Less Click, More Quick: Unattended Installation of Galaxy's Built-in Reference Data

Daniel Blankenberg
The Galaxy Team | Penn State
http://GalaxyProject.org

Overview

Intro to Built-in Datasets

Some Problems

Data Managers

- What?
- Demo

Overview

Intro to Built-in Datasets

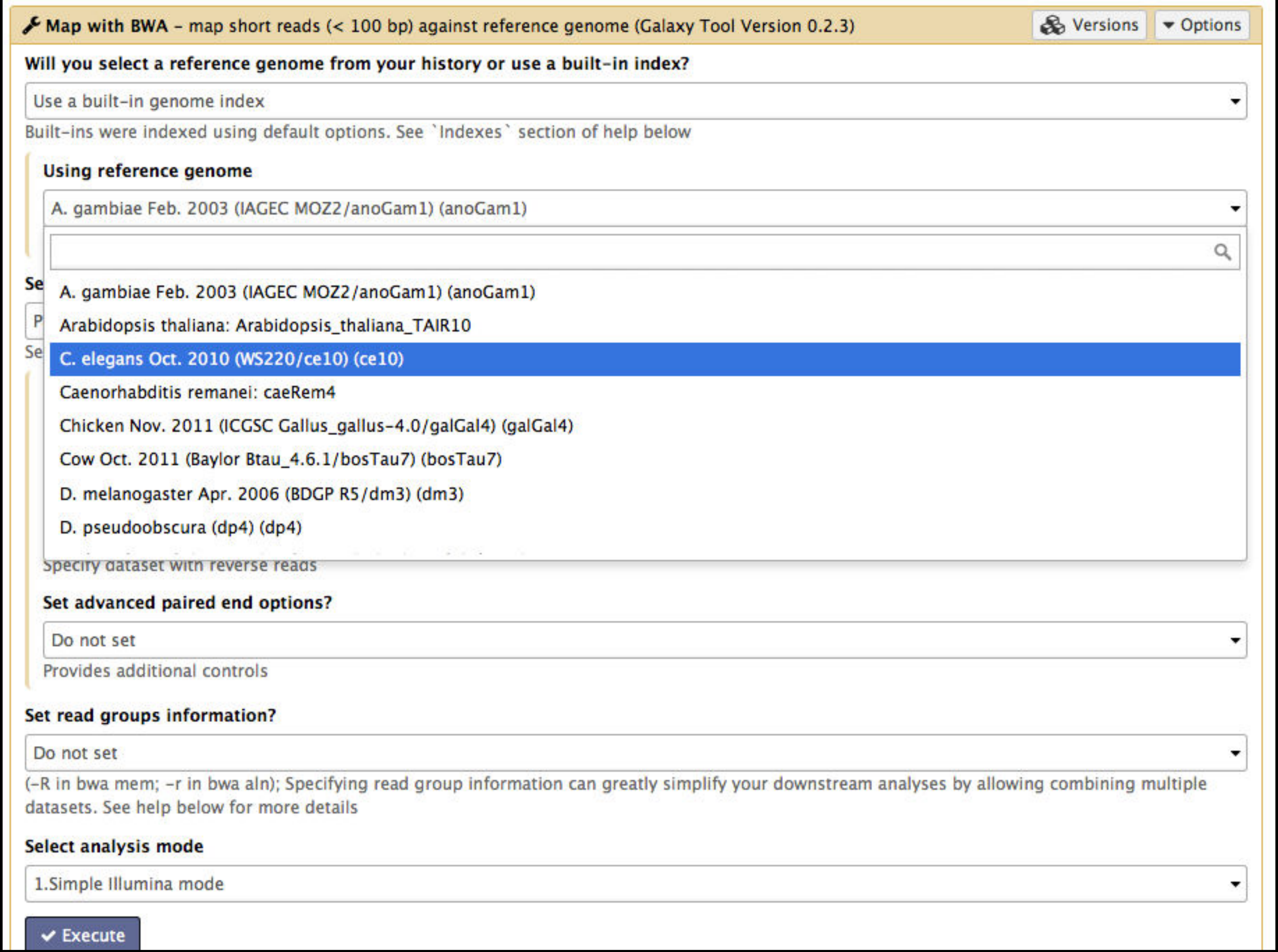
Some Problems

Data Managers

- What?
- Demo

Built-in Datasets

BWA example



Map with BWA – map short reads (< 100 bp) against reference genome (Galaxy Tool Version 0.2.3)

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

Use a built-in genome index

Built-ins were indexed using default options. See `Indexes` section of help below

Using reference genome

A. gambiae Feb. 2003 (IAGEC MOZ2/anoGam1) (anoGam1)

- A. gambiae Feb. 2003 (IAGEC MOZ2/anoGam1) (anoGam1)
- Arabidopsis thaliana: Arabidopsis_thaliana_TAIR10
- C. elegans Oct. 2010 (WS220/ce10) (ce10)
- Caenorhabditis remanei: caeRem4
- Chicken Nov. 2011 (ICGSC Gallus_gallus-4.0/galGal4) (galGal4)
- Cow Oct. 2011 (Baylor Btau_4.6.1/bosTau7) (bosTau7)
- D. melanogaster Apr. 2006 (BDGP R5/dm3) (dm3)
- D. pseudoobscura (dp4) (dp4)

Specify dataset with reverse reads

Set advanced paired end options?

Do not set

Provides additional controls

Set read groups information?

Do not set

(-R in bwa mem; -r in bwa aln); Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets. See help below for more details

Select analysis mode

1.Simple Illumina mode

Built-in Datasets

BWA example

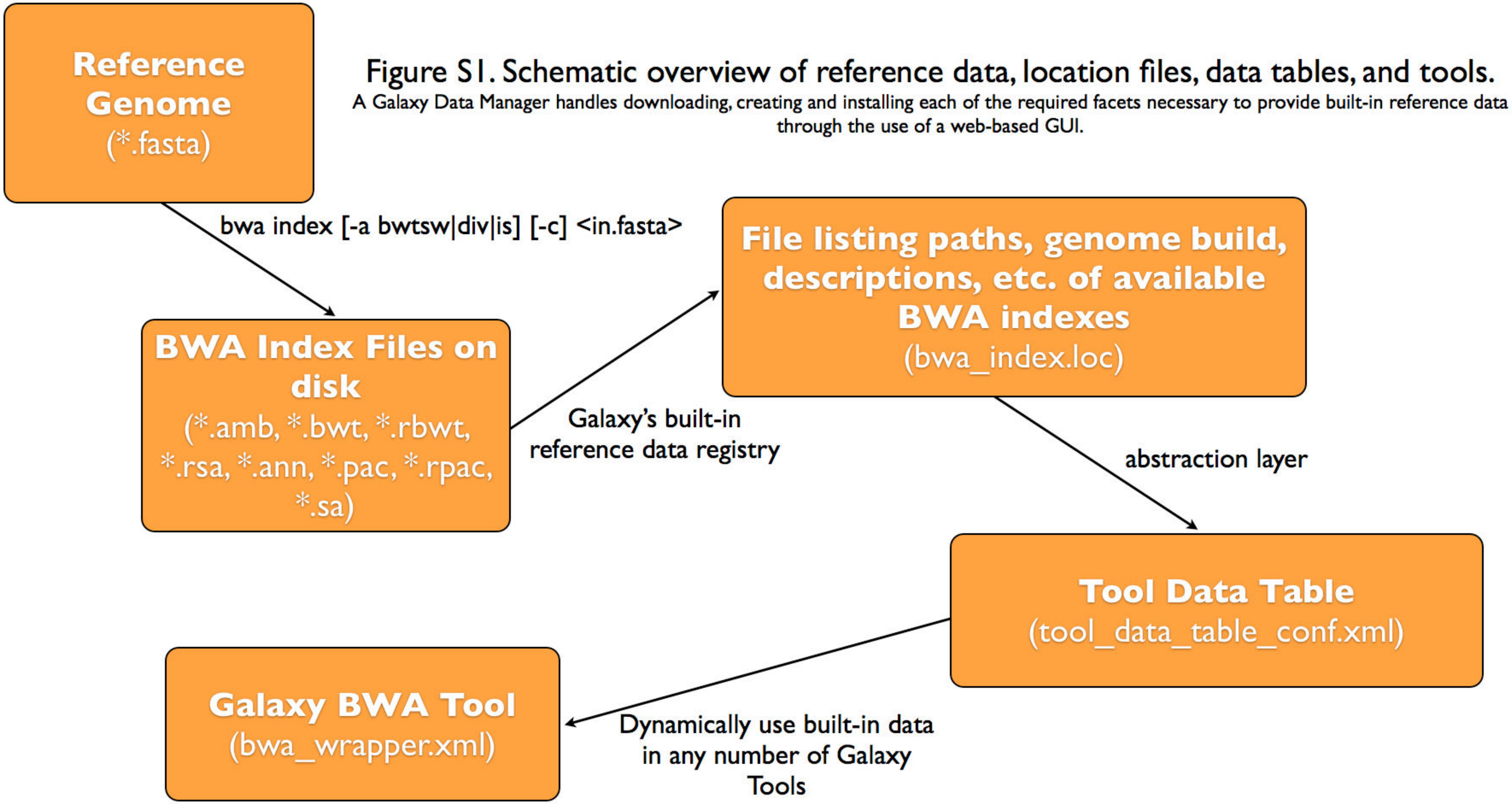


Figure S1. Schematic overview of reference data, location files, data tables, and tools.

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.

