RNA-Seq in Galaxy: Tuxedo protocol

Igor Makunin, UQ RCC, QCIF
Acknowledgments

Genomics Virtual Lab: gvl.org.au
Galaxy for tutorials: galaxy-tut.genome.edu.au
Galaxy Australia: galaxy-aust.genome.edu.au

Contributors and participants:
Plan for today

Galaxy

Data types used in RNA-Seq analysis

RNA-Seq practical

Galaxy workflow
High-throughput sequencing

Big scale sequencing
• 100,000,000s sequences, or reads, per experiment
• sequencing of a (random) library
• low cost per nucleotide

Popular technologies:
• illumina
• ion / proton
• PacBio

Emerging technologies
• Oxford Nanopore MinION

Analysis of NGS data
Big datasets
Computationally intensive
Dedicated tools and data types
Extensive use of public data

Computational resources
Tools
Storage
Public data

Knowledge and skills
Galaxy: how does it look like

Upload
Working window
History menu
Galaxy history system

- **History menu**
  - **Refresh**
  - **History menu**

- **Number**
- **Name**
  - **1: 1.bed**
  - **View**
  - **Edit**
  - **Delete**

- **Description**
- **File Format**
  - **format: bed, database: ?**
  - **uploaded bed file**
  - **Database/Genome**
  - **Tags & Annotation**

- **Actions**
- **External Display Apps**
- **Data peek**

**History**

- **State Views**
  - **5 shown**
  - **20.5 KB**
  - **5: paused**
  - **4: error**
  - **3: ok**
  - **2: running**
  - **1: queued**

Galaxy servers:
usegalaxy.org
usegalaxy.eu
galaxy-tut.genome.edu.au
galaxy-aust.genome.edu.au

Advantage of the registration:
• *access to histories over long time*
• *multiple histories*
• *ability to use Galaxy from different devices*
• *bigger quotas (on some servers)*
• *ftp*

• **Independent registration on every Galaxy server**
• **Different tools, different user policy**
• **Data can be moved between Galaxy servers**
Galaxy Australia

galaxy-aust.genome.edu.au

Worker nodes:
16 CPUs, 64 GB RAM

49 Tb Volume storage (user data)

Designed for a genome scale research
>1,600 registered users

Up to 16 CPUs 60 GB RAM per job
Up to 12 concurrent jobs per user
Up to 1 Tb per user

Galaxy-qld: Australian users

Jobs per day

Less jobs on weekends
Tuxedo protocol

GVLS Basic RNA-Seq Galaxy tutorial

Condition A
- Reads (FASTQ)
- Mapped reads (BAM)
- Tables of read counts

Condition B
- Reads (FASTQ)
- Mapped reads (BAM)
- Tables of read counts

Reference genome
Map to reference genome & identify splice sites

Annotated transcriptome
Summarise expression counts across genes, exons, CDS, TSS

Normalisation within and between sample groups

Differential expression statistical testing

List of differentially expressed genes/isoforms/TSS

Visualise alignment with IGV

FASTA

FASTQ

BAM

GFF

Tophat

Cuffdiff

Genome browser
Terminology: *read* is a sequence with quality score values produced by a sequencing machine.

Common output format: *FASTQ* compressed with gzip, e.g. SRR3145_1.fq.gz

Single-end reads vs. paired-end reads:

- Multiple reads in a single FASTQ file
- Each read is described by four lines

Example:

```plaintext
@SRR3145.19 ILLUMINA-C32_FC:3:1:80:12/1
TAGCAGCACATCATGGTTTACATCGTATGC+
IIHIDIIIIIIIIIIIIIIIIIIIIIDGIB
```

Name always starts with @
Sequence
Always starts with +; may have name
Encoded Phred quality score
FASTQ Phred quality score

A Phred quality score is a measure of the quality of the identification for every nucleotide.

\[ Q_{\text{sanger}} = -10 \log_{10} p \]

