GCC2015 Galaxy Hackathons

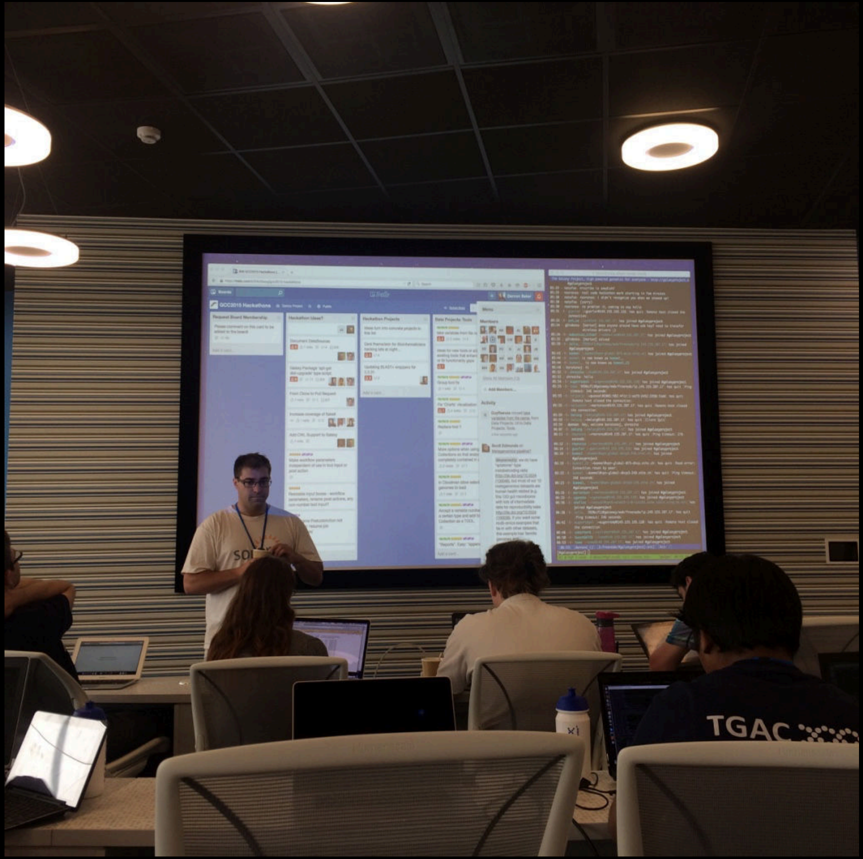
GCC Hackathon (Code and Data!) Organizers

Making Connections, Providing Footholds

- Community Building
- Bandwidth is important
- Great opportunity for ‘fun’ projects

By the Numbers

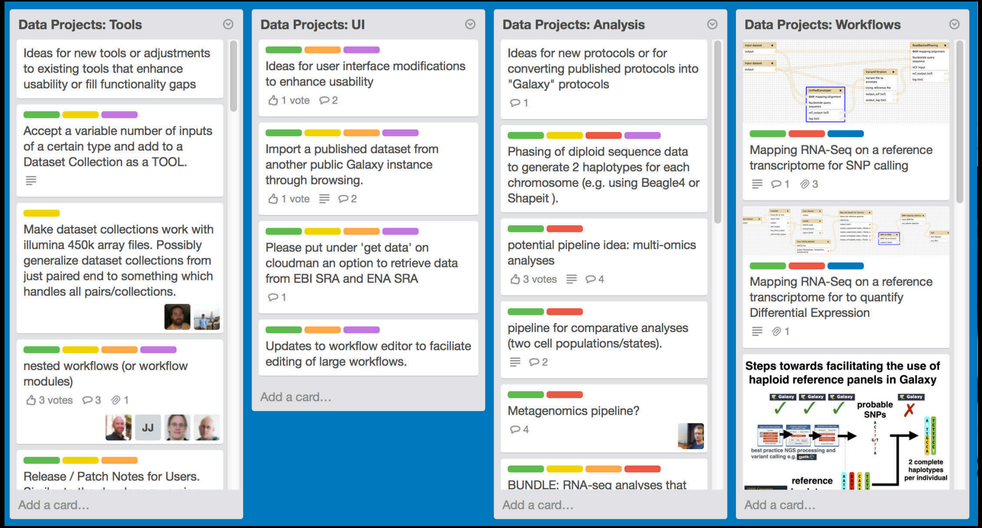
- >50 people
- 2 days
- 1 trello board
- 6 major projects
- 20 pizzas
- 163 cups of coffee



Main Themes

- Workflows
- Tools
- API
- Code Cleanup
- Smashing Eggs

bit.ly/hackboard



Data Projects: Tools

Ideas for new tools or adjustments to existing tools that enhance usability or fill functionality gaps

Accept a variable number of inputs of a certain type and add to a Dataset Collection as a TOOL.

Make dataset collections work with illumina 450k array files. Possibly generalize dataset collections from just paired end to something which handles all pairs/collections.

nested workflows (or workflow modules)
3 votes 3 1

Release / Patch Notes for Users.

Add a card...

Data Projects: UI

Ideas for user interface modifications to enhance usability
1 vote 2

Import a published dataset from another public Galaxy instance through browsing.
1 vote 2

Please put under 'get data' on cloudman an option to retrieve data from EBI SRA and ENA SRA
1

Updates to workflow editor to faciliate editing of large workflows.

Add a card...

Data Projects: Analysis

Ideas for new protocols or for converting published protocols into "Galaxy" protocols
1

Phasing of diploid sequence data to generate 2 haplotypes for each chromosome (e.g. using Beagle4 or Shapeit).

potential pipeline idea: multi-omics analyses
3 votes 4

pipeline for comparative analyses (two cell populations/states).
2

Metagenomics pipeline?
4

BUNDLE: RNA-seq analyses that

Add a card...

Data Projects: Workflows

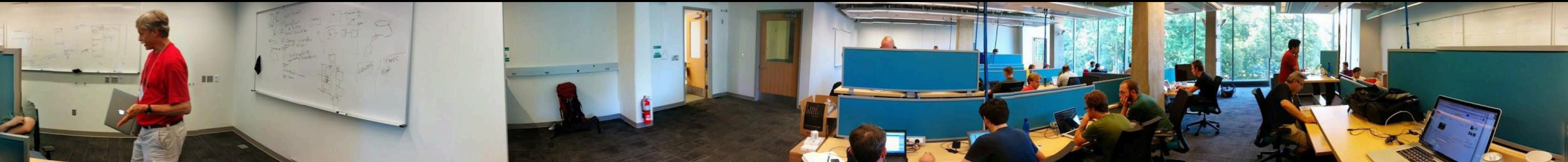
Mapping RNA-Seq on a reference transcriptome for SNP calling
1 3

Mapping RNA-Seq on a reference transcriptome for to quantify Differential Expression
1

Steps towards facilitating the use of haploid reference panels in Galaxy

Add a card...

JJ
Galaxy
Galaxy
Galaxy
probable SNPs
Galaxy
best practice NGS processing and variant calling e.g. gatk
G/T
2 complete haplotypes per individual
reference



Nested Workflows

- Goal: Enable nesting of workflows inside of other workflows.
- Steps
  a. construct tests of workflows as we want them to appear (both from the API and as unit tests)
  b. refactor tool logic to a new AbstractExecutable object and reimplement tool or workflow-specific logic

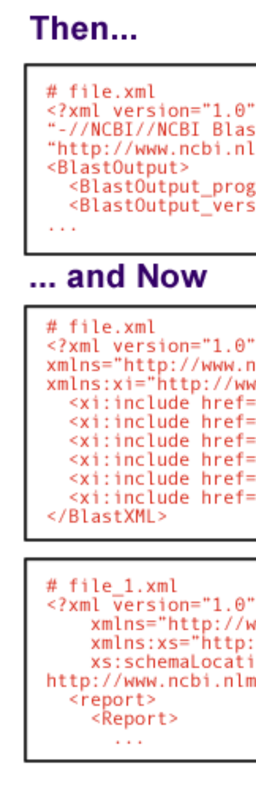
Work is happening at: https://github.com/nebiolabs/galaxy

(Peter van Huesden, Marius van den Beek, Brad Langhorst)

Tools

- BLAST+ XML2 Datatype
  a. NCBI Released a new output format: 14
  b. New challenges - produces multiple files
  c. Working on a Composite datatype to capture the relationships between these files

(Peter Cock, Carrie Ganote, Dave Bouvier; See https://github.com/peterjc/galaxy_blast)



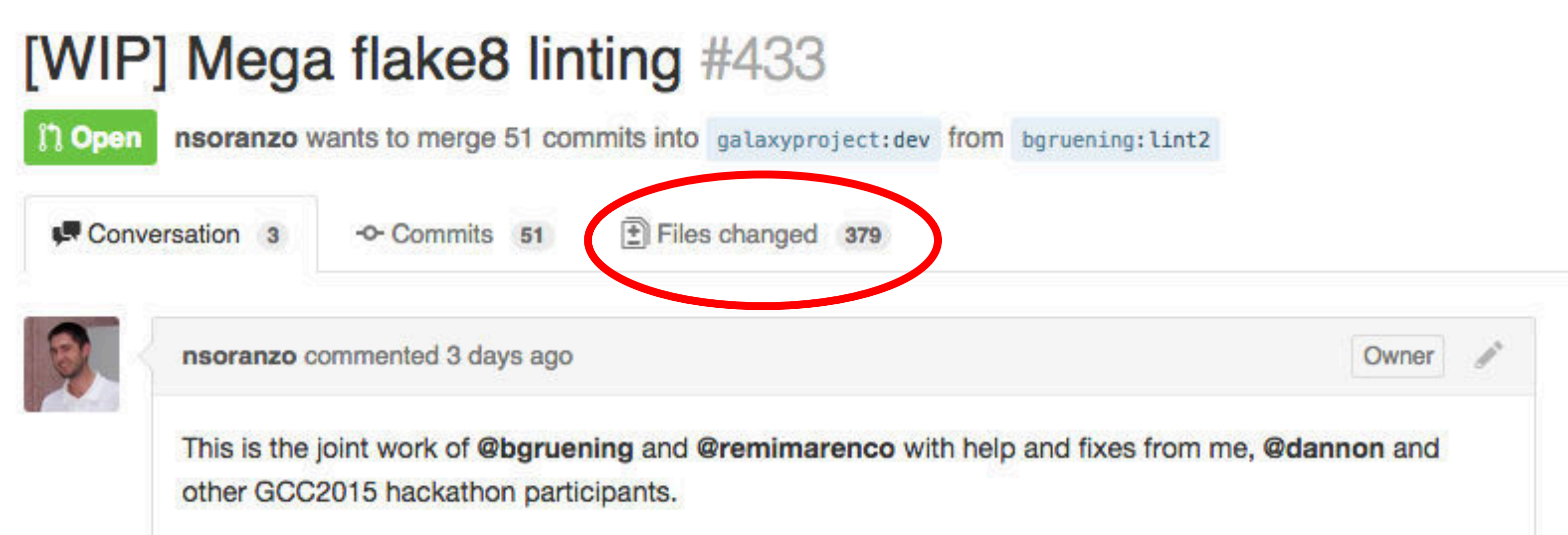
Then...

```
# file.xml
<?xml version="1.0"
“-//NCBI//NCBI Blas
“http://www.ncbi.nl
<BlastOutput>
  <BlastOutput_prog
  <BlastOutput_vers
...
```

... and Now

```
# file.xml
<?xml version="1.0"
xmlns="http://www.n
xmlns:xi="http://ww
  <xi:include href=
  <xi:include href=
  <xi:include href=
  <xi:include href=
  <xi:include href=
  <xi:include href=
</BlastXML>
```

```
# file_1.xml
<?xml version="1.0"
    xmlns="http://w
    xmlns:xs="http:
    xs:schemaLocati
http://www.ncbi.nlm
  <report>
    <Report>
      ...
```



[WIP] Mega flake8 linting #433

nsoranzo wants to merge 51 commits into `galaxyproject:dev` from `bgruening:lint2`

nsoranzo commented 3 days ago

This is the joint work of **@bgruening** and **@remimarenco** with help and fixes from me, **@dannon** and other GCC2015 hackathon participants.

Data Libraries API and Beta UI Tweaks

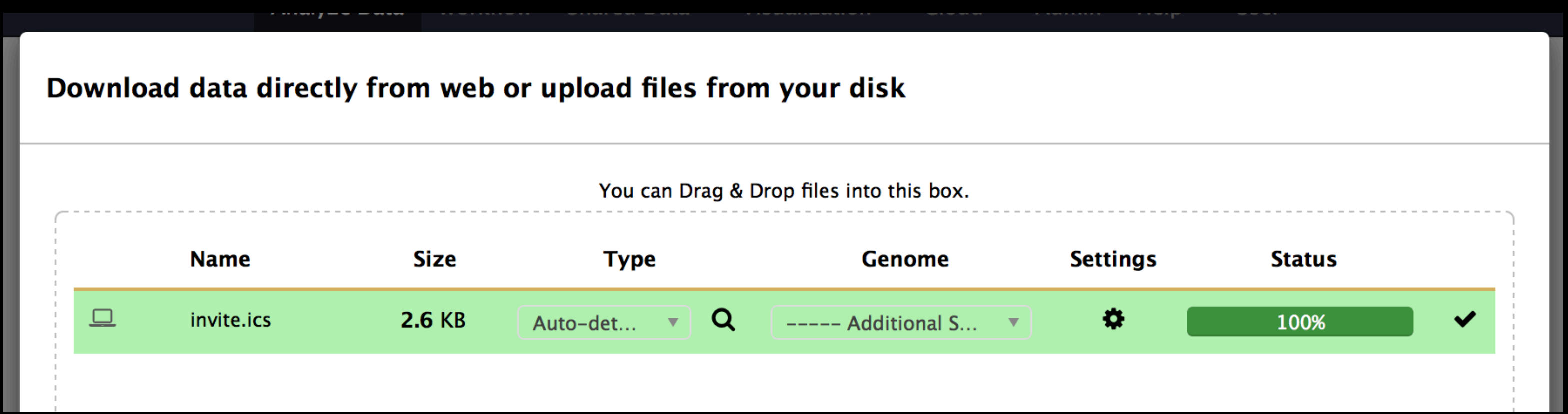
- Add the ability to import whole folders to a history
- Add ability to recursively download whole folders
- Fix Galaxy internal StreamBall module so it will create file archives with duplicated filenames properly

https://github.com/galaxyproject/galaxy/pull/426

Made by: Eric Enns - @EJEnns, Martin Čech

Composite Uploads

- Goal: Enable the upload of composite datasets.
- Steps
  a. construct UI to select composite dataset components
  b. submit selected files to API and execute the upload tool



Download data directly from web or upload files from your disk

You can Drag & Drop files into this box.

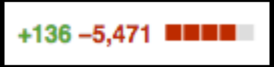
	Name	Size	Type		Genome	Settings	Status	
💻	invite.ics	2.6 KB	Auto-det... ▼	🔍	----- Additional S... ▼	⚙	100%	✓

Smashing Eggs

Eggs provide Galaxy framework dependencies

Why Smash?

- Outdated format



+136 −5,471 🟥🟥🟥🟥⬜

- Replace archaic, fragile download and build code with standard *pip*-based installation

Still WIP but we’re close:

github.com/galaxyproject/galaxy/pull/428

Data Hack

Purpose

- Bring together end-users & coders
- Identify issues with Galaxy usability and other bottlenecks from an **end-user perspective**
- Generation of "Best Practices" analysis workflows
- Wrap tools and pipelines to simplify and address common data manipulations
- Plan to publish (*GigaScience*) our initial workflow solutions along with a cloudshare with these workflows and example data

Data Hack

(GIGA)^n SCIENCE

- *Results*
- “Best Practices” analysis workflows for RNA-seq for multiple use cases, variant calling mini-pipeline
- New tools wrapped: Beagle, Impute and StepIt and pipelines for variant calling
- Identified & fixed: broken tools, new capabilities, wrapping existing tools

This will continue in a more formalized form with the formation of a “GalaxyScientists” group to provide a synthesized view from the end-user community for new functionalities, tools and capabilities.

Johns Hopkins University Data Science Specialization Program

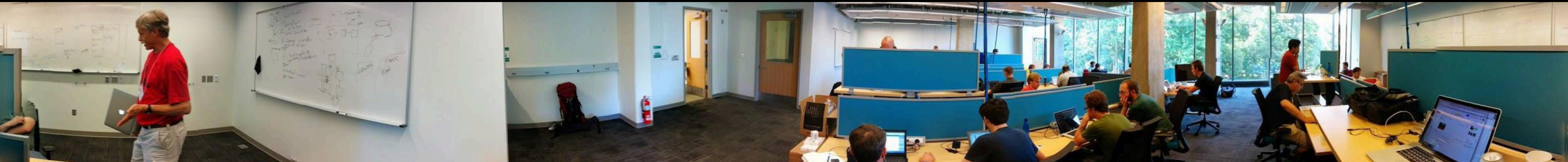
Thanks!

- Johns Hopkins Data Science

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

- TGAC - The Genome Analysis Centre

TGAC The Genome Analysis Centre™



The End