Managing Galaxy's Built-in Data

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

Intro to Built-in Datasets

A problem

Data Managers
- What?
- Demo
Overview

Intro to Built-in Datasets

A problem

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
Built-in Datasets

BWA example

Reference Genome (*.fasta)

bwa index [-a bwtsw|div|is] [-c] <in.fasta>

BWA Index Files on disk (*.amb, *.bwt, *.rbwt, *.rsa, *.ann, *.pac, *.rpac, *.sa)

File listing paths, genome build, descriptions, etc. of available BWA indexes (bwa_index.loc)

abstraction layer

Galaxy BWA Tool (bwa_wrapper.xml)

Tool Data Table (tool_data_table_conf.xml)
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>
  <when value="history">
    <param name="ownFile" type="data" format="fasta" metadata_name="dbkey" label="Select a reference genome">
    </param>
  </when>
</conditional>
```
Built-in Datasets

tool_data_table_conf.xml

```
<table name="bwa_indexes" comment_char="#">
    <columns>value, dbkey, name, path</columns>
    <file path="tool-data/bwa_index.loc" />
</table>
```
Built-in Datasets

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 james  universe  836134  2005-09-13  10:12 phiX.fa.amb
-rw-r--r--  1 james  universe  527388  2005-09-13  10:12 phiX.fa.ann
-rw-r--r--  1 james  universe  269808  2005-09-13  10:12 phiX.fa.bat
... etc...
```

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  Canonical /depot/data2/galaxy/hg18/base/hg18canon.fa
hg18full  hg18  Full /depot/data2/galaxy/hg18/base/hg18full.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
bacisSubt  bacisSubt  Bacillus subtilis subsp. subtilis str. 168: bacisSubt  /galaxy/data/microbes/bacisSubt/bwa_index/bacisSubt.fa
... etc...
```
Overview

Intro to Built-in Datasets

A problem

Data Managers
  ✷ What?
  ✷ Demo
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

Aarti
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 up when checking 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.

Check the name of your ref file. If it is hg19.fa, then modify loc file as "hg19.fa".

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 noexpandtab".

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/NGS20Local20Setup

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 "hg19 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.

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

A problem

Data Managers
- What?
- Demo
Overview

Intro to Built-in Datasets

A problem

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 Tool Shed
Data Managers

- **Flexible** Framework
  - not just Genomic data
  - Interactively Run Data Managers through UI
  - Workflow compatible
  - API

- Examples:
  - Fetching Genome (FASTA) sequences
  - Building short read mapper indexes for genomes
Overview

Intro to Built-in Datasets

A problem

Data Managers
- What?
- Demo
Data Manager Demo

- **Fetch the Genome Sequence** for sacCer2
  - UCSC as the source
  - Install fetching tool from Tool Shed
  - all_fasta table is populated automatically

- **Build BWA indexes** for sacCer2
  - Install indexing tool from Tool Shed
  - Build indexes
  - bwa_index table is populated automatically

- **Align some reads** to the newly added reference genome
The default **CloudMan** instance **comes with sacCer2** pre-installed, but I deleted it.

The **sequencing reads** to be aligned with BWA were created locally from the reference genome before uploading.

Setup Galaxy **admin account** already

Make Your Own

http://wiki.galaxyproject.org/Admin/Tools/DataManagers/HowTo/Define

Several examples available in the test Tool Shed (search for “data_manager”)
The Galaxy Team

http://wiki.galaxyproject.org/GalaxyTeam