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)

Galaxy's built-in reference data registry

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

abstraction layer

Tool Data Table
(tool_data_table_conf.xml)

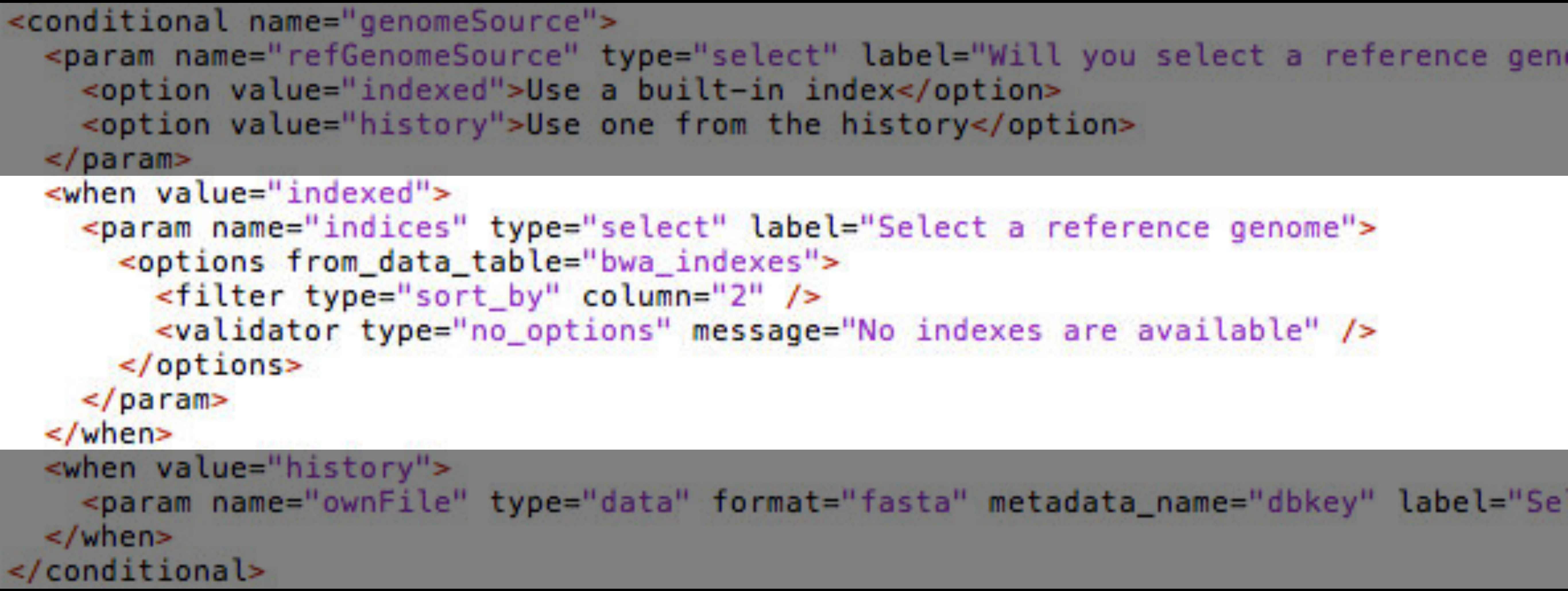
Dynamically use built-in data in any number of Galaxy Tools

Galaxy BWA Tool
(bwa_wrapper.xml)

Blankenberg, et al. Wrangling Galaxy's reference data. Bioinformatics. 2014 Jul 1.;30(13):1917-9.

Built-in Datasets

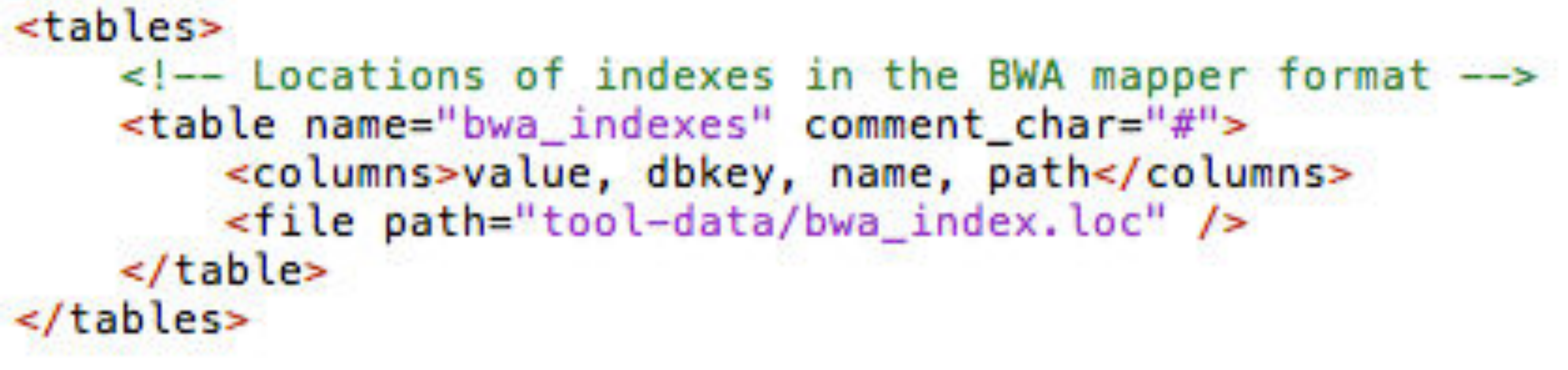
bwa_wrapper.xml



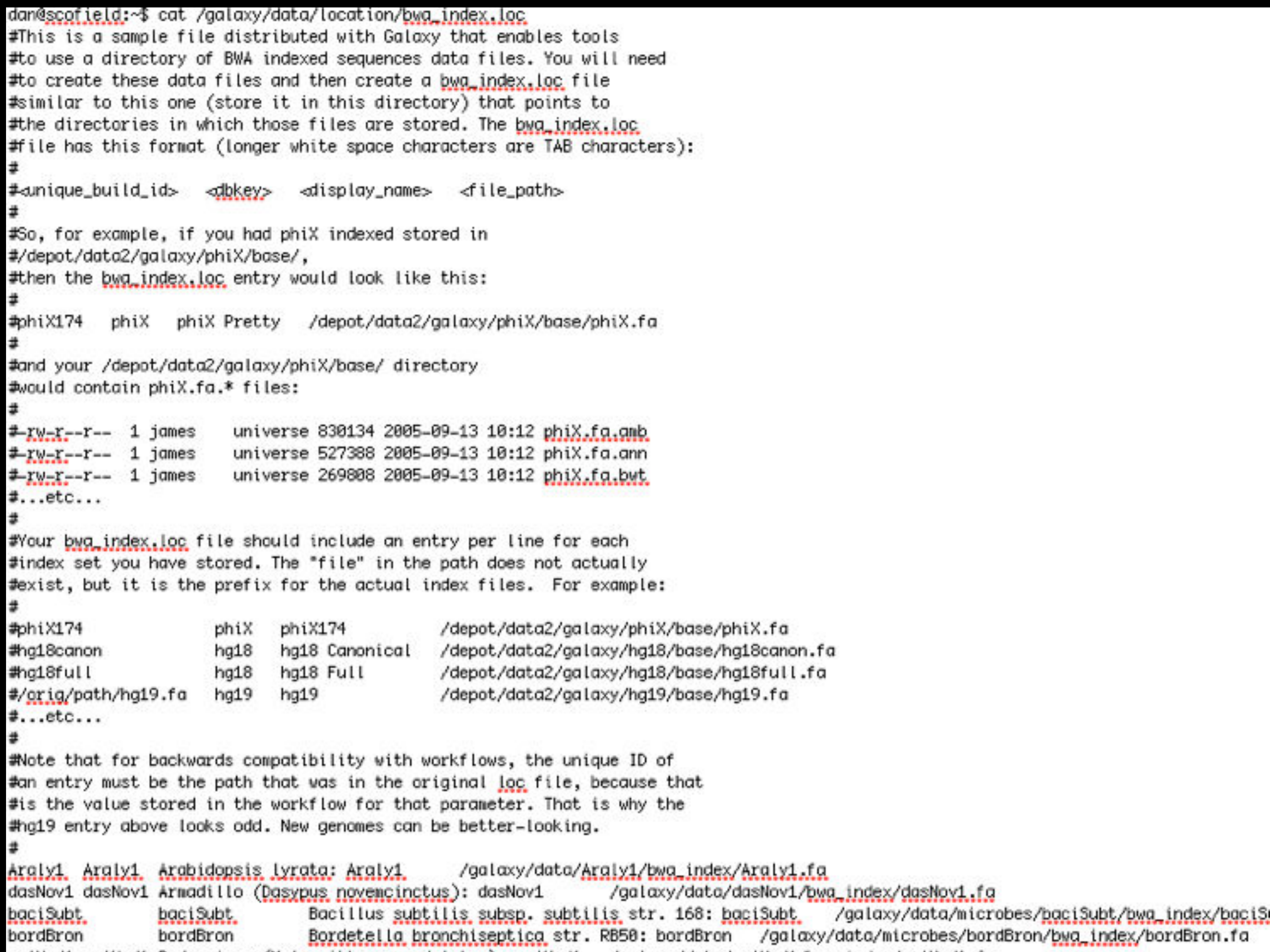
```
<conditional name="genomeSource">
  <param name="refGenomeSource" type="select" label="Will you select a reference gen
    <option value="indexed">Use a built-in index</option>
    <option value="history">Use one from the history</option>
  </param>
  <when value="indexed">
    <param name="indices" 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="Se
  </when>
</conditional>
```

