

Building More Powerful Galaxy Workflows with Dataset Collections

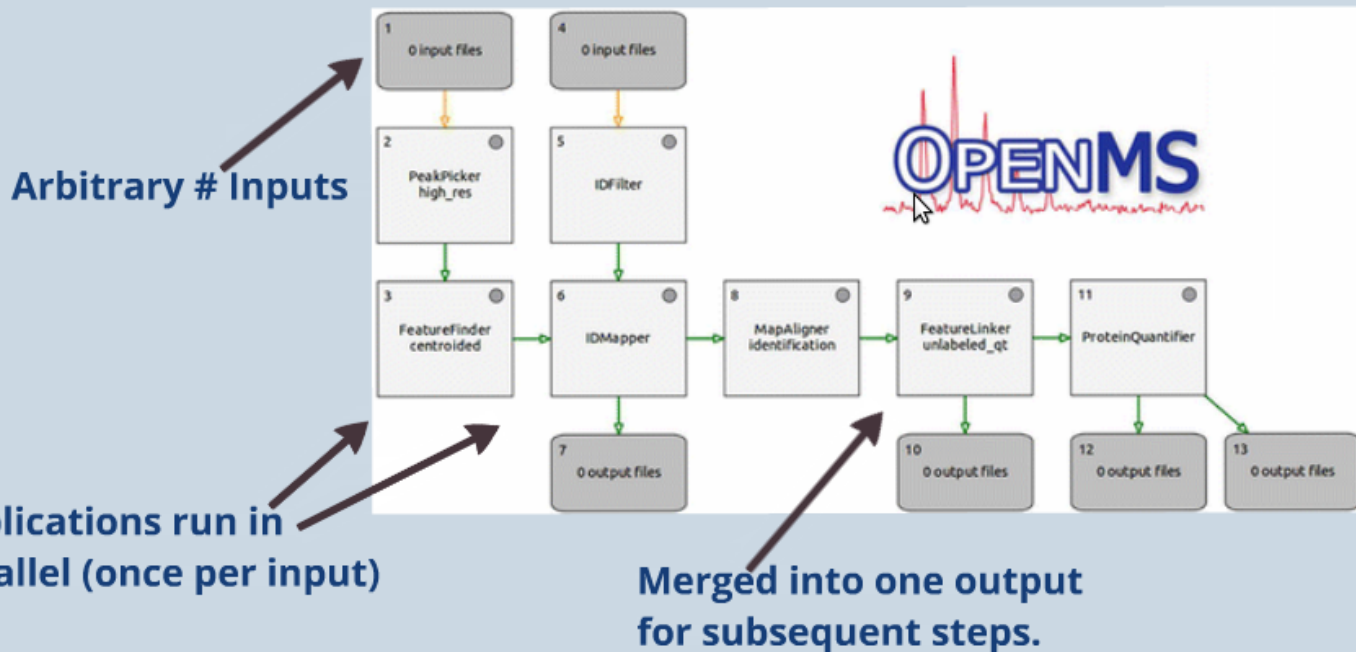
Progress and Plans

John Chilton and the Galaxy Team

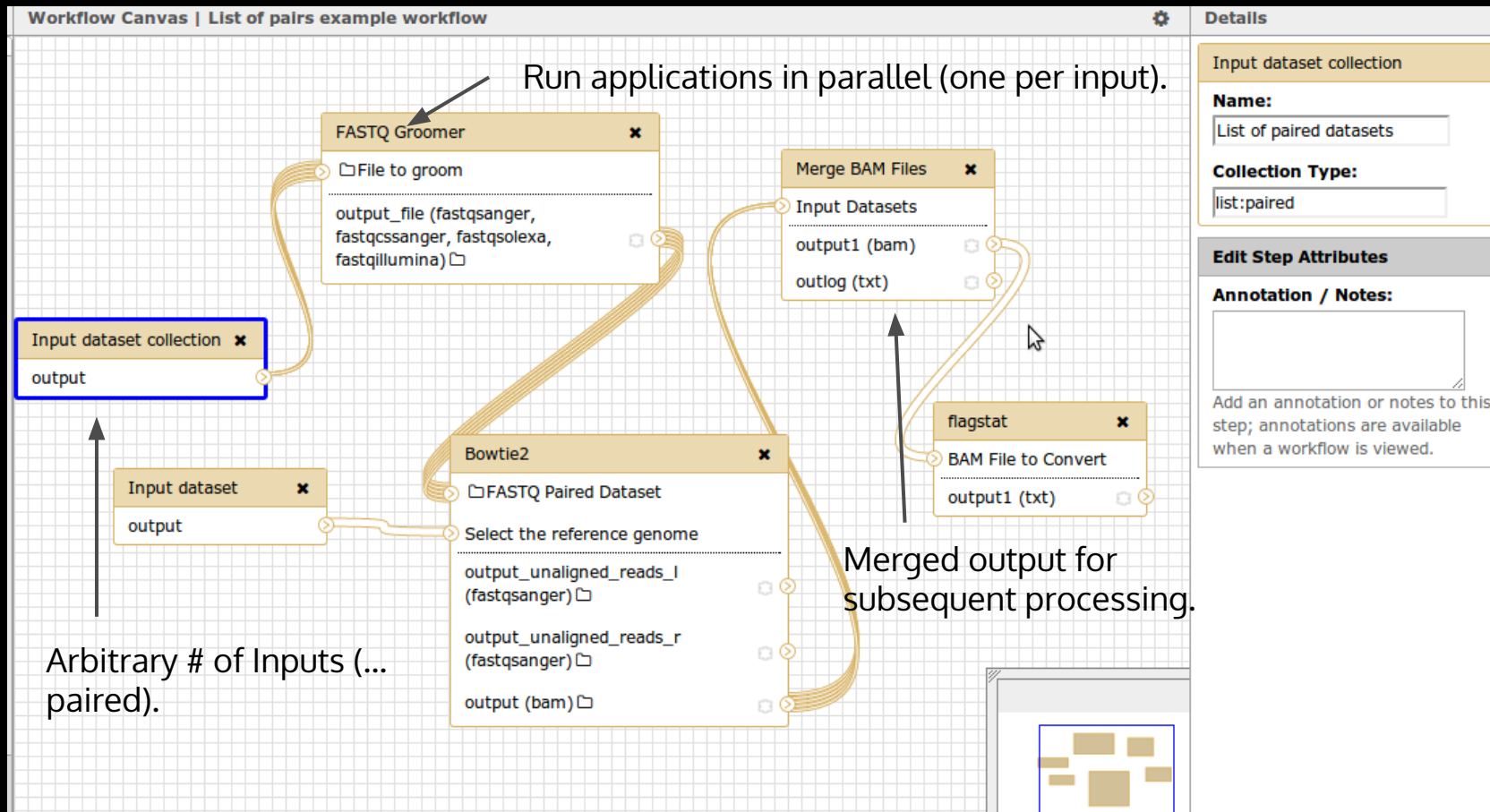
<http://bit.ly/gcc2014workflows>

John @ GCC 2012, 2013 - Workflows... not good enough!