Range: ~0 to ~40

Phred 10: accuracy 90%
Phred 20: accuracy 99%
Phred 30: accuracy 99.9%
Phred 40: accuracy 99.99%

Values are encoded by characters

Advantage: a single character is used instead of a two-digit number

Quality + Offset
39 + 33 = 72
ASCII(72): H

@S391 ILLUMINA_FC:3:80:12/1
TAGCAGCACATCATGGTTTAC
+
IIHIDIIIIIIIIIIIIIIIIIHIIH
<table>
<thead>
<tr>
<th>Decimal</th>
<th>Hex</th>
<th>Char</th>
<th>Decimal</th>
<th>Hex</th>
<th>Char</th>
<th>Decimal</th>
<th>Hex</th>
<th>Char</th>
<th>Decimal</th>
<th>Hex</th>
<th>Char</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>[NULL]</td>
<td>32</td>
<td>20</td>
<td>[SPACE]</td>
<td>64</td>
<td>40</td>
<td>@</td>
<td>96</td>
<td>60</td>
<td>`</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>[START OF HEADING]</td>
<td>33</td>
<td>21</td>
<td>!</td>
<td>65</td>
<td>41</td>
<td>A</td>
<td>97</td>
<td>61</td>
<td>a</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>[START OF TEXT]</td>
<td>34</td>
<td>22</td>
<td>&quot;</td>
<td>66</td>
<td>42</td>
<td>B</td>
<td>98</td>
<td>62</td>
<td>b</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>[END OF TEXT]</td>
<td>35</td>
<td>23</td>
<td>#</td>
<td>67</td>
<td>43</td>
<td>C</td>
<td>99</td>
<td>63</td>
<td>c</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>[END OF TRANSMISSION]</td>
<td>36</td>
<td>24</td>
<td>$</td>
<td>68</td>
<td>44</td>
<td>D</td>
<td>100</td>
<td>64</td>
<td>d</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>[ENQUIRY]</td>
<td>37</td>
<td>25</td>
<td>%</td>
<td>69</td>
<td>45</td>
<td>E</td>
<td>101</td>
<td>65</td>
<td>e</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>[ACKNOWLEDGE]</td>
<td>38</td>
<td>26</td>
<td>&amp;</td>
<td>70</td>
<td>46</td>
<td>F</td>
<td>102</td>
<td>66</td>
<td>f</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>[BEL]</td>
<td>39</td>
<td>27</td>
<td>'</td>
<td>71</td>
<td>47</td>
<td>G</td>
<td>103</td>
<td>67</td>
<td>g</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>[BACKSPACE]</td>
<td>40</td>
<td>28</td>
<td>(</td>
<td>72</td>
<td>48</td>
<td>H</td>
<td>104</td>
<td>68</td>
<td>h</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>[HORIZONTAL TAB]</td>
<td>41</td>
<td>29</td>
<td>)</td>
<td>73</td>
<td>49</td>
<td>I</td>
<td>105</td>
<td>69</td>
<td>i</td>
</tr>
<tr>
<td>11</td>
<td>B</td>
<td>[VERTICAL TAB]</td>
<td>43</td>
<td>2B</td>
<td>+</td>
<td>75</td>
<td>4B</td>
<td>K</td>
<td>107</td>
<td>6B</td>
<td>k</td>
</tr>
<tr>
<td>12</td>
<td>C</td>
<td>[FORM FEED]</td>
<td>44</td>
<td>2C</td>
<td>,</td>
<td>76</td>
<td>4C</td>
<td>L</td>
<td>108</td>
<td>6C</td>
<td>l</td>
</tr>
<tr>
<td>13</td>
<td>D</td>
<td>[CARRIAGE RETURN]</td>
<td>45</td>
<td>2D</td>
<td>-</td>
<td>77</td>
<td>4D</td>
<td>M</td>
<td>109</td>
<td>6D</td>
<td>m</td>
</tr>
<tr>
<td>14</td>
<td>E</td>
<td>[SHIFT OUT]</td>
<td>46</td>
<td>2E</td>
<td>.</td>
<td>78</td>
<td>4E</td>
<td>N</td>
<td>110</td>
<td>6E</td>