Built-in Datasets

tool_data_table_conf.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>
```



```
dan@scofield:~$ cat /galaxy/data/location/bwa_index.loc
#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 830134 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.bwt
#...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   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.
#
Araly1  Araly1  Arabidopsis lyrata: Araly1      /galaxy/data/Araly1/bwa_index/Araly1.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/baciS
bordBron        bordBron        Bordetella bronchiseptica str. RB50: bordBron   /galaxy/data/microbes/bordBron/bwa_index/bordBron.fa
```

Overview

Intro to Built-in Datasets

Some Problems

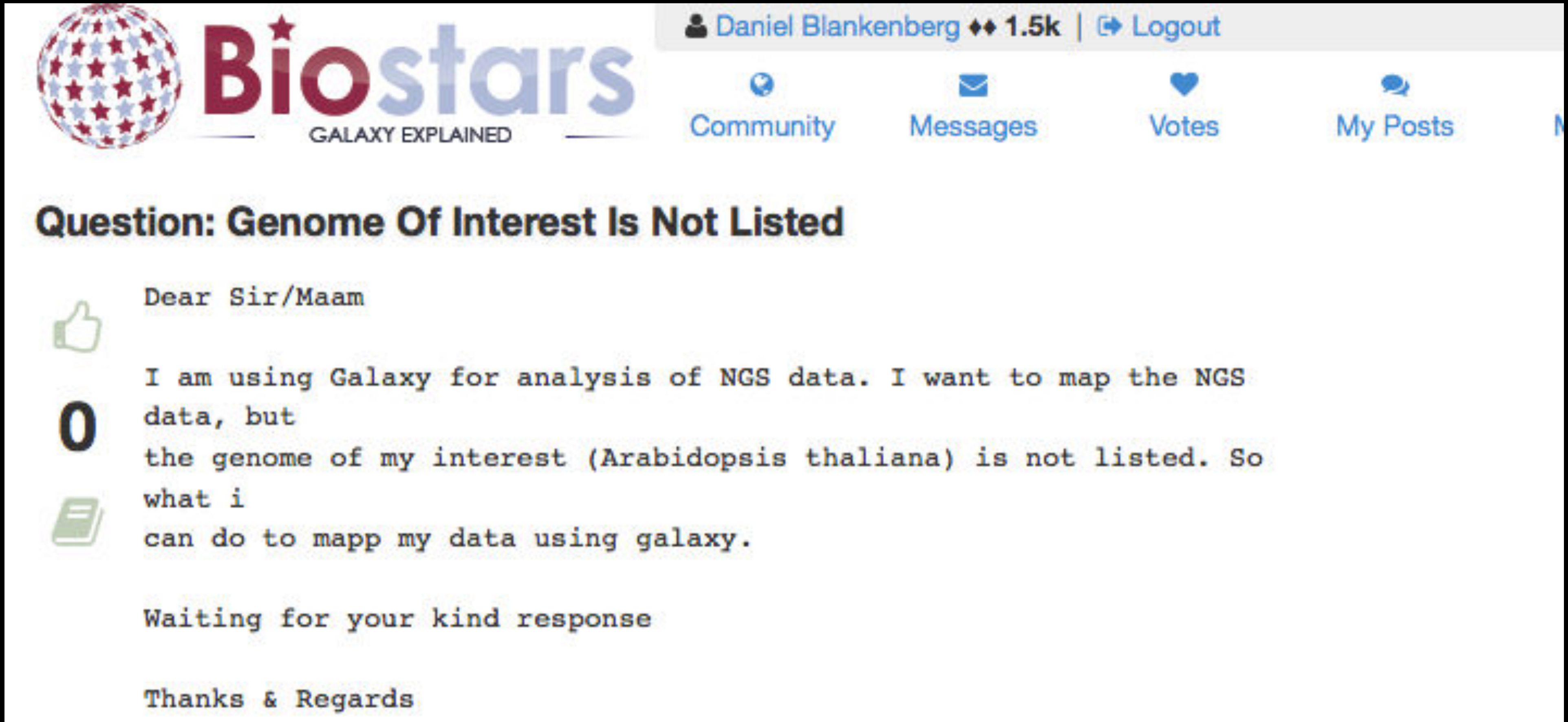
Data Managers

- What?
- Demo

Some Problems

- Time consuming and prone to Error
    - Manual process
- Administrator needs to know how to update each type of reference data
    - Format of reference Data
    - Format of Location (.loc) file
- Tools using References from the user’s History are slow

Reference not available in local instance



Question: Genome Of Interest Is Not Listed

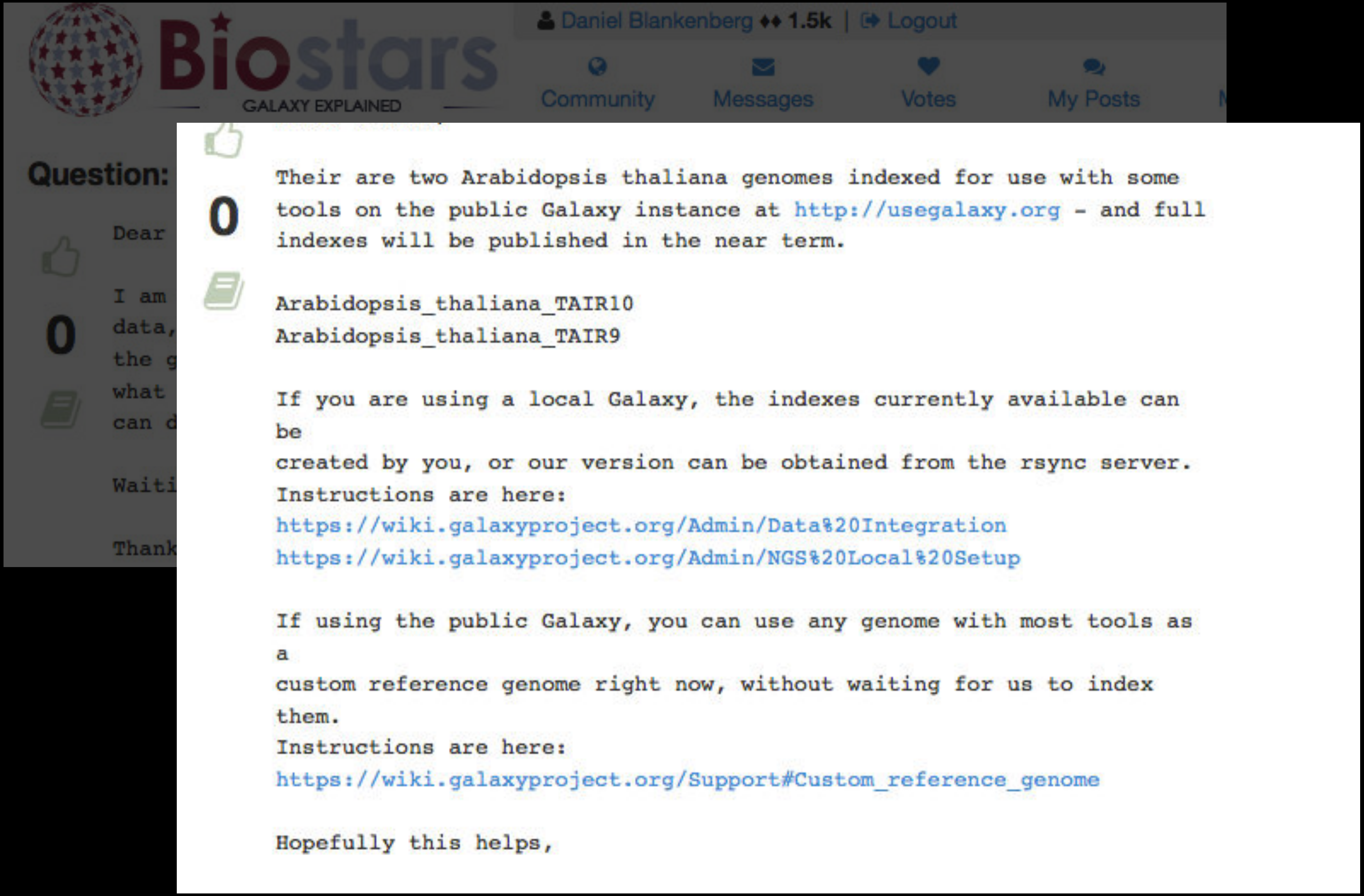
Dear Sir/Maam

I am using Galaxy for analysis of NGS data. I want to map the NGS data, but the genome of my interest (Arabidopsis thaliana) is not listed. So what i can do to mapp my data using galaxy.

Waiting for your kind response

Thanks & Regards

Build or Rsync



Their are two Arabidopsis thaliana genomes indexed for use with some tools on the public Galaxy instance at http://usegalaxy.org - and full indexes will be published in the near term.

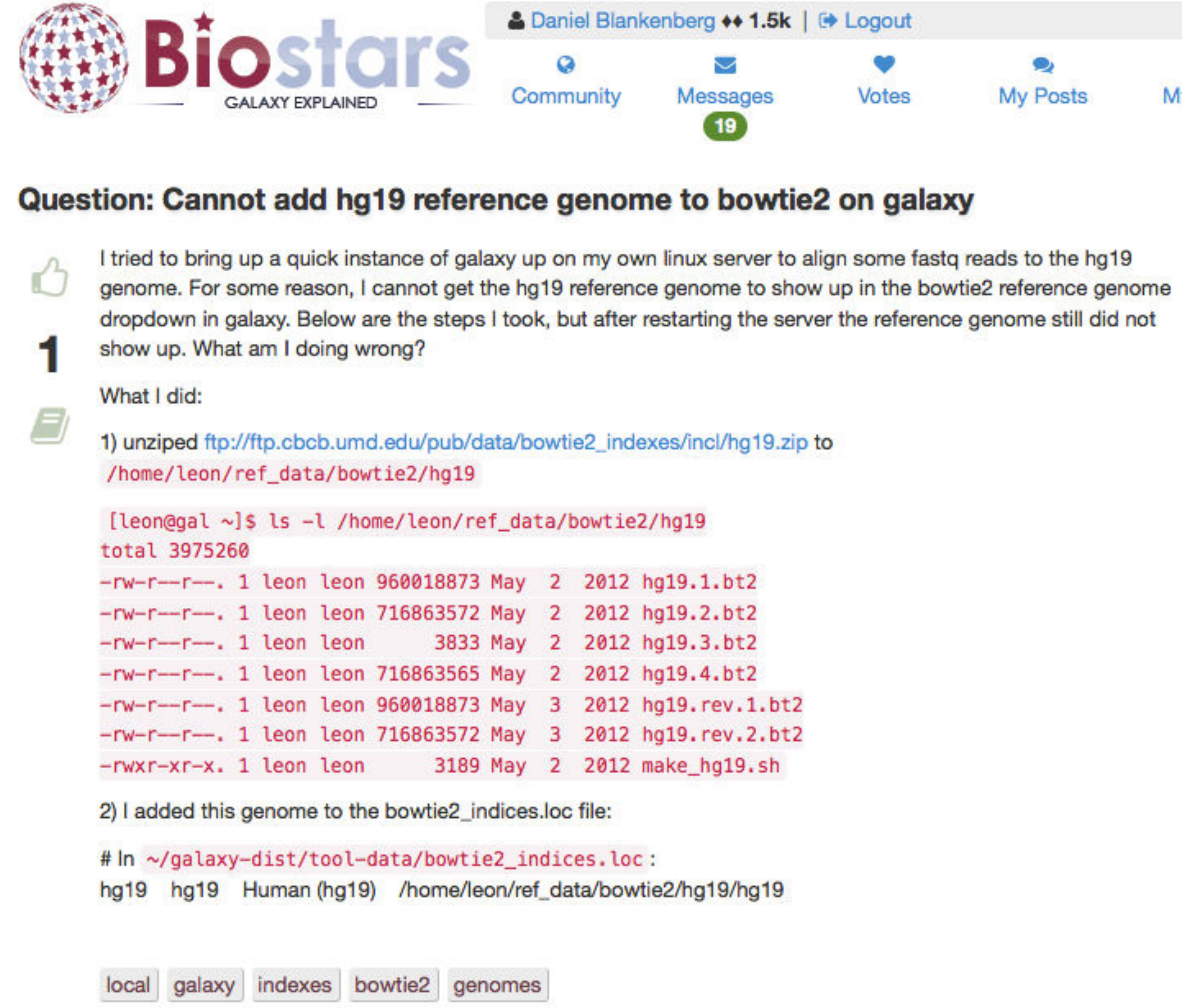
Arabidopsis_thaliana_TAIR10
Arabidopsis_thaliana_TAIR9

If you are using a local Galaxy, the indexes currently available can be created by you, or our version can be obtained from the rsync server.
Instructions are here:
https://wiki.galaxyproject.org/Admin/Data%20Integration
https://wiki.galaxyproject.org/Admin/NGS%20Local%20Setup

If using the public Galaxy, you can use any genome with most tools as a custom reference genome right now, without waiting for us to index them.
Instructions are here:
https://wiki.galaxyproject.org/Support#Custom_reference_genome

Hopefully this helps,

Have data, modified .loc file, still not working



Question: Cannot add hg19 reference genome to bowtie2 on galaxy

1

I tried to bring up a quick instance of galaxy up on my own linux server to align some fastq reads to the hg19 genome. For some reason, I cannot get the hg19 reference genome to show up in the bowtie2 reference genome dropdown in galaxy. Below are the steps I took, but after restarting the server the reference genome still did not show up. What am I doing wrong?

What I did:

1) unziped ftp://ftp.cbcb.umd.edu/pub/data/bowtie2_indexes/incl/hg19.zip to `/home/leon/ref_data/bowtie2/hg19`

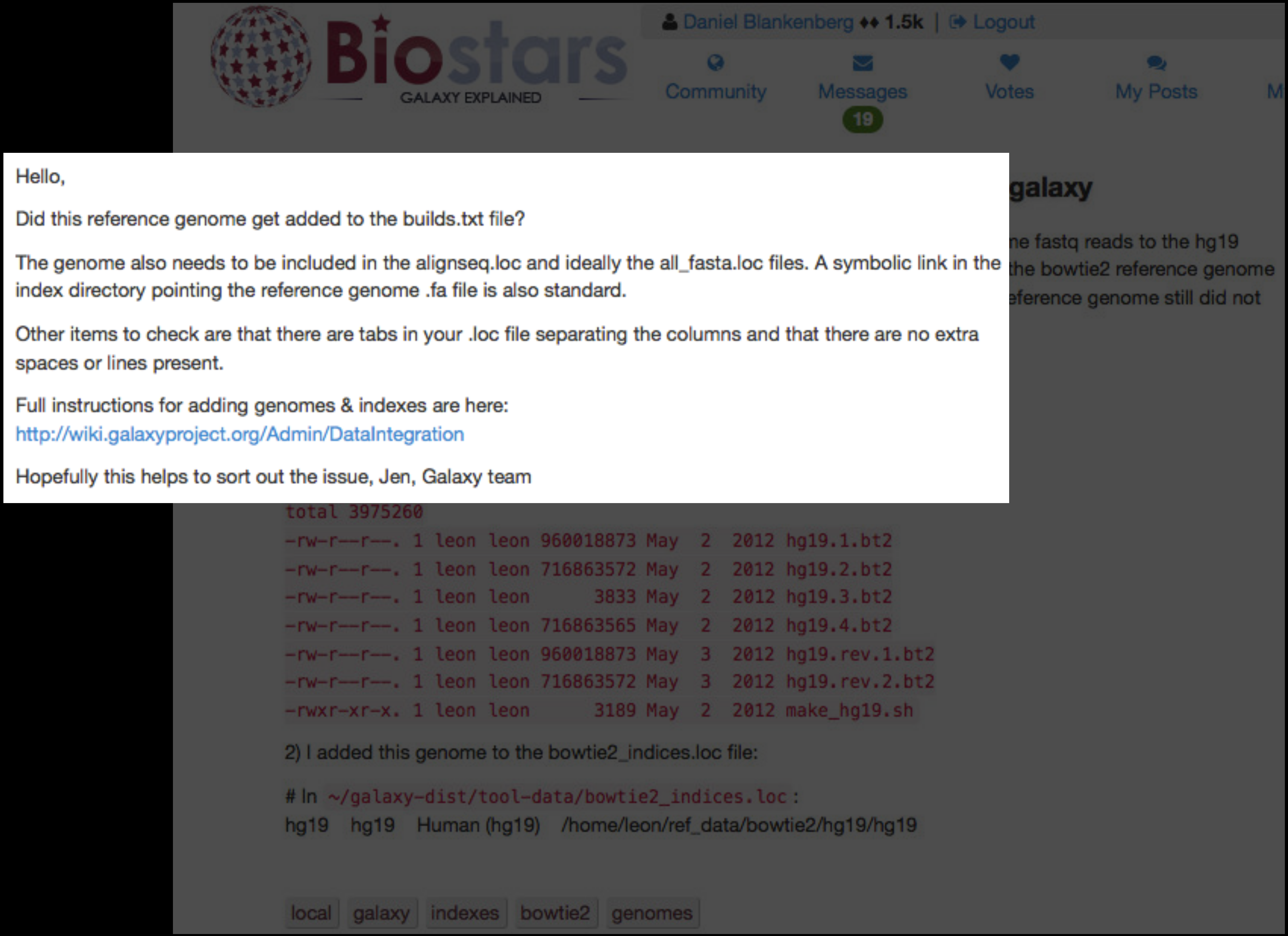
```
[leon@gal ~]$ ls -l /home/leon/ref_data/bowtie2/hg19
total 3975260
-rw-r--r--. 1 leon leon 960018873 May  2  2012 hg19.1.bt2
-rw-r--r--. 1 leon leon 716863572 May  2  2012 hg19.2.bt2
-rw-r--r--. 1 leon leon      3833 May  2  2012 hg19.3.bt2
-rw-r--r--. 1 leon leon 716863565 May  2  2012 hg19.4.bt2
-rw-r--r--. 1 leon leon 960018873 May  3  2012 hg19.rev.1.bt2
-rw-r--r--. 1 leon leon 716863572 May  3  2012 hg19.rev.2.bt2
-rwxr-xr-x. 1 leon leon      3189 May  2  2012 make_hg19.sh
```

2) I added this genome to the bowtie2_indices.loc file:

In `~/galaxy-dist/tool-data/bowtie2_indices.loc`:

```
hg19    hg19    Human (hg19)    /home/leon/ref_data/bowtie2/hg19/hg19
```

local galaxy indexes bowtie2 genomes



Hello,

Did this reference genome get added to the builds.txt file?

The genome also needs to be included in the alignseq.loc and ideally the all_fasta.loc files. A symbolic link in the index directory pointing the reference genome .fa file is also standard.

Other items to check are that there are tabs in your .loc file separating the columns and that there are no extra spaces or lines present.

Full instructions for adding genomes & indexes are here:
http://wiki.galaxyproject.org/Admin/DataIntegration

Hopefully this helps to sort out the issue, Jen, Galaxy team

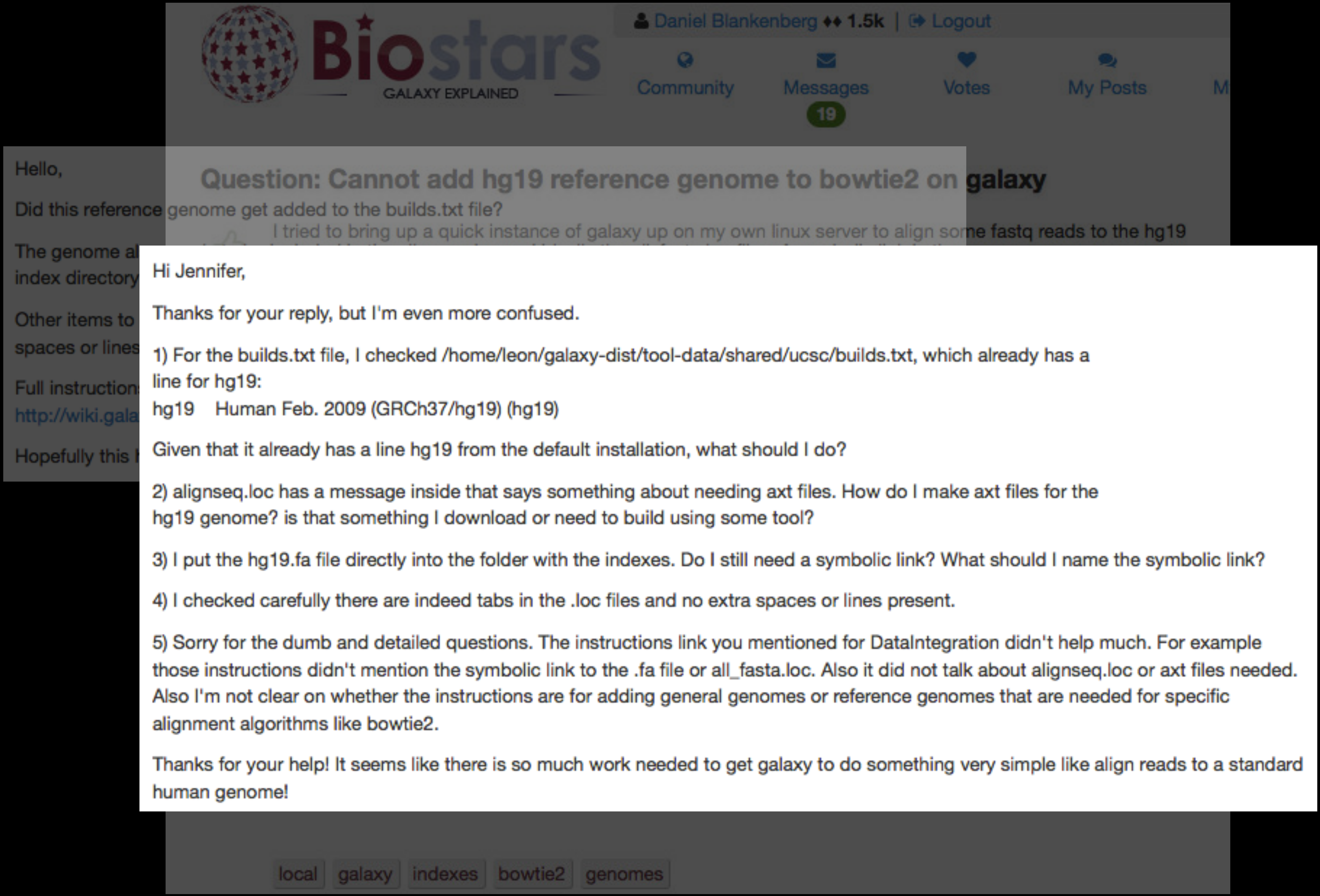
galaxy

ne fastq reads to the hg19
the bowtie2 reference genome
eference genome still did not