"An Automated Pipeline for High-Throughput Label-Free Quantitative Proteomics
(J. Proteome Res., 2013, PMID: 23391308)."



More Powerful Workflows



API First Development

Initial work focused on building an API for creating and *using* dataset collections.

Upshot - API is *richer* than UI currently (especially in stable).

bioblend contains high-level functionality for creating and “viewing” collections in different ways.

Collection Types











































Currently two supported type pseudo-plugins - "list" and "paired".

- Lists can contain arbitrary number of named elements
- Pairs contain a "forward" and "reverse" element.

Types can be combined to build nested types - for instance "list:paired" describes a list of paired datasets.

Upload Some Data...

Download data directly from web or upload files from your disk

Name	Size	Type	Genome	Settings	Status
 M236C4-ch_1.fq	45.4 MB	fastqsanger 	 unspecified (?) 		100% 
 M236C4-ch_2.fq	45.4 MB	fastqsanger 	 unspecified (?) 		100% 
 M486C2-ch_1.fq	46.9 MB	fastqsanger 	 unspecified (?) 		100% 
 M486C2-ch_2.fq	46.9 MB	fastqsanger 	 unspecified (?) 		100% 
 SC14-ch_1.fq	74.4 MB	fastqsanger 	 unspecified (?) 		100% 
 SC14-ch_2.fq	74.4 MB	fastqsanger 	 unspecified (?) 		100% 
 sequence.fasta	16.9 KB	fasta 	 unspecified (?) 		100% 

You can Drag & Drop files into this box.

Choose local file

Paste/Fetch data

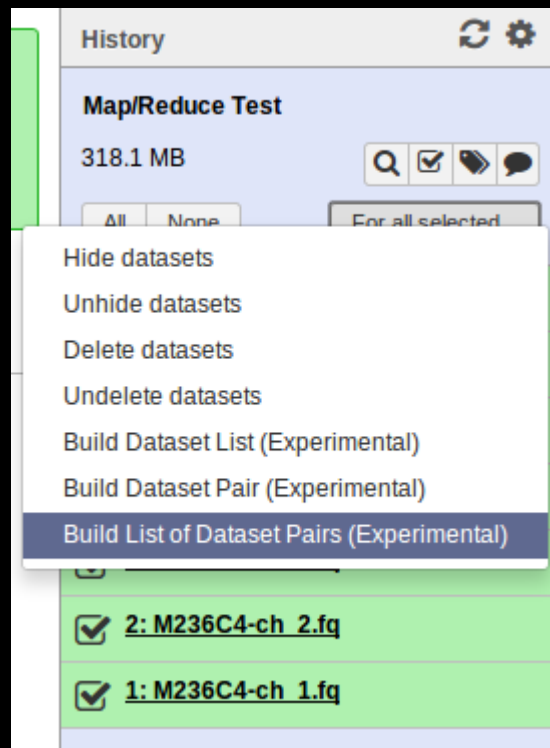
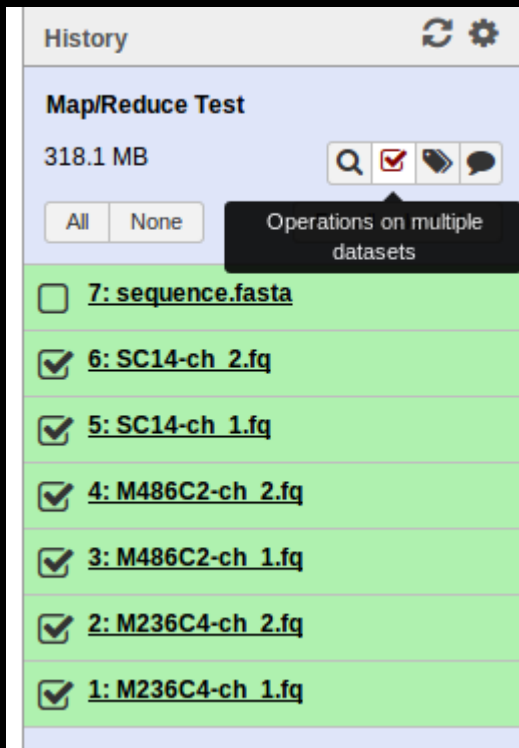
Start

Pause

Reset

Close

Select the Pairs



Create a Collection...

[Analyze Data](#) [Workflow](#) [Shared Data](#) [Visualization](#) [Help](#) [User](#)

Create a collection of paired datasets

Collections of paired datasets are ordered lists of dataset pairs (often forward and reverse reads) that can be passed to tools and workflows in order to have analyses done on the entire group. This interface allows yo... [More help](#)

0 unpaired forward - (0 filtered out)

[Choose filters](#) [Clear filters](#)

0 unpaired reverse - (0 filtered out)

3 paired [Unpair all](#)




M236C4-ch_1.fq →	M236C4	← M236C4-ch_2.fq	
M486C2-ch_1.fq →	M486C2	← M486C2-ch_2.fq	
SC14-ch_1.fq →	SC14	← SC14-ch_2.fq	

Name:

[Cancel](#) [Create list](#)

Collection Mapping (1 / 3)

FASTQ Groomer (version 1.0.4)

File to groom:   

6: SC14-ch_2.fq

Input FASTQ quality scores type:

Sanger & Illumina 1.8+

Advanced Options:

Hide Advanced Options

Execute

Tool consumes a FASTQ file.

-List of Paired Datasets

-Individual FASTQ datasets.





What it does

This tool offers several conversions options relating to the FASTQ format.

When using *Basic* options, the output will be *sanger* formatted or *cssanger* formatted (when the input is Color Space Sanger).

When converting, if a quality score falls outside of the target score range, it will be coerced to the closest available value (i.e. the minimum or maximum).

