Streamlining Access to Reference Datasets

Daniel Blankenberg
The Galaxy Team
http://UseGalaxy.org
The Galaxy Team

Enis Afgan
Dannon Baker
Dan Blankenberg
Dave Bouvier
Marten Čech
John Chilton
Dave Clements
Nate Coraor
Carl Eberhard
Jeremy Goecks
Sam Guerler
Jen Jackson
Ross Lazarus
Anton Nekrutenko
Nick Stoler
James Taylor
Greg Von Kuster

http://wiki.galaxyproject.org/GalaxyTeam
Overview

Intro to Built-in Datasets

Some Problems

Data Managers
- What?
- Demo
Overview

Intro to Built-in Datasets

Some Problems

Data Managers
- What?
- Demo
Built-in Datasets

BWA example

Map with BWA for Illumina (version 1.2.3)

Will you select a reference genome from your history or use a built-in index?:

Use a built-in index

Select a reference genome:

- Arabidopsis lyrata: Araly1
- Armadillo (Dasypus novemcinctus): dasNov1
- Bacillus subtilis subsp. subtilis str. 168: baciSubt
- Bordetella bronchiseptica str. RB50: bordBron
- Budgerigar (Melopsittacus undulatus): melUnd1
- Burkholderia pseudomallei 1106a: burkPseu_1106A
- Burkholderia pseudomallei 1710b: 13954
- Burkholderia pseudomallei 668: 13953
- Burkholderia pseudomallei K96243: 178

BWA produces SAM with several lines of header information

Execute
Figure S1. Schematic overview of reference data, location files, data tables, and tools. A Galaxy Data Manager handles downloading, creating and installing each of the required facets necessary to provide built-in reference data through the use of a web-based GUI.

Built-in Datasets

bwa_wrapper.xml

```xml
<conditional name="genomeSource">
  <param name="refGenomeSource" type="select" label="Will you select a reference genome?">
    <option value="indexed">Use a built-in index</option>
    <option value="history">Use one from the history</option>
  </param>
  <when value="indexed">
    <param name="indices" type="select" label="Select a reference genome">
      <options from_data_table="bwa_indexes">
        <filter type="sort_by" column="2" />
        <validator type="no_options" message="No indexes are available" />
      </options>
    </param>
  </when>
</conditional>

<when value="history">
  <param name="ownFile" type="data" format="fasta" metadata_name="dbkey" label="Select your own file">
  </param>
</when>
```
Built-in Datasets

tool_data_table_conf.xml

```xml
<tables>
    <!-- Locations of indexes in the BWA mapper format -->
    <table name="bwa_indexes" comment_char="#">
        <columns>value, dbkey, name, path</columns>
        <file path="tool-data/bwa_index.loc" />
    </table>
</tables>
```
# This is a sample file distributed with Galaxy that enables tools
to use a directory of BWA indexed sequences data files. You will need
to create these data files and then create a bwa_index.loc file
similar to this one (store it in this directory) that points to
the directories in which those files are stored. The bwa_index.loc
file has this format (longer white space characters are TAB characters):

#<unique_build_id> <dbkey> <display_name> <file_path>

# So, for example, if you had phiX indexed stored in
#/depot/data2/galaxy/phiX/base/,
then the bwa_index.loc entry would look like this:

phiX174 phiX phiX Pretty /depot/data2/galaxy/phiX/base/phiX.fa

# and your /depot/data2/galaxy/phiX/base/ directory
# would contain phiX.fa.* files:

rw-r-r-- 1 janes universe 830134 2005-09-13 10:12 phiX.fa.amb
rw-r-r-- 1 janes universe 527388 2005-09-13 10:12 phiX.fa.ann
rw-r-r-- 1 janes universe 269808 2005-09-13 10:12 phiX.fa.bwt

# Your bwa_index.loc file should include an entry per line for each
# index set you have stored. The "file" in the path does not actually
# exist, but it is the prefix for the actual index files. For example:

#phiX174 phiX phiX174 /depot/data2/galaxy/phiX/base/phiX.fa
#hg18canon hg18 hg18 Canonical /depot/data2/galaxy/hg18/base/hg18canon.fa
#hg18full hg18 hg18 Full /depot/data2/galaxy/hg18/base/hg18full.fa
#/orig/path/hg19.fa hg19 hg19 /depot/data2/galaxy/hg19/base/hg19.fa

# etc...

# Note that for backwards compatibility with workflows, the unique ID of
# an entry must be the path that was in the original loc file, because that
# is the value stored in the workflow for that parameter. That is why the
# hg19 entry above looks odd. New genomes can be better-looking.

Arabidopsis lyrata: Arabidopsis /galaxy/data/Arabidopsis/bwa_index/Arabidopsis.fa
dasNov1 dasNov1 Armadillo (Dasypus novemcinctus): dasNov1 /galaxy/data/dasNov1/bwa_index/dasNov1.fa
baciSubt baciSubt Bacillus subtilis subsp. subtilis str. 168: baciSubt /galaxy/data/microbes/baciSubt/bwa_index/baciSubt.fa
Overview

Intro to Built-in Datasets

Some Problems

Data Managers
  ✷ What?
  ✷ Demo
Some Problems

Time consuming

~30 minutes for workshop to add one BWA index

Administrator needs to know how to update each type of reference data

Format of reference Data

Format of Location (.loc) file
Hi,

We have a local install of galaxy and I’m trying to add the reference index files for bwa using the information provided in the following link

http://wiki.g2.bx.psu.edu/Admin/NGS%20Local%20Setup

I have modified the bwa_index.loc file present in the ./tool-data directory by adding the path to where the index is on our server (Also attached). However, even after restarting the server, the reference genome does not show when choosing the “use a built-in index option”. I’m not sure whether the loc file is correctly created and whether any other configuration file needs to be changed/updated. Help in the matter greatly appreciated.

Thanks,

Aarti
Hi,

We have a local install of galaxy and I’m trying to add the reference index files for bwa using the information provided in the following link

http://wiki.g2.bx.psu.edu/Admin/NGS%20Local%20Setup

Hi Aarti,

Check the name of your ref file. If it is hg19.fa, then modify loc file as "hg19  hg19  HG19_BWA   /root/Ref_INDEX/HG19BWAIndex/base/hg19.fa"

Avik Datta
Hi,

We have a local install of galaxy and I’m trying to add the reference index files for bwa using the information provided in the following link

http://wiki.g2.bx.psu.edu/Admin/NGS%20Local%20Setup

Hi Aarti,

I have modified the bwa_index.loc file present in the ..tool-data directory by adding the path to where the index is on our server (Also attached). However, even after restarting the server, the reference genome does not show when choosing the “use a built-in index option”. I’m not sure whether the loc file is correctly created and whether any other configuration file needs to be changed/updated.

Help is appreciated.

Aarti

Also make sure you are using TABs to separate the fields in the .loc file, this has bitten me several time in the past. My vim config places 4 spaces instead of TAB, to deactivate this option you can do ":set nobreaktab".

Hope it helps,

Carlos
Hi,

We have a local install of galaxy and I’m trying to add the reference index files for bwa using the information provided in the following link

http://wiki.g2.bx.psu.edu/Admin/NGS%20Local%20Setup

Hi Aarti,

I have modified the bwa_index.loc file present in the ..tool-data directory by adding the path to where the index is on our server (Also attached). However, even after restarting the server, the reference genome does not show when choosing the “use a built-in index option”. I’m not sure whether the loc file is correctly created and whether any other configuration file needs to be changed/updated.

Help is greatly appreciated.

Also make sure you are using TABs to separate the fields in the .loc file, this has helped me remove the warning in the log.

Hello Carlos,

Thanks a lot for the tip. The tab trick has fixed the problem.

