Galaxy Workshop

University of Illinois
16-17 October 2012

Dave Clements
Emory University

http://galaxyproject.org/

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Roy J. Carver Biotechnology Center

Institute for Genomic Biology

NCSA

Galaxy

Acknowledgements

Radhika Khetani
Christopher Fields
Luigi Marini
Victor Jongeneel

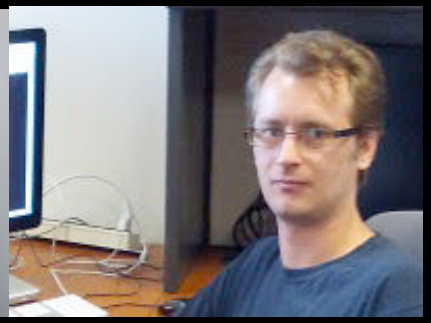
HPCBio
IGB
NCSA
Joy J. Carver Biotechnology Center
University of Illinois


Enis Afgan


Guru Ananda


Dannon Baker


Dan Blankenberg


Dave Bouvier


Dave Clements


Nate Coraor


Carl Eberhard


Jeremy Goecks


Nuwan Goonasekera


Jen Jackson


Greg von Kuster


Ross Lazarus


Rémi Marenco


Scott McManus


Anton Nekrutenko

James Taylor

The Galaxy Team

http://galaxyproject.org/wiki/GalaxyTeam

Agenda: Day 1

Welcome
Galaxy @ UIUC
Basic Analyses with Galaxy
Basic Analysis into Reusable Workflows
ChIP-Seq Example
Galaxy Project Overview

Coffee breaks somewhere in there

On Wiki: Documents/Presentations/2012_UIUC...

Goals for this workshop

1. Introduce Galaxy
2. Introduce Common Bioinformatics Formats
3. Hands-on experience:
   - Load and integrate data from online resources
   - Perform bioinformatics analysis with Galaxy
   - Save, share, describe and publish your analysis
   - Visualize your results

This workshop will not cover details of how the tools are implemented or new algorithm designs or which assembler or mapper or ... is best for you.

Agenda: Day 1

Welcome

Galaxy @ UIUC

Basic Analyses with Galaxy

Basic Analysis into Reusable Workflows

ChIP-Seq Example

Agenda: Day 1

Welcome

Galaxy @ UIUC

Basic Analyses with Galaxy

Basic Analysis into Reusable Workflows

ChIP-Seq Example

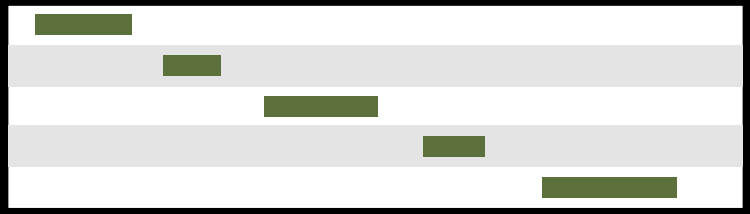
Hands On: Basic Analysis

On pig chromosome 18,
which coding exons have the most
repeats in them?

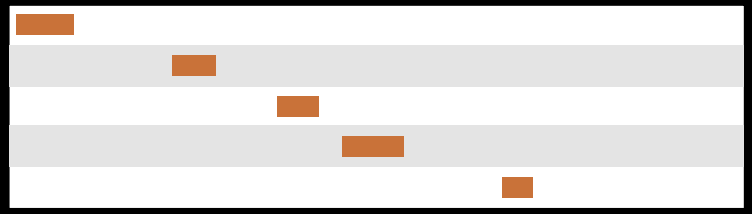
(~ http://usegalaxy.org/galaxy101)

Repetitious Pigs: A Rough Plan

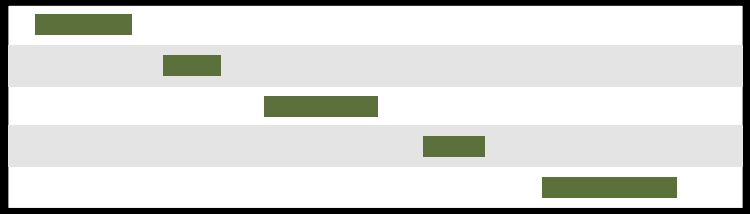
- Get some data (and explain BED)
  - Coding exons on chromosome 18
  - Repeats on chromosome 18
- Mess with it (and explain Galaxy operations)
  - Identify which exons have repeats
  - Count repeats per exon
- Visualize our results