History

Map/Reduce Test
318.1 MB    

8: Paired mt Datasets

7: sequence.fasta

6: SC14-ch_2.fq

5: SC14-ch_1.fq

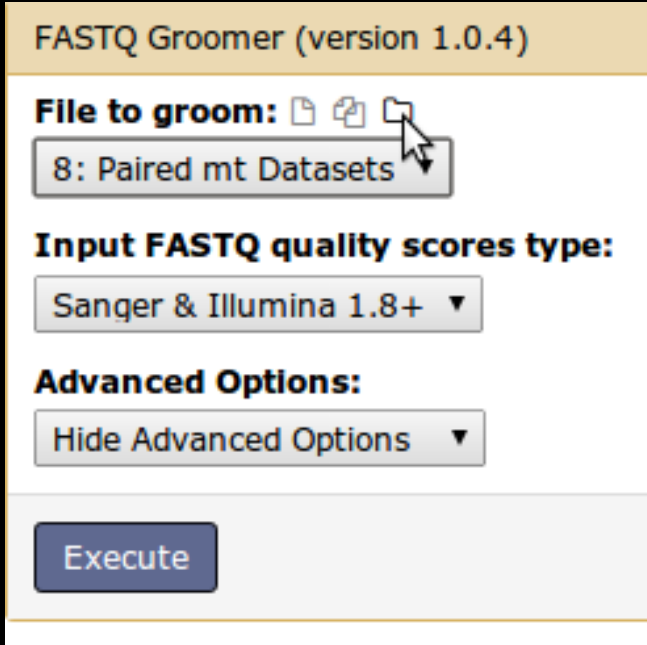
4: M486C2-ch_2.fq

3: M486C2-ch_1.fq





2: M236C4-ch_2.fq

1: M236C4-ch_1.fq

Collection Mapping (2 / 3)



FASTQ Groomer (version 1.0.4)

File to groom:    

8: Paired mt Datasets

Input FASTQ quality scores type:

Sanger & Illumina 1.8+ ▼

Advanced Options:

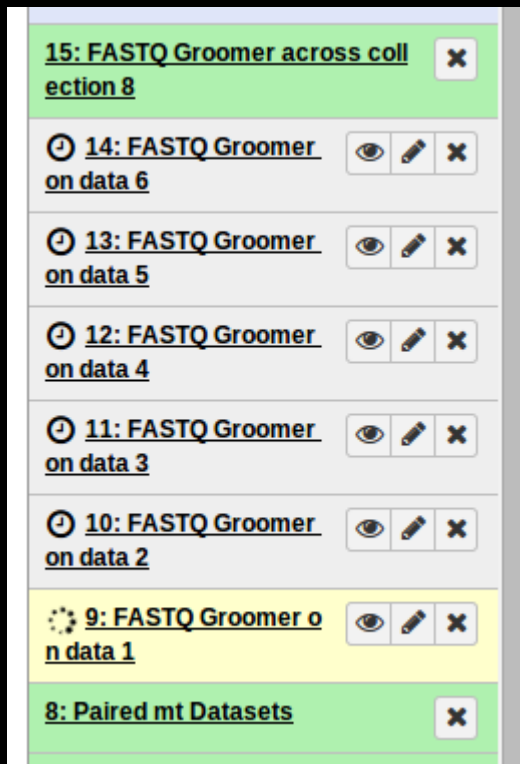
Hide Advanced Options ▼

Execute

Collection map icon replaces input options with valid collections.

Runs tool over every dataset in list of pairs and produces groomed list of pairs.

Collection Mapping (3 / 3)



15: FASTQ Groomer across collection 8

14: FASTQ Groomer on data 6

13: FASTQ Groomer on data 5

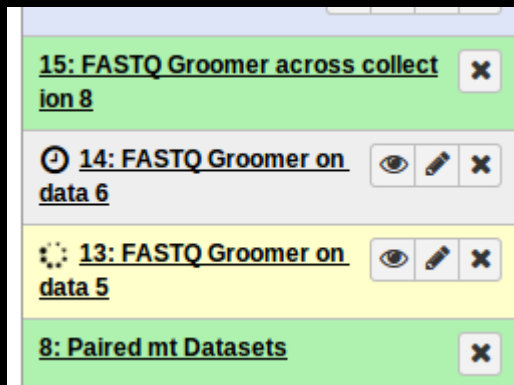
12: FASTQ Groomer on data 4

11: FASTQ Groomer on data 3

10: FASTQ Groomer on data 2

9: FASTQ Groomer on data 1

8: Paired mt Datasets

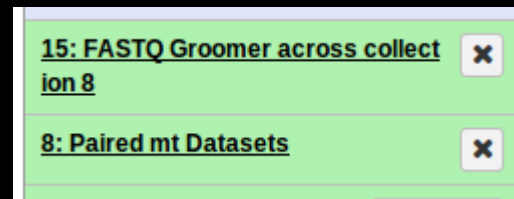


15: FASTQ Groomer across collection 8

14: FASTQ Groomer on data 6

13: FASTQ Groomer on data 5

8: Paired mt Datasets



15: FASTQ Groomer across collection 8

8: Paired mt Datasets

Like hiding workflow datasets, they are visible initially and **hidden after completion** (only collection remains visible).

Collection always green regardless of contents (**stateless**).

Need to do better on both points... not scalable enough.

Sample Tracking: Identifiers + Indices

Paired mt Datasets

list:paired collection

Element - 0:M236C4 (paired collection)

Element - 0:forward

hda - M236C4-ch_1.fq

Element - 1:reverse

hda - M236C4-ch_2.fq

Element - 1:M486C2 (paired collection)

Element - 0:forward (hda)

hda - M486C2-ch_1.fq

Element - 1:reverse (hda)

hda - M486C2-ch_2.fq

...

FASTQ Groomer across collection 8

list:paired collection

Element - 0:M236C4 (paired collection)

Element - 0:forward

hda - *FASTQ Groomer on data 1*

Element - 1:reverse

hda - *FASTQ Groomer on data 2*

Element - 1:M486C2 (paired collection)

Element - 0:forward (hda)

hda - *FASTQ Groomer on data 3*

Element - 1:reverse (hda)

hda - *FASTQ Groomer on data 4*

...

Mapping over collections -
dataset naming is normal,
but new collection created
with identical tree structure
and element identifiers
preserved.

Subcollection Mapping

Bowtie2 (version 0.2)

Is this library mate-paired?:

Paired-end Dataset

FASTQ Paired Dataset:

Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Minimum insert size for valid paired-end alignments:

0

Maximum insert size for valid paired-end alignments:

250

Write unaligned reads to separate file(s):

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

Use one from the history

Built-ins were indexed using default options

Select the reference genome:

7: sequence.fasta

History

Map/Reduce Test

636.1 MB

15: FASTQ Groomer across collection 8

8: Paired mt Datasets

7: sequence.fasta

6: SC14-ch 2.fq

5: SC14-ch 1.fq

4: M486C2-ch 2.fq

3: M486C2-ch 1.fq

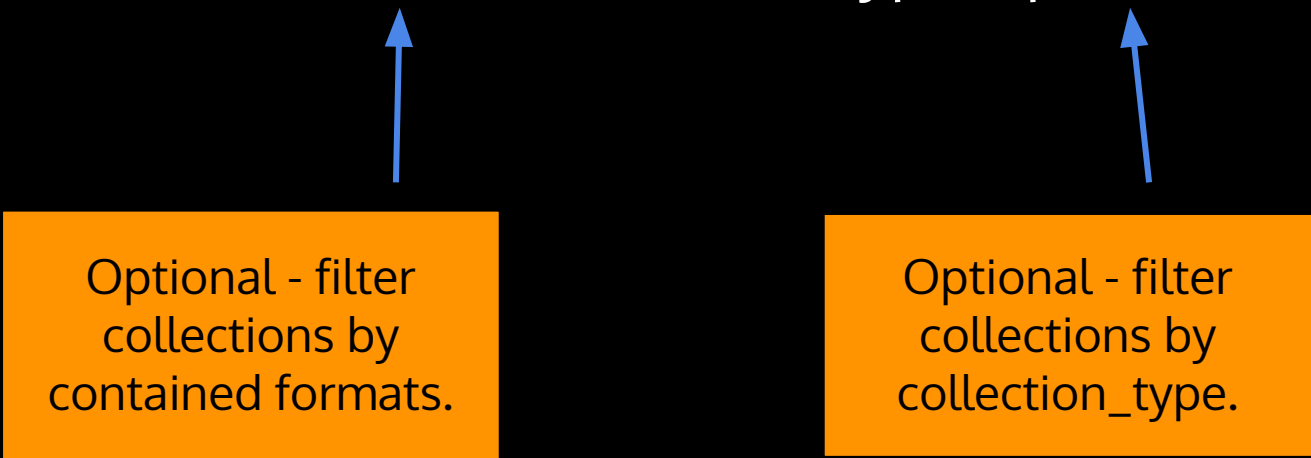
2: M236C4-ch 2.fq

1: M236C4-ch 1.fq

Bowtie2 wrapper modified with option to take in a paired dataset instead of two separate datasets.

Tool Parameters - Tool XML

```
<param name="collect_param1" type="data_collection"  
      format="bam" collection_type="paired" />
```



Optional - filter
collections by
contained formats.

Optional - filter
collections by
collection_type.

Tool Parameters - Cheetah-isms

Common paired data idiom:

```
bowtie $collect_param.forward $collect_param.reverse
```

Common list data idiom:

```
#for $f in $collect_param# $f #end for#
```

-or-

```
#for $name in $collect_param.keys()# $f[$name] #end for#
```

Nested data:

```
#for $f in $collect_param# $f.is_collection ...
```

Tool Parameters - Testing

```
<test>
```

```
  <param name="collect_param">
```

```
    <collection type="paired">
```

```
      <element name="forward" value="simple_line.txt" />
```

```
      <element name="reverse" value="simple_line_alternative.txt" />
```

```
    </collection>
```

```
  </param>
```



```
...
```


Subcollection Mapping

Bowtie2 (version 0.2)

Is this library mate-paired?:

Paired-end Dataset

FASTQ Paired Dataset:  

15: FASTQ Groomer across collection 8

Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Minimum insert size for valid paired-end alignments:

0

Maximum insert size for valid paired-end alignments:

250

Write unaligned reads to separate file(s):

History

Map/Reduce Test
636.1 MB

15: FASTQ Groomer across collection 8

8: Paired mt Datasets

7: sequence.fasta

6: SC14-ch 2.fq

5: SC14-ch 1.fq

4: M486C2-ch 2.fq

Map/Reduce Test
636.1 MB

19: Bowtie2 across collection 15

18: Bowtie2 on data 7, data 9, and others: aligned reads

17: Bowtie2 on data 7, data 9, and others: aligned reads

16: Bowtie2 on data 7, data 9, and others: aligned reads

15: FASTQ Groomer across collection 8

Subcollection Mapping (Identifiers)

Paired mt Datasets

list:paired collection

Element - 0:M236C4 (paired collection)

Element - 0:forward

hda - M236C4-ch_1.fq

Element - 1:reverse

hda - M236C4-ch_2.fq

Element - 1:M486C2 (paired collection)

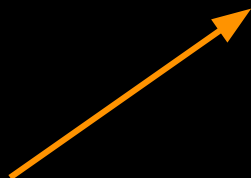
Element - 0:forward (hda)

hda - M486C2-ch_1.fq

Element - 1:reverse (hda)

hda - M486C2-ch_2.fq

...



Bowtie 2 across collection 13

list collection

Element - 0:M236C4

hda - *Bowtie 2 on data 9 and data 10*

Element - 1:M486C2

hda - *Bowtie 2 on data 11 and data 12*

...



Reducing Collections

Merge BAM Files (version 1.1.2)



Name for the output merged bam file:


This name will appear in your history so use it to re

Merge all component bam file headers into th

☒

Control the MERGE_SEQUENCE_DICTIONARIES fla
important metadata

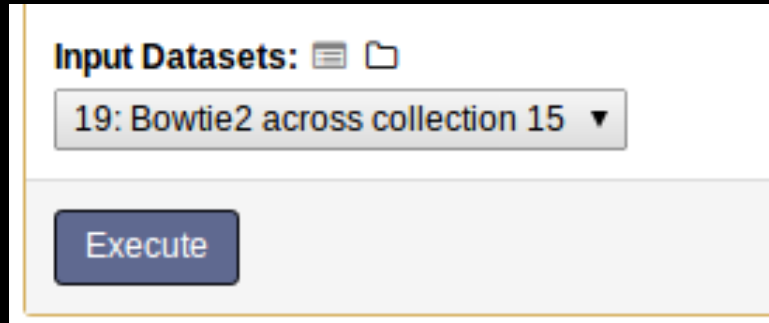
Input Datasets:  



Execute

Modified “*Merge BAM Files*” tool to use **multiple input data** parameter instead of two input parameters and a repeat block.

Reducing Collections



Can dynamically **substitute** collection for the multiple selection of datasets.

Handful of Reduction Tools...

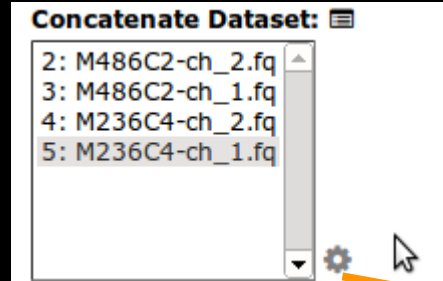
A handful of **reduction tools need to be updated** (so will **tools consuming pairs**). Using multiple input data parameters instead of repeat parameters will still allow these tools to work with uncollected dataset.

repeat blocks - while cumbersome - allow duplicated entries & control of order. Multiple input data parameters should be enhanced to have same control.

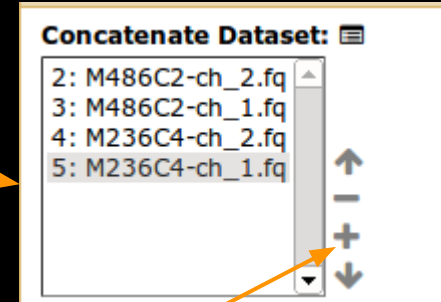
Plan: Multiple-Data Improvements

Enhance multiple input data parameters to **allow control of order and repeated entries**.

All the ease of multiple data inputs with actually greater versatility than placing simple data inputs into repeat blocks.



Mock Up



An advanced "add to selection" modal would provide interesting room to grow - options for importing library datasets, digging into collections, etc....