<td>n</td>
</tr>
<tr>
<td>15</td>
<td>F</td>
<td>[SHIFT IN]</td>
<td>47</td>
<td>2F</td>
<td>/</td>
<td>79</td>
<td>4F</td>
<td>O</td>
<td>111</td>
<td>6F</td>
<td>o</td>
</tr>
<tr>
<td>16</td>
<td>10</td>
<td>[DATA LINK ESCAPE]</td>
<td>48</td>
<td>30</td>
<td>0</td>
<td>80</td>
<td>50</td>
<td>P</td>
<td>112</td>
<td>70</td>
<td>p</td>
</tr>
<tr>
<td>17</td>
<td>11</td>
<td>[DEVICE CONTROL 1]</td>
<td>49</td>
<td>31</td>
<td>1</td>
<td>81</td>
<td>51</td>
<td>Q</td>
<td>113</td>
<td>71</td>
<td>q</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>[DEVICE CONTROL 2]</td>
<td>50</td>
<td>32</td>
<td>2</td>
<td>82</td>
<td>52</td>
<td>R</td>
<td>114</td>
<td>72</td>
<td>r</td>
</tr>
<tr>
<td>19</td>
<td>13</td>
<td>[DEVICE CONTROL 3]</td>
<td>51</td>
<td>33</td>
<td>3</td>
<td>83</td>
<td>53</td>
<td>S</td>
<td>115</td>
<td>73</td>
<td>s</td>
</tr>
<tr>
<td>20</td>
<td>14</td>
<td>[DEVICE CONTROL 4]</td>
<td>52</td>
<td>34</td>
<td>4</td>
<td>84</td>
<td>54</td>
<td>T</td>
<td>116</td>
<td>74</td>
<td>t</td>
</tr>
<tr>
<td>21</td>
<td>15</td>
<td>[NEGATIVE ACKNOWLEDGE]</td>
<td>53</td>
<td>35</td>
<td>5</td>
<td>85</td>
<td>55</td>
<td>U</td>
<td>117</td>
<td>75</td>
<td>u</td>
</tr>
<tr>
<td>22</td>
<td>16</td>
<td>[SYNCHRONOUS IDLE]</td>
<td>54</td>
<td>36</td>
<td>6</td>
<td>86</td>
<td>56</td>
<td>V</td>
<td>118</td>
<td>76</td>
<td>v</td>
</tr>
<tr>
<td>23</td>
<td>17</td>
<td>[ENG OF TRANS. BLOCK]</td>
<td>55</td>
<td>37</td>
<td>7</td>
<td>87</td>
<td>57</td>
<td>W</td>
<td>119</td>
<td>77</td>
<td>w</td>
</tr>
<tr>
<td>24</td>
<td>18</td>
<td>[CANCEL]</td>
<td>56</td>
<td>38</td>
<td>8</td>
<td>88</td>
<td>58</td>
<td>X</td>
<td>120</td>
<td>78</td>
<td>x</td>
</tr>
<tr>
<td>25</td>
<td>19</td>
<td>[END OF MEDIUM]</td>
<td>57</td>
<td>39</td>
<td>9</td>
<td>89</td>
<td>59</td>
<td>Y</td>
<td>121</td>
<td>79</td>
<td>y</td>
</tr>
<tr>
<td>26</td>
<td>1A</td>
<td>[SUBSTITUTE]</td>
<td>58</td>
<td>3A</td>
<td>:</td>
<td>90</td>
<td>5A</td>
<td>Z</td>
<td>122</td>
<td>7A</td>
<td>z</td>
</tr>
<tr>
<td>27</td>
<td>1B</td>
<td>[ESCAPE]</td>
<td>59</td>
<td>3B</td>
<td>;</td>
<td>91</td>
<td>5B</td>
<td>[</td>
<td>123</td>
<td>7B</td>
<td>{</td>
</tr>
<tr>
<td>28</td>
<td>1C</td>
<td>[FILE SEPARATOR]</td>
<td>60</td>
<td>3C</td>
<td>&lt;</td>
<td>92</td>
<td>5C</td>
<td>\</td>
<td>124</td>
<td>7C</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>1D</td>
<td>[GROUP SEPARATOR]</td>
<td>61</td>
<td>3D</td>
<td>=</td>
<td>93</td>
<td>5D</td>
<td>]</td>
<td>125</td>
<td>7D</td>
<td>}</td>
</tr>
<tr>
<td>30</td>
<td>1E</td>
<td>[RECORD SEPARATOR]</td>
<td>62</td>
<td>3E</td>
<td>&gt;</td>
<td>94</td>
<td>5E</td>
<td>^</td>
<td>126</td>
<td>7E</td>
<td>~</td>
</tr>
<tr>
<td>31</td>
<td>1F</td>
<td>[UNIT SEPARATOR]</td>
<td>63</td>
<td>3F</td>
<td>?</td>
<td>95</td>
<td>5F</td>
<td>_</td>
<td>127</td>
<td>7F</td>
<td>[DEL]</td>
</tr>
</tbody>
</table>
Phred quality score encoding