Regards,
Aarti
Other concerns

Accessible?
- Manually download genome FASTA files
- Download, compile, run bwa index; which options?

Reproducible?
- Only if the person performing manual steps keeps good notes

Transparent?
- Send email to sysadmin asking for notes

Need to restart Galaxy server when new entries are added
Overview

Intro to Built-in Datasets

Some Problems

Data Managers
- What?
- Demo
Overview

Intro to Built-in Datasets

Some Problems

Data Managers
  ✦ What?
  ✦ Demo
Data Managers

Allows for the creation of built-in (reference) data

underlying data

data tables

*.loc files

Specialized Galaxy tools that can only be accessed by an admin

Defined locally or installed from ToolShed
Data Managers

**Flexible** Framework

not just Genomic data

Interactively Run Data Managers through UI

Workflow compatible

API

Examples:

Adding New genome builds (dbkeys)

Fetching Genome (FASTA) sequences

Building short read mapper indexes for genomes
Special class of Galaxy tool

```
<tool id="data_manager_fetch_genome_all_fasta" name="Reference Genome" version="0.0.1" tool_type="manage_data">
  <outputs>
    <data name="out_file" format="data_manager_json"/>
  </outputs>
</tool>
```

Writes a JSON description of new data table entries as content of tool output file

```
{  
  "data_tables":{  
    "all_fasta": [  
      {  
        "path":"sacCer2.fa",  
        "dbkey":"sacCer2",  
        "name":"S. cerevisiae June 2008 (SGD/sacCer2) (sacCer2)",  
        "value":"sacCer2"  
      }  
    ]  
  }  
}
```

This creates a new entry in the Tool Data Table:

```
<table>
<thead>
<tr>
<th>&lt;unique_build_id&gt;</th>
<th>&lt;dbkey&gt;</th>
<th>&lt;display_name&gt;</th>
<th>&lt;file_path&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>sacCer2</td>
<td>sacCer2</td>
<td>S. cerevisiae June 2008 (SGD/sacCer2) (sacCer2)</td>
<td>/Users/dan/galaxy-central/tool-data/sacCer2/seq/sacCer2.fa</td>
</tr>
</tbody>
</table>
```

Where the sacCer2.fa file was placed by the tool in the output file’s extra_files_path
data_manager entry inside <data_managers> tag in data_mananger_conf.xml

informs Galaxy about which data tables to expect for new entries special handling of provided JSON values and files
Data Managers: Configuration

`enable_data_manager_user_view` allows non-admin users to view the available data that has been managed.

`data_manager_config_file` defines the local xml file to use for loading the configurations of locally defined data managers.

`shed_data_manager_config_file` defines the local xml file to use for saving and loading the configurations of locally defined data managers.

`galaxy_data_manager_data_path` defines the location to use for storing the files created by Data Managers. When not configured it defaults to the value of `tool_data_path`.
Overview

Intro to Built-in Datasets

Some Problems

Data Managers
- What?
- Demo
Data Manager Demo

- **Fetch the Genome Sequence** for sacCer3
  - UCSC as the source
  - Install fetching tool from ToolShed
  - define new Genome build / dbkey
  - all.fasta & __dbkeys__ tables are populated automatically

- **Build BWA indexes** for sacCer3
  - Install indexing tool from ToolShed
  - Build indexes
  - bwa_index table is populated automatically

- **Align** some reads to the newly added reference genome
Data Manager Demo: Full Disclosure

Fresh Install of galaxy-central stable

- Setup Galaxy admin account already
- Configured tool_dependency_dir
- The sequencing reads are a small subset from SRR507778, originally downloaded from EBI SRA.
- Installed BWA mapper

http://gcc2014.dblankenberg.org/
Make Your Own

Documentation
https://wiki.galaxyproject.org/Admin/Tools/DataManagers/

Several examples available in the ToolShed (search for “data_manager”)

Training Day Exercises