```
total 3975260
-rw-r--r--. 1 leon leon 960018873 May  2  2012 hg19.1.bt2
-rw-r--r--. 1 leon leon 716863572 May  2  2012 hg19.2.bt2
-rw-r--r--. 1 leon leon      3833 May  2  2012 hg19.3.bt2
-rw-r--r--. 1 leon leon 716863565 May  2  2012 hg19.4.bt2
-rw-r--r--. 1 leon leon 960018873 May  3  2012 hg19.rev.1.bt2
-rw-r--r--. 1 leon leon 716863572 May  3  2012 hg19.rev.2.bt2
-rwxr-xr-x. 1 leon leon      3189 May  2  2012 make_hg19.sh
```

2) I added this genome to the bowtie2_indices.loc file:

In `~/galaxy-dist/tool-data/bowtie2_indices.loc` :
hg19   hg19   Human (hg19)   /home/leon/ref_data/bowtie2/hg19/hg19



Question: Cannot add hg19 reference genome to bowtie2 on galaxy

I tried to bring up a quick instance of galaxy up on my own linux server to align some fastq reads to the hg19

Hello,

Did this reference genome get added to the builds.txt file?

The genome al
index directory

Other items to
spaces or lines

Full instruction
http://wiki.gala

Hopefully this h

Hi Jennifer,

Thanks for your reply, but I'm even more confused.

1) For the builds.txt file, I checked /home/leon/galaxy-dist/tool-data/shared/ucsc/builds.txt, which already has a line for hg19:
hg19 Human Feb. 2009 (GRCh37/hg19) (hg19)

Given that it already has a line hg19 from the default installation, what should I do?

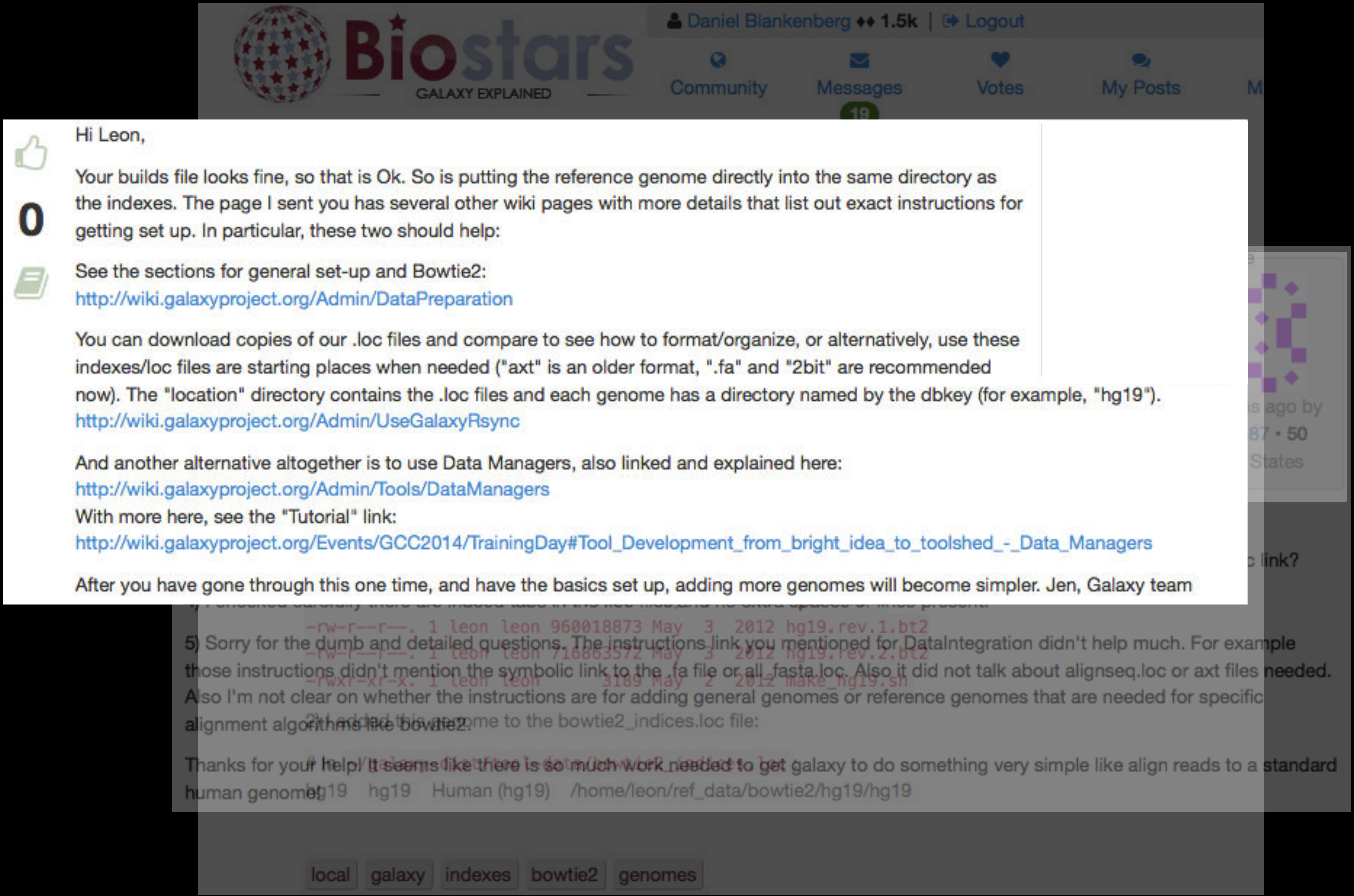
2) alignseq.loc has a message inside that says something about needing axt files. How do I make axt files for the hg19 genome? is that something I download or need to build using some tool?

3) I put the hg19.fa file directly into the folder with the indexes. Do I still need a symbolic link? What should I name the symbolic link?

4) I checked carefully there are indeed tabs in the .loc files and no extra spaces or lines present.

5) Sorry for the dumb and detailed questions. The instructions link you mentioned for DataIntegration didn't help much. For example those instructions didn't mention the symbolic link to the .fa file or all_fasta.loc. Also it did not talk about alignseq.loc or axt files needed. Also I'm not clear on whether the instructions are for adding general genomes or reference genomes that are needed for specific alignment algorithms like bowtie2.

Thanks for your help! It seems like there is so much work needed to get galaxy to do something very simple like align reads to a standard human genome!



Hi Leon,

Your builds file looks fine, so that is Ok. So is putting the reference genome directly into the same directory as the indexes. The page I sent you has several other wiki pages with more details that list out exact instructions for getting set up. In particular, these two should help:

See the sections for general set-up and Bowtie2:
http://wiki.galaxyproject.org/Admin/DataPreparation

You can download copies of our .loc files and compare to see how to format/organize, or alternatively, use these indexes/loc files are starting places when needed ("axt" is an older format, ".fa" and "2bit" are recommended now). The "location" directory contains the .loc files and each genome has a directory named by the dbkey (for example, "hg19").
http://wiki.galaxyproject.org/Admin/UseGalaxyRsync

And another alternative altogether is to use Data Managers, also linked and explained here:
http://wiki.galaxyproject.org/Admin/Tools/DataManagers
With more here, see the "Tutorial" link:
http://wiki.galaxyproject.org/Events/GCC2014/TrainingDay#Tool_Development_from_bright_idea_to_toolshed_-_Data_Managers

After you have gone through this one time, and have the basics set up, adding more genomes will become simpler. Jen, Galaxy team