Extract a Workflow

The interface displays a workflow extraction process. On the left, a list of steps is shown, each with a checkbox to include it in the workflow. The central panel lists the extracted steps, each with a checkbox to treat it as an input dataset. The right-hand panel provides details for each step, including its name and a set of icons (eye, pencil, and X) for editing or deleting the step.

Dataset Collection Creation
Dataset collection created in a way not compatible with workflows

FASTQ Groomer
☒ Include "FASTQ Groomer" in workflow

Bowtie2
☒ Include "Bowtie2" in workflow

Merge BAM Files
☒ Include "Merge BAM Files" in workflow

flagstat
☒ Include "flagstat" in workflow

7: M486C2
☐ Treat as input dataset

8: Paired mt Datasets
☒ Treat as input dataset

13: FASTQ Groomer across collection 8

16: Bowtie2 across collection 13

17: NewBam.bam

18: NewBam_Merge BAM Files.log

19: flagstat on data 17

19: flagstat on data 17

18: NewBam_Merge BAM Files.log

17: NewBam.bam

16: Bowtie2 across collection 13

13: FASTQ Groomer across collection 8

8: Paired mt Datasets

7: M486C2

6: M236C4

5: sequence.fasta

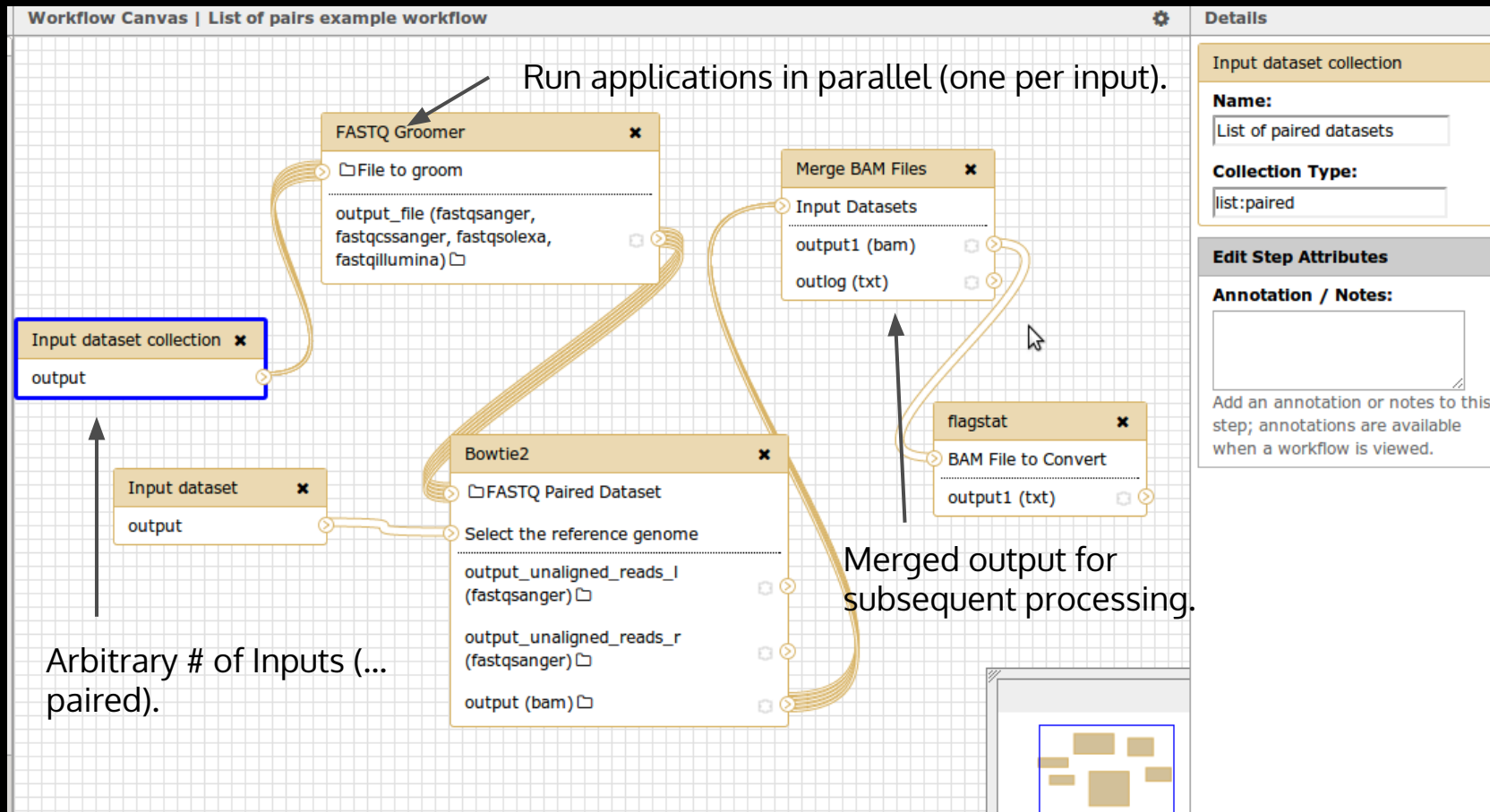
4: M486C2-ch 2.fg

3: M486C2-ch 1.fg

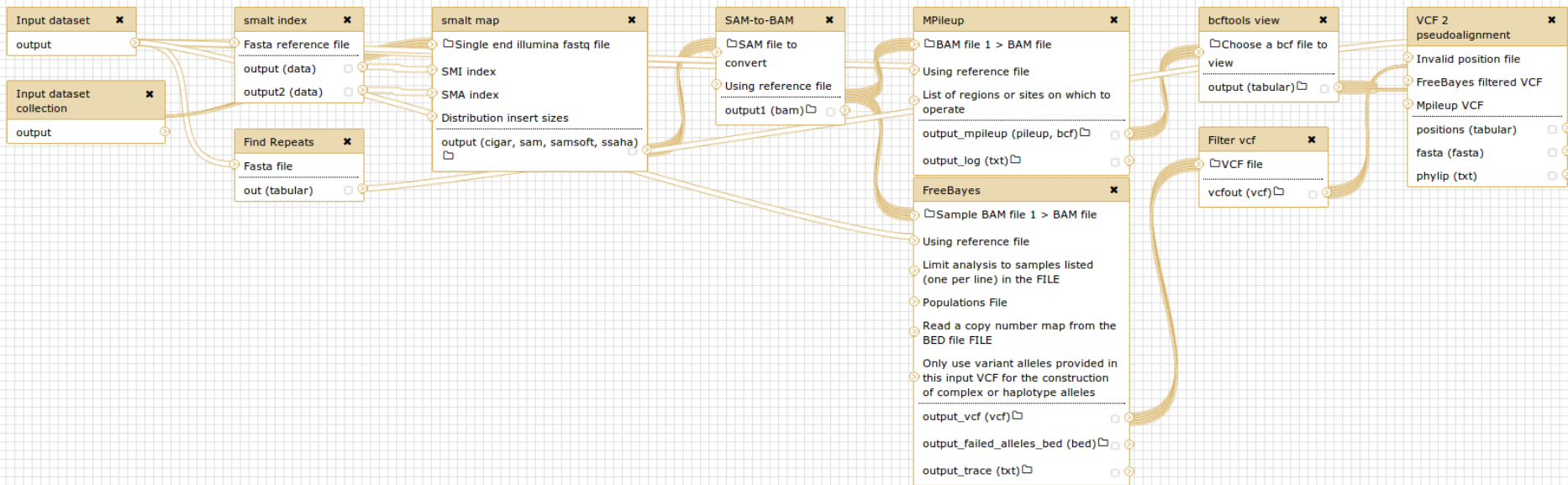
2: M236C4-ch 2.fg

1: M236C4-ch 1.fg

More Powerful Workflows



More workflows...