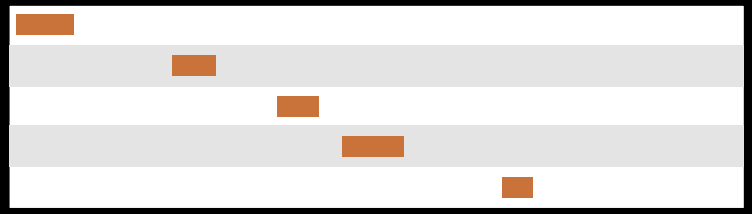
Exons, from UCSC



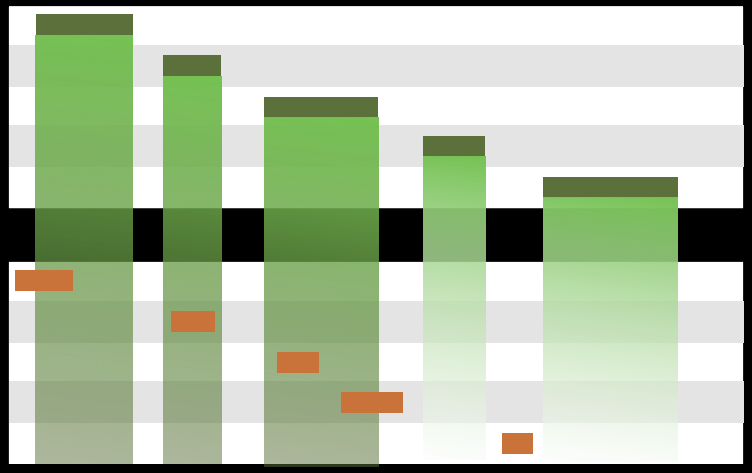
Repeats, from UCSC



Exons, from UCSC



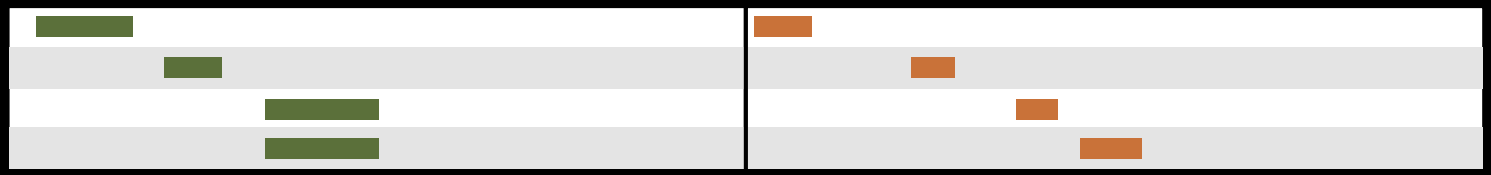
Repeats, from UCSC

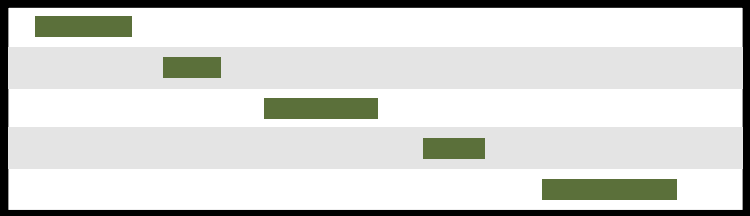


Exons, from UCSC

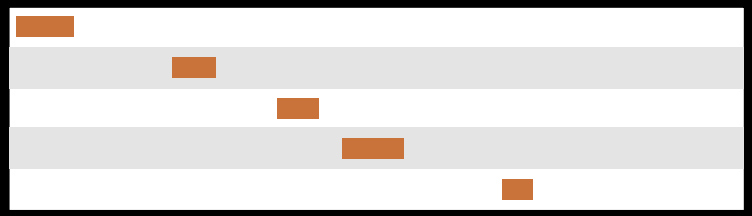
Repeats, from UCSC

Overlap pairings

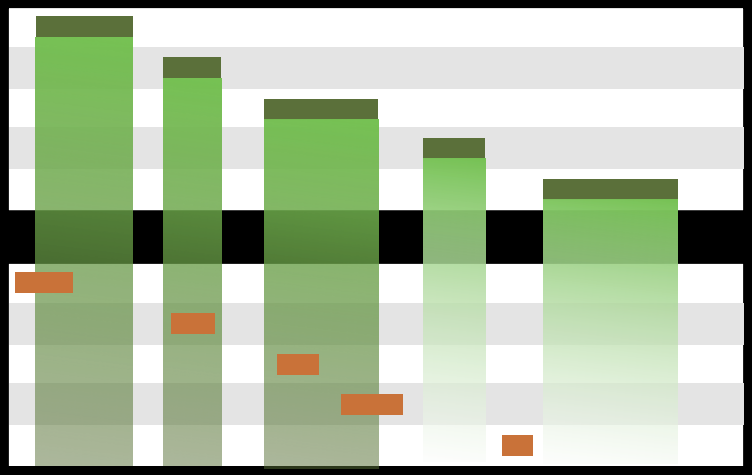




Exons, from UCSC



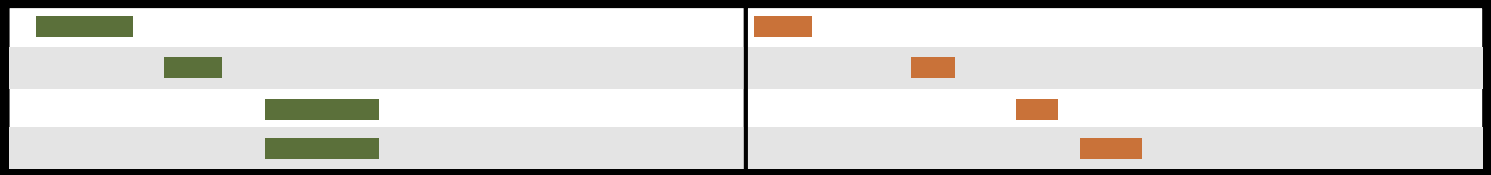
Repeats, from UCSC



Exons, from UCSC

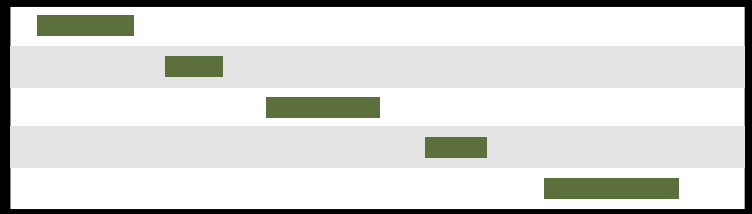
Repeats, from UCSC

Overlap pairings



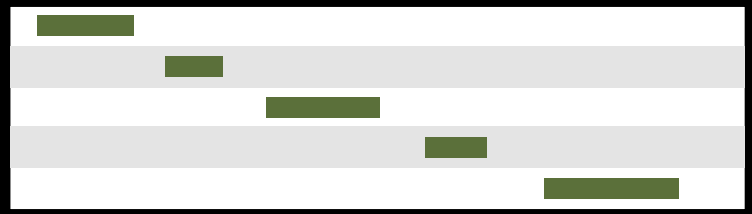
Exon overlap counts

Exon overlap counts



Exons, from UCSC

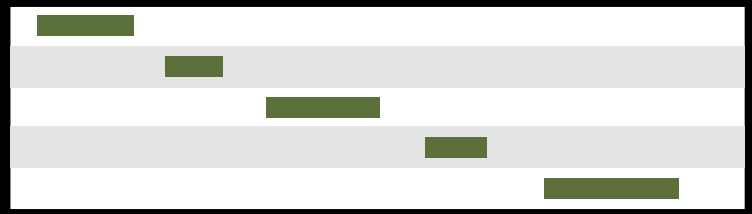
Exon overlap counts



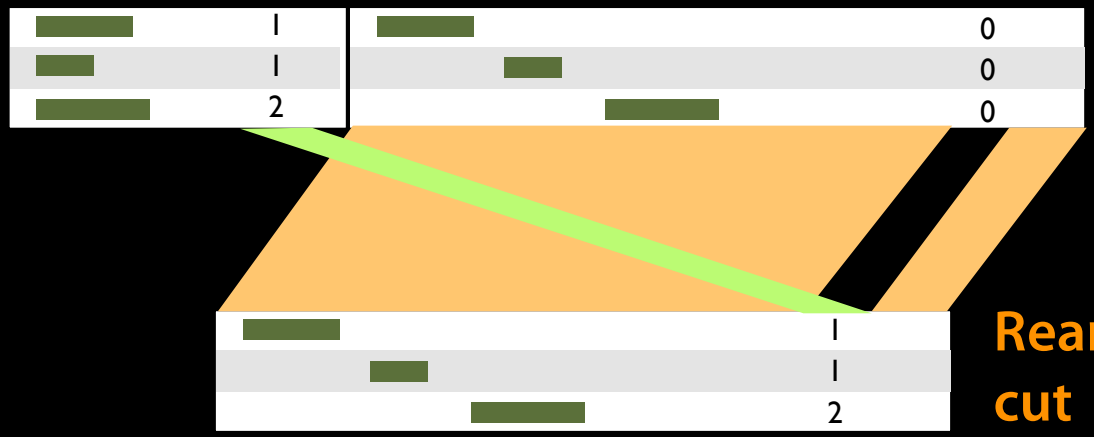
Exons, from UCSC

Join on exon name

Exon overlap counts



Exons, from UCSC



Join on exon name

Rearrange columns w/ cut

Agenda: Day 1

Welcome
Galaxy @ UIUC
Basic Analyses with Galaxy
Basic Analysis into Reusable Workflows
ChIP-Seq Example

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

Reuse: Data & Analyses

Histories: Data

Datasets from previous histories can be imported into current one.

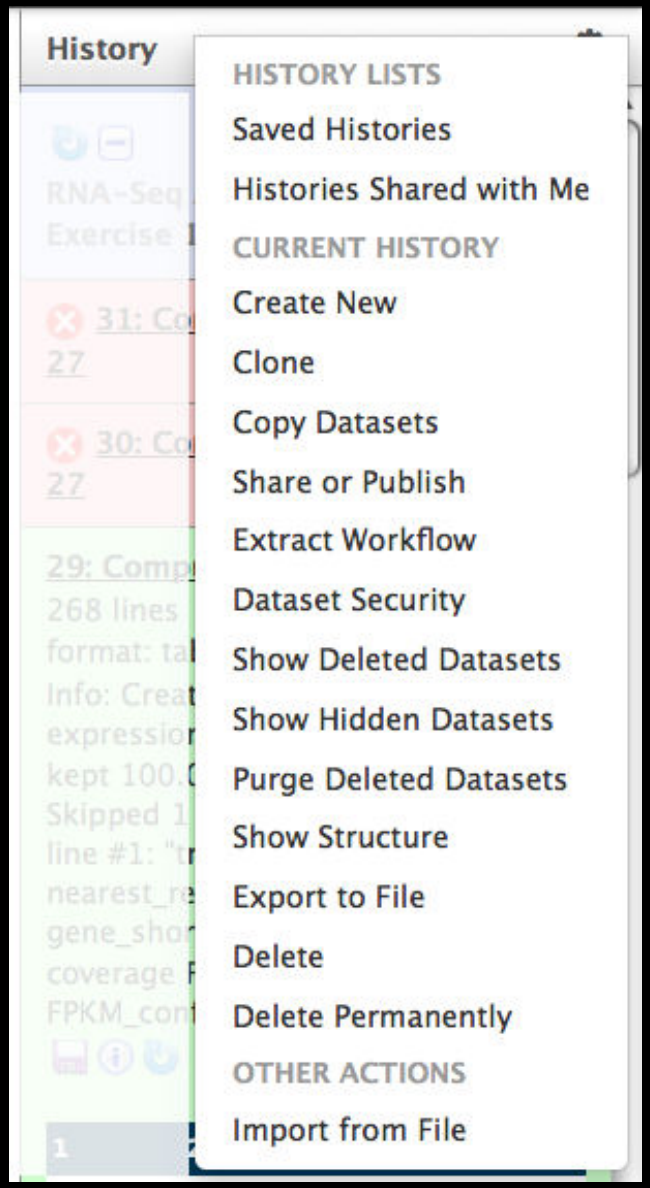
Resume any previous history

Current history can be cloned

Workflows: Analyses

Can be extracted from any history

Allows you rerun analysis with different inputs, settings



Repetitious Pigs *History* → Reusable *Workflow?*

- The analysis we just finished was about
  - Pig chromosome 18
  - Overlap between exons and repeats
- But, ...
  - there is nothing inherently in the analysis about pigs, chromosomes, exons or repeats
  - It is a series of steps that sets the score of one set of features to the number of overlaps each feature has in the other set of features.

Reuse: Create a generic *Overlap* Workflow

Extract Workflow from history

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

Run / test it

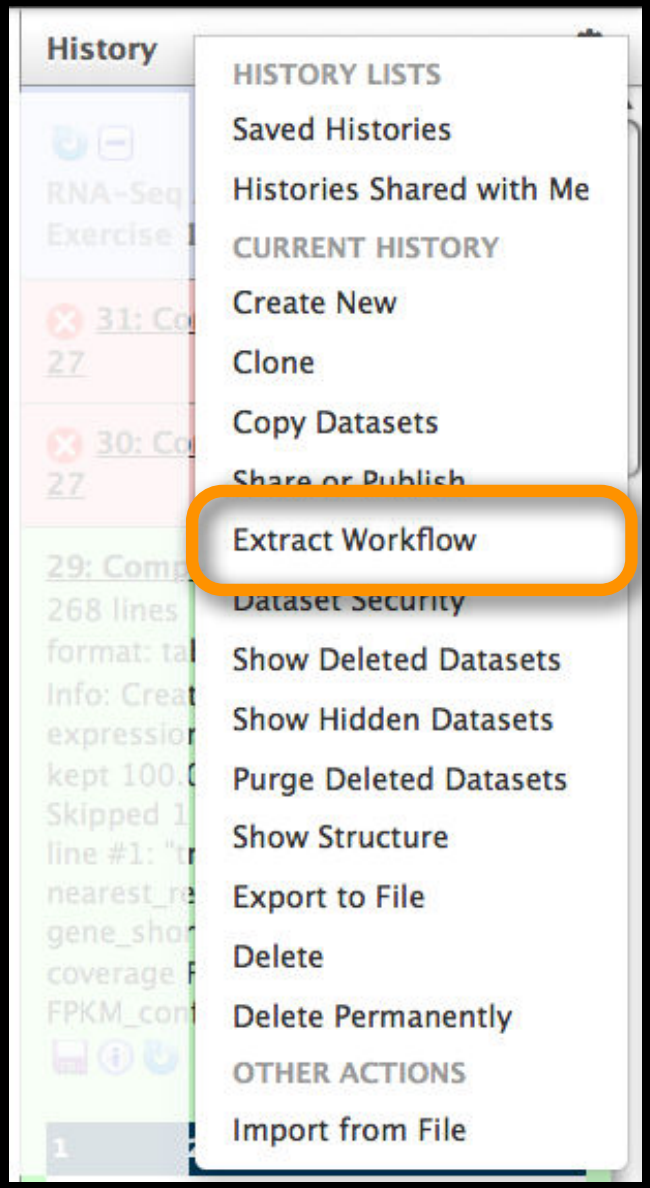
Guided: rerun with same inputs

On your own:

Count # CpG islands overlapping with each exon. Did that work?