```
-rw-r--r--. 1 leon leon 960018873 May  3  2012 hg19.rev.1.bt2
-rw-r--r--. 1 leon leon 716863572 May  3  2012 hg19.rev.2.bt2
-rwxr-xr-x. 1 leon leon      3189 May  2  2012 make_hg19.sh
```

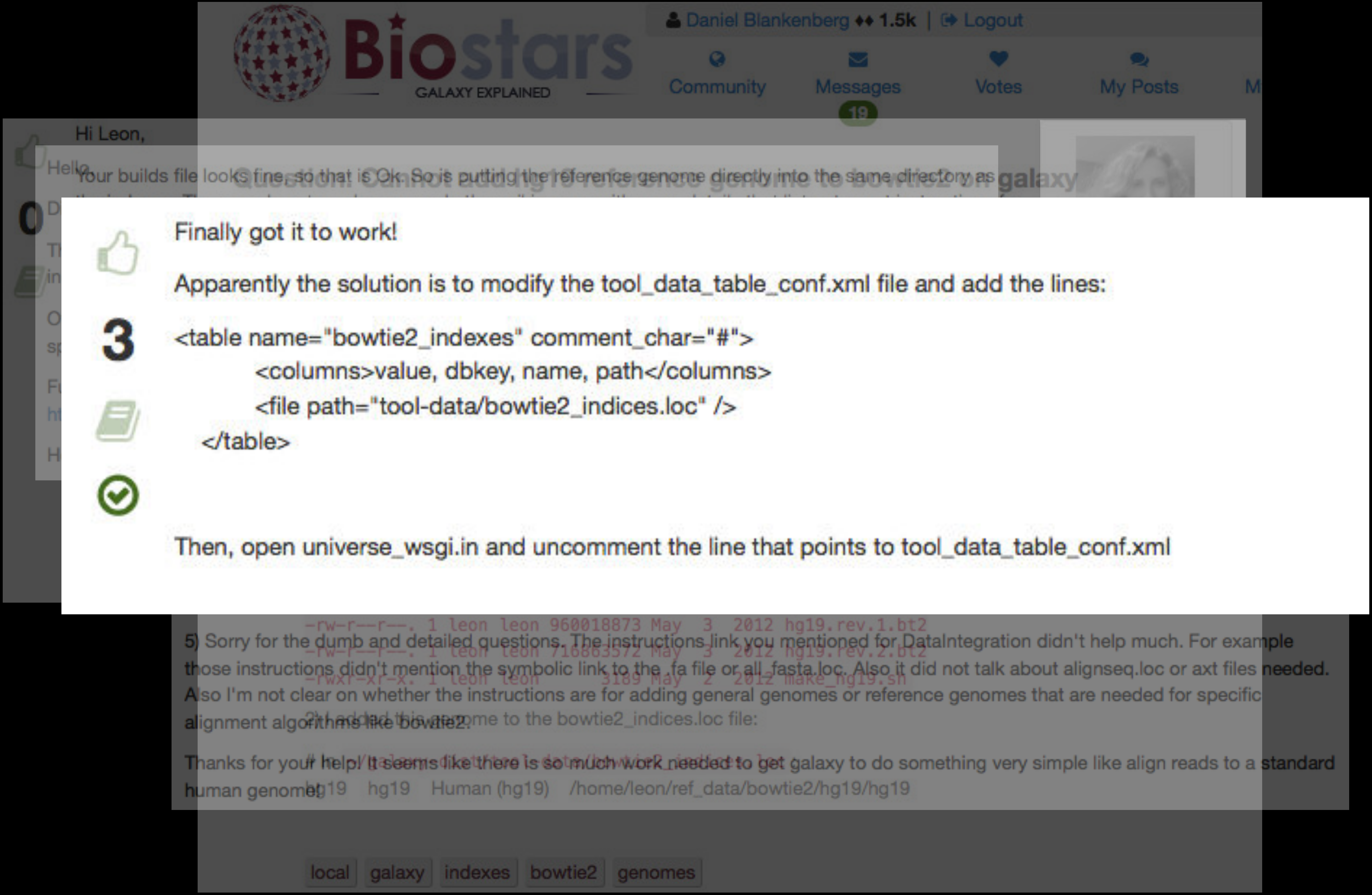
2) I added this genome to the bowtie2_indices.loc file:

```
hg19    hg19    Human (hg19)    /home/leon/ref_data/bowtie2/hg19/hg19
```

5) Sorry for the dumb and detailed questions. The instructions link you mentioned for DataIntegration didn't help much. For example those instructions didn't mention the symbolic link to the .fa file or all_fasta.loc. Also it did not talk about alignseq.loc or axt files needed. Also I'm not clear on whether the instructions are for adding general genomes or reference genomes that are needed for specific alignment algorithms like bowtie2?

Thanks for your help! It seems like there is so much work needed to get galaxy to do something very simple like align reads to a standard human genome.

Needed to uncomment Data Table in tool_data_table_conf.xml



Hi Leon,

Your builds file looks fine, so that is Ok. So is putting the reference genome directly into the same directory as

3

Finally got it to work!

Apparently the solution is to modify the tool_data_table_conf.xml file and add the lines:

```
<table name="bowtie2_indexes" comment_char="#">
         <columns>value, dbkey, name, path</columns>
         <file path="tool-data/bowtie2_indices.loc" />
   </table>
```

Then, open universe_wsgi.in and uncomment the line that points to tool_data_table_conf.xml