Core phylogenomics SNP pipeline by Aaron Petkau, Gary Van Domselaar, Philip Mabon, and Lee Katz.
Worked 208 single end reads producing 1469 datasets
Galaxy took 10 minutes to schedule workflow.

Plans - Improvements to Builder

Iteration 2 - <https://trello.com/c/8hEO00xj>

Regex filters, more assistance, allow reordering

Iteration 3 - <https://trello.com/c/LLk9ICvM>

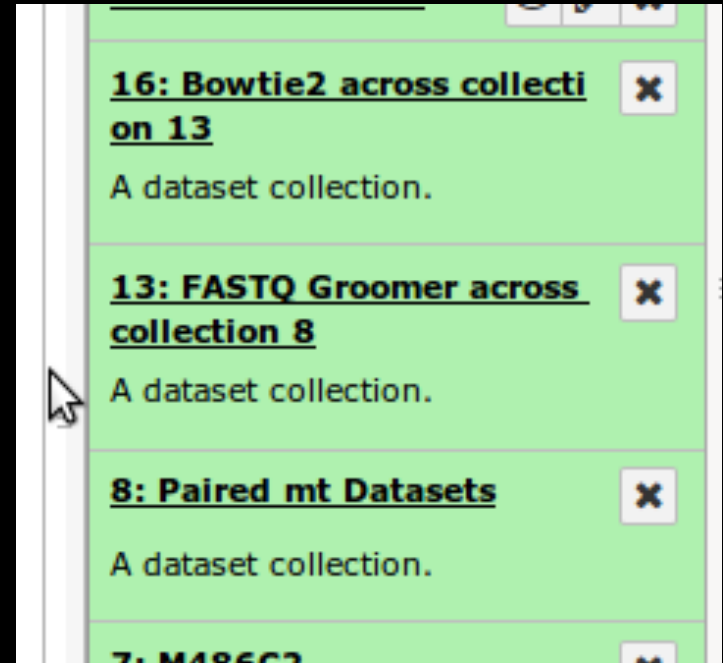
Batch renaming, dataset info on click, hide original datasets.

Plans - More Options in History Panel

<https://trello.com/c/hnmVWKlB>

Currently can hide, delete, and see name.

Cannot rename, rerun, see type, see contents, see/add annotations, see/add tags, download, etc...



Plans - UI for Uploading Collections

<https://trello.com/c/ZAXwWOZ2>

Incorporate collection builder when uploading files (or vise versa).

Plans - UI for Viewing Collections

<https://trello.com/c/PVdbbpQS>

Plans - Store Collections in Data Libraries

<https://trello.com/c/3axmjaxE>

Plans - Improved Reductions

<https://trello.com/c/lp5YmA1O>

Improvements to multiple data parameters described earlier and/or ability to reduce across repeat statements.

Plans - Filtering Collections

<https://trello.com/c/ryKJrsYc>

Main Goal: Filter out the **failed** datasets and keep going.

Would like more **general filters** - filter on metadata (*file size, number of sequences, etc...*)

Needs to be trackable so can extract and execute in workflows. May require delayed workflow evaluation.

Plans - Output Collections

<https://trello.com/c/KXjp6lIn>

Use Cases:

- $1 \rightarrow N$ (metagenomics, splitting)
- $N \rightarrow N$ (normalization across files)

Progress on tool running was made at hackathon (thanks JJ and Carrie) - workflows will be challenging (ever more bookkeeping for editor).

Plans - Rerun Tools / Resuming Workflows

<https://trello.com/c/lxVJy7fs>

Plans - Update and Add New Tools

<https://trello.com/c/lxVJy7fs>

- Paired-end mappers (bowtie, etc...)
- Concatenate Datasets
- Merge Bam
- Many sorts of interesting tabular operations to merge datasets (also using element identifiers).
- etc...

Toward 10,000 samples (beyond collections)

- Optimize database interactions, tool execution.
- Move workflow scheduling into own process, optimize.
- Differentiate between cluster failures and tool failures.
 - Retry later on cluster failures.
 - Retry on different cluster or with different resource params on failures.
- Optimize disk usage - streaming
- More diverse and bigger compute and storage
 - Separate metadata calculation out into its own "job"
 - XSEDE
 - More portable dependency management (docker, nix, tool shed installs without galaxy)

Docker... Docker... Docker...



docker

https://github.com/jmchilton/gcc2014_demo

The **Galaxy Team**

Thanks!



