modENCODE Galaxy: Uniform ChIP-Seq Processing Tools for modENCODE and ENCODE Data

Quang M Trinh
Ontario Institute for Cancer Research
qtrinh@oicr.on.ca

Outline

- **M**odel **O**rganism **ENC**yclopedia **O**f **D**NA **E**lements (modENCODE) project & mandates for the modENCODE **D**ata **C**oordinating **C**enter (DCC)

- modENCODE data & Galaxy on Amazon Cloud

- Uniform Processing/Peak calling pipeline for modENCODE & ENCODE (**ENC**yclopedia **O**f **D**NA **E**lements) data using Galaxy

Model **O**rganism **ENC**yclopedia **O**f **D**NA **E**lements (modENCODE) Project

- Funding by the **N**ational **I**nstitutes of **H**ealths (NIH)
  - http://www.genome.gov/modencode/

- Aim of modENCODE is to provide a comprehensive encyclopedia of functional genomics for both worm and fly
  - 11 groups of data providers
  - 1 analysis center
  - 1 **D**ata **C**oordinating **C**enter (DCC)

Mandates for the DCC

- *Collect*, *validate*, and *release* data submitted from the 11 groups of data providers
- *Collection*
  - Data upload via a website or an ftp site
- *Validation*
  - uses controlled vocabularies to describe data and metadata
  - QC to ensure consistency and completeness of submission
  - Integrates data
- *Release*
  - Over 10 TB of data publicly available on faceted browser, modmine, and clouds (Amazon & Bionimbus)

modENCODE Data on Amazon Cloud

- The entire set of modENCODE data is on Amazon Cloud as a list of snapshots

- Custom modENCODE **A**mazon **M**achine **I**mage (AMI) with the entire data pre-mounted for convenience.
  - Users can also select and mount any of the snapshots
    - automated

- Step-by-step instructions on how to use the custom AMI or how to mount modENCODE data snapshots
  - http://data.modencode.org/modencode-cloud.html

Main Challenges With Accessing the Entire modENCODE Data Set

- Downloading the entire data set (over 10TB) from Amazon Cloud will take a while
  - Additional local disks & computing resources are needed

- Tools for analysis
  - Setup tools locally will also take a while

Our Solution: modENCODE Galaxy on Amazon Cloud

- Bring tools and analysis to our data on Amazon Cloud

- Build and integrate tools and workflows to Galaxy on Amazon Cloud
  - Automate Galaxy launching on Amazon Cloud and installations of modENCODE tools on Galaxy and Galaxy cluster

modENCODE Galaxy on Amazon Cloud

- Put together by our co-op students
  - Ravpreet Setia, Fei-Yang (Arthur) Jan, Ziru Zhou, Karming Chu
- https://github.com/modENCODE-DCC/Galaxy
  - Scripts to launch Galaxy and install tools and their dependencies
  - Peak calling and QC tools
    - SPP, macs2, peak ranger, and bamedit
  - Workflows
    - Uniform processing/peak calling pipeline for modENCODE and ENCODE data
      - Worm, fly, human, and mouse
  - Enable users to import modENCODE data directly from the faceted browser to Galaxy
  - Step-by-step documentations

Simple Steps to Launch modENCODE Galaxy & Installations of Tools

- Setup Amazon credentials and environments (one time)
- Setup Galaxy config.txt (one time)
- Launch Galaxy on Amazon Cloud
  - *bin/modENCODE_galaxy_create.pl config.txt*
- Setup Galaxy Cluster using CloudMan console
- Setup modENCODE tools for Galaxy
  - Install tools in parallel using *bin/auto_install.pl*

Setup Amazon Credentials and Environments (env.sh)

```
# set JAVA_HOME
export JAVA_HOME=your_JAVA_HOME_PATH

# set your AWS credentials
export AWS_ACCESS_KEY=your_AWS_ACCESS_KEY_ID
export AWS_SECRET_KEY=your_AWS_SECRET_KEY

##############################################
# no changes are needed below this line
##############################################

# set EC2_HOME and add $EC2_HOME/bin to $PATH
export EC2_HOME=`pwd`/external_tools/ec2-api-tools-1.6.1.4
export PATH=$PATH:$EC2_HOME/bin
```