```
-rw-r--r--. 1 leon leon 960018873 May  3  2012 hg19.rev.1.bt2
-rw-r--r--. 1 leon leon 716863572 May  3  2012 hg19.rev.2.bt2
-rwxr-xr-x. 1 leon leon      3189 May  2  2012 make_hg19.sh
```

2) I added this genome to the bowtie2_indices.loc file:

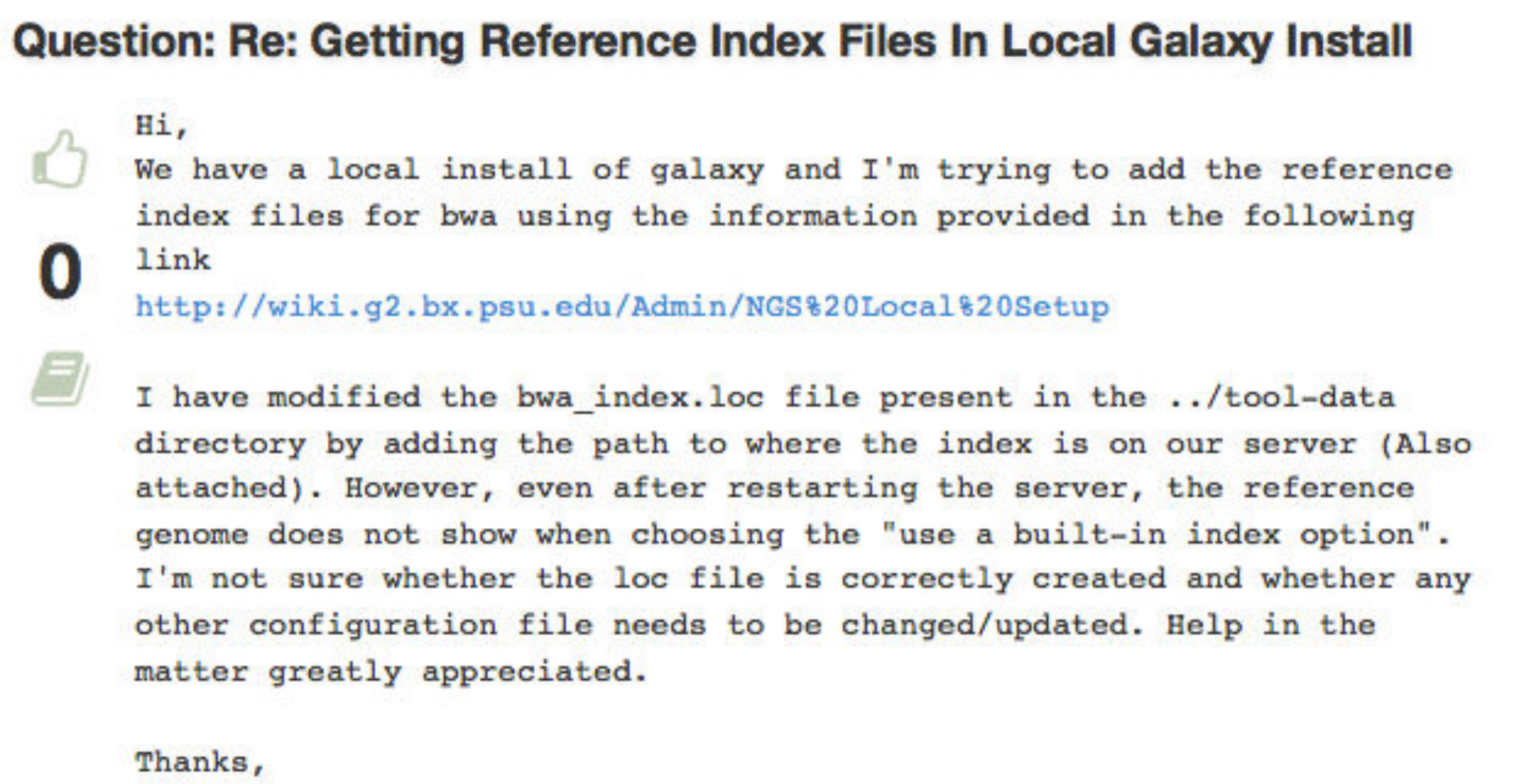
```
hg19    hg19    Human (hg19)    /home/leon/ref_data/bowtie2/hg19/hg19
```

5) Sorry for the dumb and detailed questions. The instructions link you mentioned for DataIntegration didn't help much. For example those instructions didn't mention the symbolic link to the .fa file or all_fasta.loc. Also it did not talk about alignseq.loc or axt files needed. Also I'm not clear on whether the instructions are for adding general genomes or reference genomes that are needed for specific alignment algorithms like bowtie2?

Thanks for your help! It seems like there is so much work needed to get galaxy to do something very simple like align reads to a standard human genome!

local galaxy indexes bowtie2 genomes

Followed setup directions, built index files, modified .loc file, and tool-data table exists

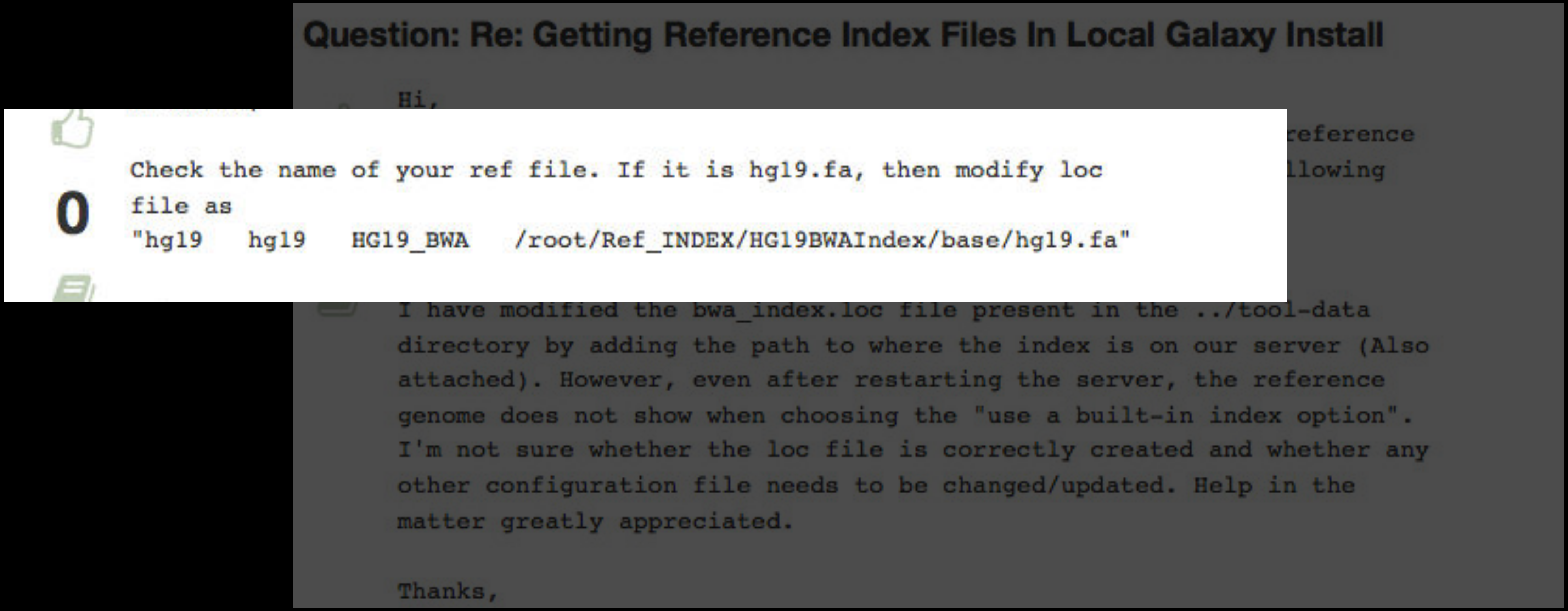


Question: Re: Getting Reference Index Files In Local Galaxy Install

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,



Question: Re: Getting Reference Index Files In Local Galaxy Install

Hi,

reference

llowing

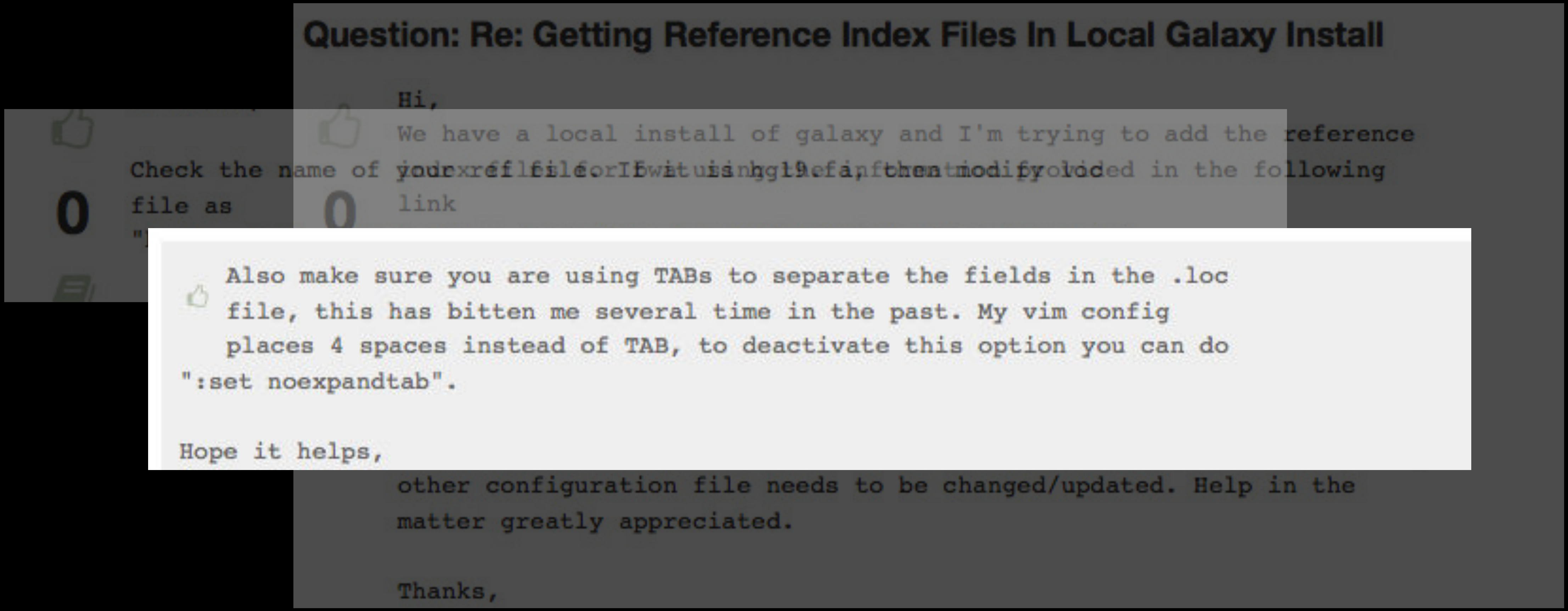
I have modified the bwa_index.loc file present in the ../tool-data directory by adding the path to where the index is on our server (Also attached). However, even after restarting the server, the reference genome does not show when choosing the "use a built-in index option". I'm not sure whether the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

Thanks,

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"
```

Did you use TABs?