Offset 33 - Sanger
Offset 64 - old illumina

Qual. = 40
Offset = 33
40+33 = 73
ASCII(73): I

Source: https://en.wikipedia.org/wiki/FASTQ_format
FASTQ quality score in Galaxy

Many old illumina datasets have a proprietary data encoding (offset 64)
Currently most NGS datasets use the Sanger encoding (offset 33)

**Galaxy**
By default Galaxy assign ‘fastq’ data type to uploaded FASTQ files. In this case the offset is not specified, and many tools do not recognize the data

*fastqillumina* – old illumina quality score encoding (offset 64, illumina 1.3+)
*fastqsanger* – new illumina 1.8+ / Sanger quality score encoding
Some tools in Galaxy now work only with *fastqsanger* datatype

**Solution:**
- specify *fastqsanger* or *fastqillumina* datatype during upload
- change the format via Attributes > Datatype
- use **NGS: QC and manipulation** > **FASTQ Groomer** tool
Tuxedo protocol

GVL Basic RNA-Seq Galaxy tutorial

Condition A
- Reads (FASTQ)
- Mapped reads (BAM)
- Tables of read counts

Condition B
- Reads (FASTQ)
- Mapped reads (BAM)
- Tables of read counts

Reference genome
- Map to reference genome & identify splice sites

Annotated transcriptome
- Summarise expression counts across genes, exons, CDS, TSS
- Normalisation within and between sample groups

Differential expression statistical testing

List of differentially expressed genes/isoforms/TSS

Visualise alignment with IGV
- Tophat
- Cuffdiff

FASTA
- Genome browser

FASTQ
- BAM
Reference genomes

Genome Reference Consortium: ... a consensus representation of the genome.

FASTA format

The human reference sequence GRCh37 (hg19) contains the mitochondrial genome, 22 autosomes, chrX, chrY, 9 haplotype chromosomes, 39 unplaced contigs, and 20 unlocalized contigs.

Genomes are big. GRCh38.p10 total non-N bases: 3,080,585,178

Genomes may have many assembly versions (releases, build): mm9, mm10

Use the same assembly version for the reference sequence and gene annotations.

Order of sequences / contigs might be important for some tools.

“chr1” and “1” are not identical for some tools.

http://hgdownload.cse.ucsc.edu/gbdb/hg19/html/description.html
**Gene annotations**

**Coordinate-based**: linked to a particular genome assembly, e.g., hg19

GFF (General Feature Format) format consists of **one line per feature**, each containing 9 columns of data, plus optional track definition lines. Popular versions: GTF(=GFF2), GFF3

tab-separated fields

<table>
<thead>
<tr>
<th>seqid</th>
<th>type</th>
<th>start</th>
<th>end</th>
<th>strand</th>
<th>attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ctg123</td>
<td>mRNA</td>
<td>1300</td>
<td>9000</td>
<td>+</td>
<td>ID=mRNA0001;Name=sonichedgehog</td>
</tr>
<tr>
<td>ctg123</td>
<td>exon</td>
<td>1300</td>
<td>1500</td>
<td>+</td>
<td>ID=exon00001;Parent=mRNA0001</td>
</tr>
<tr>
<td>ctg123</td>
<td>exon</td>
<td>1050</td>
<td>1500</td>
<td>+</td>
<td>ID=exon00002;Parent=mRNA0001</td>
</tr>
<tr>
<td>ctg123</td>
<td>exon</td>
<td>3000</td>
<td>3902</td>
<td>+</td>
<td>ID=exon00003;Parent=mRNA0001</td>
</tr>
<tr>
<td>ctg123</td>
<td>exon</td>
<td>5000</td>
<td>5500</td>
<td>+</td>
<td>ID=exon00004;Parent=mRNA0001</td>
</tr>
<tr>
<td>ctg123</td>
<td>exon</td>
<td>7000</td>
<td>9000</td>
<td>+</td>
<td>ID=exon00005;Parent=mRNA0001</td>
</tr>
</tbody>
</table>

The first line must be a comment that identifies the version

```
##gff-version 3
```

http://asia.ensembl.org/info/website/upload/gff3.html
**Intervals**

**Coordinate-based**: linked to a particular genome assembly, *e.g.*, hg19

BED format, up to 12 columns of data (UCSC Table Browser), plus optional track header lines.

Tab-separated fields

**GFF3**

```plaintext
##gff-version 3
ctg123 . mRNA 1300 9000 . + . ID=mrna0001;Name=sonichedgehog
```

**BED**

```plaintext
ctg123 1299 9000 sonichedgehog . +
```

- `chrom`: chromosome
- `chromStart`: start position
- `chromEnd`: end position
- `name`: feature name
- `score`: score
- `strand`: strand (+ or -)
- `0-based` or `1-based`
Aligners map reads to a reference sequence.

Aligners use proprietary index files for mapping.

```
bwa index hg19.fa
```

Only for BWA  Only for hg19

Galaxy-qld provides indices for several genome assemblies (hg19, hg38, mm9, mm10 etc.)

Galaxy users also can use a custom reference sequence for alignment. In this situation the aligner creates indices in a temporary working directory for every job.

Contact Galaxy-qld admins if you plan to run many alignment jobs with a custom genome. We can add genome indices to the server.
Alignments: SAM and BAM

50x coverage of the human genome with read length 100 bp:
~1,500,000,000 reads
Compressed size of such alignment can be > 100 Gb.

SAM: *Sequence Alignment/Map*. Plain text format.
BAM: binary (compressed) version of the alignment format.

SAM coordinates are 1-based
BAM coordinates are 0-based

BAMs are indexed for rapid access. Useful for alignment visualization.

**It is always good to have a header!**

```
@HD   VN:1.0   SO:queryname
@RG   ID:igGroup   SM:igSmpl   LB:igL1   PL:ILLUMINA
@SQ   SN:chr2L   LN:23011544
@PG   ID:TopHat   VN:2.0.14
          CL:/mnt/galaxy/tools/tophat/2.0.14/iuc/package_tophat_2_0_14/536f7bb5616d/bin/tophat --num-threads 5 ....
```

Read groups

Can handle multiple samples in alignment
SAM format

```
Coor  12345678901234  567890123456789012345
ref   AGCATGTTAGATAA**GATAGCTGTCAGTAGGCAGTCAGCGCCAT
+r001/1  TTAGATAAAGGATA*CTG
+r002   aaaAGATAAA*GGATA
+r003   gcctaAGCTAA
+r004   ATAGCT..............TCAGC
-r003   ttagctTAGGC
-r001/2  CAGCGGCCAT
```

11 mandatory columns and optional fields with the TAG:TYPE:VALUE format
Visualization of BAMs

Galaxy servers can act as a track hub

It is possible to add multiple tracks: BAMs, gene annotations, known variants...
Genome browsers

**Integrative Genomics Viewer, IGV**
Efficient genome viewer developed by the Broad Institute. Installable on personal computers. Can add a custom genome.

