DNA Sequence Bioinformatics Analysis with the Galaxy Platform

University of São Paulo, Brazil
28 July - 1 August 2014

Dave Clements
Johns Hopkins University

Robson Francisco de Souza
University of São Paulo

José Belazario
University of São Paulo

The Week's Agenda

Mon	Introductions: Cloud Computing, Nuvem Cloud, Basic Analysis in Galaxy
Tues	Workflows, Sharing, Quality Control, ChIp-Seq, Genome Assembly Concepts
Wed	Genome Assembly, RNA-Seq
Thur	SNP and Variant Calling
Fri	Intro to Command Line, Genome Annotation using MAKER, CloudMan and AWS

bit.ly/gxyusp2014

Monday's Agenda

9:00 Introduction to Cloud Computing and Using the Nuvem Cloud

20 minute Break at around 10:20

12:00 Lunch

2:00 Galaxy Project Introduction

2:15 Introduction to using Galaxy

20 minute Break at around 3:20

5:00 Done

Monday's Agenda

9:00 Introduction to Cloud Computing and Using the Nuvem Cloud

20 minute Break at around 10:20

12:00 Lunch

2:00 Galaxy Project Introduction

2:15 Introduction to using Galaxy

20 minute Break at around 3:20

5:00 Done

English and Português ...

Are two *excellent languages!*

But, I only speak one of them.

If I start to speak English too quickly,
or if I am not clear, then please tell me.

If that doesn't work ...

Se eu fizer isso de novo, em seguida, começar a falar comigo em muito rápidos Português.

I will slow down

Monday's Agenda

9:00 Introduction to Cloud Computing and Using the Nuvem Cloud

20 minute Break at around 10:20

12:00 Lunch

2:00 Galaxy Project Introduction

2:15 Introduction to using Galaxy

20 minute Break at around 3:20

5:00 Done

Goals

Provide a solid foundation in bioinformatic analysis.

Demonstrate how Galaxy can help you explore and learn options, perform analysis, and then share, repeat, and reproduce your analyses.

Not Goals

This workshop will *not* cover

- details of how tools are implemented, or
- new algorithm designs, or
- which assembler or mapper or peak caller or ... is best for you.

While this workshop does cover ChIP-Seq, RNA-Seq, assembly, ... you won't be an expert in any of these by the end of the week.

What is Galaxy?

A free (for everyone) web server

Open source software

These options result in several ways to use Galaxy

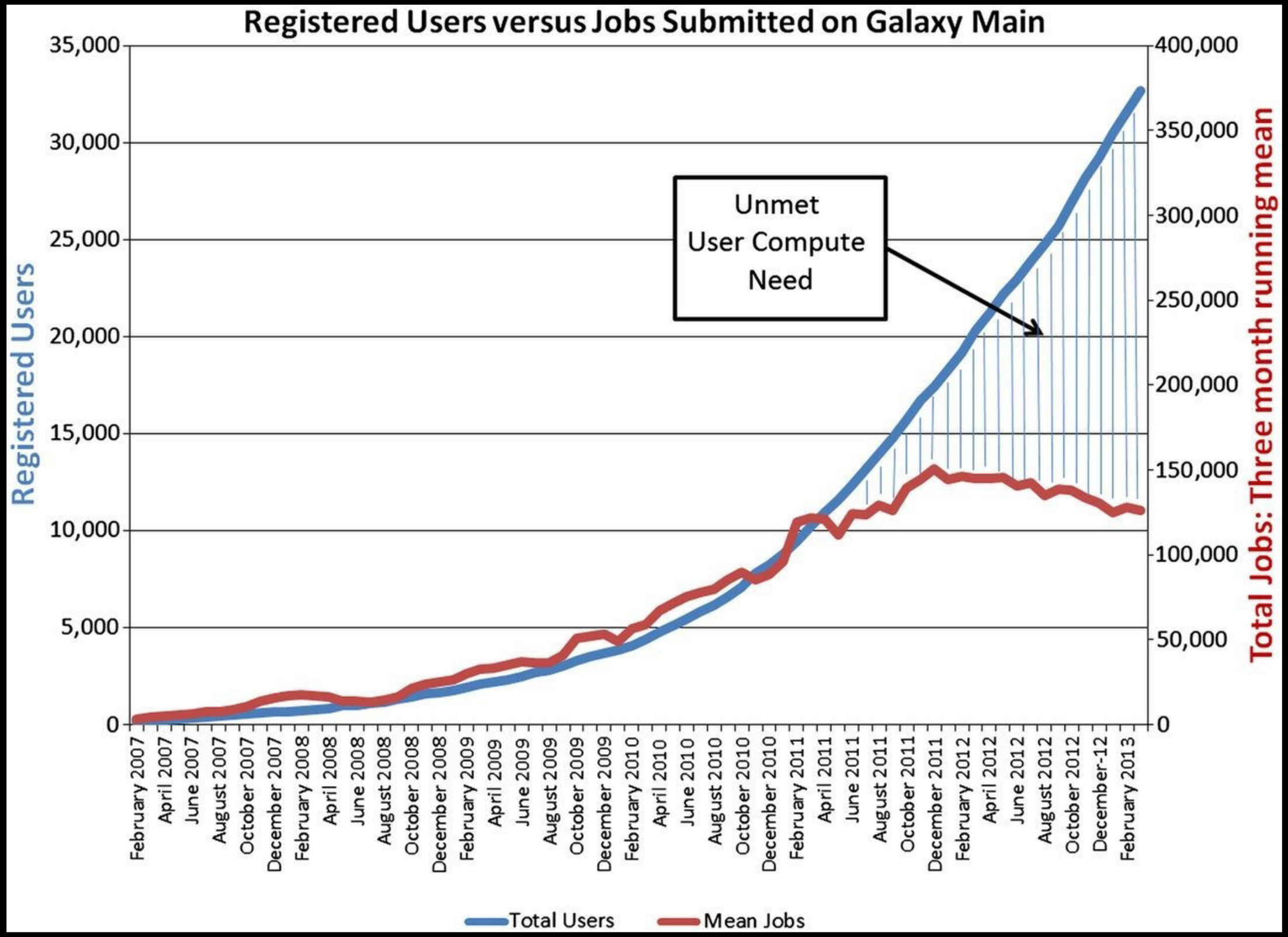
http://galaxyproject.org

Galaxy is available ...

As a free (for everyone) web server integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage

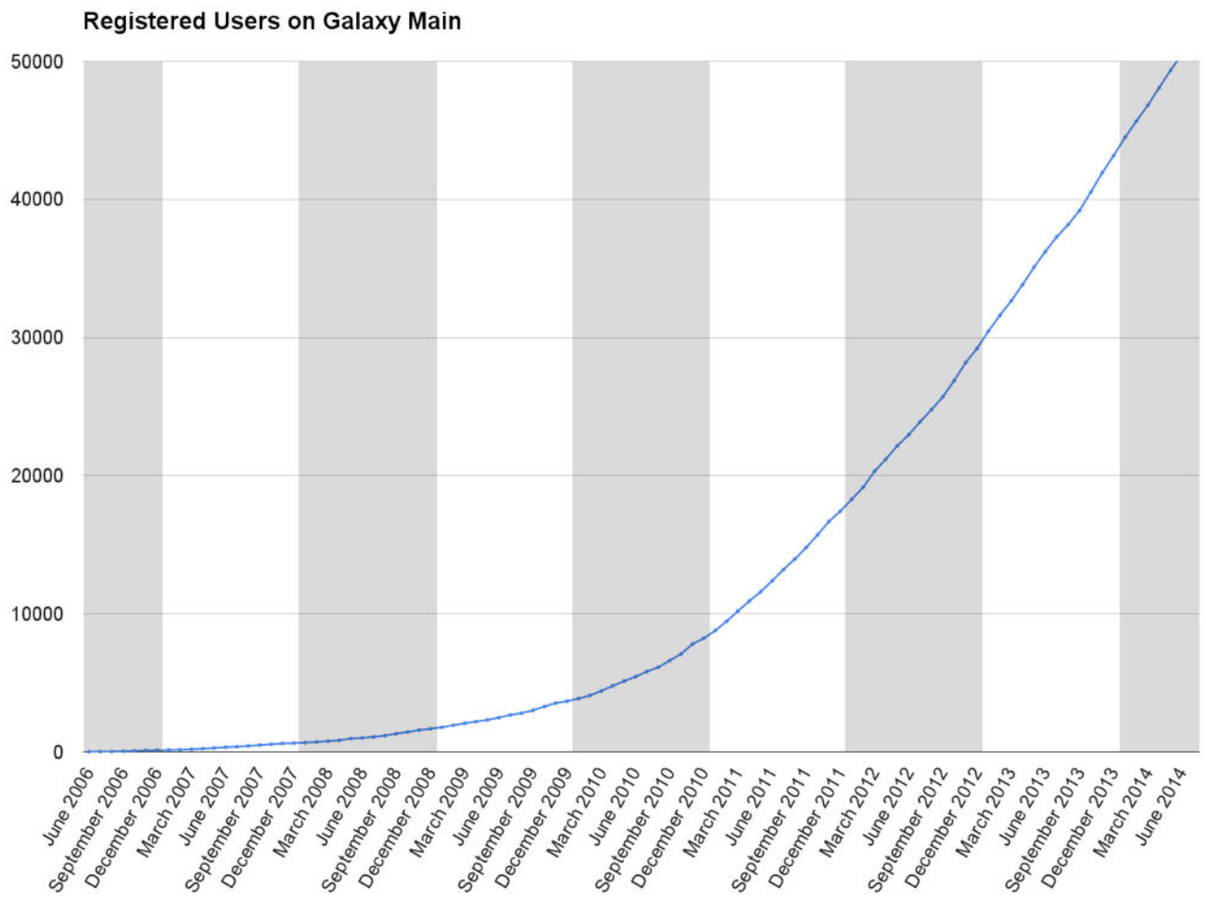
http://usegalaxy.org

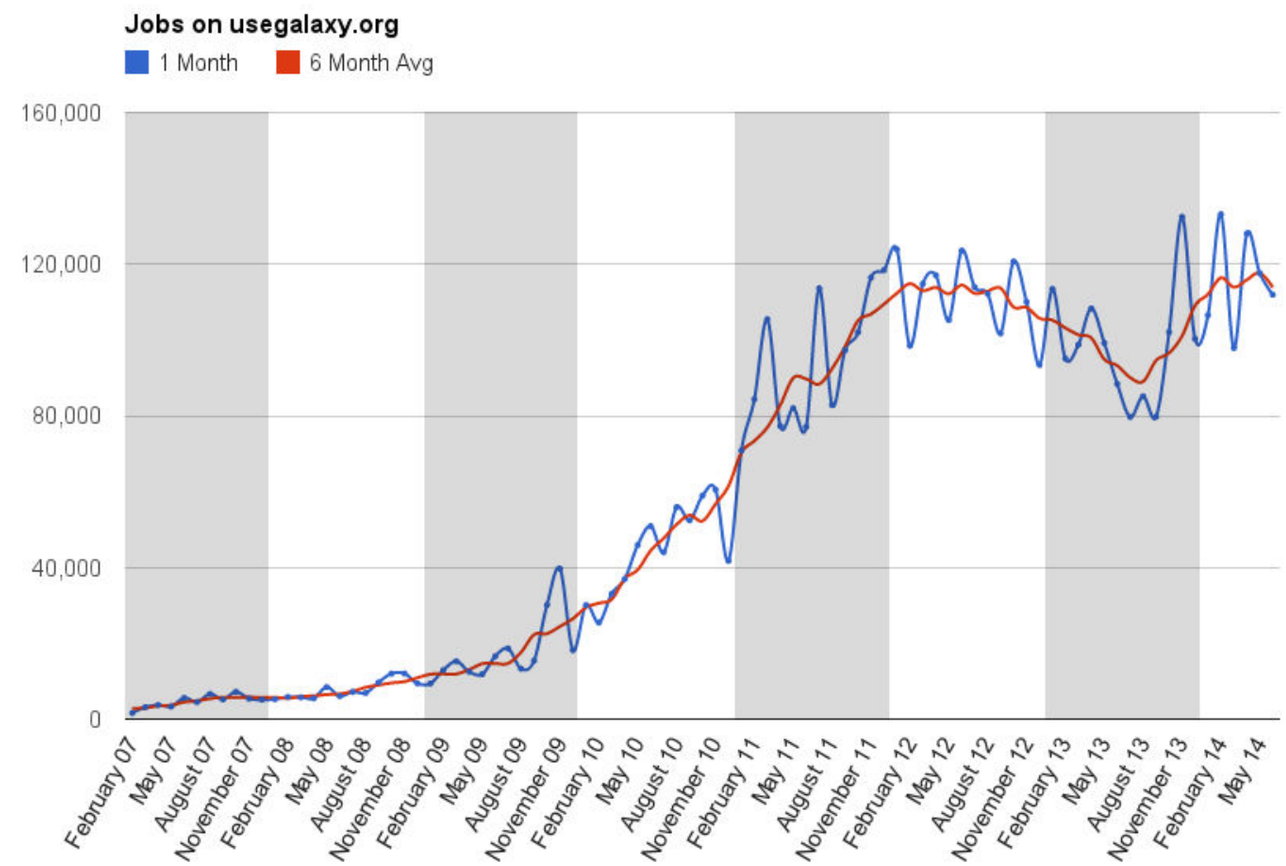
However, ***a centralized solution cannot support the different analysis needs of the entire world.***



Leveraging the national cyberinfrastructure for biomedical research

LeDuc, *et al. J Am Med Inform Assoc doi:10.1136/amiajnl-2013-002059*





And those trends have continued

bit.ly/gxyStats

Galaxy is available ...

- As a free (for everyone) web service

  http://usegalaxy.org

- As open source software

  http://getgalaxy.org

It is installed in locations around the world

Galaxy is available ...

- As a free (for everyone) web servi
  http://usegalaxy.org
- As open source software
  http://getgalaxy.org
- *On the Cloud*

We are using this today.

https://wiki.uspdigital.usp.br/nuvem/
http://aws.amazon.com/education
http://globus.org/

http://wiki.galaxyproject.org/Cloud

Galaxy is available: With Commercial Support

A ready-to-use appliance
(BioTeam)

Cloud-based solutions
(ABgenomica, AIS, Appistry, GenomeCloud)

Consulting & Customization
(Arctix, BioTeam, Deena Bioinformatics)

SLIPSTREAM APPLIANCE GALAXY EDITION
Galaxy made easy.

abgenomica

GenomeCloud

Appistry

AIS™ ENTERPRISE-CLASS DATA CENTER SERVICES

Galaxy Project: Further reading & Resources

http://galaxyproject.org

http://usegalaxy.org

http://getgalaxy.org

http://wiki.galaxyproject.org/Cloud

http://bit.ly/gxychoices

Monday's Agenda

9:00 Introduction to Cloud Computing and Using the Nuvem Cloud

20 minute Break at around 10:20

12:00 Lunch

2:00 Galaxy Project Introduction

2:15 Introduction to using Galaxy

20 minute Break at around 3:20

5:00 Done

Basic Analysis

Which genes have most overlapping Repeats?

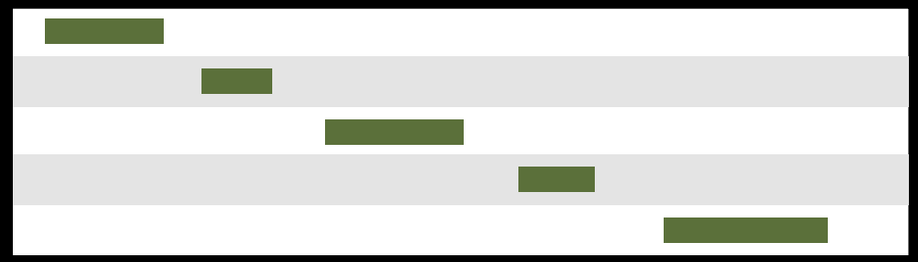
Use Human, HG19, Chromosome 22

(~ http://usegalaxy.org/galaxy101)

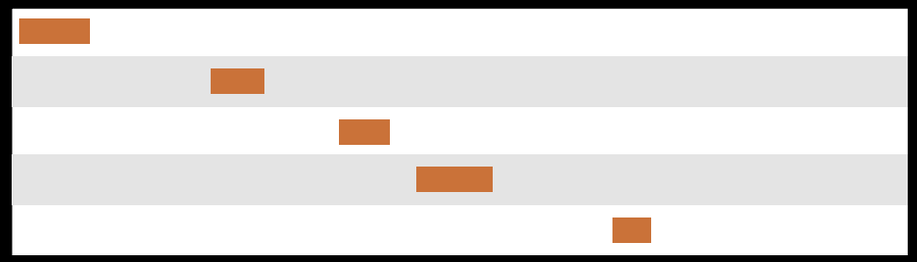
Genes & Repeats: A General Plan

- Get some data
  - Get Data → UCSC Table Browser
- Identify which genes/exons have Repeats
- Count Repeats per exon
- Visualize, save, download, ... exons with most Repeats

(~ http://usegalaxy.org/galaxy101)

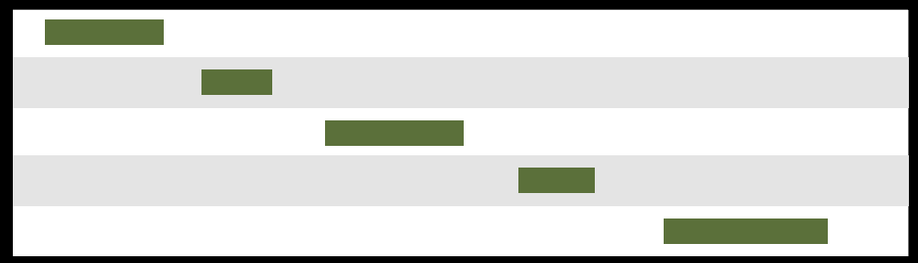


Exons

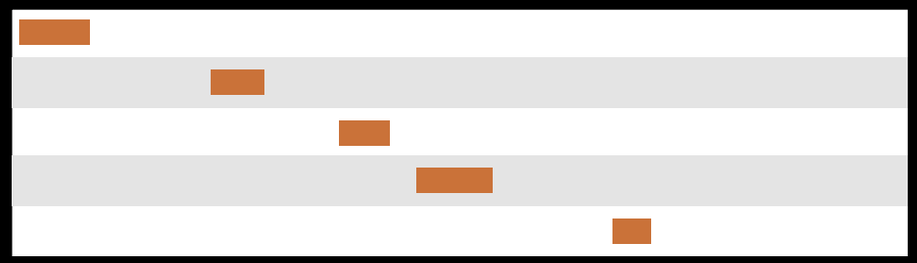


Repeats

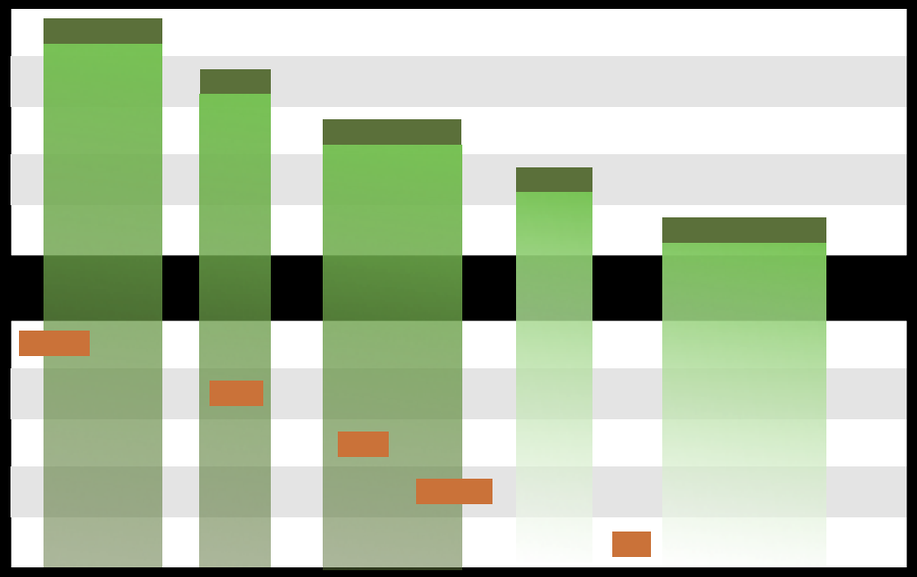
(Identify which genes/exons have Repeats)



Exons



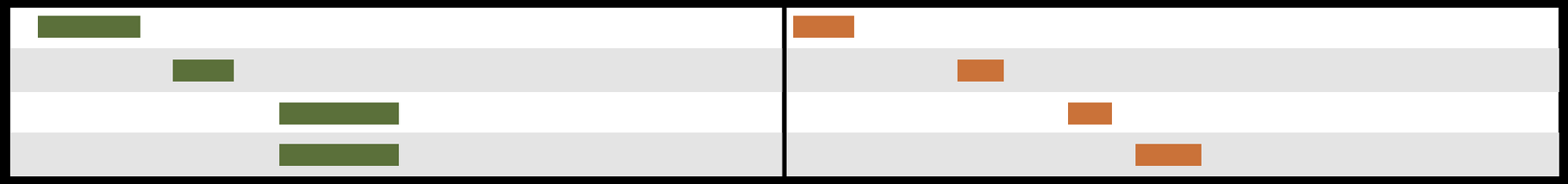
Repeats



Exons

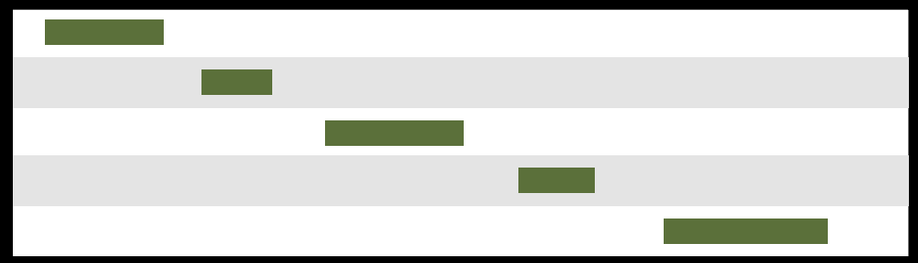
Repeats

Overlap pairings

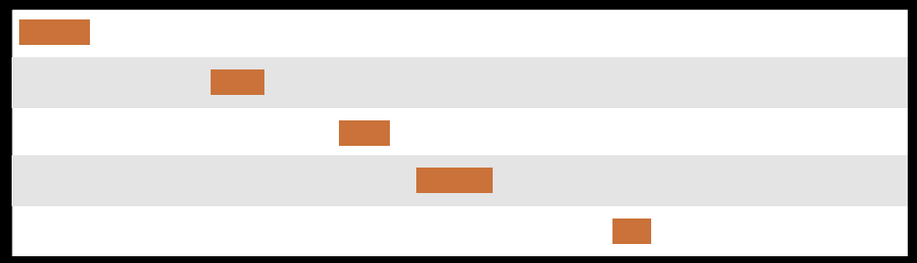


Operate on Genomic Intervals → **Join**

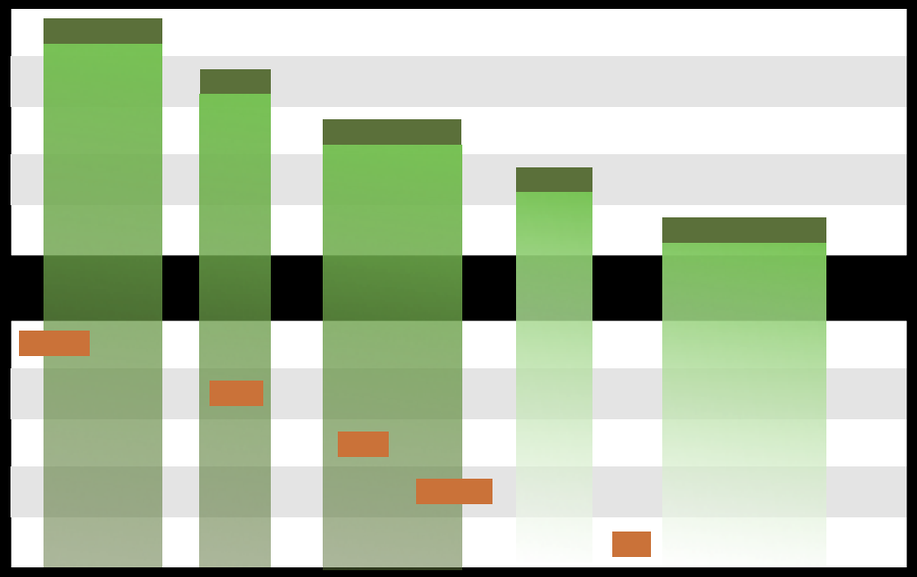
(Identify which genes/exons have Repeats)



Exons



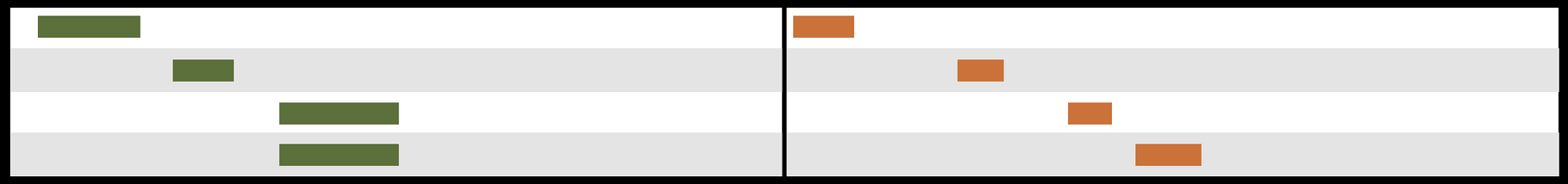
Repeats



Exons

Repeats

Overlap pairings

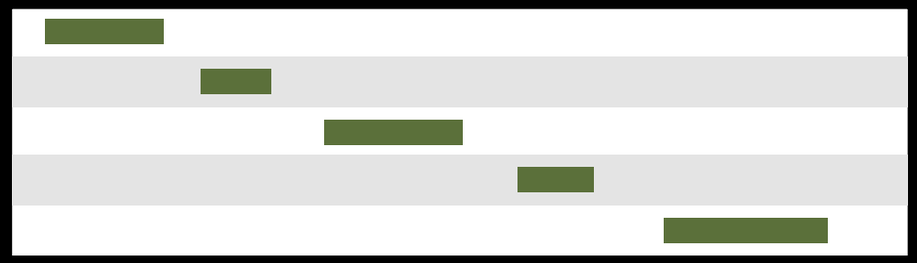


Exon overlap counts

Join, Subtract, and Group → Group

(Count Repeats per exon)

Exon overlap counts

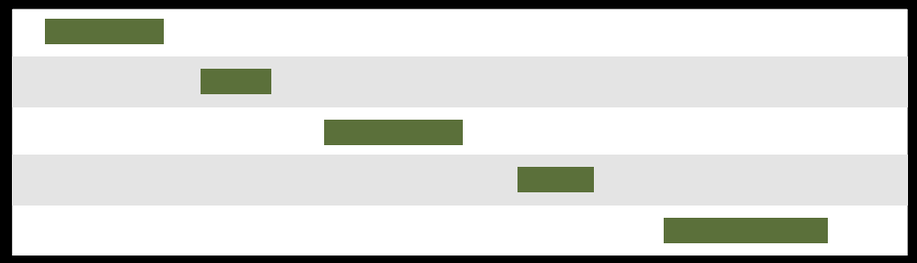


Exons

We've answered our question, but we can do better.

Incorporate the overlap count with rest of Exon information

Exon overlap counts

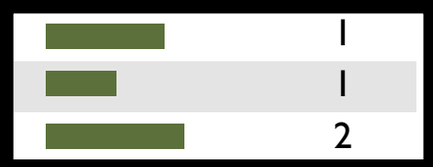


Exons

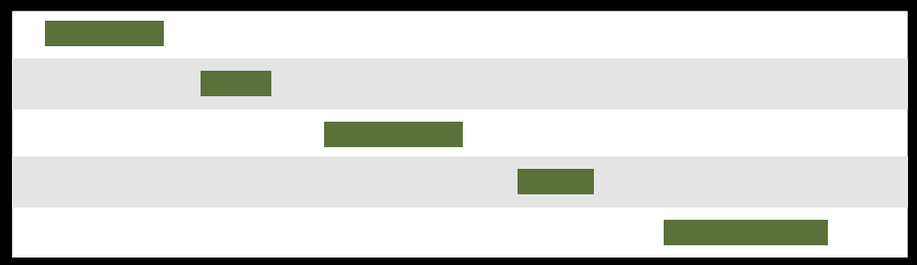
Join on exon name

Join, Subtract, and Group → Join

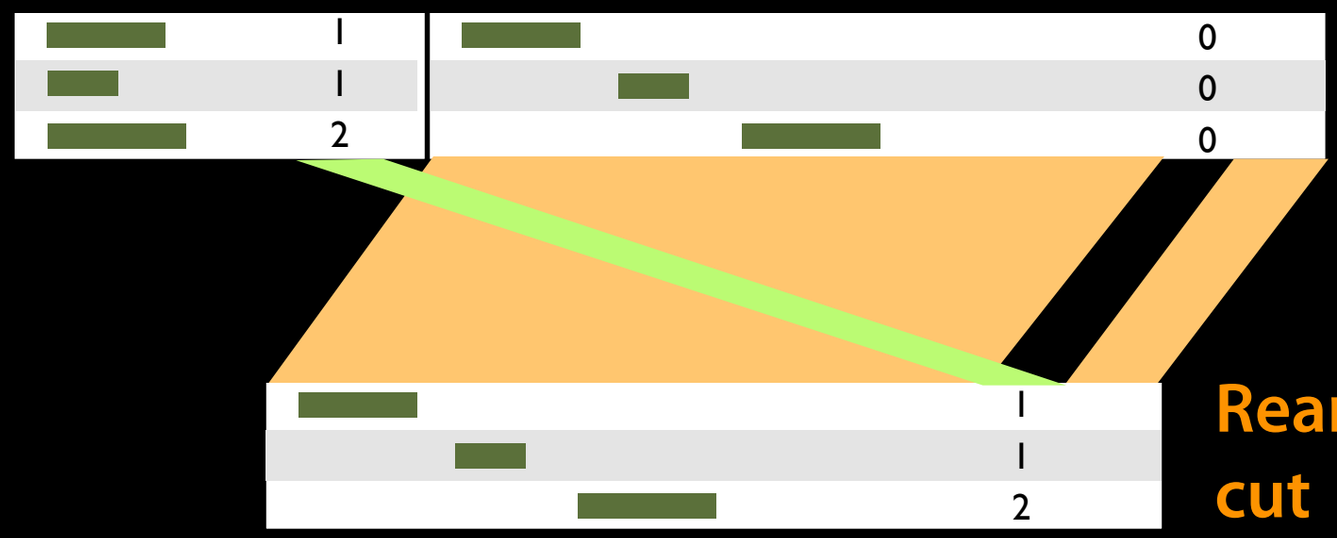
(Incorporate the overlap count with rest of Exon information)



Exon overlap counts



Exons



Join on exon name

Rearrange columns w/ cut

Text Manipulation → Cut

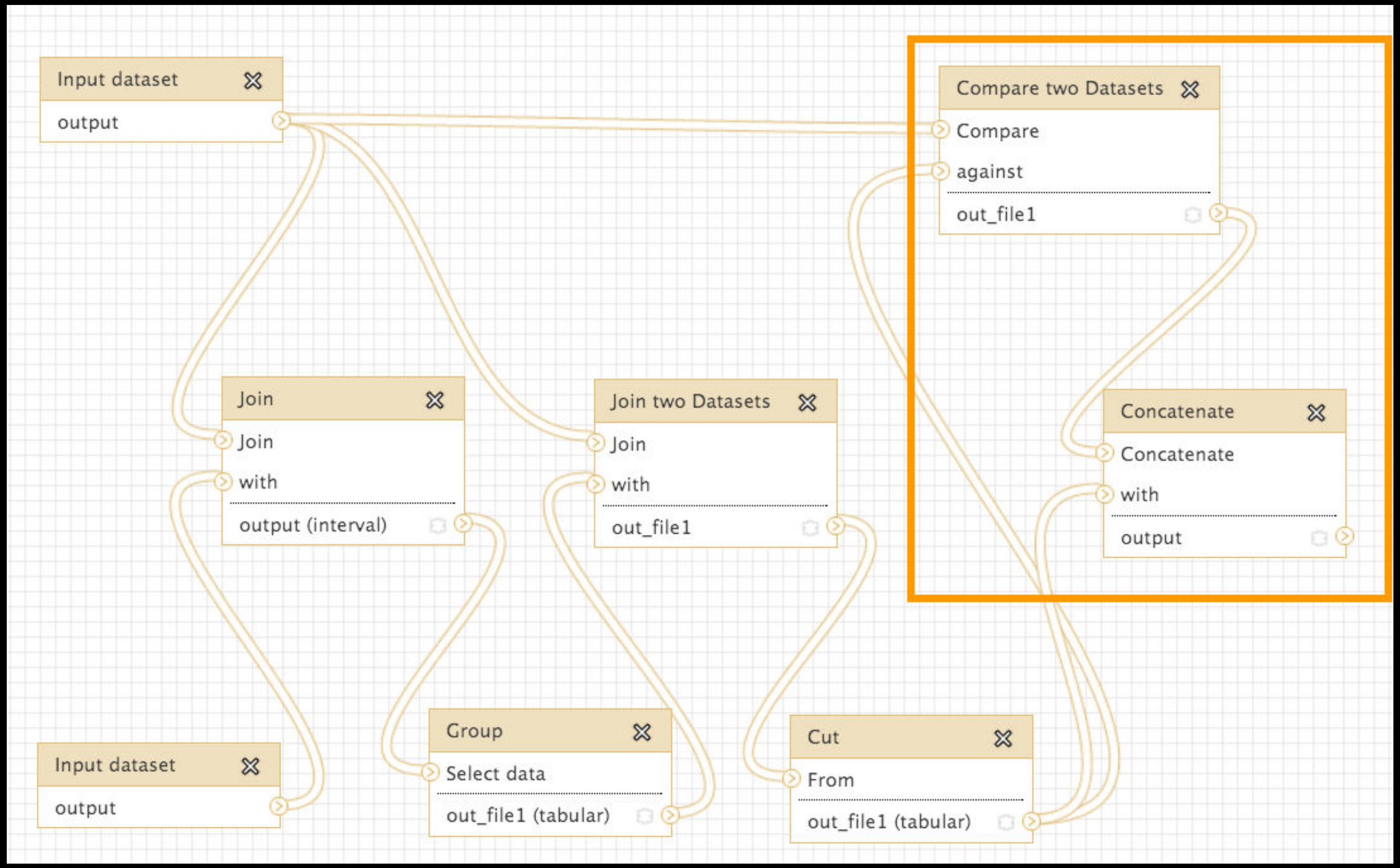
(Incorporate the overlap count with rest of Exon information)

Genes & Repeats: Exercise

Include exons with no overlaps in final output.
Set the score for these to 0.

Everything you need will be in the toolboxes we used in the first Exon-Repeats exercise.

One Possible Solution



Solution from Stanford Kwenda and Caron Griffiths in Pretoria.

Takes advantage of the fact that Exons already have 0 scores.

Basic Analysis: Further reading & Resources

http://usegalaxy.org/galaxy101

https://vimeo.com/76343659

Monday's Agenda

9:00 Introduction to Cloud Computing and Using the Nuvem Cloud

20 minute Break at around 10:20

12:00 Lunch

2:00 Galaxy Project Introduction

2:15 Introduction to using Galaxy

20 minute Break at around 3:20

5:00 Done

Thanks

Dave Clements

Galaxy Project

Johns Hopkins University

outreach@galaxyproject.org

Some Galaxy Terminology

Dataset:

Any input, output or intermediate set of data + metadata

History:

A series of inputs, analysis steps, intermediate datasets, and outputs

Workflow:

A series of analysis steps
Can be repeated with different data

Exons and Repeats *History* → Reusable *Workflow?*

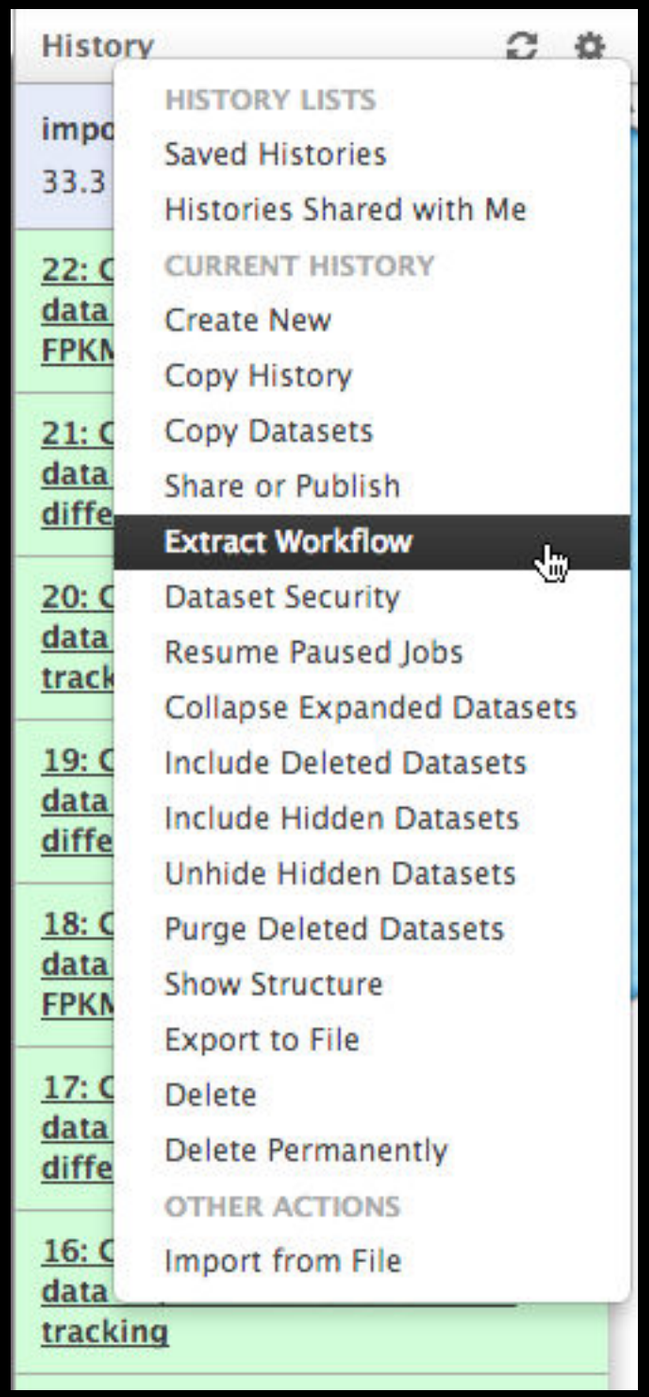
- The analysis we just finished was about
  - Human chr22
  - Overlap between exons and Repeats
- But, ...
  - there is nothing inherent in the analysis about humans, exons or repeats
  - It is a series of steps that sets the score of one set of features to the number of overlaps from another set of features.

Create a Workflow from a History

Extract Workflow from history

Create a workflow from this history.
Edit it to make some things clearer.

(cog) → Extract Workflow



Run / test it

Guided: rerun with same inputs

Did that work?

On your own:

Count # of exons in each Repeat

Did that work? *Why not?*

Edit workflow: doc assumptions

More Galaxy Terminology

Share:

Make something available to someone else

Publish:

Make something available to everyone

Galaxy Page:

Analysis documentation within Galaxy; easy to embed any Galaxy object

Let's all share...

Sharing & Publishing enables Reproducibility

Reproducibility: Everybody talks about it, but ...

Galaxy aims to push the goal of reproducibility from the bench to the bioinformatics realm

All analysis in Galaxy is recorded without any extra effort from the user.

Histories, workflows, visualizations and ***pages*** can be shared with others or published to the world.

Sharing & Publishing enables Reproducibility

CSH PRESS GENOME RESEARCH

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]

OPEN ACCESS ARTICLE

This Article

Published in Advance October 9, 2009, doi: 10.1101/gr.094508.109

- » Abstract *Free*

Current Issue

October 2010, 20 (10)

Sharing & Publishing enables Reproducibility

CSH PRESS GENOME RESEARCH

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]

OPEN ACCESS ARTICLE

This Article

Published in Advance October 9, 2009, doi: 10.1101/gr.094508.109

- » Abstract *Free*

Current Issue

October 2010, 20 (10)

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.]

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Help | User | Using

Published Pages | aun1 | Windshield Splatter

Windshield splatter analysis with the Galaxy metagenomic pipeline: A live supplement

SERGEI KOSAKOVSKY POND[1,2,*], SAMIR WADHAWAN[3,6*], FRANCESCA CHIAROMONTE[4], GURUPRASAD ANANDA[1,3], WEN-YU CHUNG[1,3,7], JAMES TAYLOR[1,5], ANTON NEKRUTENKO[1,3] and THE GALAXY TEAM[1*]

Correspondence should addressed to SKP, JT, or AN.

How to use this document

This document is a live copy of supplementary materials for the manuscript. It provides access to the **exact** analyses and workflows discussed in the paper, so you can play with them by re-running, changing parameters, or even applying them to your own data. Specifically, we provide the two histories and one workflow found below. You can view these items by clicking on their name to expand them. You can also import these items into your Galaxy workspace and start using them; click on the green plus to import an item. To import workflows you must create a Galaxy account (unless you already have one) - a hassle-free procedure where you are only asked for a username and password.

This is the Galaxy history detailing the comparison of our pipeline to MEGAN:

Galaxy History | Galaxy vs MEGAN
Comparison of Galaxy vs. MEGAN pipeline.

This is the Galaxy history showing a generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3A**):

Galaxy History | metagenomic analysis

This is the Galaxy workflow for generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3B**):

Galaxy Workflow | metagenomic analysis
Generic workflow for performing a metagenomic analysis on NGS data.

Accessing the Data

Windshield Splatter datasets analyzed in this manuscript can be accessed through this Galaxy Library. From

About this Page

Author

aun1

Related Pages

All published pages
Published pages by aun1

Rating

Community
(6 ratings, 5.0 average)

Tags

Community:
paper | galaxy | megan

http://usegalaxy.org/u/aun1/p/windshield-splatter

Sharing for Galaxy Administrators Too

Data Libraries

Make data easy to find

Genome Builds

Care about a particular subset of life?

Galaxy Tool Shed

Wrapping tools and datatypes