Question: Re: Getting Reference Index Files In Local Galaxy Install

Hi,

We have a local install of galaxy and I'm trying to add the reference

in the following link

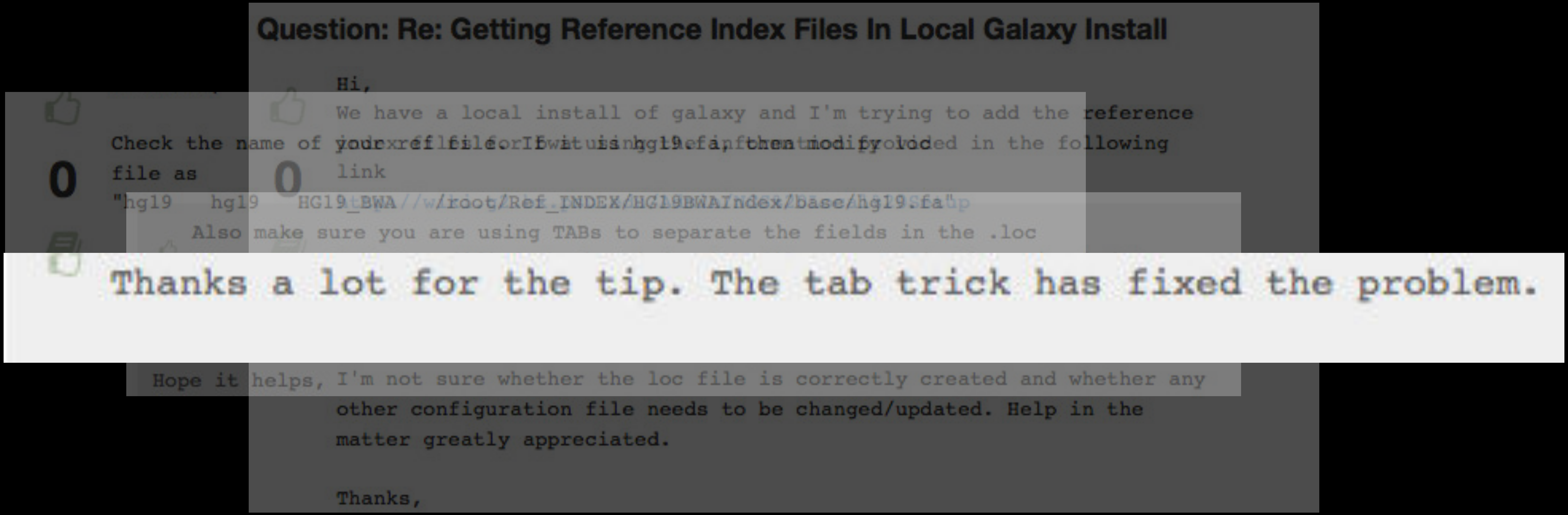
other configuration file needs to be changed/updated. Help in the matter greatly appreciated.

Thanks,

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

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,



Question: Re: Getting Reference Index Files In Local Galaxy Install

Hi,

We have a local install of galaxy and I'm trying to add the reference

link

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,

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"

Also make sure you are using TABs to separate the fields in the .loc

Hope it helps,

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

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? (No — There is a reload button)

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:

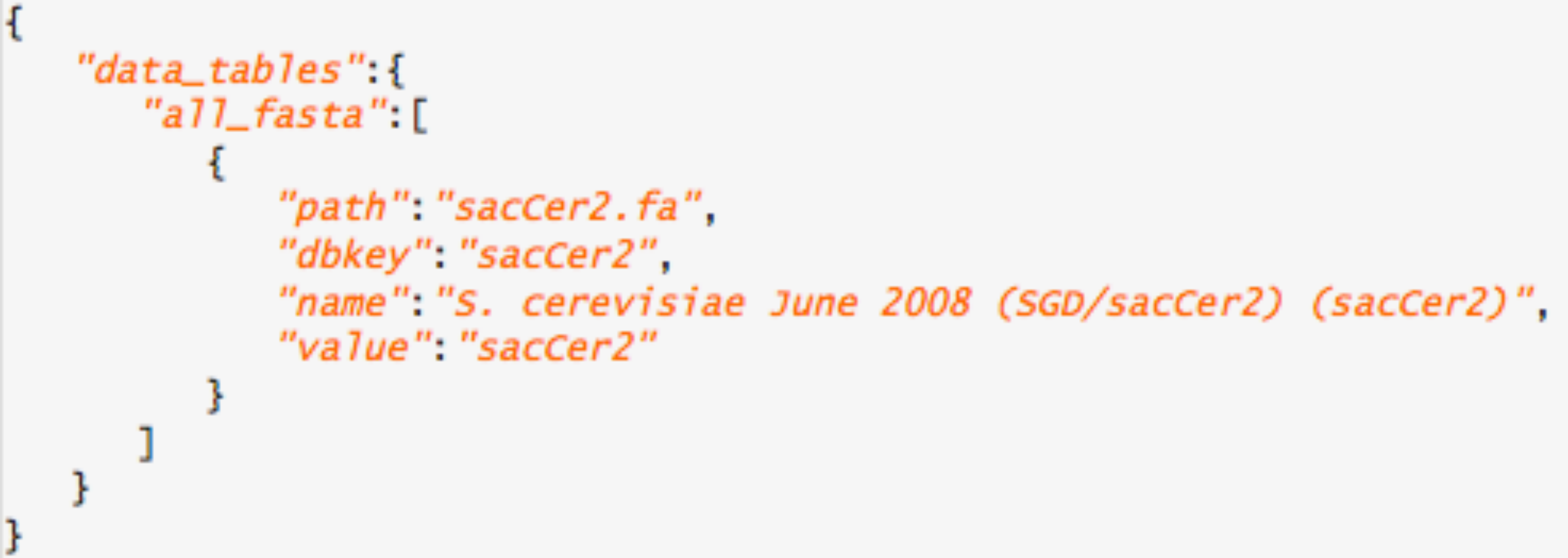
- Adding New genome builds (dbkeys)
- Fetching Genome (FASTA) sequences
- Building short read mapper indexes for genomes

Special class of Galaxy tool

```
<tool id="data_manager_fetch_genome_all_fasta" name="Reference Genome" version="0.0.1" tool_type="manage_data">
```

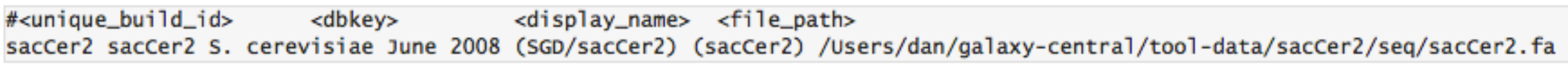
```
<outputs>
    <data name="out_file" format="data_manager_json"/>
</outputs>
```

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:



```
#<unique_build_id>      <dbkey>         <display_name>   <file_path>
sacCer2 sacCer2 S. cerevisiae June 2008 (SGD/sacCer2) (sacCer2) /Users/dan/galaxy-central/tool-data/sacCer2/seq/sacCer2.fa
```

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_manager_conf.xml

```
<data_manager tool_file="data_manager/bwa_index_builder.xml" id="bwa_index_builder" version="0.0.1">
    <data_table name="bwa_indexes">
        <output>
            <column name="value" />
            <column name="dbkey" />
            <column name="name" />
            <column name="path" output_ref="out_file" >
                <move type="directory" relativize_symlinks="True">

                    <target base="${GALAXY_DATA_MANAGER_DATA_PATH}">${dbkey}/bwa_index/${value}</target>
                </move>
                <value_translation>${GALAXY_DATA_MANAGER_DATA_PATH}/${dbkey}/bwa_index/${value}/${path}</value_translation
                <value_translation type="function">abspath</value_translation>
            </column>
        </output>
    </data_table>
</data_manager>
```

informs Galaxy about

- which data tables to expect for new entries
- special handling of provided JSON values and files

Data Managers: Configuration

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

Three Short Demos

Use admin interface to install genome and bwa indexes for sacCer2.

Use Rsync Data Manager tool to grab a copy of Data in use at usegalaxy.org

Fetch genome and build multiple indexes using a single command.

http://gcc2015.dblankenberg.org/

Data Manager Demo 1

- Fetch the Genome Sequence for sacCer2
  - UCSC as the source
  - Fetching tool installed from ToolShed
  - all_fasta table is populated automatically
- Build BWA indexes for sacCer2
  - Indexing tool installed from ToolShed
  - Build indexes
  - bwa_index table is populated automatically
- Align some reads to the newly added reference genome

Demo 2

The Reference Data that I want is available at usegalaxy.org can’t I just copy that?

Demo 2

The Reference Data that I want is available at usegalaxy.org can't I just copy that?

Yes, we have an Rsync server available. You can download indexes using the rsync command and update your .loc files.

Demo 2

The Reference Data that I want is available at usegalaxy.org can't I just copy that?

Yes, we have an Rsync server available. You can download indexes using the rsync command and update your .loc files.

What is an rsync?

Demo 2

The Reference Data that I want is available at usegalaxy.org can’t I just copy that?

Yes, we have an Rsync server available. You can download indexes using the rsync command and update your .loc files.

What is an rsync?

There is a tool for that.

Demo 3

I need a bunch of reference for each tool that are not available at usegalaxy.org, thats a lot of clicking.

Demo 3

One command to fetch and build them all.

```
run_data_managers.py -h
Usage: run_data_managers.py [options]

Options:
  -h, --help            show this help message and exit
  -s GALAXY_SERVER, --galaxy_server=GALAXY_SERVER
                        Admin user's API key
  -k API_KEY, --api_key=API_KEY
                        Admin user's API key
  -d DBKEY, --dbkey=DBKEY
                        DBkey to use for reference Data
  -n DBKEY_NAME, --name=DBKEY_NAME
                        Display Name to use for reference Data
  -u UCSC_DBKEY, --ucsc_dbkey=UCSC_DBKEY
                        UCSC DBkey to use for reference Data
retrieval
  --do_not_remove_workflow
                        Keep workflow created, after execution
```

Two Example Commands

```
python scripts/data_managers/run_data_managers.py -s http://gcc2015.dblankenberg.org -k 460c850ab4395a6262b0ea46a47d0d5f -d funYeast -n "New fun yeast" -u sacCer2
```

```
python scripts/data_managers/run_data_managers.py -s http://gcc2015.dblankenberg.org -k 460c850ab4395a6262b0ea46a47d0d5f -d hg17 -u hg17 --do_not_remove_workflow
```

Data Manager Demo: Full Disclosure

Fresh instance in cloud from launch.usegalaxy.org

- Updated to latest galaxy dev
- Setup Galaxy admin account already
- Configured tool_dependency_dir
- The sequencing reads are a small subset from SRR507778, originally downloaded from EBI SRA.

Make Your Own

Documentation

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

Several examples available in the ToolShed (Look in "Data Manager" section)