Evolution of Galaxy:

A rough timeline
mais avant de commencer...
Avant 2005...
The legend goes like this: in 1980 Webb Miller and Eugene Myers asked Ross Hardison if there was anything interesting to do in biology...
...by early 2000’s the big data in biology was genomic sequences and alignments. Penn State was central in developing alignment tools.

Webb Miller  
Ross Hardison
...by early 2000’s the big data in biology was genomic sequences and alignments. Penn State was central in developing alignment tools
The basic question in the *early* 2000’s was:

*What is aligned to what and does it overlap with anything interesting?*
GALA, a Database for Genomic Sequence Alignments and Annotations

Belinda Giardine, Laura Elnitski, Cathy Riemer, Izabela Makalowska, Scott Schwartz, Webb Miller and Ross C. Hardison

Departments of Computer Science and Engineering, Biochemistry and Molecular Biology, Biology, and Huck Institute for Life Sciences, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

We have developed a relational database to contain whole genome sequence alignments between human and mouse with extensive annotations of the human sequence. Complex queries are supported on recorded features, both directly and on proximity among them. Searches can reveal a wide variety of relationships, such as finding all genes expressed in a designated tissue that have a highly conserved noncoding sequence 5’ to the start site. Other examples are finding single nucleotide polymorphisms that occur in conserved noncoding regions upstream of genes and identifying CpG islands that overlap the 5’ ends of divergently transcribed genes. The database is available online at http://globin.cse.psu.edu/ and http://bio.cse.psu.edu/.
GALA enabled query annotation information from the human genome, alongside alignments with the mouse genome, integrated with the UCSC browser, and allowed building up set queries using the results of previous queries (the birth of the History system)
Can GALA be extended to other analyses?
Tools (D.1, D.2):
- Specialized data retrieval
- Neutral evolution rate
- $K_A$ and $K_S$ estimation
- Orthology

Operations (D.1):
- ° Set operations
- ° Proximity
- ° Clustering

Galaxy core (D.1):
- Result BED file
- MySQL database (history)

Data Sources (D.1):
- UCSC Table Browser
- ENCODEdb
- NCBI Entrez
- User data

Output

User Interface
Tools (D.1, D.2):
- Specialized data retrieval
- Neutral evolution rate
- $K_A$ and $K_S$ estimation
- Orthology

Operations (D.1):
- Set operations
- Proximity
- Clustering

User Interface

Galaxy core (D.1):
- Result BED file
- MySQL database (history)

Data Sources (D.1):
- UCSC Table Browser
- ENCODEdb
- NCBI Entrez
- User data

Output
Galaxy as a PERL script (~50,000 lines)
We threw the first one away (quickly) and rewrote from scratch in Python. At this point we made several key design decisions that (in hindsight) determined whether we would succeed or fail. (We got very lucky.)
1. No longer store data in a database, but in flat files in various common formats

This meant existing tools could be integrated easily because they did not need to change the data formats they work with or interact with a database.

It also meant that when high-throughput sequence data suddenly came along (2005), we were prepared to deal with data at that scale easily.
2. Rather than build new analysis tools in the system, build an abstract configuration driven interface to command line tools

We did this to make our lives easier, we had many analysis tools lying around that we didn’t want to rewrite for Galaxy

But this was equally appealing to other developers who could now easily make their tools available to Biologists
Early pythonic Galaxy circa 2005

Basic Statistics:
- Histogram: histogram.tool
- Scatter Plot: scatter.tool
- Filtering: filtering.tool
- Correlation: correlation.tool
- Region Length: region_length.tool
- Score distribution: scoreGraph.tool

Operations:
- Complement: complement.tool
- Restrict: restrict.tool
- Merge overlapping regions: merge.tool
- Cluster: cluster.tool
- Union: union.tool
- Intersect: intersect.tool
- Subtract: subtract.tool
- Proximity: proximity.tool
- Join Lists: joinLists.tool
- Vicinity: vicinity.tool
- Join Same Coordinates Region: joinSameCoor.tool

Sequence Tools:
- Extract sequences: fasta-subseq-wrapper.tool
- Extract blastZ alignments: extractAxt_wrapper.tool

Data Sources:
- UCSC query: ucs.cool.tool
- Genbank: genbank.tool
- Encode DB: encodedb.tool
- Featured datasets: import.tool

Format Converters:
- BED and xBED converter: bed_convert.tool
Dan Blankenberg and the letter $P$
3. Make the entire stack self-contained, allowing a complete Galaxy to be setup on most systems in minutes

We primarily did this to engage tool developers, making it as easy as possible to develop new tool wrappers for contribution

We envisioned those tools would all be made available through the main Galaxy service

But it also provided a scaling strategy, making it easy for sites to run their own Galaxy
4. Open-source *and* openly developed from the first commit

Provide everything we do under a liberal open-source license (no copyleft), and only support open-source tools on the main instance

Our primary development repository is exposed to the public, initially hosted by us but later moved out third parties (bitbucket.org, and then github.com)

The software is distributed only through version control, with a rapid release cycle (at least monthly)
Connection with UCSC
Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

<table>
<thead>
<tr>
<th>clade:</th>
<th>Mammal</th>
<th>genome:</th>
<th>Human</th>
<th>assembly:</th>
<th>Dec. 2013 (GRCh38/hg38)</th>
</tr>
</thead>
<tbody>
<tr>
<td>group:</td>
<td>Genes and Gene Predictions</td>
<td>track:</td>
<td>GENCODE v24</td>
<td>add custom tracks</td>
<td>track hubs</td>
</tr>
<tr>
<td>table:</td>
<td>knownGene</td>
<td>region:</td>
<td>genome</td>
<td>position</td>
<td>chr1:11102837-11267747</td>
</tr>
<tr>
<td>identifiers (names/accessions):</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>paste list</td>
</tr>
<tr>
<td>filter:</td>
<td>create</td>
<td>intersection:</td>
<td>create</td>
<td></td>
<td>correlation:</td>
</tr>
<tr>
<td>output format:</td>
<td>BED - browser extensible data</td>
<td>Send output to:</td>
<td>Galaxy</td>
<td>GREAT</td>
<td>GenomeSpace</td>
</tr>
<tr>
<td>output file:</td>
<td></td>
<td>file type returned:</td>
<td>plain text</td>
<td>gzip compressed</td>
<td>(leave blank to keep output in browser)</td>
</tr>
</tbody>
</table>

To reset all user cart settings (including custom tracks), click here.
Genome sequencing in microfabricated high-density picolitre reactors

The basic question in the late 2000’s becomes: *What would happen if I sequence the s****t out of anything?*

*For metagenomic studies this was, in fact, precisely the question asked.*
### Perform Actions

- **Basic Statistics**
  - **Histogram** - builds histogram for any numeric column
  - **Scatter Plot** - builds scatterplot for any numeric column
  - **Filtering** - filters data on any column using simple conditional expressions
  - **Correlation** - computes Pearson’s correlation between any two numerical columns
  - **Region Length** - computes length of bed intervals
  - **Score distribution** - display the score distribution of a selected score name

- **Operations**
- **Sequence Tools**
- **Data Sources**
- **Format Converters**
- **Upload Data**
No workflows?! I thought Galaxy was a workflow system...
2010- The modern Galaxy era
Galaxies on private clouds

Galaxies on public clouds

http://usegalaxy.org

Private Tool Sheds

private Galaxy installations
There was no reason to let it last this long and get this bad.
IT'S DANGEROUS TO GO ALONE! TAKE THIS.

BIOCONDA®
Tool and dependency binaries, built in minimal environment with controlled libs

Container defines minimum environment

Virtual machine controls kernel and apparent hardware environment

Increasingly precise environment control
Best thing about the introduction of the ToolShed: Birth of the Intergalactic Utilities Commission
Galaxy-WORKshop
Tour De France
2012

28th of May:
Perpignan, hosted by
Chimie RAGO CNRS

30th of May:
Lyon, hosted by
LBBE CNRS

1st of June:
Paris, hosted by
Institut Curie

Workshop for Biologists & Workshop for Developers
Presented by:
- Nate Coraor, Anton Nekrutenko & James Taylor -
Also this happened...
Meanwhile...
The great flood of 2012
The great flood of 2012

Your data here
...In which Anton almost loses his job
Stability and sustainability crisis!
...In which

Save = Galaxy main
The community established itself and the evolutionary timeline accelerated! The only bad thing about it is that it hard to put things in chronological order from memory without using git log.
This included many things covered today and tomorrow including:

- Visualizations beyond trackster
- Expansion beyond genomics
- Massive tool suite contributions and updates
- Interactive environments
- Training & Tours
- … uhhhh … so much more
More today and tomorrow!

Stay tuned.
evolution of biology
1. **A massive, dead disk galaxy in the early Universe.**
   PMID: 28640271

2. **The RNA workbench: best practices for RNA and high-throughput sequencing**
   PMID: 28582575
The problem is that PubMed indexes *Nature* and *Science* that are scientific journals with a broad subject coverage.
Being in sciences let’s invent an index:

\[ G_i = \text{true Galaxy pubs/} \text{false Galaxy pubs} \]

- **2005:** \( G_i = \frac{1}{15} \)
- **2017:** \( G_i = \frac{11}{14} \)
but seriously...
Number of papers per year

- 2005: 10
- 2006: 40
- 2007: 100
- 2008: 200
- 2009: 300
- 2010: 400
- 2011: 500
- 2012: 600
- 2013: 700
- 2014: 800
- 2015: 900
- 2016: 1000
from the beginning we tried to focus on biology
A framework for collaborative analysis of ENCODE data: Making large-scale analyses biologist-friendly

Daniel Blankenberg, James Taylor, Ian Schenck, Jianbin He, Yi Zhang, Matthew Ghent, Narayanan Veeraraghavan, Istvan Albert, Webb Miller, Kateryna D. Makova, Ross C. Hardison, and Anton Nekrutenko

Center for Comparative Genomics and Bioinformatics, Huck Institutes of the Life Sciences, Penn State University, University Park, Pennsylvania 16802, USA

The standardization and sharing of data and tools are the biggest challenges of large collaborative projects such as the Encyclopedia of DNA Elements (ENCODE). Here we describe a compact Web application, Galaxy2ENCEODE, that effectively addresses these issues. It provides an intuitive interface for the deposition and access of data, and features a vast number of analysis tools including operations on genomic intervals, utilities for manipulation of multiple sequence alignments, and molecular evolution algorithms. By providing a direct link between data and analysis tools, Galaxy2ENCEODE allows addressing biological questions that are beyond the reach of existing software. We use Galaxy2ENCEODE to show that the ENCODE regions contain >2000 unannotated transcripts under strong purifying selection that are likely functional. We also show that the ENCODE regions are representative of the entire genome by estimating the rate of nucleotide substitution and comparing it to published data. Although each of these analyses is complex, none takes more than 15 min from beginning to end. Finally, we demonstrate how new tools can be added to Galaxy2ENCEODE with almost no effort. Every section of the manuscript is supplemented with QuickTime screencasts. Galaxy2ENCEODE and the screencasts can be accessed at http://g2.bx.psu.edu.

[Supplemental material is available online at www.genome.org and http://g2.bx.psu.edu.]
...we even invented a study just to do analyses in Galaxy
“Just so you know, you've got a lot of really rare specimens preserved here”
Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond, 1,2,6,9 Samir Wadhawan, 3,6,7 Francesca Chiaromonte, 4
Guruprasad Ananda, 1,3 Wen-Yu Chung, 1,3,8 James Taylor, 1,5,9 Anton Nekrutenko, 1,3,9
and The Galaxy Team 1

1 http://galaxyproject.org; 2 Division of Infectious Diseases, Division of Biomedical Informatics, School of Medicine University of California San Diego, San Diego, California 92103, USA; 3 Huck Institute for the Life Sciences, Penn State University, University Park, Pennsylvania 16803, USA; 4 Department of Statistics, Penn State University, University Park, Pennsylvania 16803, USA; 5 Departments of Biology and Mathematics & Computer Science, Emory University, Atlanta, Georgia 30322, USA
... and
evolution of community
In 2006 Ross Lazarus was Galaxy community
He suggested having a conference
A conferences needed T-shirts
VIVA LA EVOLUCIÓN
GALAXY
http://main.g2.bx.psu.edu/

DON'T DO GENOMICS ALONE!
share analyses with GALAXY
http://g2.bx.psu.edu

WWW.SMD?
growlessly appendages...

usegalaxy.org
Hans-Rudolf Hotz: La connexion européenne!
First Galaxy Developer Community Conference
Björn Grüning is connected to Internet directly

(probably born with 802.11 circuitry. His particular hardware version lacks sleep functionality)
And now the meeting starts...
Community activity grows immensely...
Contributors with ≥ 10 commits excluding galaxy team
Contributors with $\geq 10$ commits excluding galaxy team
Contributors with ≥ 10 commits excluding galaxy team
Looking forward...
evolution across analyses scales
Analysis Scale

(exploratory) | Analysis Process Phase | (batch)
Analysis Scale

2006 Galaxy:
Batch analysis of 10s of datasets
(exploratory)
Analysis Scale

Analysis Process Phase

2008 Galaxy:
Workflows: 100s of datasets
(exploratory)

Analysis Process Phase
(batch)
Analysis Scale

Analysis Process Phase

10s, batch
100s, batch

2017 Galaxy:
10k - 100k datasets

(exploratory) Analysis Process Phase (batch)
Analysis Scale

10s, batch
100s, batch
100k, batch

(exploratory) Analysis Process Phase (batch)
We need better ways to look at, think about, and manage datasets and the 100k scale. At some point users no longer care about seeing the individual history, workflow, just specific results. New: many workflow view, for monitoring the execution of many workflows in parallel. New: reports — generate summaries of executing workflows, multiple workflows, from user templates with continuous updates.
Analysis Scale

Analysis Process Phase

(exploratory)  Analysis Process Phase  (batch)
Analysis Scale

Interactive Environments:
10s of datasets, *ad hoc* analyses

(exploratory)  Analysis Process Phase  (batch)
Analysis Scale

Analysis Process Phase

( exploratory )

ad hoc, more flexible

10s, batch

100s, batch

100k, batch

( batch )

?
Analysis Scale

- **Analysis Process Phase**
  - **Phase (exploratory)**
  - **Phase (batch)**

Visualization and analytics
- 10s of datasets, *highly interactive*
Analysis Scale

- **visual exploration**
- **ad hoc, more flexible**
- **10s, batch**
- **100s, batch**
- **100k, batch**

( exploratory )  **Analysis Process Phase**  ( batch )
We need to support exploratory data analysis even more than we do now

Dataset complexity, heterogeneity, dimensionality and all only increasing

The analysis decision process requires more support for data exploration, both visual and interactive data manipulation
The future Galaxy needs to scale seamlessly across the data analysis process…

…supporting analysts as they transition from exploratory, to batch, to high-throughput
At either end of the spectrum, there are common themes.

The future Galaxy embraces *real time* and continuous communication. From exploratory analysis to batch job tracking to automatic reports, Galaxy needs to be responsive and informative.

The future Galaxy is increasingly interactive. The future Galaxy better supports transitions between analysis modes.
So what is the Galaxy team?
Viva contributors!

Thank you for contributing to Galaxy IUC tools.
The future of this project depends solely on the community, its openness, and continuing outreach!
Thank you!

Now please stay for talks that actually contain useful information!