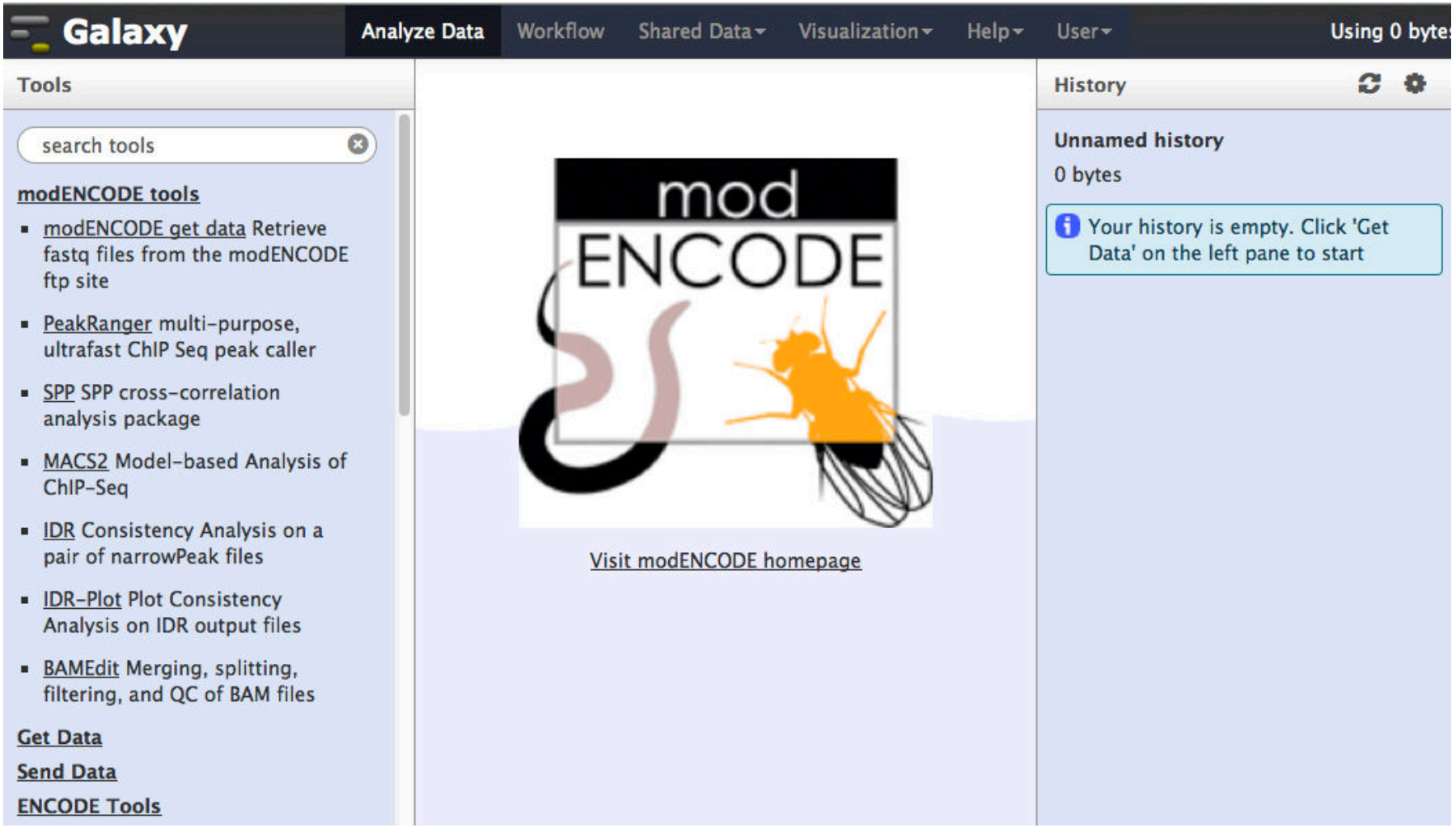
Setup Configurations (config.txt)

```
# CloudMan password.  Default is 'galaxy_123'.
CLOUDMAN_PASSWORD: galaxy_123

#
# key pair, security group, instance name, cluster name for Galaxy instance
KEY_PAIR: YOUR_NAME_modENCODE_Galaxy_Key
SECURITY_GROUP: YOUR_NAME_modENCODE_Galaxy_Group
INSTANCE_NAME: YOUR_NAME_modENCODE_Galaxy_Instance
CLUSTER_NAME: YOUR_NAME_modENCODE_Galaxy_Cluster

#
# Galaxy Amazon Machine Image (AMI) ID
# Also, see http://wiki.g2.bx.psu.edu/CloudMan/AWS/GettingStarted
AMI: ami-da58aab3
#
# Galaxy instance type and where to launch Galaxy
INSTANCE_TYPE: m1.medium
REGION: us-east-1
AVAILABILITY_ZONE: us-east-1a
```