**UCSC Genome Browser**
A big server in the US. Table Browser for data analysis (intersection) Support data export to Galaxy Custom sessions (can save your tracks) liftOver tool Public track hubs
RNA-Seq with the Cufflinks package

GVL Basic RNA-Seq Galaxy tutorial

Condition A
- Reads (FASTQ)
- Mapped reads (BAM)
- Tables of read counts

Map to reference genome & identify splice sites

Condition B
- Reads (FASTQ)
- Mapped reads (BAM)
- Tables of read counts

Reference genome
- Annotated transcriptome
- Summarise expression counts across genes, exons, CDS, TSS
- Normalisation within and between sample groups
- Differential expression statistical testing
- List of differentially expressed genes/isoforms/TSS

D. melanogaster
- Two conditions
- Three replicates
- Data for chr4

Tophat
Visualise alignments
Cuffdiff
Data manipulation
Setup for the workshop

Genomics Virtual Lab
Taking the IT out of Bioinformatics

Contents
- Learn Galaxy
- Learn GenomeSpace
- RNA Seq
- Variant Calling
- Assembly
- ChIP-Seq
- Metagenomics
- Amplicons
- Microbial genomics

Basic Galaxy tutorial

RNA-seq DGE Basic Tutorial

Tuxedo Protocol Tutorial

GVL website: gvl.org.au

Register on Galaxy-tut: galaxy-tut.genome.edu.au
Galaxy is a workflow engine

A Galaxy workflow is a series of tools and dataset actions that run in sequence as a batch operation.

Select tool or input dataset

Add name, comments

Input

Noodle

Tool box

Email notification
## Galaxy workflow

### Your workflows

<table>
<thead>
<tr>
<th>Name</th>
<th># of Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>transcript_assembly_with_Trinity</td>
<td>3</td>
</tr>
<tr>
<td>align reads and sort SAM on queryname</td>
<td>6</td>
</tr>
<tr>
<td>Sort SAM file by queryname</td>
<td>5</td>
</tr>
<tr>
<td>GenomeSpaceTest</td>
<td>8</td>
</tr>
<tr>
<td>Copy of 'filter-sort-cut-RNAChipInt'</td>
<td>3</td>
</tr>
<tr>
<td>four steps '<a href="mailto:fatima.naim@qut.edu.au">fatima.naim@qut.edu.au</a>'</td>
<td>4</td>
</tr>
</tbody>
</table>

### Workflows shared with you by others

<table>
<thead>
<tr>
<th>Name</th>
<th>Owner</th>
<th># of Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>16S metagenomic (RDP, genus level, Krona)</td>
<td><a href="mailto:vebaev@gmail.com">vebaev@gmail.com</a></td>
<td>19</td>
</tr>
<tr>
<td>filter-sort-cut-RNAChipInt</td>
<td><a href="mailto:fatima.naim@qut.edu.au">fatima.naim@qut.edu.au</a></td>
<td>6</td>
</tr>
</tbody>
</table>
Create a Galaxy workflow

Your workflows

Create new workflow
Upload or import workflow

From scratch

From history

HISTORY LISTS
Saved Histories
Histories Shared with Me
HISTORY ACTIONS
Create New
Copy History
Share or Publish
Show Structure
Extract Workflow
Exercise

We will create a Galaxy workflow for RNA-Seq analysis without replicates: 
*tophat2 > Cuffdiff > Filter*
Acknowledgments

Genomics Virtual Lab: gvl.org.au
Galaxy for tutorials: galaxy-tut.genome.edu.au
Galaxy Australia: galaxy-aust.genome.edu.au

Contributors and participants: