Exploit your Data

Galaxy

and the

GMOD Tool Suite

http://galaxyproject.org  http://gmod.org
Introduction

Galaxy
  Worked example
  Deployment Options
  Community

GMOD
  Software
  Community
Goal for this workshop

Give you some idea what these open source tools can do, and how you might use them in your research.

This workshop will not cover details of how the tools are implemented or new algorithm designs.
Are these tools timely?
Are these tools timely?

Yes.

Next Generation Genomics: World Map of High-throughput Sequencers
Nick Loman, James Hadfield

http://pathogenomics.bham.ac.uk/hts/
Introduction

Galaxy
- Worked example
- Deployment Options
- Community

GMOD
- Software
- Community
Galaxy URLs to Remember

http://galaxyproject.org
http://usegalaxy.org
http://getgalaxy.org
http://usegalaxy.org/galaxy101
What is Galaxy?

- **A free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- **An analysis and data integration** tool
- **Open source software** that makes integrating your own tools and data and customizing for your own site simple
- A part of **GMOD**
Enable accessible, reproducible, and transparent computational biomedical research.
On human chromosome 22, which coding exons have the most known SNPs?

http://usegalaxy.org
Galaxy: A Rough Plan

- Get some data
  - Coding exons on chromosome 22
  - SNPs on chromosome 22
- Mess with it
  - Identify which exons have SNPs
  - Count number of SNPs in each of those exons.
  - Identify exons with most SNPs

http://usegalaxy.org/galaxy101
Dozens of tools for different NGS applications packaged with Galaxy
Enable accessible, reproducible, and transparent computational biomedical research.
Demo: Reproducibility and Transparency

http://usegalaxy.org
Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond1,2,6,9, Samir Wadhawan3,6,7, Francesca Chiaromonte4, Guruprasad Ananda1,3, Wen-Yu Chung1,3,8, James Taylor1,5,9, Anton Nekrutenko1,3,9 and The Galaxy Team1

Abstract

How many species inhabit our immediate surroundings? A straightforward collection technique suitable for answering this question is known to anyone who has ever driven a car at highway speeds. The windshield of a moving vehicle is subjected to numerous insect strikes and can be used as a collection device for representative sampling. Unfortunately the analysis of biological material collected in that manner, as with most metagenomic studies, proves to be rather demanding due to the large number of required tools and considerable computational infrastructure. In this study, we use organic matter collected by a

Footnotes

[Supplemental material is available online at http://www.genome.org. All data and tools described in this manuscript can be downloaded or used directly at http://galaxyproject.org. Exact analyses and workflows used in this paper are available at http://usegalaxy.org/u/aun1/p/windshield-splatter.]
Agenda

Introduction

Galaxy
  Worked example
  Deployment Options
  Community

GMOD
  Software
  Community
Galaxy main site
http://usegalaxy.org

• Public web site, anybody can use

• Hundreds of tools

• Persistent

• ~500 new users per month, ~100 TB of user data, ~130,000 analysis jobs per month, every month is our busiest month ever...
But, it’s a big world

Main has lots of tools, storage, processor, users, ...

- But **not all tools** - there are thousands and adding new tools is not taken lightly

- But **not infinite storage and processors** - main will continue to be maintained and enhanced, but with use limits and storage quotas

A centralized solution cannot scale to meet data analysis demands of the whole world
Scaling Galaxy

So much data:

- Encourage local Galaxy instances and Galaxy on the cloud. Support increasingly decentralized model and *improve access to existing resources*

So many tools and workflows:

- Focus on building infrastructure to allow community to integrate and share tools, workflows, and best practices
Local Galaxy Instances
http://getgalaxy.org

Galaxy is designed for local installation and customization

- Easily integrate new tools
- Easy to deploy and manage on nearly any (Unix) system
- *Just download and run, completely self-contained!* *

* Some assembly required.†
† But not much.‡
‡ And help is on the way.
Local Galaxy Instances
http://getgalaxy.org

Galaxy is designed for local installation and customization

- Easily integrate new tools
- Easy to deploy and manage on nearly any (Unix) system
- *Just download and run, completely self-contained!* *

* Some assembly required. †
† But not much. ‡
‡ And help is on the way.
Public Galaxy Servers
http://galaxyproject.org/PublicGalaxyServers

Interested in:

- ChIP-chip and ChIP-seq?
  - ✓ Cistrome
- Statistical Analysis?
  - ✓ Genomic Hyperbrowser
- Sequence and tiling arrays?
  - ✓ Oqtans
- Text Mining?
  - ✓ DBCLS Galaxy
- Reasoning with ontologies?
  - ✓ GO Galaxy
- Internally symmetric protein structures?
  - ✓ SymD
Scale up on existing resources

- Move intensive processing (tool execution) to other hosts
- Frees up the application server to serve requests and manage jobs
- Utilize existing resources
- Supports any batch scheduler that supports DRMAA (most of them)
- All levels of job running and scheduling are pluggable
Galaxy Cloud
http://usegalaxy.org/cloud

- On-demand resource acquisition fits well with the irregular resource needs of many labs working with sequence data
- Our goal is to approach the ease of use of a “software as a service” solution while maintaining the flexibility and control of an infrastructure based solution
Using Amazon EC2: Startup in 3 steps
Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application will allow you to manage this cloud and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be add and remove additional services as well as ‘worker’ nodes on which jobs are run.

Terminate cluster  Add nodes  Remove nodes  Access Galaxy

Status

Cluster name: ttt
Disk status: 0 / 0 (0%)
Worker status: Idle: 0 Available: 0 Requested: 0
Service status: Applications • Data •

Cluster status log
Can use like any other Galaxy instance, with additional compute nodes acquired and released (automatically) in response to usage
Tool installation and configuration, image creation, etc, all **completely automated and extensible**

Cloud instances include all tools available in main Galaxy *and more*

Same automation approach can be used for configuring tool dependencies for a local Galaxy

VM image with just tools available, currently at [http://usegalaxy.org/vm](http://usegalaxy.org/vm)
Why we love clouds and cloud-like things:

Reasonably cost effective and efficient (elasticity + autoscaling definitely save money)

Analysis costs are more directly quantifiable

Infrastructure as an abstraction + standard APIs for provisioning reduces risk of vendor lock-in

Virtualization makes so many things easier
Some future challenges

- Capturing and automatically deploying tool dependencies, automatic tool acquisition in Galaxy instances
- Better interfaces for highly parallel analysis (e.g. running the same workflow across 192 individuals)
- Various workflow engine improvements, partial data streaming, combined experimental/computational workflows
Galaxy Tool Shed vision

- Allow users to share “suites” containing tools, datatypes, workflows, sample data, and automated installation scripts for tool dependencies
- Integration with Galaxy instances to automate tool installation and updates
Galaxies on private clouds

Galaxies on public clouds

http://usegalaxy.org

http://usegalaxy.org/community

public and private Galaxy installations

private & public Tool Sheds

http://usegalaxy.org/community

1 2 3 ∞
<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Repositories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly</td>
<td>Tools for working with assemblies</td>
<td>12</td>
</tr>
<tr>
<td>Computational chemistry</td>
<td>Tools for use in computational chemistry</td>
<td>2</td>
</tr>
<tr>
<td>Convert Formats</td>
<td>Tools for converting data formats</td>
<td>12</td>
</tr>
<tr>
<td>Data Source</td>
<td>Tools for retrieving data from external data sources</td>
<td>3</td>
</tr>
<tr>
<td>Fasta Manipulation</td>
<td>Tools for manipulating fasta data</td>
<td>17</td>
</tr>
<tr>
<td>Graphics</td>
<td>Tools producing images</td>
<td>7</td>
</tr>
<tr>
<td>Next Gen Mappers</td>
<td>Tools for the analysis and handling of Next Gen sequencing data</td>
<td>22</td>
</tr>
<tr>
<td>Ontology Manipulation</td>
<td>Tools for manipulating ontologies</td>
<td>3</td>
</tr>
<tr>
<td>SAM</td>
<td>Tools for manipulating alignments in the SAM format</td>
<td>7</td>
</tr>
<tr>
<td>Sequence Analysis</td>
<td>Tools for performing Protein and DNA/RNA analysis</td>
<td>41</td>
</tr>
<tr>
<td>SNP Analysis</td>
<td>Tools for single nucleotide polymorphism data such as WGA</td>
<td>4</td>
</tr>
<tr>
<td>Statistics</td>
<td>Tools for generating statistics</td>
<td>8</td>
</tr>
<tr>
<td>Text Manipulation</td>
<td>Tools for manipulating data</td>
<td>14</td>
</tr>
<tr>
<td>Visualization</td>
<td>Tools for visualizing data</td>
<td>8</td>
</tr>
<tr>
<td>Name</td>
<td>Synopsis</td>
<td>Revision</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------------------------------------------------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>abyss toolsuite</td>
<td>This suite contains Abyss and Abyss-PE config files and wrappers for Galaxy</td>
<td>0.92636934a189</td>
</tr>
<tr>
<td>agile wrapper</td>
<td>Quickly match reads to a reference genome or sequence file</td>
<td>0:d6a426afa46</td>
</tr>
<tr>
<td>asdf</td>
<td>asdf</td>
<td>-1:0000000000000</td>
</tr>
<tr>
<td>assemblystats</td>
<td>Summarise an assembly (e.g. N50 metrics)</td>
<td>0.6544228ea290</td>
</tr>
<tr>
<td>bam_to_bigwig</td>
<td>Generate BigWig coverage files from BAM files. Allows gapped reads to be split (useful for RNA-Seq).</td>
<td>5:5b40b93e9ae3</td>
</tr>
</tbody>
</table>
Repository revision

[Revision: 2:bb1847435ec1]
repository tip
Select a revision to inspect and download versions of tools from this repository.

clustalomega

Clone this repository:
hg clone http://toolshed.g2.bx.psu.edu/repos/clustalomega/clustalomega

Name:
clustalomega

Synopsis:
multiple sequence alignment program for proteins

Detailed description:
Clustal Omega is a general purpose multiple sequence alignment program for proteins. It produces high quality alignments and is efficient for large datasets.

Revision:
2:bb1847435ec1

Owner:
clustalomega

Times downloaded:
39

Preview tools and inspect metadata by tool version

<table>
<thead>
<tr>
<th>name</th>
<th>description</th>
<th>version</th>
<th>requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustal Omega</td>
<td>multiple sequence alignment program for proteins</td>
<td>1.0.2</td>
<td>none</td>
</tr>
</tbody>
</table>
Clustal Omega is a general purpose multiple sequence alignment (MSA) program for proteins. It produces high quality MSAs and is capable of handling data-sets of hundreds of thousands of sequences in reasonable time.

In default mode, users give a file of sequences to be aligned and these are clustered to produce a guide tree and this is used to guide a "progressive alignment" of the sequences. There are also facilities for aligning existing alignments to each other, aligning a sequence to an alignment and for using a hidden Markov model (HMM) to help guide an alignment of new sequences that are homologous to the sequences used to make the HMM. This latter procedure is referred to as "external profile alignment" or EPA.

Clustal-Omega uses HMMs for the alignment engine, based on the HHalign package from Johannes Soding [1]. Guide trees are optionally made using mBed [2] which can cluster very large numbers of sequences in O(N*\log(N)) time. Multiple alignment then proceeds by aligning larger and larger alignments using HHalign, following the clustering given by the guide tree.

In its current form Clustal-Omega can only align protein sequences but not DNA/RNA sequences. It is envisioned that DNA/RNA will become available in a future version.

A full version of these instructions is available at [http://www.clustal.org/](http://www.clustal.org/)

This is a beta version of Clustal Omega. Bugs should be reported to clustalw@ucd.ie

A standalone version of Clustal Omega for Linux/Windows/Mac is available from [http://www.clustal.org/](http://www.clustal.org/)


Galaxy Community

- Tool Shed
- Local Public Installs
- Mailing Lists (very active)
- Screencasts
- Events Calendar, News Feed
- Community Wiki
- Annual Community Meeting
  - Summer 2012 in Chicago

http://galaxyproject.org/wiki/Get%20Involved
Try it now: http://UseGalaxy.org

Develop and deploy: http://GetGalaxy.org

Supported by the NHGRI (HG005542, HG004909, HG005133), NSF (DBI-0850103), Penn State University, Emory University, and the Pennsylvania Department of Public Health

http://GalaxyProject.org
What is GMOD?

- A set of interoperable open-source software components for visualizing, annotating, integrating and querying, and analyzing biological data.

- An active community of developers and users asking diverse questions and facing common challenges with their biological data.

http://gmod.org
Who uses GMOD?

Plus hundreds, if not thousands, of others
GMOD Server Requirements

- GMOD is not a hosted solution
  - *Usually*
- Server
  - Most use Linux or other Unix variant
- GMOD System Administrator
  - Understands Linux package management, scripting, command line interfaces, relational databases, ...
  - Grad/Undergrad, half time when starting up

http://gmod.org/wiki/Computing_Requirements
Agenda

Introduction

Galaxy
- Worked example
- Deployment Options
- Community

GMOD
- Software
- Community
GMOD Software

Visualization
- **GBrowse & JBrowse** genomic browsers
- **CMap & GBrowse_syn** for comparative genomics

Annotation
- **Maker & Apollo**

Data Integration and Querying
- **BioMart, InterMine, Chado**

Analysis
- **Galaxy, Ergatis, ISGA**
Visualization: GBrowse

- Visualize features, SNPs, quantitative data, NGS data (uses SAMtools)
- Semantic zooming
- Custom tracks, shared tracks, DAS
- Highly customisable by both admin and end-user

GBrowse:

Genome Browser

ChIP-Seq data

Visualization by Computational Biology Research Group @ Oxford.
GBrowse: Short Reads
GBrowse

HapMap
Allele Frequencies

hapmap.org
GBrowse

Work by Yi-Hsin Erica Tsai & Ben Faga, using PhyloGeoViz
GBrowse or JBrowse?

**GBrowse**
- Widely used
- Robust features & ecosystem
- Familiar interface
- Configuration is trial and error
- Requires more server
- Slower

**JBrowse**
- Limited features
- Unfamiliar
- Lots of future development
- Configuration is simpler
- Requires less server
- Much Faster
CMap

- Comparative map viewer
- Data type agnostic. Can show linkage, physical, deletion, QTL, ....
- Anything that is points or regions on a line

**GBrowse_syn**

- Comparative genomics browser
- Shows a “reference” compared to 1 or more others
- Built on GBrowse and can show any GBrowse-based annotations

McKay, et al., Using the Generic Synteny Browser (GBrowse_syn), *Current Protocols in Bioinformatics*, Unit 9.12, September 2010
Raw genomic sequences

- Mask repeats (RepeatMasker and Tandem Repeats Finder, nmerge, etc.)
- Identify orthologous regions (MERCATOR, orthocluster, etc.)
- Nucleotide-level alignment (PECAN, MAVID, etc.)
- Further processing (GERT)

GBrowse_syn
GBrowse_syn: Duplications
GBrowse_syn: Base level alignments
GBrowse_syn: Visual feedback on assembly
Q: CMap or GBrowse_syn?
A: Not an either/or choice

<table>
<thead>
<tr>
<th>CMap</th>
<th>GBrowse_syn</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Typically used when you have non-sequence based markers</td>
<td>• Used when to show multiple alignment data</td>
</tr>
<tr>
<td>• Popular in the plant community</td>
<td>• Can be orthology, region, or sequence based alignments</td>
</tr>
<tr>
<td></td>
<td>• Exploits GBrowse data sources and infrastructure</td>
</tr>
</tbody>
</table>
Annotation: MAKER

- Genome annotation pipeline for creating gene models
- Produces GFF3 and FASTA which can be loaded into GBrowse, JBrowse, Apollo, Chado, Galaxy, BioMart, InterMine, ...
- Incorporates SNAP, RepeatMasker, exonerate, BLAST, Augsustus, FGENESH, GeneMark, MPI
- Can also map existing annotation on to new assemblies and update existing annotations with new evidence
- Also available as a hosted service.

Cantarel, et al., MAKER: An Easy-to-use Annotation Pipeline Designed for Emerging Model Organism Genomes, Genome Research 2008 18(1) 188-96.
Apollo: Manual genome annotation editor

Lewis et al., Apollo: a sequence annotation editor., Genome Biology 2002, 3(12)
Data Integration & Querying: BioMart

Data Integration & Querying: InterMine

FlyMine is back! As a supplement to another grant the NHGRI have said they will provide support to keep FlyMine going until 2011.

**Data Categories**
Select a category to see more information about the data sets included. Each category includes associated templates and lists.

- Genomics
- Comparative Genomics
- Proteins
- Protein Structure
- Interactions
- Gene Ontology
- Gene Expression
- Transcriptional Regulation

**Templates**
Templates are predefined queries, each has a simple form and a description. You can edit templates in the QueryBuilder, if you log in you can create new templates yourself.

Example templates (196 total):

- Gene --> Orthologues.

**Lists**
You can run queries on whole lists of data. Create lists from the results of a query or by uploading identifiers. Click on a list to view graphs and summaries in a list analysis page, if you log in you can save lists permanently.
Data Integration & Querying: Chado

- Chado is GMOD’s core database schema
- A blueprint for organizing biological data
- Modular and extensible.
- Modules for IDs, sequence features, ontologies, attribution, ...
- Database backing many web sites from FlyBase to ParameciumDB
Agenda

Introduction

Galaxy
  Worked example
  Deployment Options
  Community

GMOD
  Software
  Community
GMOD Community

- Mailing lists (very active)
- Annual week long course
- Tutorials
- Events Calendar, News Feed
- Community Wiki
- Semi-Annual Community Meetings
  - Tomorrow! Toronto
  - April 2012, Washington, DC

http://gmod.org/
GMOD Contributors

Scott Cain
Carson Holt
Brian Osborne
Ed Lee
Stephen Ficklin
Mitch Skinner
Ian Holmes
Gos Micklem
Barry Moore
Dave Emmert
Alex Kalderimis
Steve Taylor
Dorie Main
Brian O’Connor
Sook Jung

Sheldon McKay
Lincoln Stein
Jim Hu
Daniel Renfro
Todd Vision
Nathan Liles
Chris Hemmerich
James Taylor
LBCC Adv. Design
Seth Redmond
Chris Fields
Ben Faga
Joshua Orvis
Junjun Zhang
Naama Menda

Hilmar Lapp
Dan Blankenberg
John Aikman
Robert Buels
Richard Smith
Ken Youens-Clark
Nicole Washington
Alexie Papanicolaou
Josh Goodman
Yuri Bendana
Jason Stajich
Suzi Lewis
Mark Yandell
Don Gilbert
Meg Staton

Plus hundreds, if not thousands, of others
Thanks

CSIRO
Dr Jen Taylor
Dr Annette McGrath
Dr Alexie Papanicolaou

AMATA 2011 Organizing Committee