On your own:

Count # of exons in each repeat
Did that work? *Why not?*
Edit workflow: doc assumptions



Hands On: Basic Analysis ...
A Simple Change ...

On pig chromosome 18,
which coding exons (GTF format)
have the most repeats (BED format)
in them?

Repetitious Pigs: GTF and BED

- Get the GTF from UCSC
  - *Hmm*: There is no "coding exons" choice w/ GTF

- Points you may eventually ponder
  - Do we care about *coding exons* versus *exons?*
  - Do we care about *exon names, gene names, transcript names, or just coordinates?*
  - *Can the same approach even work with GTF?*

Agenda: Day 1

Welcome
Galaxy @ UIUC
Basic Analyses with Galaxy
Basic Analysis into Reusable Workflows
ChIP-Seq Example

ChIP-Seq Exercise

- Identify zinc-finger CTCF transcription factor tags in mouse
- Exercise and data from
  - Hillman-Jackson, *et al.*, "Using Galaxy to Perform Large-Scale Interactive Data Analyses" *Curr. Protoc. Bioinform*. 38:10.5.1-10.5.47;
  - ENCODE transcription factor binding experiment: http://bit.ly/QmD6Nk. Raw original data generated & analyzed at Michael Snyder's lab, Stanford University, and Sherman Weissman's Lab, Yale University.
- We'll use build mm10 and datasets that have been prescreened to mostly map to chr19
- All datasets are FASTQ

What is FASTQ?

Specifies sequence (FASTA) and quality scores (PHRED)

Text format, 4 lines per entry

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

FASTQ is such a cool standard, that one version is not enough!

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.......................................................
 ........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 ...............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.......................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
   with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
   (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

http://en.wikipedia.org/wiki/FASTQ_format

ChIP-Seq Exercise: A Plan

- Get input datasets; control and tags
- Groom the datasets into FASTQSanger format
- Look at quality
- Trim as we see fit
- Map the reads to genome using Bowtie
- Call peaks with MACS (Model-based Analysis of ChIP-seq)

ChIP-Seq Exercise: A Plan

- Get input datasets; control and tags
  - Shared Data → Published Histories
    - ChIP-Seq basic datasets (clements)
      - Import History

ChIP-Seq Exercise: A Plan

- Get input datasets; control and tags
- Groom the datasets into FASTQSanger format
  - NGS: QC and manipulation → FASTQ Groomer
    - Input FASTQ quality scores type: Illumina 1.3-1.7
    - Run on both datasets

ChIP-Seq Exercise: A Plan

- Get input datasets; control and tags
- Groom the datasets into FASTQSanger format
- Look at quality: Option 1
  - NGS QC and Manipulation →
    - Compute Quality Statistics
    - Draw quality score boxplot
  - Get stats in text and graphic format
  - No control over how it is calculated or presented

ChIP-Seq Exercise: A Plan

- Get input datasets; control and tags
- Groom the datasets into FASTQSanger format
- Look at quality: Option 2
  - NGS QC and Manipulation → FastQ Summary Statistics
  - Graph / Display Data → Boxplot of quality statistics
  - Gives you a lot of control over what the box plot looks like, but no additional information

ChIP-Seq Exercise: A Plan

- Get input datasets; control and tags
- Groom the datasets into FASTQSanger format
- Look at quality: Option 3
  - NGS QC and Manipulation → Fastqc
  - Gives you a lot a lot more information but no control over how it is calculated or presented.

ChIP-Seq Exercise: A Plan

- ...
- Look at quality
- Trim as we see fit: Option 1
  - NGS QC and Manipulation → FASTQ Trimmer by column
  - Trim same number of columns from every record
  - Can specify different trim for 5′ and 3′ ends

ChIP-Seq Exercise: A Plan

- ...
- Look at quality
- ~~Trim~~ Filter as we see fit: Option 2
  - NGS QC and Manipulation → Filter FASTQ reads by quality score and length
  - Keep or discard whole reads at a time
  - Can have different thresholds for different regions of the reads.
  - Keeps original read length.

ChIP-Seq Exercise: A Plan

- Look at quality
- Trim as we see fit: Option 3
  - NGS QC and Manipulation → FASTQ Quality Trimmer by sliding window
  - Trim from both ends, using sliding windows, until you hit a high-quality section.
  - Produces variable length reads

Read length is only used for building model to predict fragment length. So if you set fragment size by yourself, it really doesn't matter how long each read is. Also, in MACS models, only 5' ends of each read (only talking about single end sequencing here), where ultrasound or enzymes cut DNA, are informative, for both fragment size prediction and peak calling. So you can still try to let MACS predict fragment size by setting a fixed read length. I think the current cross-correlation way in MACS v2 can give a more stable result than the previous way in MACS v1 just measuring distance between plus and minus read pileup summits.

Tao Liu https://groups.google.com/forum/?fromgroups=#!topic/macs-announcement/A_Rf0eQ_BLU

ChIP-Seq Exercise: A Plan

- Get input datasets; control and tags
- Groom the datasets into FASTQSanger format
- Look at quality
- Trim as we see fit
- Map the reads to genome using Bowtie
  - NGS: Mapping → Bowtie2
    - Library: Single-end
    - Run on both control and tag files
    - Use mm10 as the reference genome

ChIP-Seq Exercise: A Plan

- Get input datasets; control and tags
- Groom the datasets into FASTQSanger format
- Look at quality
- Trim as we see fit
- Map the reads to genome using Bowtie
- Call peaks with MACS (Model-based Analysis of ChIP-seq)

Model-based Analysis of ChIP-seq (MACS)

Open Access

Method

Model-based Analysis of ChIP-Seq (MACS)

Yong Zhang[¤*], Tao Liu[¤*], Clifford A Meyer[*], Jérôme Eeckhoute[†], David S Johnson[‡], Bradley E Bernstein[§¶], Chad Nusbaum[¶], Richard M Myers[¥], Myles Brown[†], Wei Li[#] and X Shirley Liu[*]

Addresses: [*]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, 44 Binney Street, Boston, MA 02115, USA. [†]Division of Molecular and Cellular Oncology, Department of Medical Oncology, Dana-Farber Cancer Institute and Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 44 Binney Street, Boston, MA 02115, USA. [‡]Gene Security Network, Inc., 2686 Middlefield Road, Redwood City, CA 94063, USA. [§]Molecular Pathology Unit and Center for Cancer Research, Massachusetts General Hospital and Department of Pathology, Harvard Medical School, 13th Street, Charlestown, MA 02129, USA. [¶]Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA, 02142, USA. [¥]Department of Genetics, Stanford University Medical Center, Stanford, CA 94305, USA. [#]Division of Biostatistics, Dan L Duncan Cancer Center, Department of Molecular and Cellular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA.

[¤] These authors contributed equally to this work.

Correspondence: Wei Li. Email: wl1@bcm.edu. X Shirley Liu. Email: xsliu@jimmy.harvard.edu

Published: 17 September 2008

Genome ***Biology*** 2008, **9**:R137 (doi:10.1186/gb-2008-9-9-r137)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/9/R137

Received: 4 August 2008
Revised: 3 September 2008
Accepted: 17 September 2008

ChIP-Seq Exercise: A Plan

- Call peaks with MACS (Model-based Analysis of ChIP-seq)
  - NGS: Peak Calling → MACS
  - Set ChIP-Seq Tag File and ChIP-Seq Control File
  - Set Effective genome size: 1.87e+9
  - Set Tag size to 36 (still correct?)
  - Set Select the regions with MFOLD: 32
  - Set Parse xls files into distinct interval files
  - Save shifted raw tag count at every bp into a wiggle file
  - Resolution for saving wiggle files: 1 (or 10?)

That's a lot of knobs to set. Get used to it.

Using MACS to Identify Peaks from ChIP-Seq Data

Jianxing Feng,[1] Tao Liu,[2] and Yong Zhang[1]

[1]School of Life Sciences and Technology, Tongji University, Shanghai, China
[2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, Massachusetts

ABSTRA

Model-bas
Shirley Li
karyotes, e
sites and I
control sa
information on how to use MACS to identify either the binding sites of a transcription factor or the enriched regions of a histone modification with broad peaks. Furthermore, the basic ideas for the MACS algorithm and its appropriate usage are discussed. *Curr. Protoc. Bioinform.* 34:2.14.1-2.14.14. © 2011 by John Wiley & Sons, Inc.

Keywords: MACS • ChIP-Seq • peak-calling • cistrome • epigenome

types of histone modifications, the distribution of reads obeys a continuous property, as the epigenetic status of nearby nucleosomes tends to be similar, usually resulting in quite broad peaks. With proper parameter settings, MACS performs well to detect histone-modification-enriched regions. Similarly, MACS can also be applied in affinity enrichment-based DNA methylation studies, such as MeDIP-Seq data.

Know what you are doing

There is no such thing (yet) as an automated gearshift in short read mapping. It is all like stick-shift driving in San Francisco. In other words = running this tool with default parameters will probably not give you meaningful results. A way to deal with this is to **understand** the parameters by carefully reading the documentation and experimenting. Fortunately, Galaxy makes experimenting easy.

Agenda: Day 2

ChIP-Seq Example, continued
RNA-Seq Example: through TopHat
Galaxy Project Overview
Persistence, Sharing, and Publishing
RNA-Seq Example: Cufflinks
Visual Analytics

Agenda: Day 2

ChIP-Seq Example, continued

RNA-Seq Example: through TopHat

Galaxy Project Overview

Persistence, Sharing, Publishing, Reproducibility

RNA-Seq Example: Cufflinks

Visual Analytics

RNA-seq Exercise

http://usegalaxy.org/u/jeremy/p/galaxy-rna-seq-analysis-exercise

http://bit.ly/gxyRNASEX

RNA-seq Exercise: A Plan

- Get input datasets; hg19, will mostly map to chr19
- Look at quality
- Trim as we see fit.
- Map the reads to the human reference using Tophat
- Run Cufflinks on Tophat output to assemble reads into transcripts

http://bit.ly/gxyRNASEX

RNA-seq Exercise: A Plan

- Get input datasets; hg19, will mostly map to chr19
  - All datasets are FASTQ and from the Body Map 2.0 project

http://bit.ly/gxyRNASEX

RNA-seq Exercise: A Plan

- Get input datasets; hg19, will mostly map to chr19
- Look at quality: Same options as for ChIP-Seq
- Trim as we see fit: Same options as for ChIP-Seq

http://bit.ly/gxyRNASEX

RNA-seq Exercise: A Plan

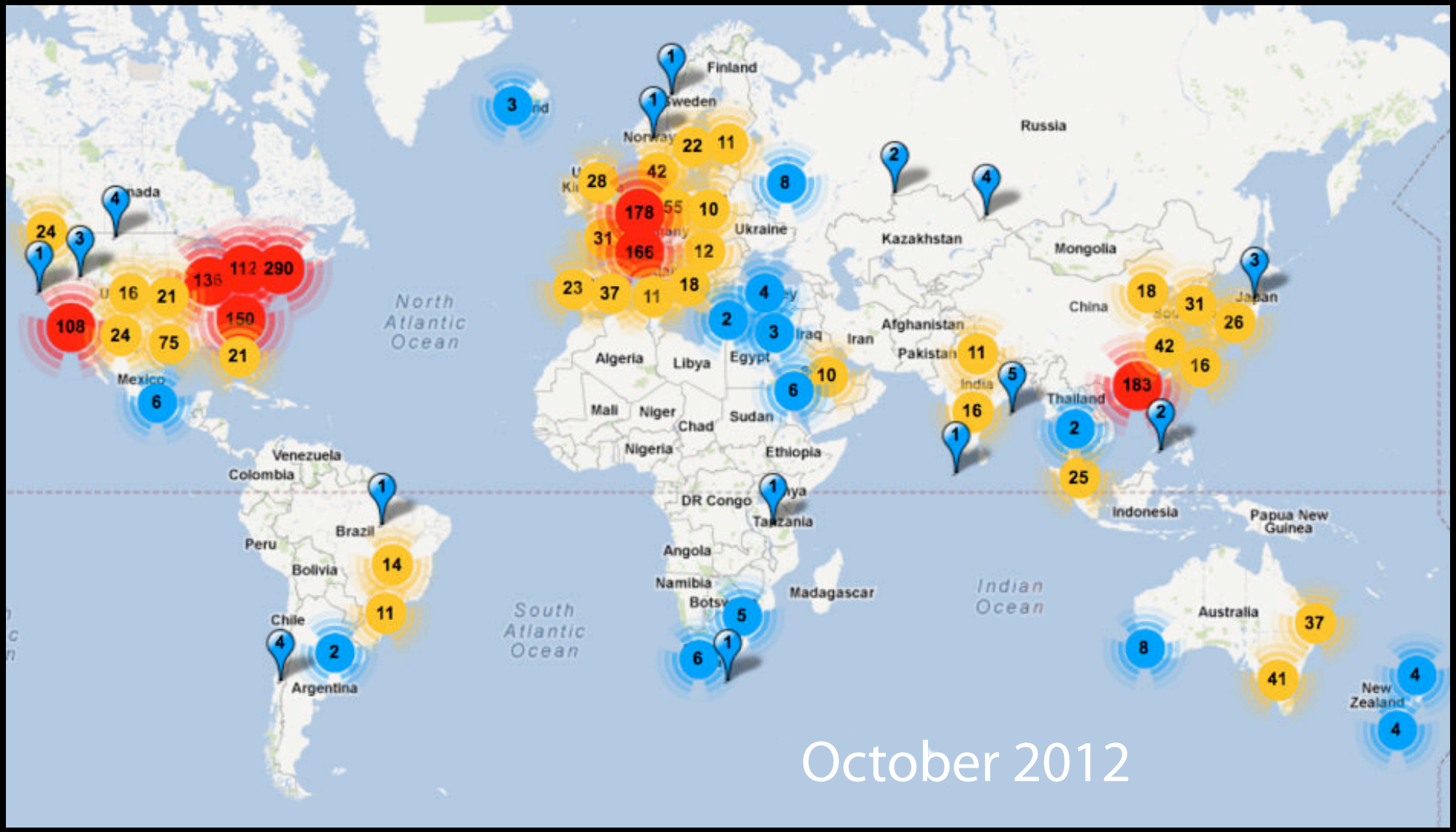
- Get input datasets; hg19, will mostly map to chr19
- Look at quality
- Trim as we see fit.
- Map the reads to the human reference using Tophat
  - *Imagine pages and pages of discussion on the intricacies and pitfalls of RNA-seq mapping here.*
  - Visualize results

http://bit.ly/gxyRNASEX

Agenda: Day 2

ChIP-Seq Example, continued
RNA-Seq Example: through TopHat
Galaxy Project Overview
Persistence, Sharing, Publishing, Reproducibility
RNA-Seq Example: Cufflinks
Visual Analytics

The Motivation Slide



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

http://omicsmaps.com

What is Galaxy?

- A **data analysis and integration** tool
- **A free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- **Open source software** that makes integrating your own tools and data and customizing for your own site simple
- These options result in several **ways to use Galaxy**

http://galaxyproject.org

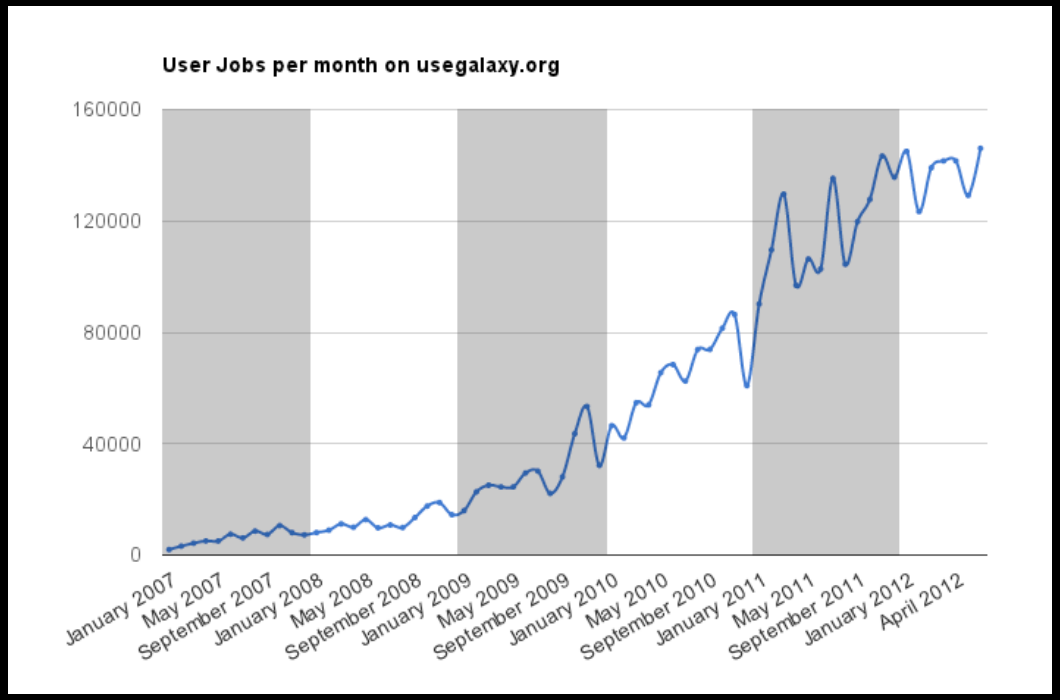
Galaxy is available ...

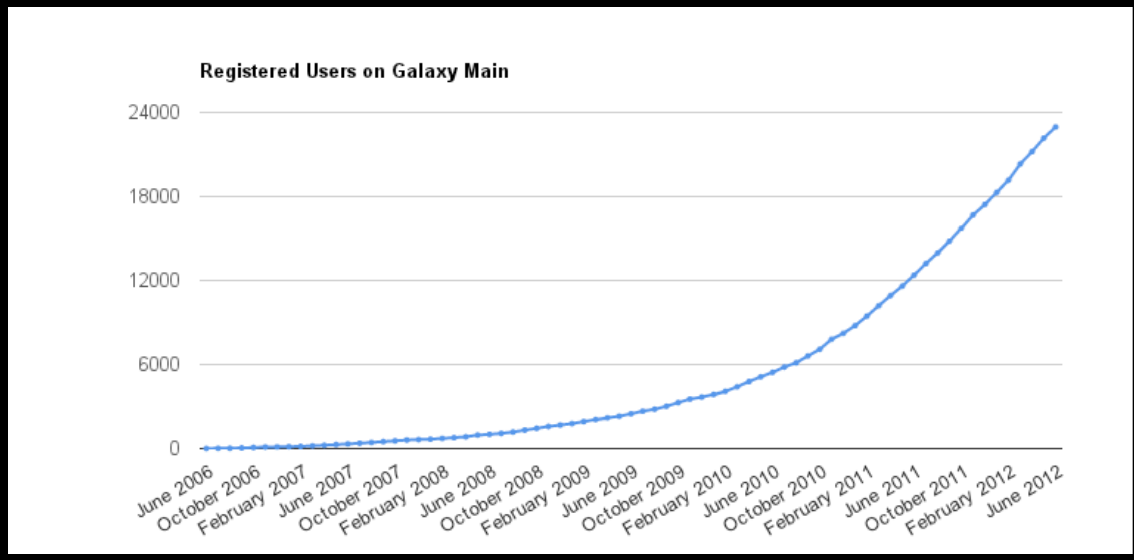
- **As a free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage

http://usegalaxy.org

http://usegalaxy.org
(a.k.a Main)

- Public web site
- Anybody can use it
- Persistent
- + 500 users / month
- ~300 TB of user data
- ~140,000 jobs / month
- Hundreds of tools ...





http://bit.ly/gxystats

usegalaxy.org: a wealth of tools

NGS: QC and manipulation

ILLUMINA DATA

- FASTQ Groomer convert between various FASTQ qual formats
- FASTQ splitter on joined pair end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

ROCHE-454 DATA

- Build base quality distributio
- Select high quality segments
- Combine FASTA and QUAL in FASTQ

AB-SOLID DATA

- Convert SOLiD output to fast
- Compute quality statistics fo SOLiD data
- Draw quality score boxplot f SOLiD data

GENERIC FASTQ MANIPULATION

- Filter FASTQ reads by quality score and length
- FASTQ Trimmer by column
- FASTQ Quality Trimmer by sliding window
- FASTQ Masker by quality sco
- Manipulate FASTQ reads on various attributes
- FASTQ to FASTA converter
- FASTQ to Tabular converter
- Tabular to FASTQ converter

FASTX-TOOLKIT FOR FASTQ DATA

- Quality format converter (ASCII Numeric)
- Compute quality statistics
- Draw quality score boxplot
- Draw nucleotides distribution chart
- FASTQ to FASTA converter
- Filter by quality
- Remove sequencing artifacts
- Barcode Splitter
- Clip adapter sequences
- Collapse sequences
- Rename sequences
- Reverse-Complement
- Trim sequences

FASTQ QC

- FastQC:Read QC reports using FastQC

NGS: Mapping

ILLUMINA

- Map with Bowtie for Illumina
- Map with BWA for Illumina

ROCHE-454

- Lastz map short reads against reference sequence
- Megablast compare short reads against htgs, nt, and wgs databases
- Parse blast XML output

AB-SOLID

- Map with Bowtie for SOLiD
- Map with BWA for SOLiD

NGS: SAM Tools

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format
- Merge BAM Files merges BAM files together
- Generate pileup from BAM dataset
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases
- flagstat provides simple stats on BAM files
- rmdup remove PCR duplicates
- MPileup SNP and indel caller
- Slice BAM by provided regions

NGS: GATK Tools (beta)

ALIGNMENT UTILITIES

- Depth of Coverage on BAM files
- Print Reads from BAM files

REALIGNMENT

- Realigner Target Creator for us in local realignment
- Indel Realigner - perform local realignment

BASE RECALIBRATION

- Count Covariates on BAM files
- Table Recalibration on BAM file
- Analyze Covariates - draw plots

GENOTYPING

- Unified Genotyper SNP and inde caller

ANNOTATION

- Variant Annotator

FILTRATION

- Variant Filtration on VCF files
- Select Variants from VCF files

VARIANT QUALITY SCORE RECALIBRATION

- Variant Recalibrator
- Apply Variant Recalibration

VARIANT UTILITIES

- Validate Variants
- Eval Variants
- Combine Variants

NGS: Indel Analysis

- Filter Indels for SAM
- Extract indels from SAM
- Indel Analysis

NGS: Peak Calling

- MACS Model-based Analysis of ChIP-Seq
- SICER Statistical approach for the Identification of ChIP-Enriched Regions
- GeneTrack indexer on a BED file
- Peak predictor on GeneTrack index

NGS: RNA Analysis

RNA-SEQ

- Tophat for Illumina Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffmerge merge together several Cufflinks assemblies
- Cuffdiff find significant changes

For example, the first 5 pages of NGS tools

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 now has job limits and storage quotas

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

Galaxy is available ...

- **As a free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- **As open source software** that makes integrating your own tools and data and customizing for your own site simple

http://getgalaxy.org

Local Galaxy Instances

- Galaxy is designed for local installation and customization
  - Easily integrate new tools
  - Easy to deploy and manage on nearly any (unix) system
  - Run jobs on existing compute clusters
- Requires an existing computational resource on which to be deployed

http://getgalaxy.org

Encourage Local Galaxy Instances

- Support **increasingly decentralized model** and *improve access to existing resources*
- Focus on building **infrastructure to enable the community to integrate and share** tools, workflows, and best practices

Galaxy Tool Shed

- Allow sites 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

toolshed.g2.bx.psu.edu

Public Galaxy Servers

http://galaxyproject.org/wiki/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

Local Galaxy Instances

- Galaxy is designed for local installation and customization
  - Easily integrate new tools
  - Easy to deploy and manage on nearly any (unix) system
  - Run jobs on existing compute clusters
- Requires an **existing computational resource** on which to be deployed

http://getgalaxy.org

Got your own cluster?

- Move tool execution to other systems
- Galaxy works with any DRMAA compliant cluster job scheduler (which is most of them).
- Galaxy is just another client to your scheduler.

Galaxy is available ...

- **As a free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- **As open source software** that makes integrating your own tools and data and customizing for your own site simple
- **On the Cloud**

http://usegalaxy.org/cloud

Galaxy CloudMan

http://usegalaxy.org/cloud

- Start with a **fully configured and populated** (tools and data) Galaxy instance.
- Allows you to scale up and down your compute assets as needed.
- Someone else manages the data center.

amazon web services™

http://aws.amazon.com/education

Step by Step Instructions on the Wiki for Amazon

Galaxy Wiki

Login | Search:

CloudMan/AWS/GettingStarted

Getting Started with Galaxy CloudMan

This page provides a step-by-step instructions on how to start your own instance of Galaxy on Amazon Web Services (AWS) Elastic Compute Cloud (EC2). More general information and instructions about Galaxy CloudMan (GC) can be found here.

AWS
Get Started
Capacity Planning
AMIs
↑ CloudMan

Contents

Step 1: One Time Amazon Setup

1. Because AWS services implement pay-as-you-go access model for compute resources, it is necessary for every user of the service to *register with Amazon*. You will need a credit card to register. (You can apply for a AWS Education Grant after you register).
2. Once your account has been approved by Amazon (note that this may take up to

Step 1 Screenshots

Instant CloudMan

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Help | User | Using 0%

Tools

search tools

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Archaea table browser
- BX main browser
- EBI SRA ENA SRA
- BioMart Central server
- GrameneMart Central server
- Flymine server
- modENCODE fly server
- modENCODE modMine server

New Cloud Cluster

Managing Data

Store, Manage, and Share data with Libraries

An in-depth tutorial

Live Quickies

0 bytes

Your history is empty. Click 'Get Data' on the left pane to start

Launch a CloudMan instance directly from Main, and transfer your current history.

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Help | User | Using 0%

Launch a Galaxy Cloud Instance

Cluster Name

Password

Key ID

Secret Key

Instance Share String (optional)

Instance Type

Large

Requesting the instance may take a moment, please be patient. Do not refresh your browser or navigate away from the page

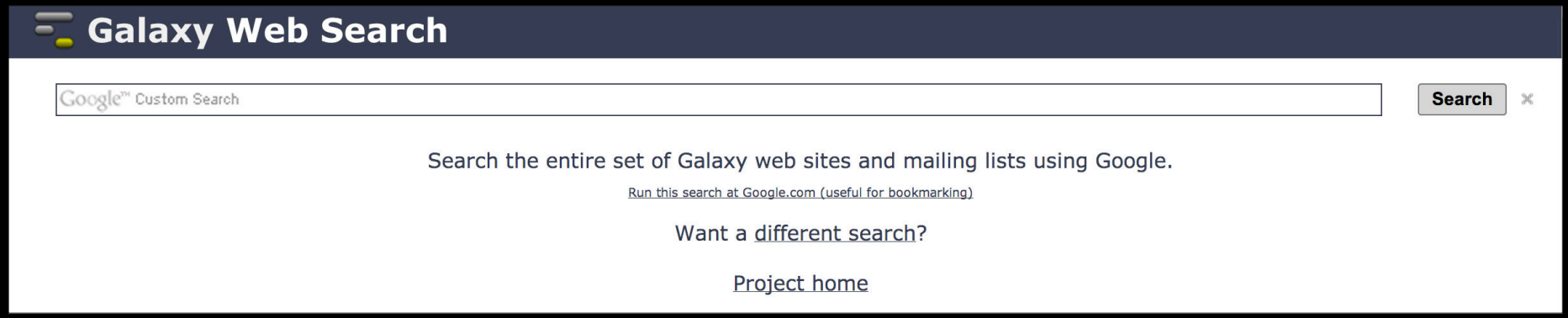
Submit

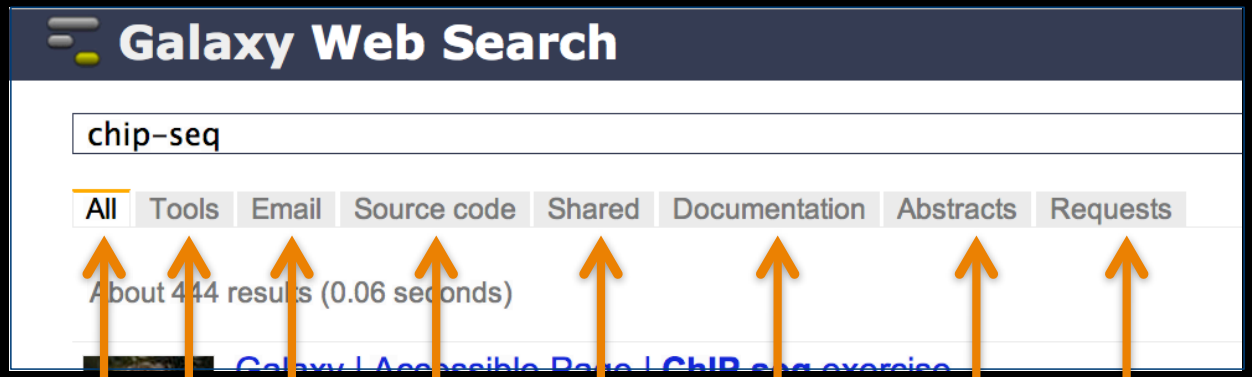
Galaxy Community

Tool Shed
Mailing Lists (very active)
Screencasts
Events Calendar, News Feed
Community Wiki
Local Public Installs
CiteULike group, Mendeley mirror
Annual Community Meting

http://galaxyproject.org/wiki

Galaxy Search: http://galaxyproject.org/search





Find

Everything on …

Tools for …

Email about …

Source code for …

Published Histories, Pages, Workflows, about …

Documentation on …

Papers using Galaxy for …

Related feature requests

http://galaxyproject.org/GCC2013

Other Upcoming Galaxy Events

Date	Topic/Event	Venue/Location	Contact
October 15-17	*Advanced NGS Course: RNA-seq data analysis*	Amsterdam Medical Centre (AMC), The Netherlands	Patrick Koks
October 18-30	*Advanced Sequencing Technologies and Applications Course*	Cold Spring Harbor Laboratory, New York, United States	Anton Nekrutenko
October 31 - November 6	*Computaional & Comparative Genomics Course*	Cold Spring Harbor Laboratory, New York, United States	William Pearson, James Taylor
October 28 - November 2	*Genomic Virtual Laboratory Workshop*	eResearch Australasia, Sydney, Australia	Enis Afgan
November 6-10	*Galaxy 101: Data Integration, Analysis and Sharing* Sold out	**American Society of Human Genetics (ASHG)**, San Francisco, California, United States	Jennifer Jackson, Jeremy Goecks
	Working with High-Throughput Data and Data Visualization Sold out		
November 12-14	*The Genome Access Course*	Cold Spring Harbor Laboratory, New York, United States	Assaf Gordon
November 13-15	*Analyse des données RNA-seq et ChIP-seq (séquençage haut-débit), à l'aide d'outils orientés vers un public de biologistes*	PRABI (Pôle Rhône-Alpes de Bioinformatique), Doua de l'Université Claude Bernard - Lyon, Lyon, France	Guy Perrière
January 14-18	**Plant and Animal Genome (PAG 2013)**	San Diego, California, United States	Dave Clements
March 2-5	*W6: Community Resource Solutions to Analyzing*	**ABRF 2013**	Dave Clements

http://galaxyproject.org/wiki/Events

Galaxy URLs to Remember

http://galaxyproject.org

http://usegalaxy.org

http://getgalaxy.org

Agenda: Day 2

ChIP-Seq Example, continued
RNA-Seq Example: through TopHat
Galaxy Project Overview
Persistence, Sharing, Publishing, Reproducibility
RNA-Seq Example: Cufflinks
Visual Analytics

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

Share:

Make something available to someone else

Publish:

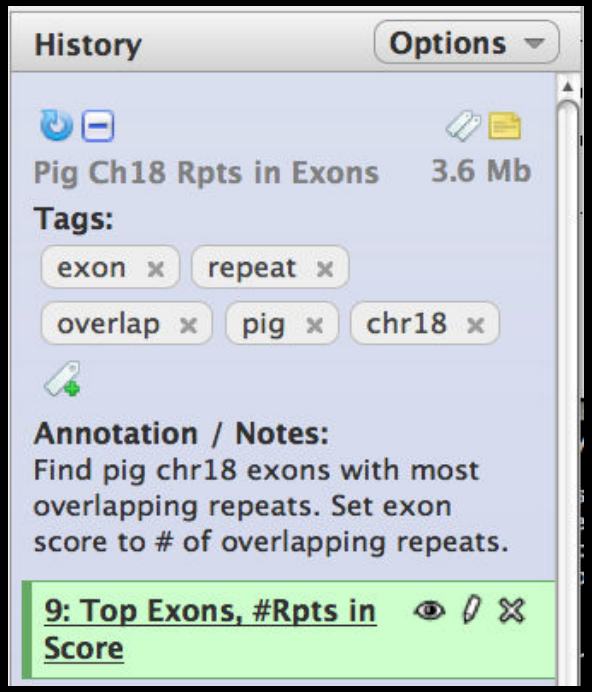
Make something available to everyone

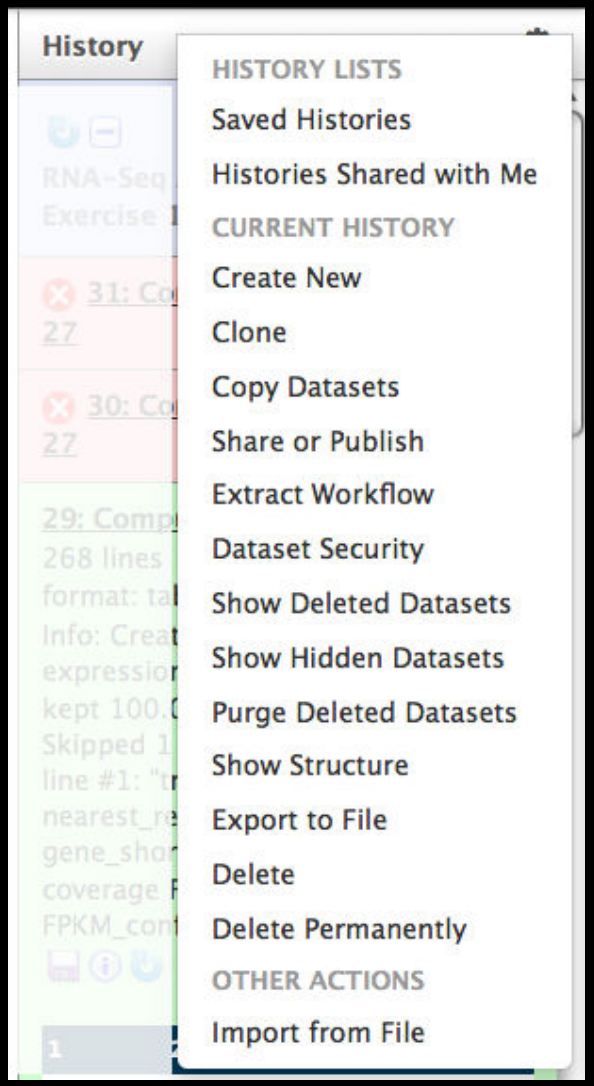
Managing Histories and Datasets

Give every **history and dataset** a **clear name**

Datasets and histories can also have annotation and tags

Each **history** has an options/actions list



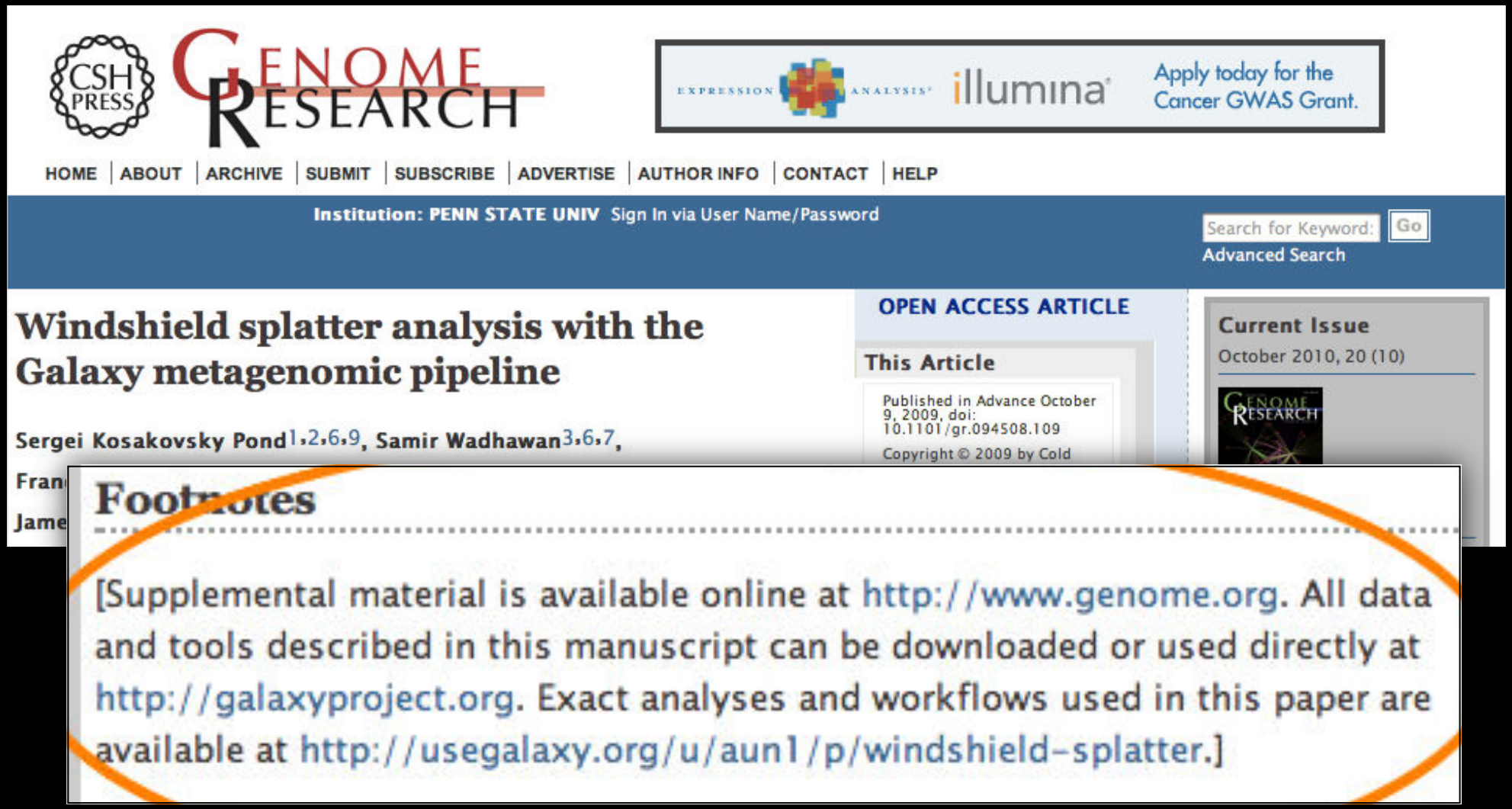


Sharing and Publishing Your Work

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

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

Sharing and Publishing Your Work



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

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

Galaxy Tool Shed

- 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

toolshed.g2.bx.psu.edu

Agenda: Day 2

ChIP-Seq Example, continued
RNA-Seq Example: through TopHat
Galaxy Project Overview
Persistence, Sharing, Publishing, Reproducibility
RNA-Seq Example: Cufflinks
Visual Analytics

RNA-seq Exercise: A Plan

- ...
- Trim as we see fit.
- Map the reads to the human reference using Tophat
- Run Cufflinks on Tophat output to assemble reads into transcripts
  - *Imagine pages and pages of discussion on the intricacies and pitfalls of RNA-seq transcript prediction here.*

http://bit.ly/gxyRNASEX

RNA-seq Exercise: A Plan

- ...
- Map the reads to the human reference using Tophat
- Run Cufflinks on Tophat output to assemble reads into transcripts
  - *Imagine pages and pages of discussion on the intricacies and pitfalls of RNA-seq transcript prediction here.*
  - Visualize it

http://bit.ly/gxyRNASEX

Two RNA-seq Papers

NATURE METHODS | REVIEW

Computational methods for transcriptome annotation and quantification using RNA-seq

Manuel Garber, Manfred G Grabherr, Mitchell Guttman & Cole Trapnell

Affiliations | **Corresponding author**

Nature Methods **8**, 469–477 (2011) | doi:10.1038/nmeth.1613
Published online 27 May 2011 | Corrected online **15 June 2011**

NATURE PROTOCOLS | PROTOCOL

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn & Lior Pachter

Affiliations | **Contributions** | **Corresponding author**

Nature Protocols **7**, 562–578 (2012) | doi:10.1038/nprot.2012.016
Published online 01 March 2012

Agenda: Day 2

ChIP-Seq Example, continued
RNA-Seq Example: through TopHat
Galaxy Project Overview
Persistence, Sharing, Publishing, Reproducibility
RNA-Seq Example: Cufflinks
Visual Analytics

Visualize

Send data results to **external** genome browsers

Trackster: Galaxy's genome browser

Galaxy

- tool integration framework
- heavy focus on usability
- sharing, publication framework

Genome Browser

- physical depiction of data
- visually identify correlations
- find interesting regions, features

Galaxy

- tool integration framework
- heavy focus on usability
- sharing, publication framework

Genome Browser

- physical depiction of data
- visually identify correlations
- find interesting regions, features

Trackster

Trackster

View your data from within Galaxy

- No data transfers to external site
- Use it locally, even without internet access

Supports common filetypes

- BAM, BED, GFF/GTF, WIG

Unique features

- custom genomes
- highly interactive

Galaxy | Published Visualization | GCC2011-1: Viewing and Navigating

http://main.g2.bx.psu.edu/u/jeremy/v/gcc2011-1-viewing-and-navigating

Galaxy

Analyze Data | Workflow | Shared Data | Visualization | Help | User

Published Visualizations | jeremy | GCC2011-1: Viewing and chr19 1,290 - 4,168,475

0 | 1,000,000 | 2,000,000 | 3,000,000 | 4,000,000

UCSC Main on Human: knownGene (chr19) — Auto (Squish)

UCSC Main on Human: all_est (chr19) — Auto (coverage histogram)

11431

UCSC Main on Human: phyloP46wayPrimates (chr19) — Histogram

1

-1

h1-hESC Tophat Mapped Reads — Auto (coverage histogram)

8732

h1-hESC Cufflinks assembled transcripts — Auto (Squish)

0 | 1,000,000 | 2,000,000 | 3,000,000 | 4,000,000

Display a menu

Galaxy | Published Visualization | GCC2011-1: Viewing and Navigating

http://main.g2.bx.psu.edu/u/jeremy/v/gcc2011-1-viewing-and-navigating

Galaxy

Analyze Data | Workflow | Shared Data | Visualization | Help | User

Published Visualizations | jeremy | GCC2011-1: Viewing and chr19 625,719 - 682,581

630,000 640,000 650,000 660,000 670,000 680,000

UCSC Main on Human: knownGene (chr19) Auto (Squish)

UCSC Main on Human: all_est (chr19) Dense

UCSC Main on Human: phyloP46wayPrimates (chr19) Histogram

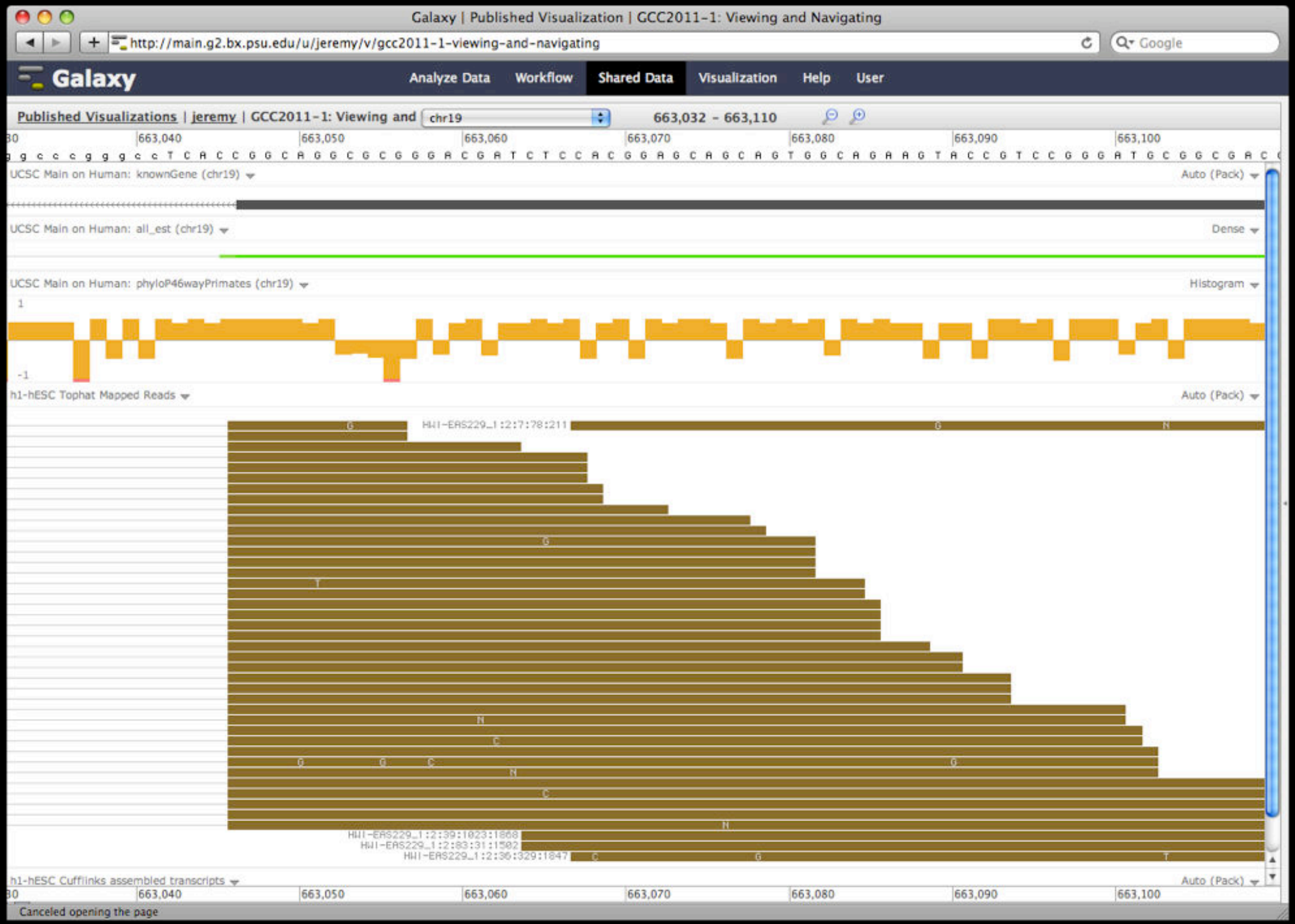
1

-1

h1-hESC Tophat Mapped Reads Auto (Squish)

630,000 640,000 650,000 660,000 670,000 680,000

Display a menu


Galaxy | Published Visualization | GCC2011-1: Viewing and Navigating
http://main.g2.bx.psu.edu/u/jeremy/v/gcc2011-1-viewing-and-navigating
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Published Visualizations | jeremy | GCC2011-1: Viewing and
chr19
663,032 - 663,110
663,040
663,050
663,060
663,070
663,080
663,090
663,100
UCSC Main on Human: knownGene (chr19)
Auto (Pack)
UCSC Main on Human: all_est (chr19)
Dense
UCSC Main on Human: phyloP46wayPrimates (chr19)
Histogram
h1-hESC Tophat Mapped Reads
Auto (Pack)
h1-hESC Cufflinks assembled transcripts
Auto (Pack)
Canceled opening the page

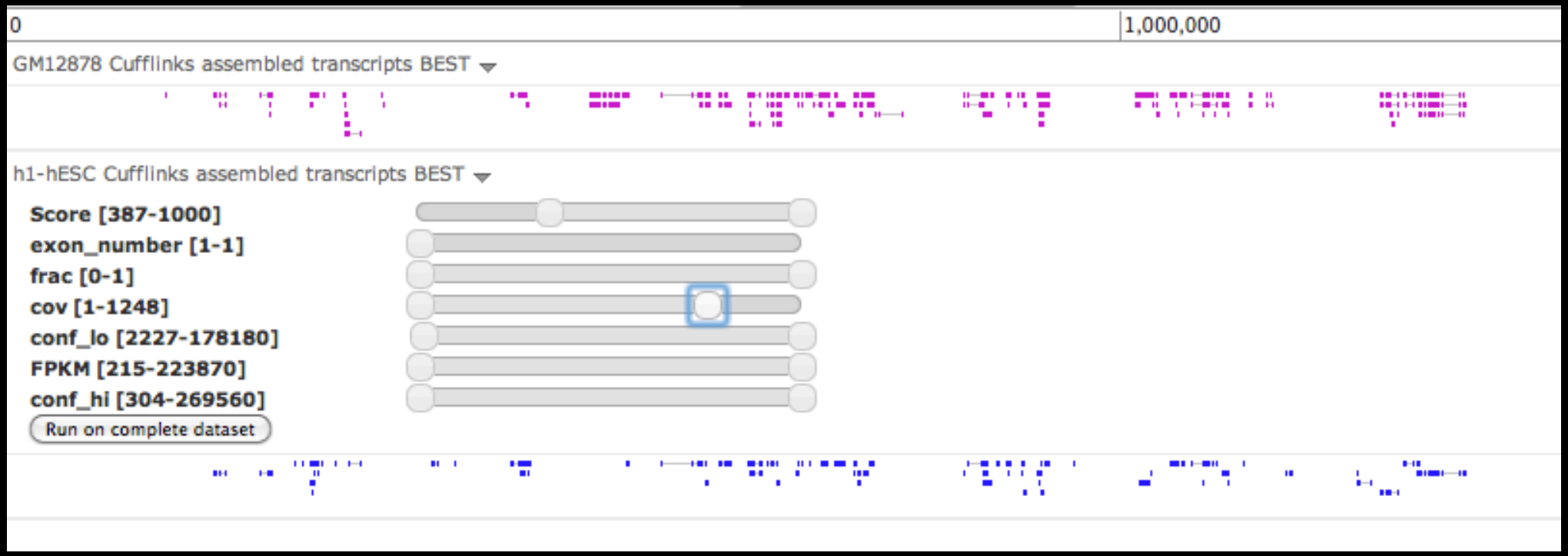
But really, why *another* genome browser

From static browsing to visual analysis

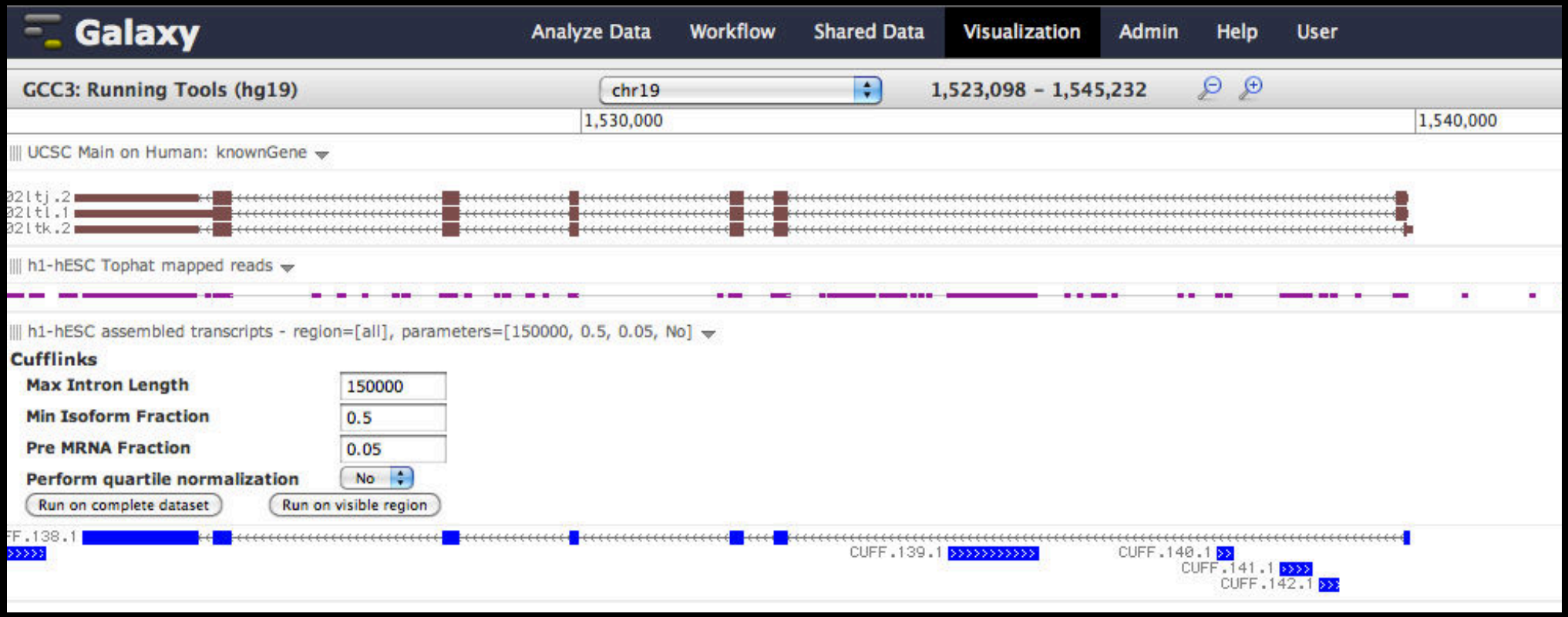
Visual feedback and experimentation needed for complex tools with many parameters

Leverage Galaxy strengths: a very sound model for abstracting interfaces to analysis tools and already integrates an enormous number

Dynamic Filtering



Integrating Tools and Visualization



Visualization: Even More

- usegalaxy.org → Shared Data → Published Visualizations
  - Don't everyone do this!
- galaxyproject.org/wiki/Events/GCC2012/Program → Session 4 → The Galaxy Visualization Framework
  - Jeremy Goecks GCC2012 presentation.
  - Basic Navigation Demo starts @ 10:40
  - Dynamic Filtering Demo starts @ 12:15
  - Circster Demo starts @ 14:10
  - Visual Analytics Demo starts @ 15:40
  - Next @

Workshop Feedback

Please help.

http://bit.ly/UIUCFeedback

Thanks

http://bit.ly/UIUCFeedback