
</tr>
<tr>
<td>XLOC_00009</td>
<td>XLOC_00009</td>
<td>ADIPOR2</td>
<td>chr12:1800246-1897845</td>
</tr>
<tr>
<td>XLOC_00013</td>
<td>XLOC_00013</td>
<td>FKBP4</td>
<td>chr12:2904107-2914587</td>
</tr>
<tr>
<td>XLOC_00014</td>
<td>XLOC_00014</td>
<td>ITFG2</td>
<td>chr12:2921786-2934237</td>
</tr>
<tr>
<td>XLOC_00016</td>
<td>XLOC_00016</td>
<td>RHNO1</td>
<td>chr12:2945981-2998691</td>
</tr>
<tr>
<td>XLOC_00021</td>
<td>XLOC_00021</td>
<td>CCND2</td>
<td>chr12:4382901-4414522</td>
</tr>
<tr>
<td>XLOC_00023</td>
<td>XLOC_00023</td>
<td>RAD51AP1</td>
<td>chr12:4647949-4669213</td>
</tr>
<tr>
<td>XLOC_00025</td>
<td>XLOC_00025</td>
<td>NDUFA9</td>
<td>chr12:4758263-4796720</td>
</tr>
</tbody>
</table>
User Interface: Histories

Cam RNA-Seq Day 2
15 shown, 15 hidden
194.2 MB

27: R3G REP3 Mapped & Filtered

26: StringTie on collection 10: Coverage
a list of datasets

25: StringTie on collection 10: Assembled transcripts
a list of datasets

18: StringTie on collection 9: Coverage
a list of datasets

17: StringTie on collection 9: Assembled transcripts
a list of datasets

10: R3G
a list of datasets
User Interface: Histories

Cam RNA-Seq Day 2
15 shown, 15 hidden
194.2 MB

Cam ChIP-Seq 1
39 shown, 6 hidden
476.9 MB

Copy of 'Cam RNA-Seq Day 2' (active items only)
34 shown, 2 deleted, 15 hidden
203.0 MB
User Interface: Data Load

Download data directly from web or upload files from your disk

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Type</th>
<th>Genome</th>
<th>Settings</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>MeOH_REP1_R1.fastq</td>
<td>28.2 MB</td>
<td>Auto-detect</td>
<td>Additional Species...</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>MeOH_REP1_R2.fastq</td>
<td>28.2 MB</td>
<td>Auto-detect</td>
<td>Additional Species...</td>
<td></td>
<td>79%</td>
</tr>
<tr>
<td>MeOH_REP2_R1.fastq</td>
<td>28.9 MB</td>
<td>Auto-detect</td>
<td>Additional Species...</td>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>MeOH_REP2_R2.fastq</td>
<td>28.9 MB</td>
<td>Auto-detect</td>
<td>Additional Species...</td>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>MeOH_REP3_R1.fastq</td>
<td>28.9 MB</td>
<td>Auto-detect</td>
<td>Additional Species...</td>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>MeOH_REP3_R2.fastq</td>
<td>28.9 MB</td>
<td>Auto-detect</td>
<td>Additional Species...</td>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>R3G_REP1_R1.fastq</td>
<td>23.5 MB</td>
<td>Auto-detect</td>
<td>Additional Species...</td>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>R3G_REP1_R2.fastq</td>
<td>23.5 MB</td>
<td>Auto-detect</td>
<td>Additional Species...</td>
<td></td>
<td>0%</td>
</tr>
</tbody>
</table>

Type (set all): Auto-detect

Genome (set all): Additional Species...
Tools: Multiple version support

Cuffdiff find significant changes in transcript expression, splicing, and promoter use (GFF3/GTF)

Transcripts
27: Cuffmerge on data 7, data 8, and others: merged ...

A transcript GFF3 or GTF file produced by cufflinks, cuffcompare, or other source.

Omit Tabular Datasets
Yes  No
Discard the tabular output.

Generate SQLite
Yes  No
Generate a SQLite database for use with cummerBund.

Input data type
SAM/BAM
CuffNorm supports either CXB (from cuffquant) or SAM/BAM input files. Mixing is not supported. Default: SAM/BAM

Condition
1: Condition
Name
MeOH
Replicates
CuffNorm supports either CXB (from cuffquant) or SAM/BAM input files. Mixing is not supported. Default: SAM/BAM

Condition

1: Condition

Name
MeOH

Replicates
5: R3G_REP2 Mapped & Filtered
4: R3G_REP1 Mapped & Filtered
3: MeOH_REP3 Mapped & Filtered
2: MeOH_REP2 Mapped & Filtered
1: MeOH_REP1 Mapped & Filtered

2: Condition

Name
R3G

Replicates
6: R3G_REP3 Mapped & Filtered
5: R3G_REP2 Mapped & Filtered
4: R3G_REP1 Mapped & Filtered
3: MeOH_REP3 Mapped & Filtered
2: MeOH_REP2 Mapped & Filtered

+ Insert Condition

Library normalization method
Using Docker for Tool Integration

Tools can now be run in Docker containers.

Tool authors may annotate Docker container ids the tool works with in the Tool XML file.

Deployers may specify Docker container ids on a per destination basis.

https://wiki.galaxyproject.org/Admin/Tools/Docker
Interactive Environments

Interactive environments such as IPython and RStudio can now be invoked from Galaxy.

Work contributed by

Björn Grüning
University of Freiburg

Eric Rashe
Texas A&M University
Interactive Environments
Interactive Environments

Install Misopy

In [1]: !pip install --user --quiet misopy

Restart the kernel (Kernel -> Restart), so that IPython finds Misopy

Get Samtools

In [1]: from urllib import urlretrieve
urlretrieve("http://depot.galaxyproject.org/package/linux/x86_64/samtools!	ar -xf samtools.tgz

The samtools binary should now be in bin/samtools

Generate Indices of the bam Files

- Now the bam files from the Galaxy history can be imported with get(history_id)
- Then index files of the bam files can be generated via samtools

In [2]: get(5)
get(6)
!mv 5 461177.bam
!mv 6 461178.bam
!bin/samtools index 461177.bam
!bin/samtools index 461178.bam
The samtools binary should now be in bin/samtools

Generate indices of the bam Files

- Now the bam files from the Galaxy history can be imported with get(history_id)
- Then index files of the bam files can be generated via samtools

In [2]:
get(1)
get(2)
!mv 1 461177.bam
!mv 2 461178.bam

!bin/samtools index 461177.bam
!bin/samtools index 461178.bam

If everything went well there should now be index files (*.bai) next to the bam files.

In [3]:
!ls *.ba*

461177.bam  461177.bam.bai  461178.bam  461178.bam.bai

Generate Index of an Annotation File

An annotation file is needed.

In [4]:
get(10)
!mv 10 dm3.ensGene.gff3
Interactive Environments

Finally save the pdf back to the Galaxy history

```
In [9]: put("output/FBgn0039909.pdf")
```
Interactive Environments

Finally save the pdf back to the Galaxy history

In [9]: put("output/FBgn0039909.pdf")
Interactive Environments

In [9]: put("output/FBgn0039909.pdf")
Join us in beautiful Bloomington, Indiana

June 28-29, 2016
What's coming: Right here! Right now!

TT16: RiboGalaxy: a platform for the alignment, analysis and visualization of ribo-seq data

Audrey Michel
Thanks

Dave Clements
Galaxy Project
Johns Hopkins University
clements@galaxyproject.org
Scalability ...

Data generation is cheap and will stay cheap. Larger / more complex analysis are being done. More and more people are running bioinformatics analysis of all complexities. Scalability haunts us. Data generation never sleeps.