New Galaxy AMI is available on June 29 – see email from Enis Afgan to galaxy-dev


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 0 byte
Tools
search tools
modENCODE tools
modENCODE get data Retrieve fastq files from the modENCODE ftp site
PeakRanger multi-purpose, ultrafast ChIP Seq peak caller
SPP SPP cross-correlation analysis package
MACS2 Model-based Analysis of ChIP-Seq
IDR Consistency Analysis on a pair of narrowPeak files
IDR-Plot Plot Consistency Analysis on IDR output files
BAMEdit Merging, splitting, filtering, and QC of BAM files
Get Data
Send Data
ENCODE Tools
mod
ENCODE
Visit modENCODE homepage
History
Unnamed history
0 bytes
Your history is empty. Click 'Get Data' on the left pane to start

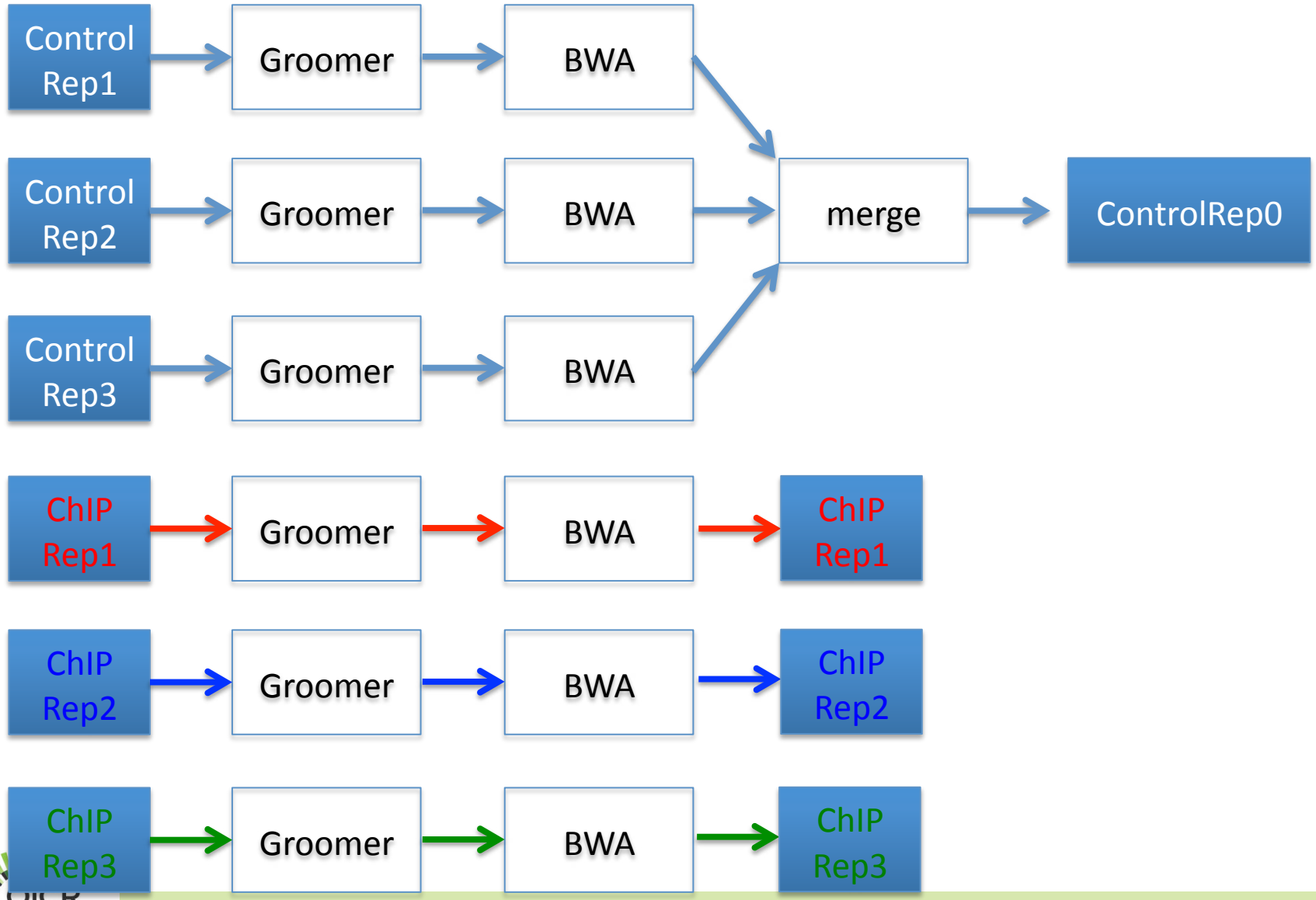
Galaxy Tool Shed

Name ↓	Synopsis	Metadata Revisions	Tools Verified	Category
bamedit	Merging, splitting, filtering, and QC of BAM files (bamedit)	15:eb166cebbe3c	no	• SAM
idr_package	consistency analysis on a pair of Peak files (idr)	20:6f6a9fbe264e	no	• Sequence Analysis
macs2	Model-based Analysis of ChIP-Seq (macs2). **NOTE: This package requires Python 2.7.X and numpy (>=1.3) installed on all cluster nodes.	16:14f378e35191	no	• Sequence Analysis
peakranger	multi-purpose, ultrafast ChIP-Seq peak caller (peakranger)	21:d6b1bb81fa6c	no	• Sequence Analysis
spp_package	cross-correlation analysis package (spp)	13:64f2784d471f	no	• Sequence Analysis

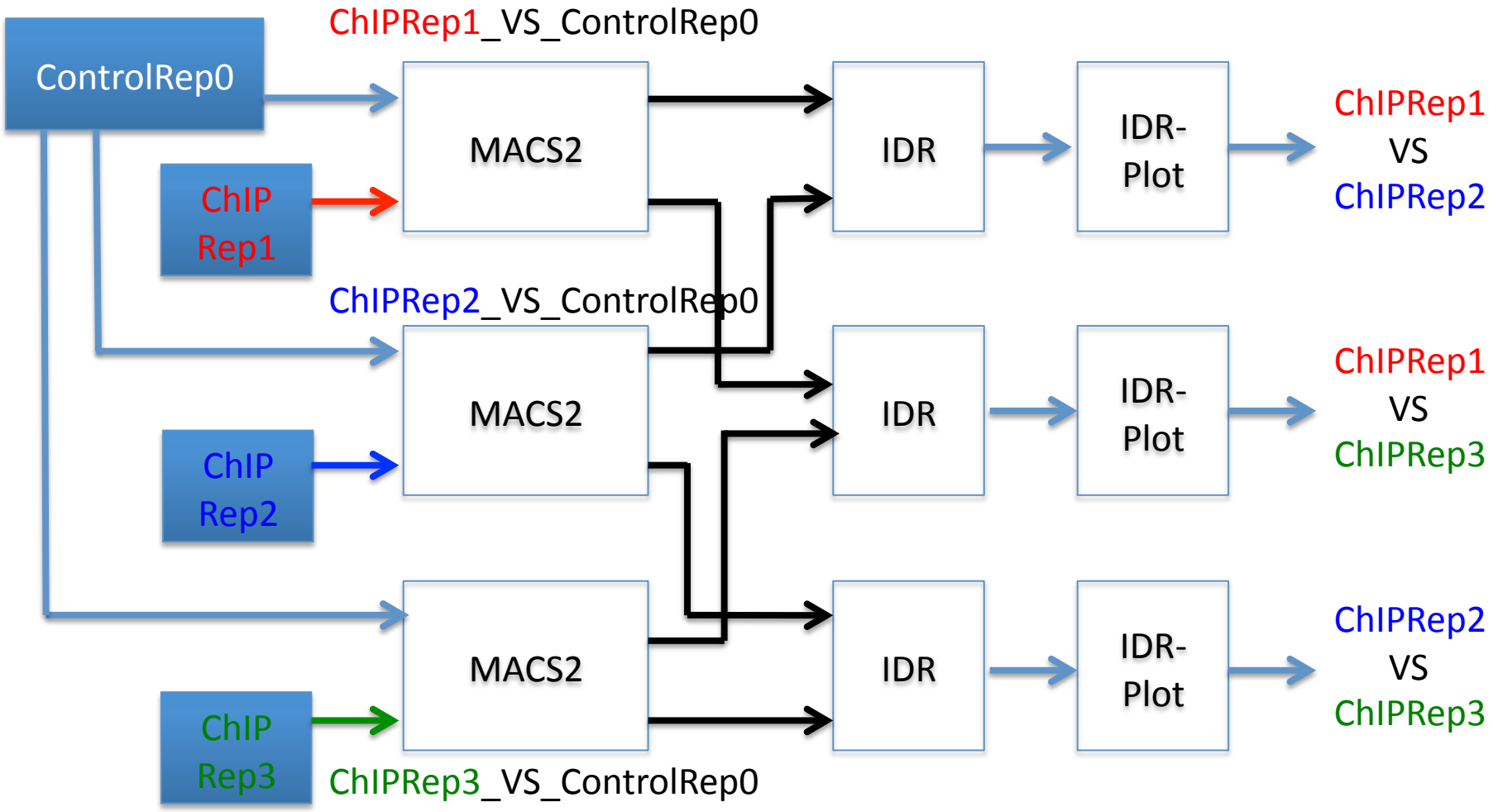
Uniform Processing/Peak Calling Pipeline

- A uniform pipeline for calling peaks and ranking reproducibility between replicates for ChIP-seq data
- Used by both modENCODE and ENCODE communities for human, mouse, worm, and fly
- Begins with raw FASTQ files and ends with peak files in BED format and pdf plots of consistency comparisons between replicates.

Uniform Processing/Peak Calling Pipeline for 3 replicates



Uniform Processing/Peak Calling Pipeline for 3 replicates (cont'd)



Uniform Processing/Peak Calling Workflows

- https://github.com/modENCODE-DCC/Galaxy/tree/master/workflows
  - 3-replicate and 2-replicate workflows

Conclusions

- Galaxy is a great platform for data analysis
- We chose Galaxy because of its availability, functionality, and ease of result reproducibility
- Integrated modENCODE tools & workflows with Galaxy on Amazon Cloud
  - Works great with the entire modENCODE data set on Amazon Cloud

- For more info, see
  - https://github.com/modENCODE-DCC/Galaxy

Acknowledgments

- Co-op students
  - Rav Setia
  - Fei-Yang (Arthur) Jen
  - Ziru Zhou
  - Karming Chu
- modENCODE DCC Data Wranglers
  - Marc Perry
  - Ellen Kephart
  - Sergio Contrino
  - Peter Ruzanov
  - Lincoln Stein (PI)

Funding provided by