Enis Afgan

Dannon Baker

Dan Blankenberg

Dave Bouvier

Marten Cech

John Chilton



Dave Clements

Nate Coraor

Carl Eberhard

Dorine Francheteau

Jeremy Goecks

Sam Guerler



Jen Jackson

Greg von Kuster

Ross Lazarus

Anton Nekrutenko

Nick Stoler

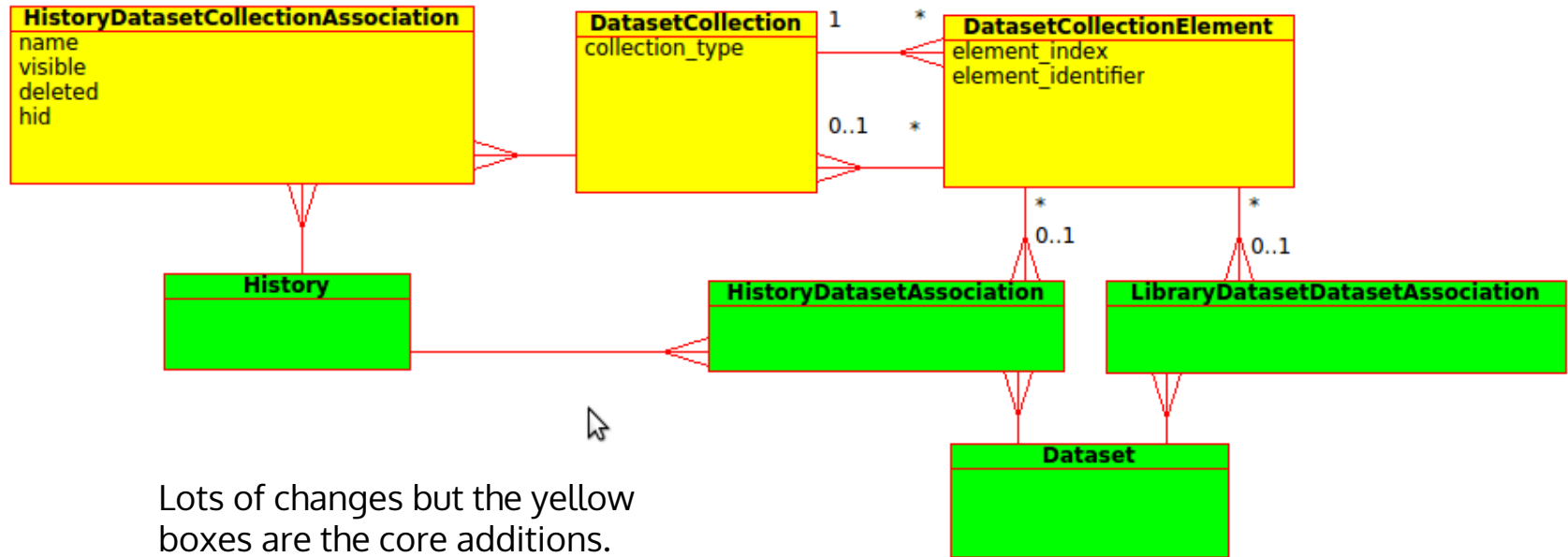
James Taylor

With special thanks to **Carl Eberhard** - for building UI powering this work, **Jeremy** and **Dannon** for scoping out initial plans, and **Nick**, **James**, **Dan**, and **Anton** for ongoing feedback.

The **Galaxy-P** grant, team, and the Minnesota Supercomputing Institute for funding development of multiple file datasets (a precursor) - with special thanks to Tim Griffin, Pratik Jagtap, Benjamin Lynch, and Anne-Françoise Lamblin.

The **Galaxy Community** for building awesome stuff with Galaxy and pushing the platform forward - especially **Philip Mabon** and **Bjoern Gruening**.

Models



Lots of changes but the yellow boxes are the core additions.

Extra Content

Plans - Other

- <https://trello.com/c/WodW2sLb>
- Subcollection mapping over multiple data parameters.
- Fix history import/export for data collections.
- Implicit conversion
- Allow batch input of collections to workflows
- HIDs of copied collections are wrong - either always copy HDAs also or reconsider naming in context of collections.

TODO:

- Screenshots of building up workflow from scratch?

Extra Slides (post presentation)...

- Comparison with multiple file datasets.

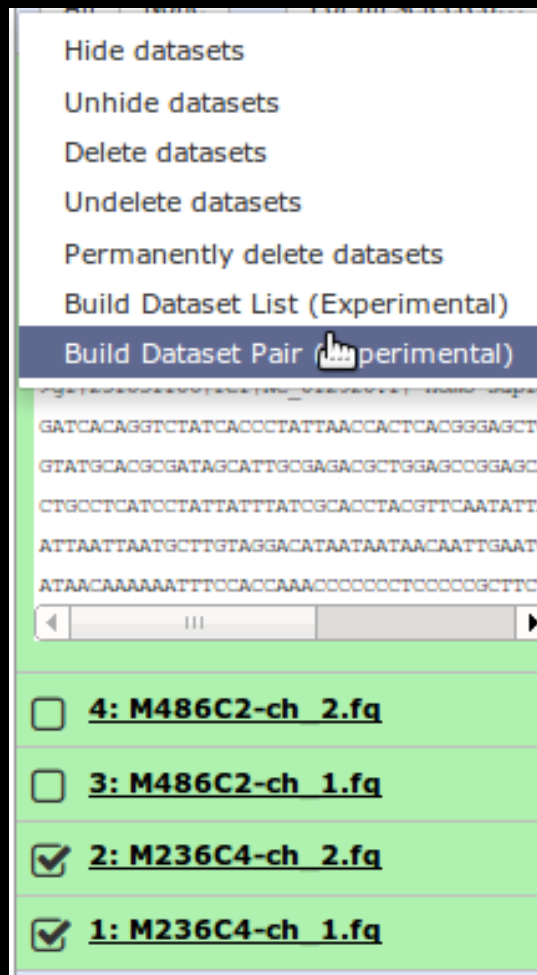
REDO Initial Screenshots with Correct History Name on Bigger Monitor.

Building Collections...

```
>>> from bioblend import galaxy
>>> gi = galaxy.GalaxyInstance(url="localhost:8080",
                               key="db53bb4500dfaeda25ceb378069b722b")
>>> hist = gi.histories.get_histories(name="Map/Reduce Test")[0]
>>> gi.histories.show_history(hist["id"], contents=True, deleted=False)
>>> pair1_id = [d for d in gi.histories.show_history(hist["id"], contents=True)
                if d["hid"] == 5][0]["id"]
>>> pair2_id = [d for d in gi.histories.show_history(hist["id"], contents=True)
                if d["hid"] == 6][0]["id"]
>>> gi.histories.update_dataset_collection(hist["id"], pair1_id, name="M236C4")
>>> gi.histories.update_dataset_collection(hist["id"], pair2_id, name="M486C2")
```

bioblend contains support for creating, reading, updating (name, annotations, etc...), and deleting history dataset collections.

<https://github.com/afgane/bioblend/commit/f8d40b687be4c699d608e930c59726793922fa0a>



Collection Mapping (1 / 3)

FASTQ Groomer (version 1.0.4)

File to groom: 4: M486C2-ch_2.fq

Input FASTQ quality scores type: Sanger & Illumina 1.8+

Advanced Options: Hide Advanced Options

Execute

What it does

This tool offers several conversions options relating to the FASTQ format.

When using *Basic* options, the output will be *sanger* formatted or *cssanger* formatted (when the input is Color Space Sanger).

When converting, if a quality score falls outside of the target

History

Map/Reduce Test

519.8 MB

8: Paired mt Datasets

7: M486C2

6: M236C4

5: sequence.fasta

4: M486C2-ch_2.fq

3: M486C2-ch_1.fq

2: M236C4-ch_2.fq

1: M236C4-ch_1.fq

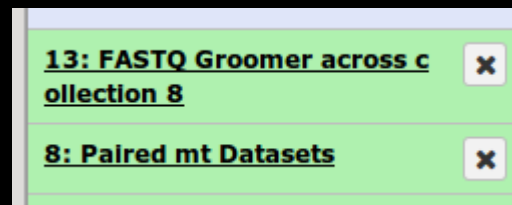
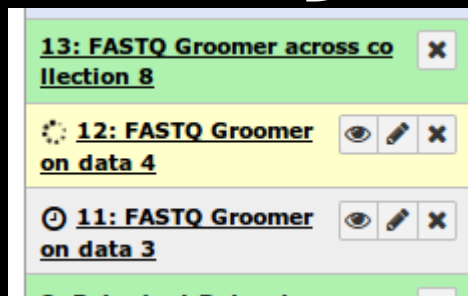
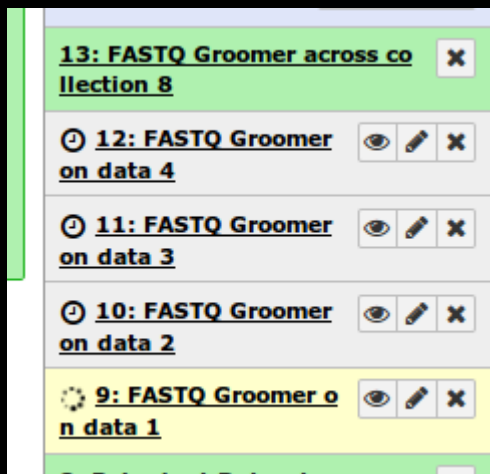
Tool consumes a FASTQ file.

-List of Paired Datasets

-Paired Datasets

-Individual FASTQ datasets.

Collection Mapping (3 / 3)



Like hiding datasets in workflow execution, datasets are visible running or queued and they are hidden after (and only collection is visible).

Collection is always green regardless of contents - is currently stateless.

Need to do a better job on both points - this is not too scalable - but it was an easy quick win.

Plans - UI for Creating Collections

<https://trello.com/c/CIIIdaxl2>

[Mockup @ mybalsamiq](#)

Create a list of paired datasets

(help text) Create a list of paired datasets by...

Forward		Reverse
Q_1	9 Pairs	Q_2
MRX3348_1.fastq	MRX3348	MRX3348_2.fastq
MRX3348_1.fastq	MRX3348	MRX3348_2.fastq
MRX3348_1.fastq	MRX3348	MRX3348_2.fastq
MRX3348_1.fastq	MRX3348	MRX3348_2.fastq
MRX3348_1.fastq	MRX3348	MRX3348_2.fastq
MRX3348_1.fastq	MRX3348	MRX3348_2.fastq
MRX3348_1.fastq	MRX3348	MRX3348_2.fastq
MRX3348_1.fastq	MRX3348	MRX3348_2.fastq
MRX3348_1.fastq	MRX3348	MRX3348_2.fastq
exp_1000.bed		data_2.fasta
yerinia_214_1.fastq		

Name of new list: My List

Cancel Create a different kind of collection Create list

The middle section is a scrollable table divided into two parts: the upper, paired section and the lower, unpaired section. Filtering only affects the unpaired section.

A: Color, background color, font, and justification can all be used to differentiate paired/unpaired.

When the user clicks on an unpaired forward then an unpaired reverse (or vice versa) a pair is created. That pair is moved to the bottom of the paired section of the table.

Each row in the 'Pairs' section of the list will have some control to unpair that pair. When clicked, the row disappears and the two files go back to the unpaired/lower section of the table in the appropriate, sorted order.

Alternately, we can send the user to a second pane (2nd 'Wizard' step) to review and re-order the final list. (An option to move back to this step should also be there)

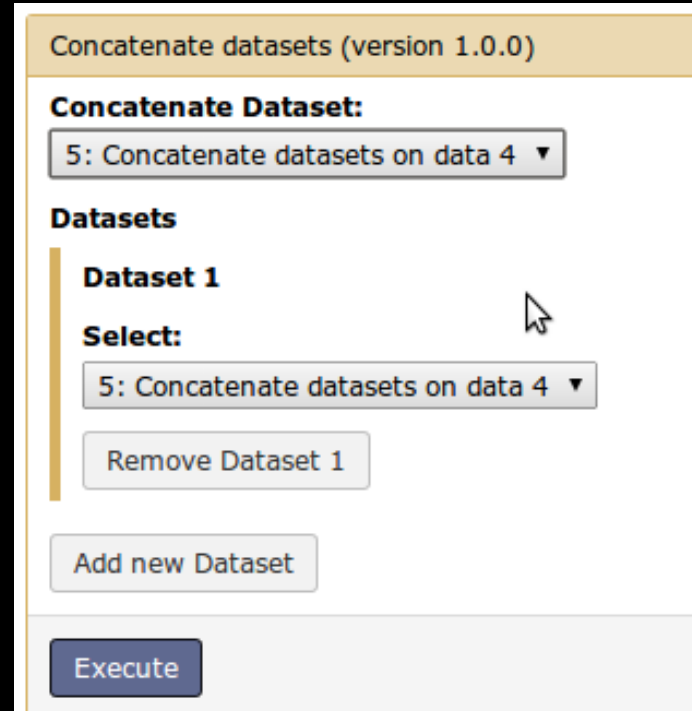
Why not repeat replacements?

In its most simple form - allowing replacement of one repeat block with a collection - this feature would be gross to implement - it would add a lot of complexity to already complex parts of Galaxy.

... and it would not work with any tools.

Concatenate (Easiest Reduction)

Not just a repeat, would need to be able to dynamically replace input + repeat to work with this. That will be ugly and will have implications all over.



The screenshot shows a web-based interface for concatenating datasets. At the top, a title bar reads "Concatenate datasets (version 1.0.0)". Below this, the section "Concatenate Dataset:" contains a dropdown menu with the text "5: Concatenate datasets on data 4". The "Datasets" section features a vertical list with "Dataset 1" highlighted. Under "Dataset 1", there is a "Select:" label and another dropdown menu with the same text "5: Concatenate datasets on data 4". To the right of this dropdown, a mouse cursor is visible. Below the dropdown is a button labeled "Remove Dataset 1". At the bottom of the dataset list is a button labeled "Add new Dataset". At the very bottom of the interface is a large blue button labeled "Execute".

Merging Bams

Second most common reduction - has two inputs and a repeat. So we need to be able to dynamically replace any number inputs and a repeat. Hmm....

Merge BAM Files (version 1.1.2)

Name for the output merged bam file:

This name will appear in your history so use it to remember what the new file is called

Merge all component bam file headers into the merged bam file:

☐

Control the MERGE_SEQUENCE_DICTIONARIES flag for Picard MergeSamFiles to preserve important metadata

First file:

with file:

Need to add more files? Use controls below.

Input Files

Add new Input Files

Execute

Merging BedGraph

Found another reduction tool on main. Multiple inputs, multiple extra options. How could this reasonably allow collection replacement at the infrastructure level.

Merge BedGraph files (version 0.1.0)

First BedGraph file:

❌ History does not include a dataset of the required format / build

Sample name:

Second BedGraph file:

❌ History does not include a dataset of the required format / build

Sample name:

Add'l BedGraph files

Add'l BedGraph files 1

BedGraph file:

Sample name: