Biologists on the cloud

our experiences using galaxy for next-gen sequencing analyses

Karen Reddy, Johns Hopkins University School of Medicine

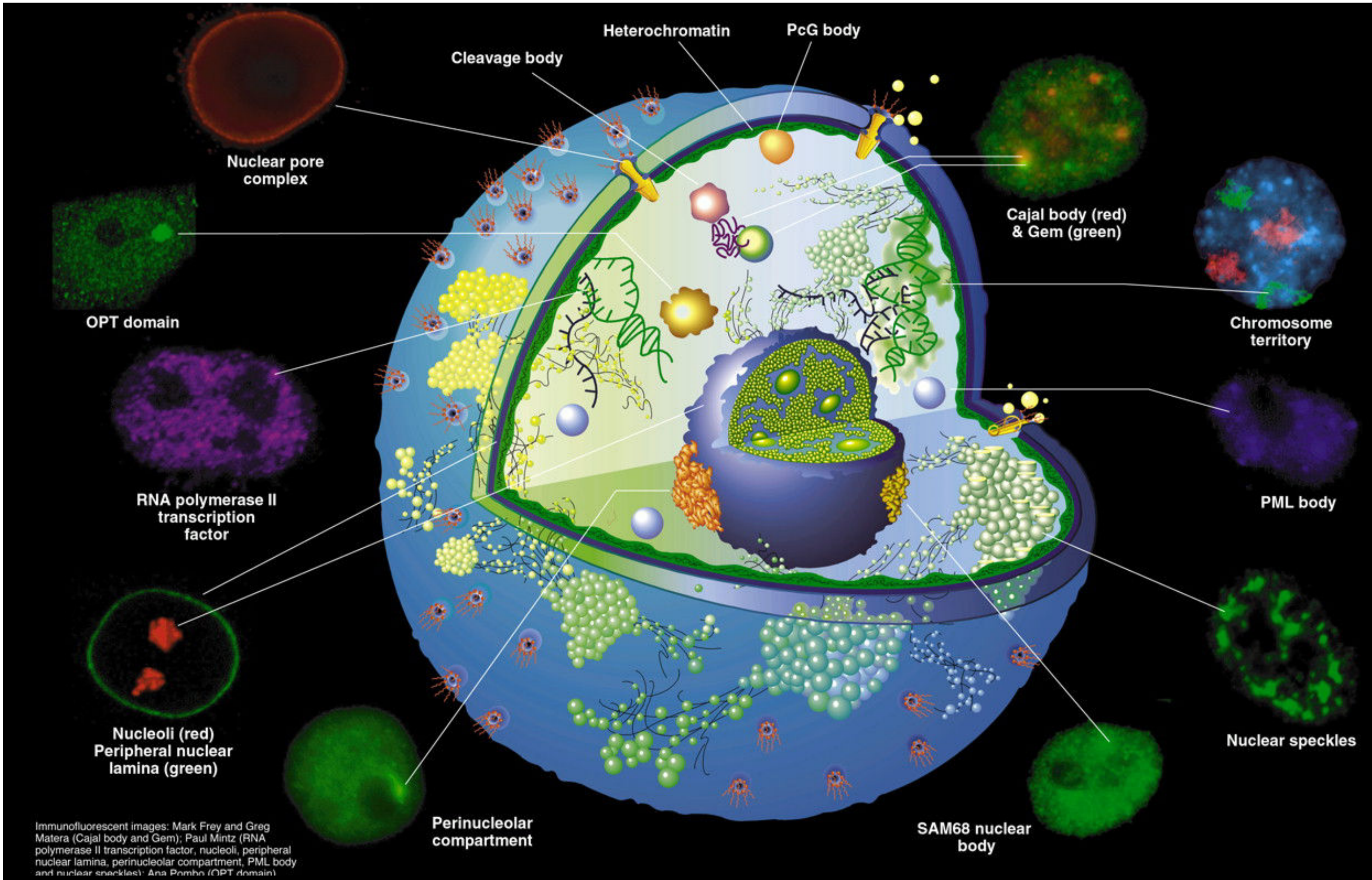
Mo Heydarian, Johns Hopkins University School of Medicine

Warning!!!

- WE ARE END USERS…
- Blinking cursors on a black screen make us break out in a cold sweat (i.e. anything remotely resembling command line)
- Putty, SOAP, keys etc may mean something different to us

Who we are:

- Interested in organization of the nucleus with respect to gene regulation
  - Epigenetics, transcription, cell biology, structure


Heterochromatin
PcG body
Cleavage body
Nuclear pore complex
Cajal body (red) & Gem (green)
OPT domain
Chromosome territory
RNA polymerase II transcription factor
PML body
Nucleoli (red) Peripheral nuclear lamina (green)
Nuclear speckles
Perinucleolar compartment
SAM68 nuclear body

Immunofluorescent images: Mark Frey and Greg Matera (Cajal body and Gem); Paul Mintz (RNA polymerase II transcription factor, nucleoli, peripheral nuclear lamina, perinucleolar compartment, PML body and nuclear speckles); Ana Pombo (OPT domain)

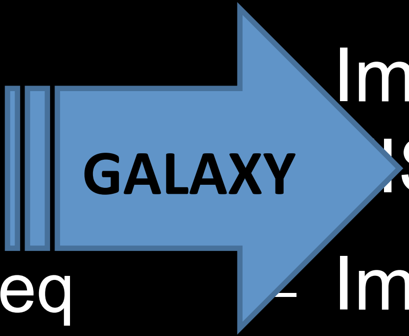
Proteins , DNA and RNA are compartmentalized

‘Nuclear domains’ Spector, Journal Cell Science

Who we are:

- Interested in organization of the nucleus with respect to gene regulation
  - Epigenetics, transcription, cell biology, structure

- Molecular approaches
  - ChIP-seq
  - Hi-C
  - DamID/DamID-seq
  - RNA-seq
  - SmallRNA-seq
  - Proteomics



validation , cellular and functional Assays

- Cellular approaches
  - Immuno-DNA/RNA-FISH
  - Immunofluorescence
  - Phenotype/ developmental assays

We need to be able to make sense of our billions of NGS reads

Why Galaxy?

We need to perform the analysis

- Lack of bioinformatician availability
- We need to understand what is being done to our data
- Create “workflows” and infrastructure for the lab (and the community)
- Additional and portable skill set for trainees

Analysis options using Galaxy

- On a cluster/local instance
  - Resources
    - Expertise ($$$)
    - Space ($$$)
    - Access
  - Command line....

- Galaxy Cloudman
  - we do not have to maintain!!
  - Portable
  - ‘pre-packaged’

Additional Benefits Using Galaxy Cloudman

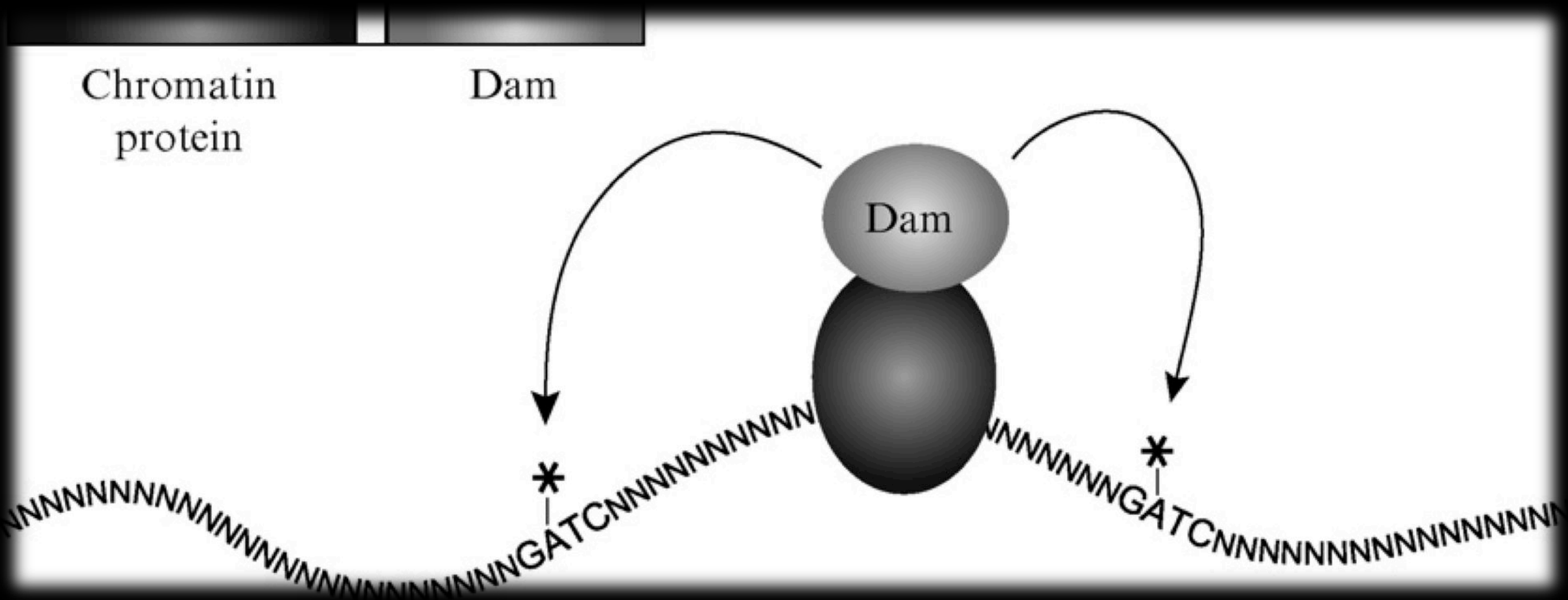
- User community (http://wiki.g2.bx.psu.edu/Learn/Screencasts)
- Graphical user interface
- Intuitive packages
- Alleviates dependency on external resources
- Scalability perfect for intermittent need and smaller laboratories
- Extensively vetted tools avaialble

Two examples

DamID-seq

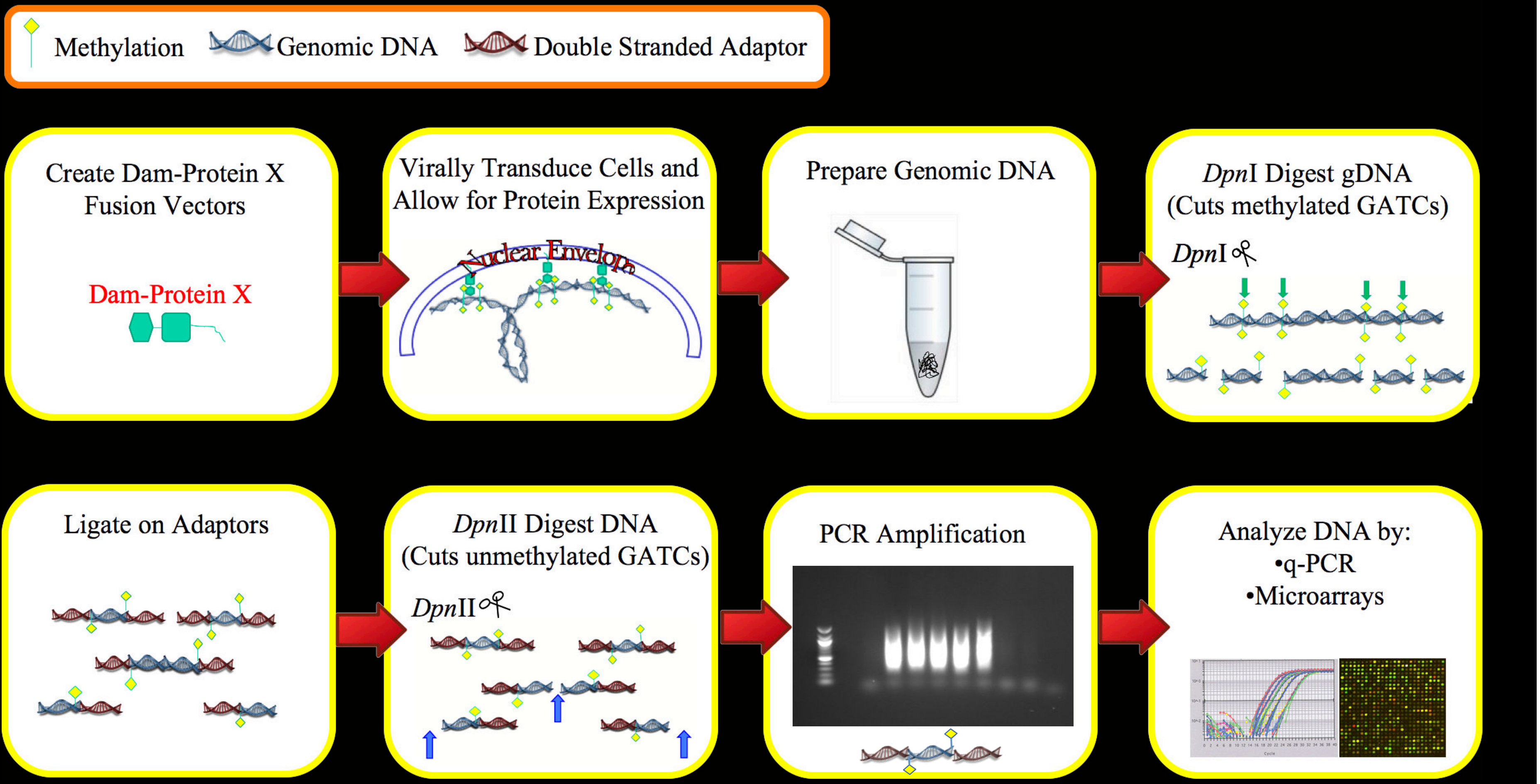
- Adapted protocol
  - No analysis pipeline existed
    - We consulted with a bioinformatician on the broad strokes
    - A workflow was created using common straightforward Galaxy tools
    - The resultant data was confirmed by independent DamID array hybridization
    - Traditional peak-calling does not work as these regions are very broad domains in the genome

Strategy: DamID


Chromatin protein
Dam
Dam
NNNNNNNNNNNNNNNNNNNNNNNNNNNGATCNNNNNNNNNNNN
*
NNNNNNNGATCNNNNNNNNNNNNNNNN
*

Greil, Moorman and van Steensel 2006

DNA Adenine Methylase Identification DamID



Methylation   Genomic DNA   Double Stranded Adaptor

Create Dam-Protein X Fusion Vectors

Virally Transduce Cells and Allow for Protein Expression

Prepare Genomic DNA

*Dpn*I Digest gDNA (Cuts methylated GATCs)

Ligate on Adaptors

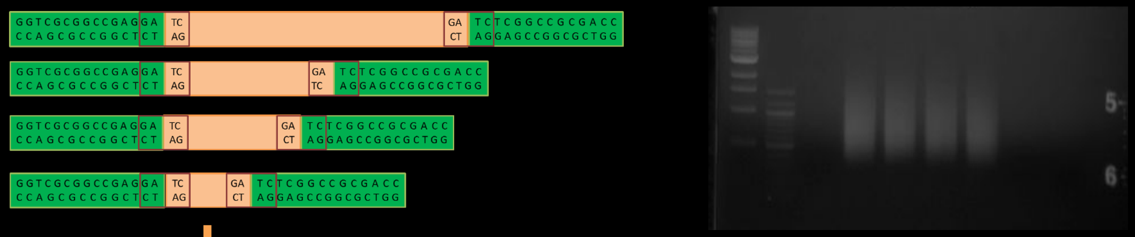
*Dpn*II Digest DNA (Cuts unmethylated GATCs)

PCR Amplification

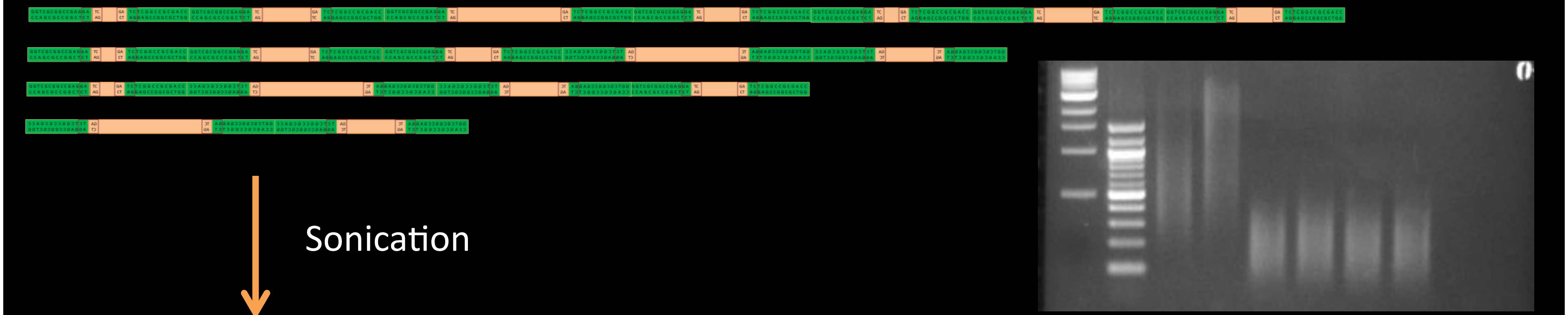
Analyze DNA by:

- q-PCR
- Microarrays

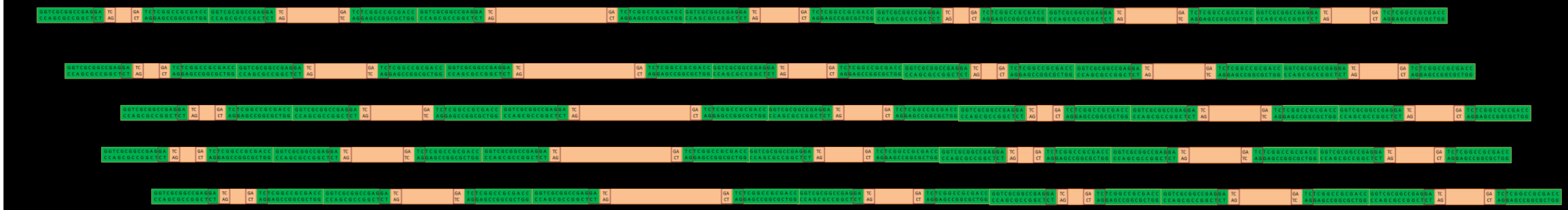
DamID-Seq : NOT DRAWN TO SCALE! Lamina Associated Domains (LADs) should be much longer than adapters! (LADs ~ 200 – 4000 bp; Adapter – 15 bp)



Randomisation: End Repair and ligate

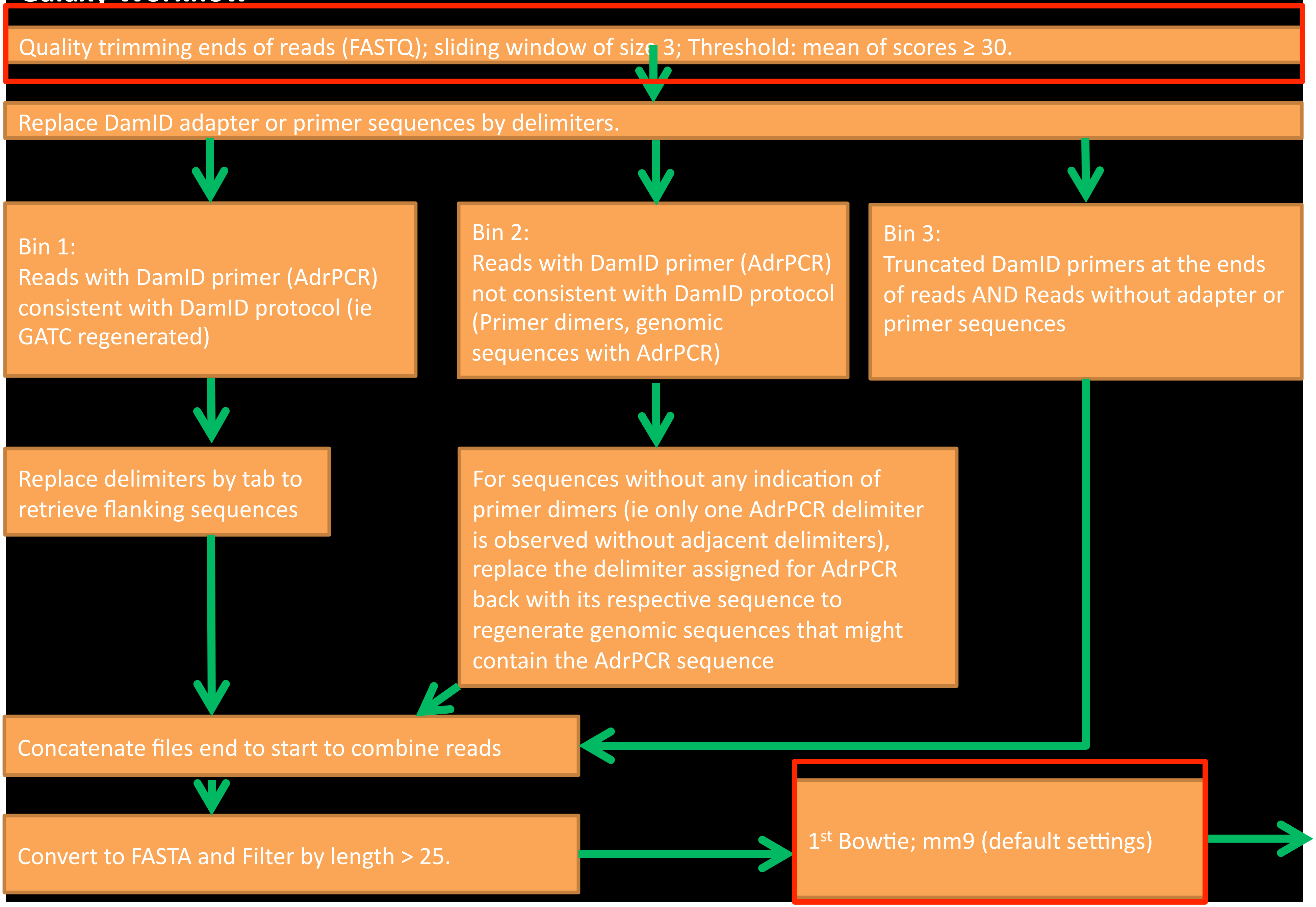


Sonication



Library Prep: End Repair, A tailing, adapter ligation, PCR -> cluster generation and sequencing

Galaxy Workflow



Quality trimming ends of reads (FASTQ); sliding window of size 3; Threshold: mean of scores ≥ 30.

Replace DamID adapter or primer sequences by delimiters.

Bin 1:
Reads with DamID primer (AdrPCR) consistent with DamID protocol (ie GATC regenerated)

Bin 2:
Reads with DamID primer (AdrPCR) not consistent with DamID protocol (Primer dimers, genomic sequences with AdrPCR)

Bin 3:
Truncated DamID primers at the ends of reads AND Reads without adapter or primer sequences

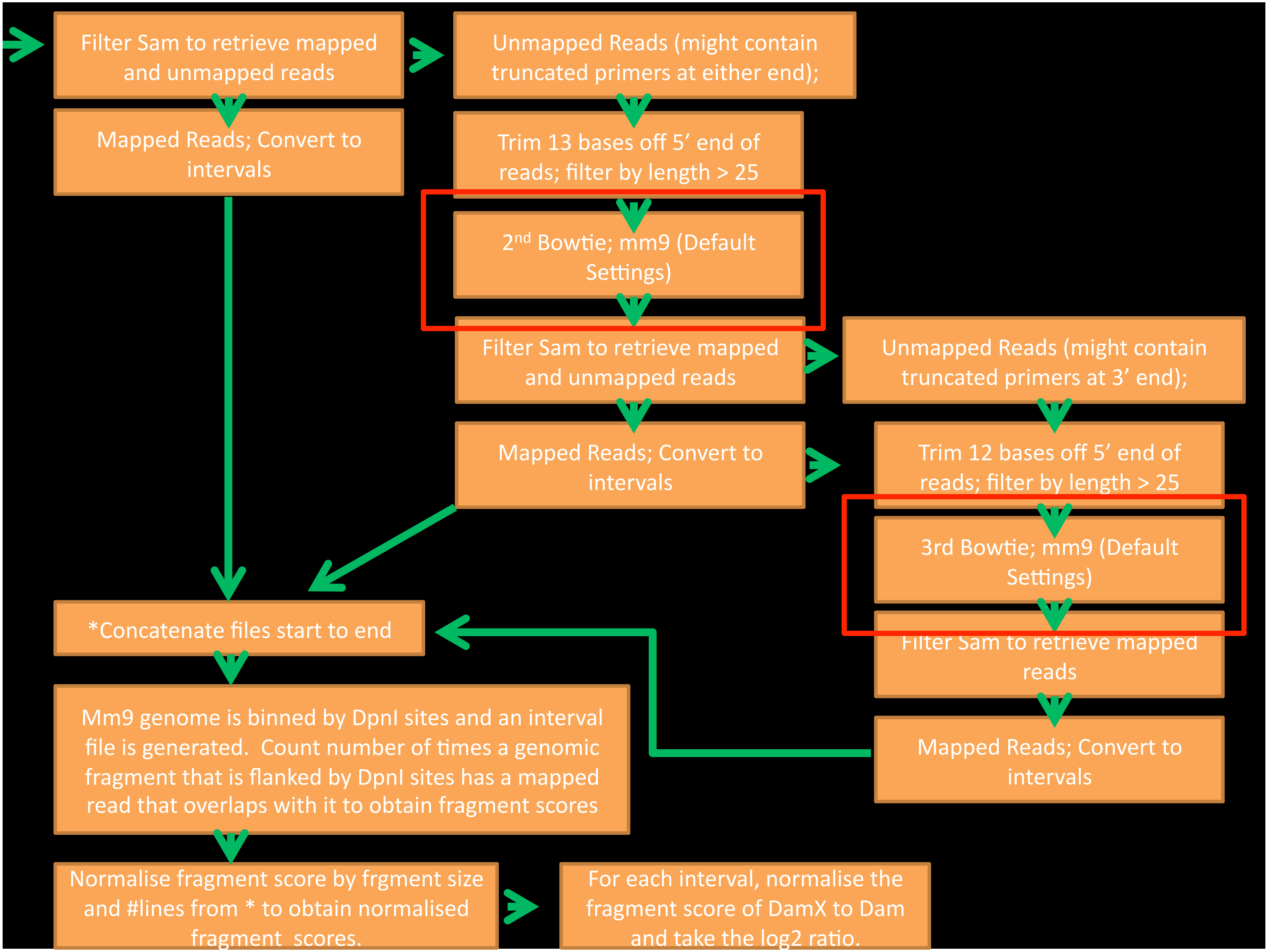
Replace delimiters by tab to retrieve flanking sequences

For sequences without any indication of primer dimers (ie only one AdrPCR delimiter is observed without adjacent delimiters), replace the delimiter assigned for AdrPCR back with its respective sequence to regenerate genomic sequences that might contain the AdrPCR sequence

Concatenate files end to start to combine reads

Convert to FASTA and Filter by length > 25.

1st Bowtie; mm9 (default settings)



Filter Sam to retrieve mapped and unmapped reads

Mapped Reads; Convert to intervals

Unmapped Reads (might contain truncated primers at either end);

Trim 13 bases off 5’ end of reads; filter by length > 25

2nd Bowtie; mm9 (Default Settings)

Filter Sam to retrieve mapped and unmapped reads

Mapped Reads; Convert to intervals

Unmapped Reads (might contain truncated primers at 3’ end);

Trim 12 bases off 5’ end of reads; filter by length > 25

3rd Bowtie; mm9 (Default Settings)

Filter Sam to retrieve mapped reads

Mapped Reads; Convert to intervals

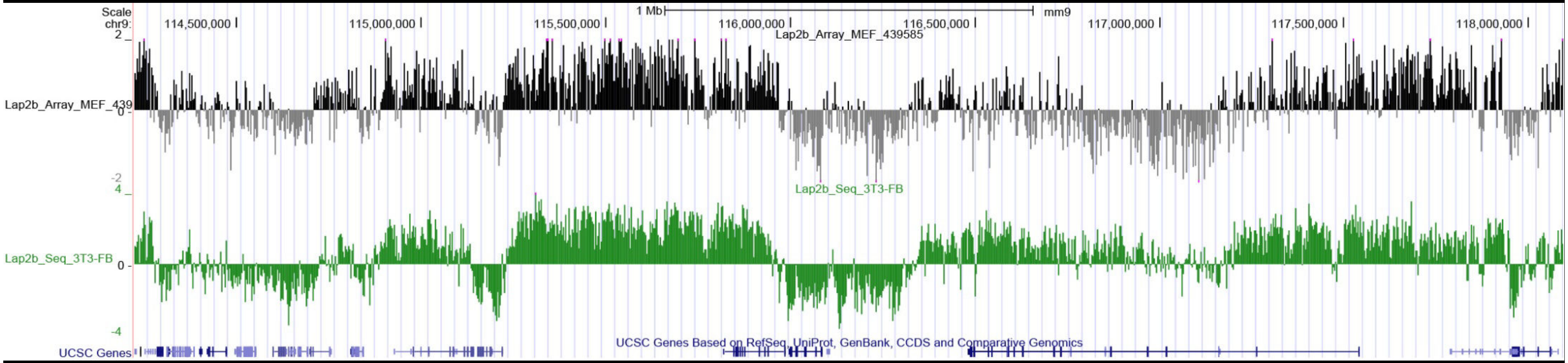
*Concatenate files start to end

Mm9 genome is binned by DpnI sites and an interval file is generated. Count number of times a genomic fragment that is flanked by DpnI sites has a mapped read that overlaps with it to obtain fragment scores

Normalise fragment score by frgment size and #lines from * to obtain normalised fragment scores.

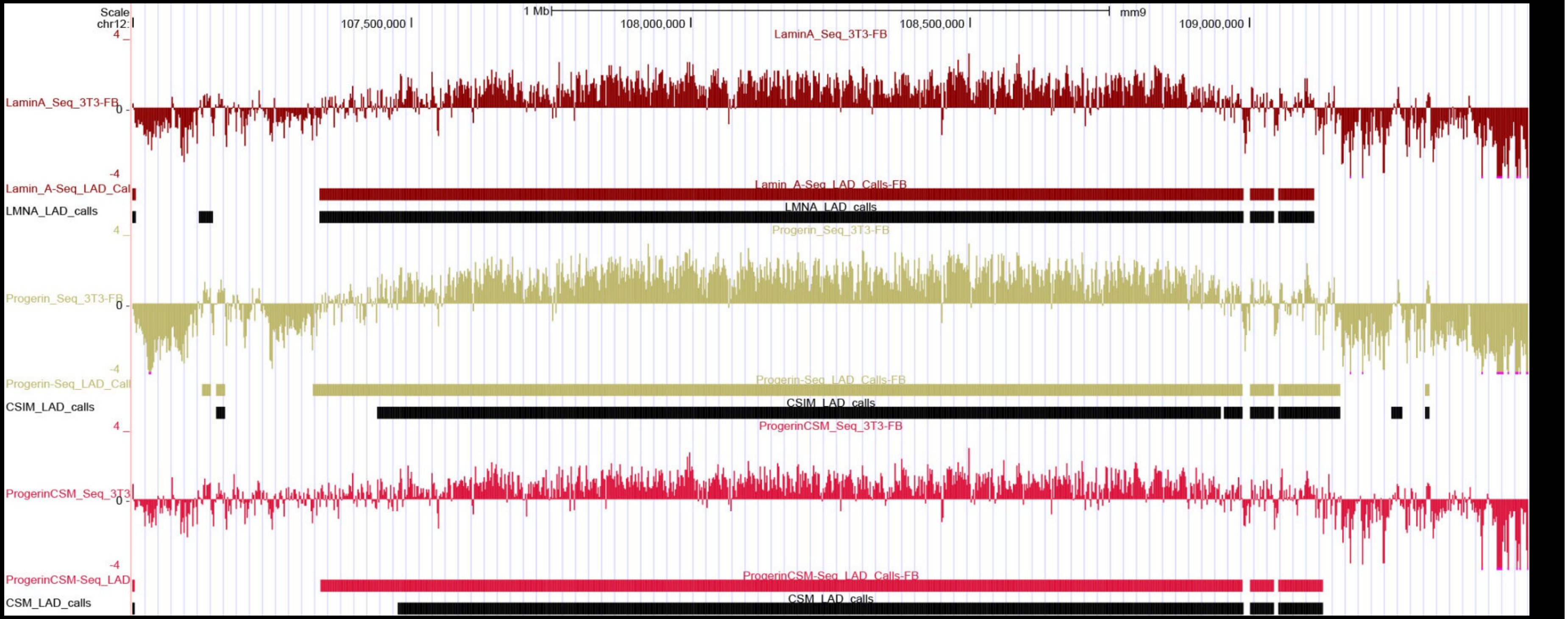
For each interval, normalise the fragment score of DamX to Dam and take the log2 ratio.

Verification of workflow/analysis by comparing DamID-seq with DamID-tiled array

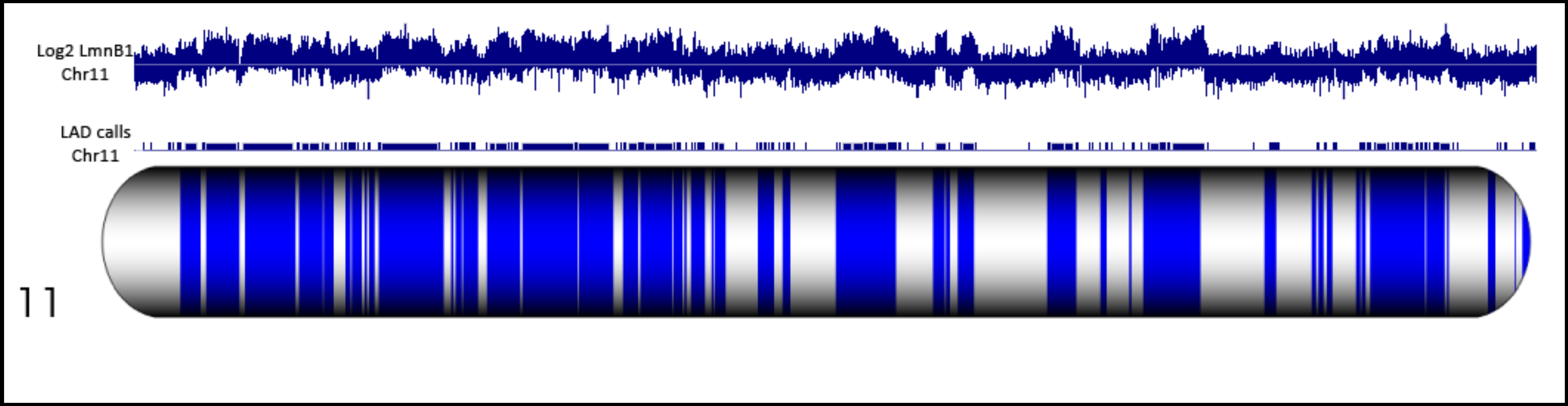


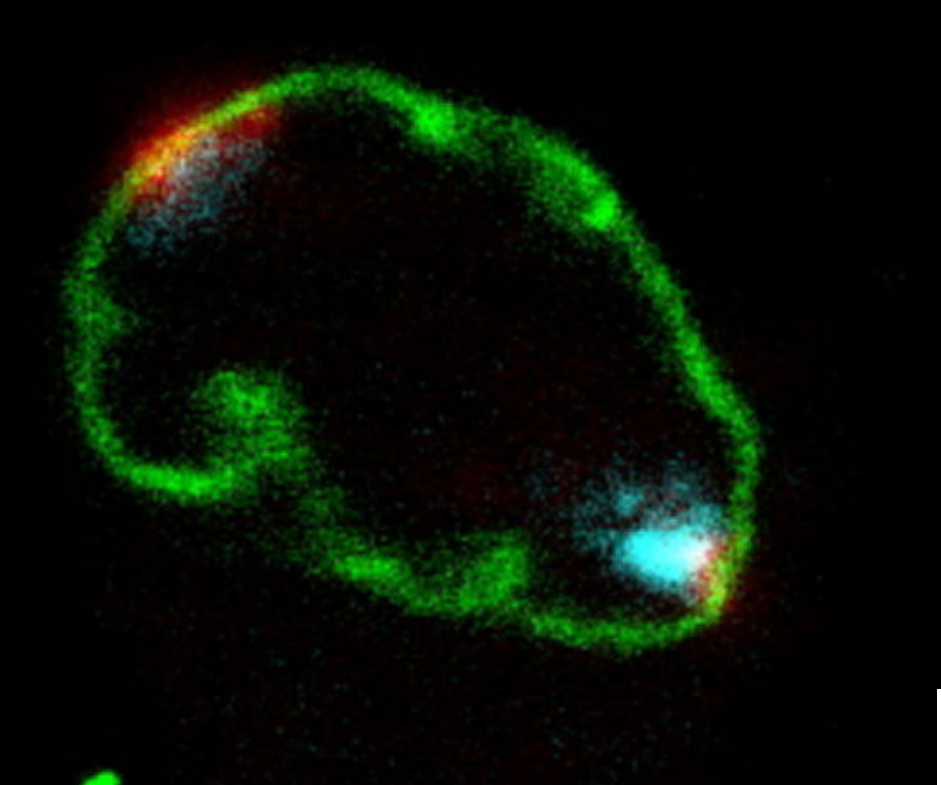
We want to do more of the ‘fun’ stuff with bioinformaticians—

- like developing algorithms to further analyze our data



Probes were made to LADed regions (red, Cy3) and unLADed regions (cyan, Cy5)



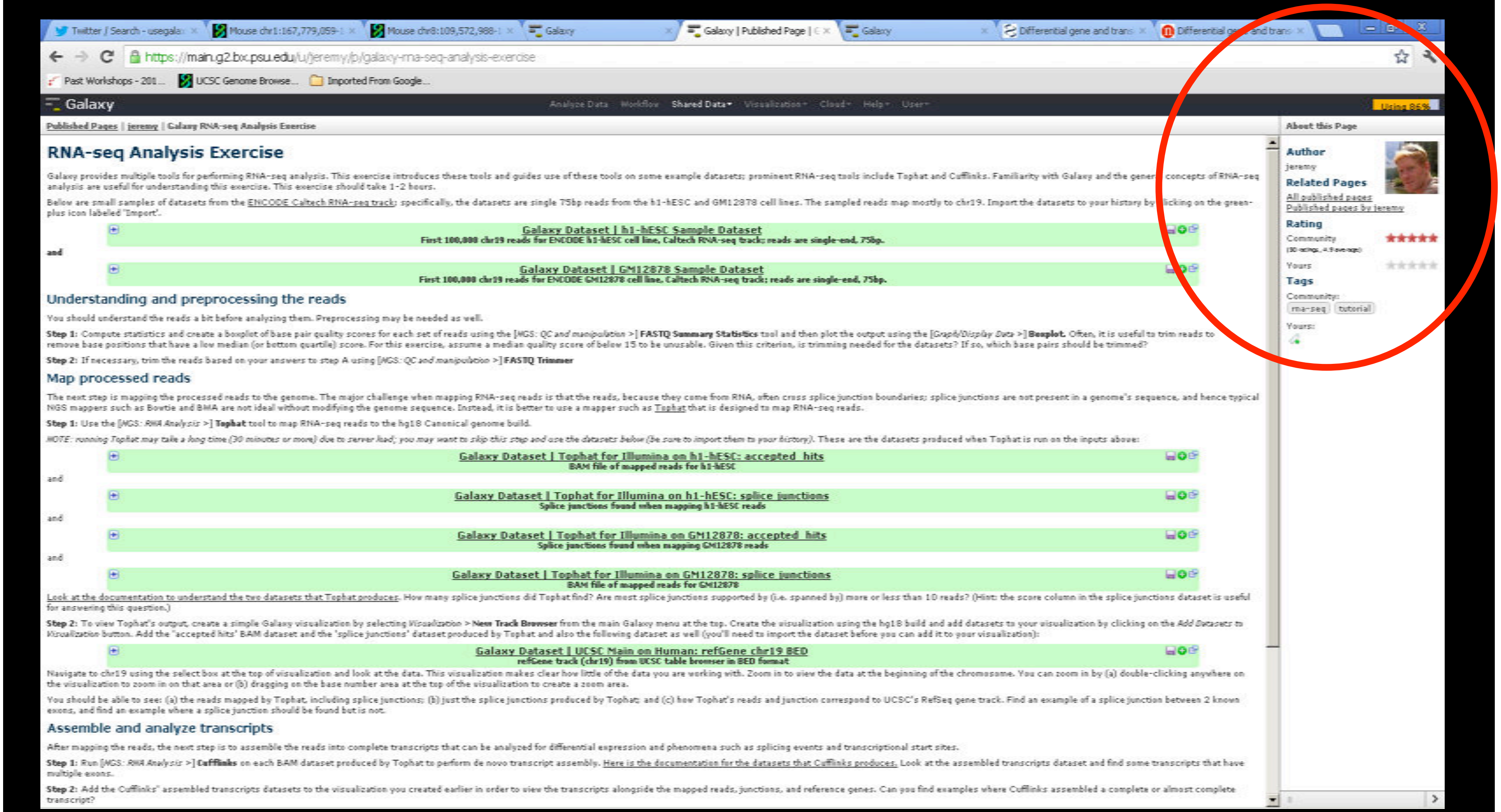


Collaboration with Agilent

RNA-seq

- We performed directional RNA-seq on mouse primary lymphoid cell lines
- Aimed to profile gene expression and discover ncRNAs
- With assistance from the Galaxy community and the literature we were able to perform the analysis on our own using the intuitive Galaxy Cloudman
  - https://main.g2.bx.psu.edu/u/jeremy/p/galaxy-rna-seq-analysis-exercise
  - Trapnell, C., et. al.. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protocols. March 1 2012.

On-line tutorials, guides and workflows invaluable



Published Pages | jeremy | Galaxy RNA-seq Analysis Exercise

RNA-seq Analysis Exercise

Galaxy provides multiple tools for performing RNA-seq analysis. This exercise introduces these tools and guides use of these tools on some example datasets; prominent RNA-seq tools include Tophat and Cufflinks. Familiarity with Galaxy and the general concepts of RNA-seq analysis are useful for understanding this exercise. This exercise should take 1-2 hours.

Below are small samples of datasets from the ENCODE Caltech RNA-seq track; specifically, the datasets are single 75bp reads from the h1-hESC and GM12878 cell lines. The sampled reads map mostly to chr19. Import the datasets to your history by clicking on the green-plus icon labeled 'Import'.

Galaxy Dataset | h1-hESC Sample Dataset
First 100,000 chr19 reads for ENCODE h1-hESC cell line, Caltech RNA-seq track; reads are single-end, 75bp.

and

Galaxy Dataset | GM12878 Sample Dataset
First 100,000 chr19 reads for ENCODE GM12878 cell line, Caltech RNA-seq track; reads are single-end, 75bp.

Understanding and preprocessing the reads

You should understand the reads a bit before analyzing them. Preprocessing may be needed as well.

Step 1: Compute statistics and create a boxplot of base pair quality scores for each set of reads using the [*NGS: QC and manipulation* >] **FASTQ Summary Statistics** tool and then plot the output using the [*Graph/Display Data* >] **Boxplot.** Often, it is useful to trim reads to remove base positions that have a low median (or bottom quartile) score. For this exercise, assume a median quality score of below 15 to be unusable. Given this criterion, is trimming needed for the datasets? If so, which base pairs should be trimmed?

Step 2: If necessary, trim the reads based on your answers to step A using [*NGS: QC and manipulation* >] **FASTQ Trimmer**

Map processed reads

The next step is mapping the processed reads to the genome. The major challenge when mapping RNA-seq reads is that the reads, because they come from RNA, often cross splice junction boundaries; splice junctions are not present in a genome's sequence, and hence typical NGS mappers such as Bowtie and BWA are not ideal without modifying the genome sequence. Instead, it is better to use a mapper such as Tophat that is designed to map RNA-seq reads.

Step 1: Use the [*NGS: RNA Analysis* >] **Tophat** tool to map RNA-seq reads to the hg18 Canonical genome build.

NOTE: running Tophat may take a long time (30 minutes or more) due to server load; you may want to skip this step and use the datasets below (be sure to import them to your history). These are the datasets produced when Tophat is run on the inputs above:

Galaxy Dataset | Tophat for Illumina on h1-hESC: accepted_hits
BAM file of mapped reads for h1-hESC

and

Galaxy Dataset | Tophat for Illumina on h1-hESC: splice junctions
Splice junctions found when mapping h1-hESC reads

and

Galaxy Dataset | Tophat for Illumina on GM12878: accepted_hits
Splice junctions found when mapping GM12878 reads

and

Galaxy Dataset | Tophat for Illumina on GM12878: splice junctions
BAM file of mapped reads for GM12878

Look at the documentation to understand the two datasets that Tophat produces. How many splice junctions did Tophat find? Are most splice junctions supported by (i.e. spanned by) more or less than 10 reads? (Hint: the score column in the splice junctions dataset is useful for answering this question.)

Step 2: To view Tophat's output, create a simple Galaxy visualization by selecting *Visualization* > **New Track Browser** from the main Galaxy menu at the top. Create the visualization using the hg18 build and add datasets to your visualization by clicking on the *Add Datasets to Visualization* button. Add the 'accepted hits' BAM dataset and the 'splice junctions' dataset produced by Tophat and also the following dataset as well (you'll need to import the dataset before you can add it to your visualization):

Galaxy Dataset | UCSC Main on Human: refGene chr19 BED
refGene track (chr19) from UCSC table browser in BED format

Navigate to chr19 using the select box at the top of visualization and look at the data. This visualization makes clear how little of the data you are working with. Zoom in to view the data at the beginning of the chromosome. You can zoom in by (a) double-clicking anywhere on the visualization to zoom in on that area or (b) dragging on the base number area at the top of the visualization to create a zoom area.

You should be able to see: (a) the reads mapped by Tophat, including splice junctions; (b) just the splice junctions produced by Tophat; and (c) how Tophat's reads and junction correspond to UCSC's RefSeq gene track. Find an example of a splice junction between 2 known exons, and find an example where a splice junction should be found but is not.

Assemble and analyze transcripts

After mapping the reads, the next step is to assemble the reads into complete transcripts that can be analyzed for differential expression and phenomena such as splicing events and transcriptional start sites.

Step 1: Run [*NGS: RNA Analysis* >] **Cufflinks** on each BAM dataset produced by Tophat to perform de novo transcript assembly. Here is the documentation for the datasets that Cufflinks produces. Look at the assembled transcripts dataset and find some transcripts that have multiple exons.

Step 2: Add the Cufflinks' assembled transcripts datasets to the visualization you created earlier in order to view the transcripts alongside the mapped reads, junctions, and reference genes. Can you find examples where Cufflinks assembled a complete or almost complete transcript?

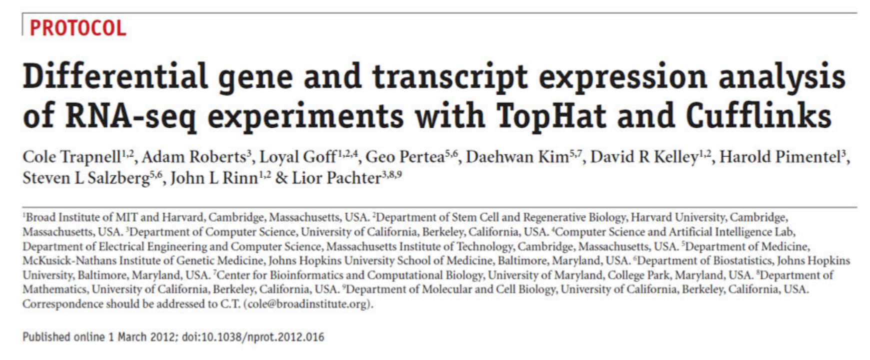
About this Page

Author
jeremy

Related Pages
All published pages
Published pages by jeremy

Rating
Community ★★★★★
(30 ratings, 4.9 average)
Yours ☆☆☆☆☆

Tags
Community:
rna-seq tutorial
Yours:



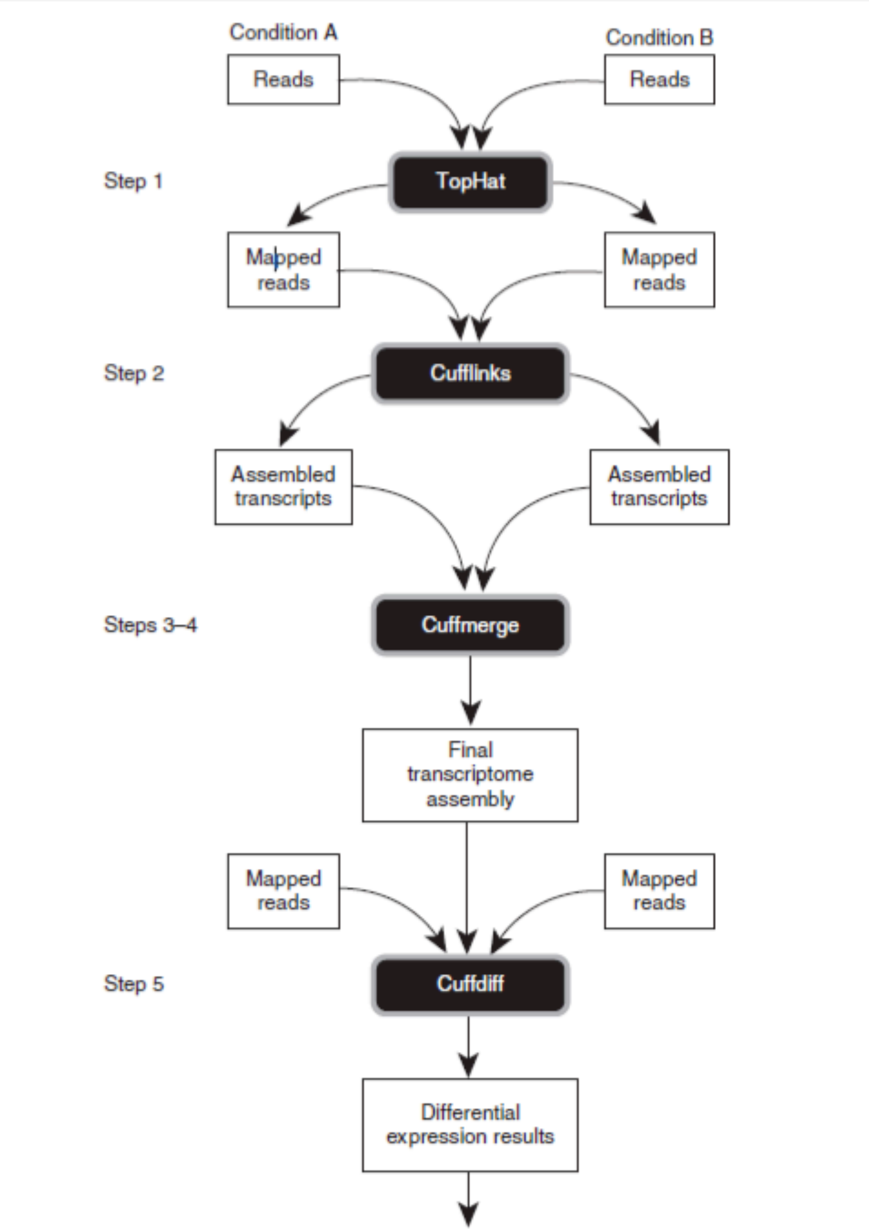
PROTOCOL

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

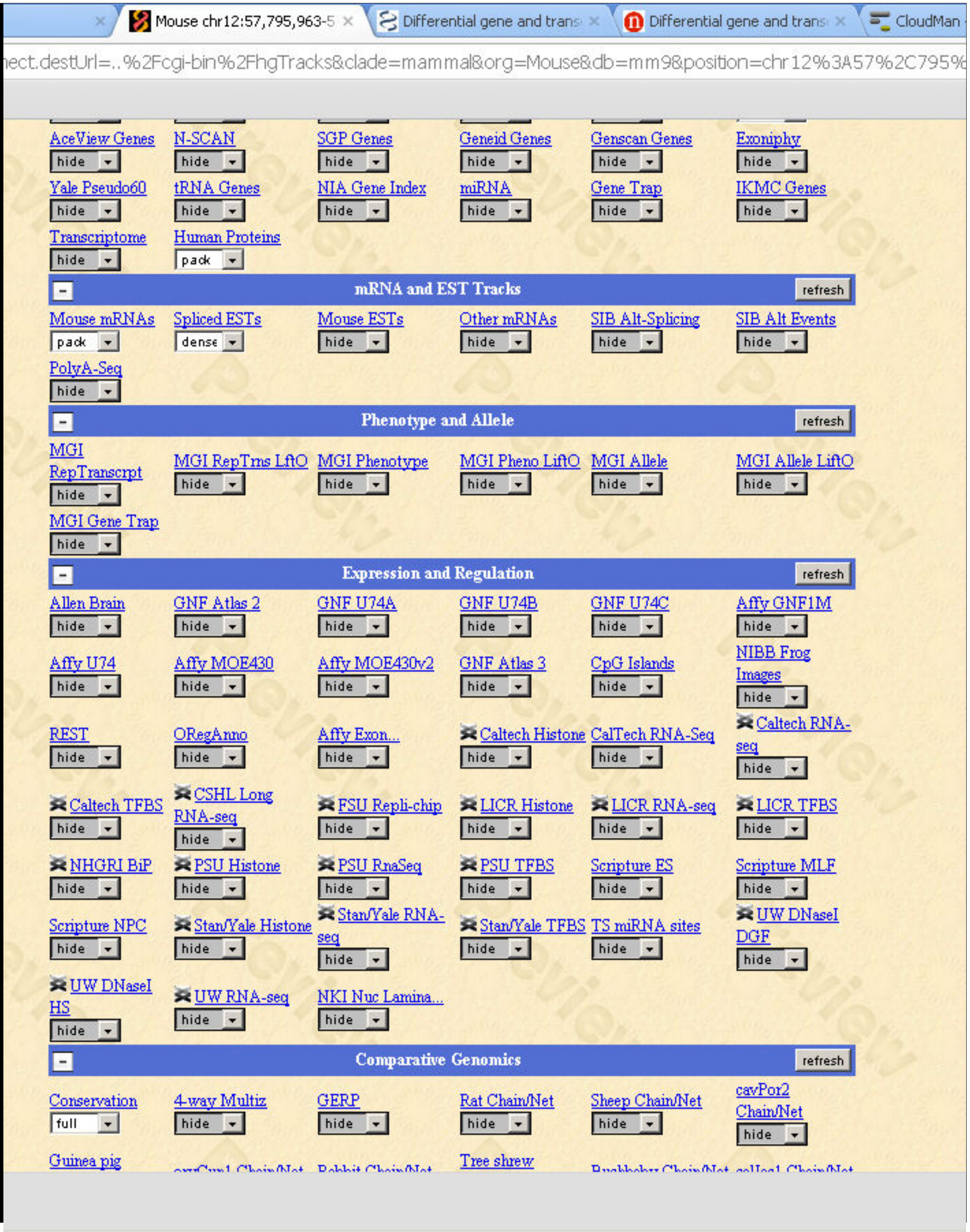
Cole Trapnell[1,2], Adam Roberts[3], Loyal Goff[1,2,4], Geo Pertea[5,6], Daehwan Kim[5,7], David R Kelley[1,2], Harold Pimentel[3], Steven L Salzberg[5,6], John L Rinn[1,2] & Lior Pachter[3,8,9]

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. [3]Department of Computer Science, University of California, Berkeley, California, USA. [4]Computer Science and Artificial Intelligence Lab, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Department of Medicine, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [6]Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland, USA. [7]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA. [8]Department of Mathematics, University of California, Berkeley, California, USA. [9]Department of Molecular and Cell Biology, University of California, Berkeley, California, USA. Correspondence should be addressed to C.T. (cole@broadinstitute.org).

Published online 1 March 2012; doi:10.1038/nprot.2012.016



Visualization-UCSC



ect.destUrl=..%2Fcgi-bin%2FhgTracks&clade=mammal&org=Mouse&db=mm9&position=chr12%3A57%2C795%

AceView Genes hide	N-SCAN hide	SGP Genes hide	Geneid Genes hide	Genscan Genes hide	Exoniphy hide
Yale Pseudo60 hide	tRNA Genes hide	NIA Gene Index hide	miRNA hide	Gene Trap hide	IKMC Genes hide
Transcriptome hide	Human Proteins pack				

mRNA and EST Tracks

Mouse mRNAs pack	Spliced ESTs dense	Mouse ESTs hide	Other mRNAs hide	SIB Alt-Splicing hide	SIB Alt Events hide
PolyA-Seq hide					

Phenotype and Allele

MGI RepTranscrpt hide	MGI RepTrns LftO hide	MGI Phenotype hide	MGI Pheno LiftO hide	MGI Allele hide	MGI Allele LiftO hide
MGI Gene Trap hide					

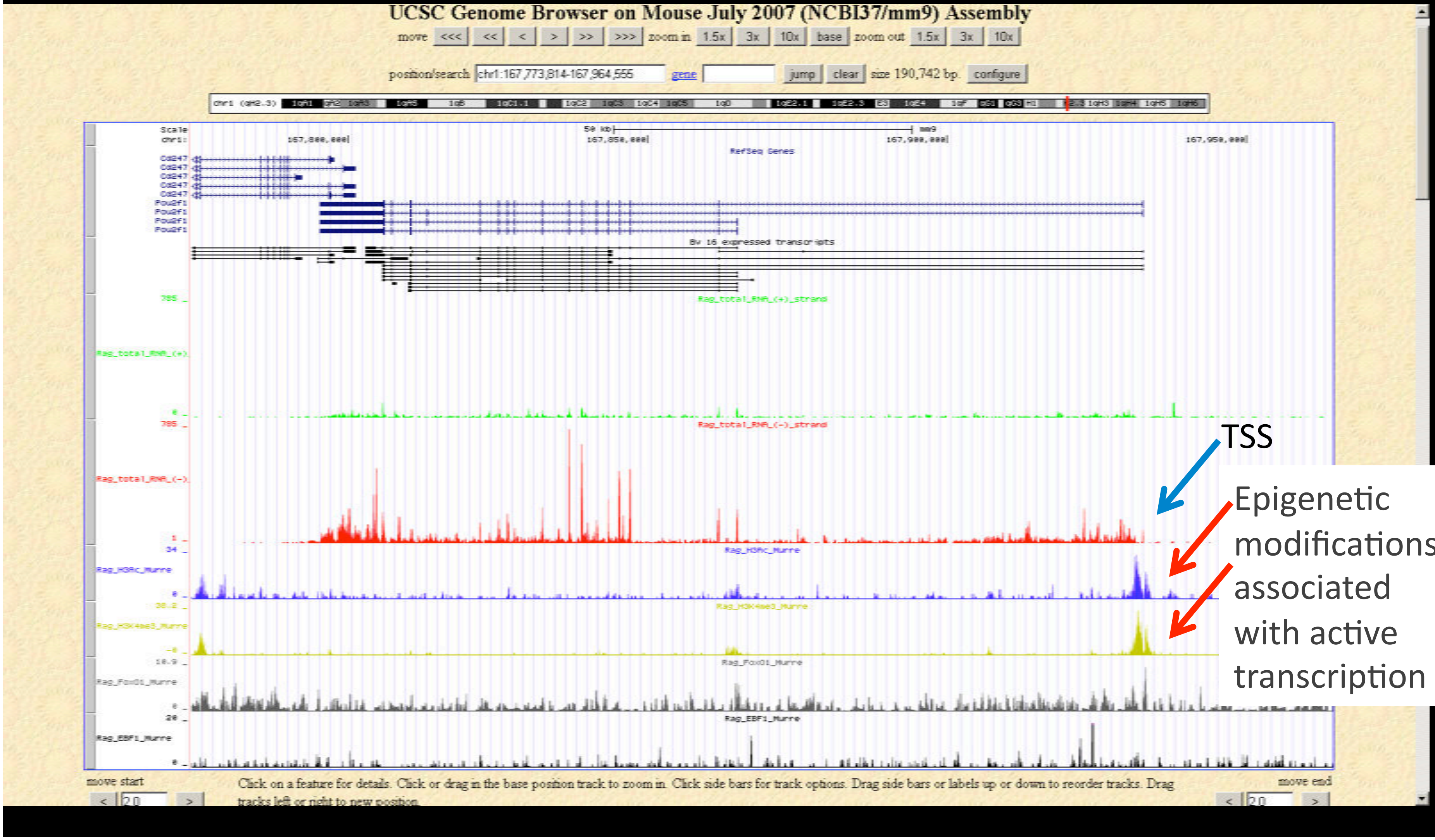
Expression and Regulation

Allen Brain hide	GNF Atlas 2 hide	GNF U74A hide	GNF U74B hide	GNF U74C hide	Affy GNF1M hide
Affy U74 hide	Affy MOE430 hide	Affy MOE430v2 hide	GNF Atlas 3 hide	CpG Islands hide	NIBB Frog Images hide
REST hide	ORegAnno hide	Affy Exon... hide	Caltech Histone hide	CalTech RNA-Seq hide	Caltech RNA-seq hide
Caltech TFBS hide	CSHL Long RNA-seq hide	FSU Repli-chip hide	LICR Histone hide	LICR RNA-seq hide	LICR TFBS hide
NHGRI BiP hide	PSU Histone hide	PSU RnaSeq hide	PSU TFBS hide	Scripture ES hide	Scripture MLF hide
Scripture NPC hide	Stan/Yale Histone hide	Stan/Yale RNA-seq hide	Stan/Yale TFBS hide	TS miRNA sites hide	UW DNaseI DGF hide
UW DNaseI HS hide	UW RNA-seq hide	NKI Nuc Lamina... hide			

Comparative Genomics

Conservation full	4-way Multiz hide	GERP hide	Rat Chain/Net hide	Sheep Chain/Net hide	cavPor2 Chain/Net hide
Guinea pig	oryCun1 Chain/Net	Rabbit Chain/Net	Tree shrew	Bushbaby Chain/Net	calJac1 Chain/Net

Visualization-UCSC



UCSC Genome Browser on Mouse July 2007 (NCBI37/mm9) Assembly

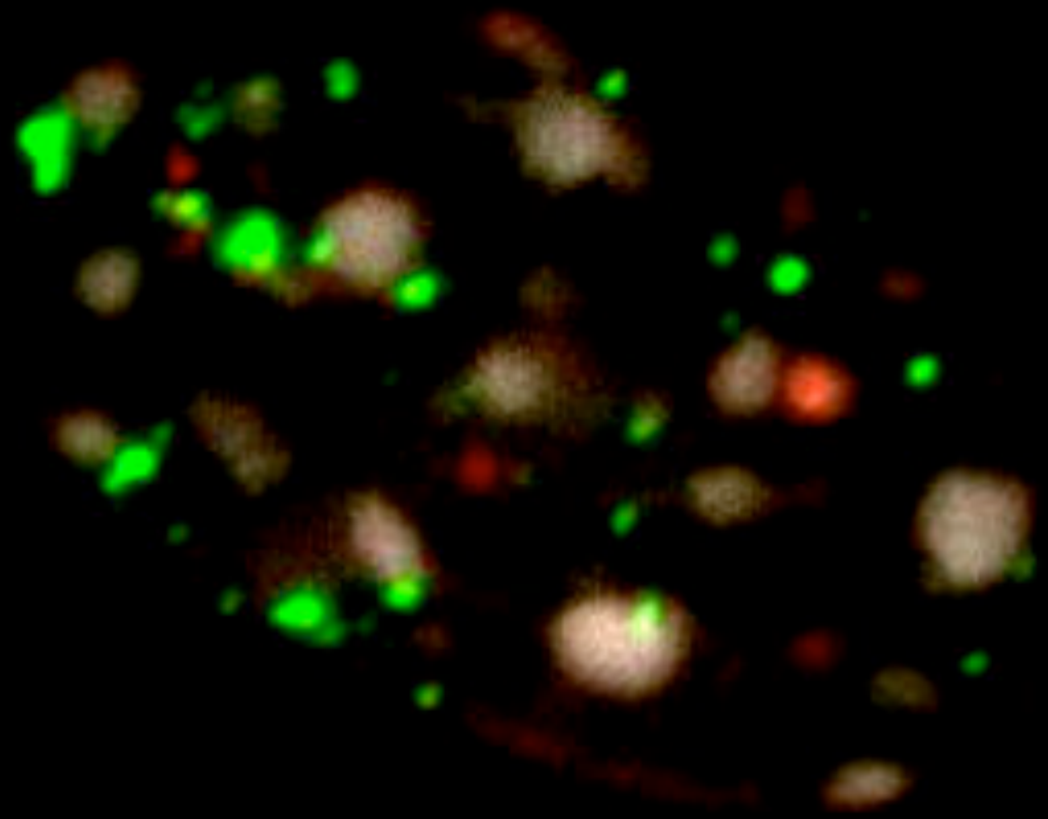
position/search chr1:167,773,814-167,964,555 size 190,742 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

TSS

Epigenetic modifications associated with active transcription

The ultimate visualization (for us)



Issues encountered on the cloud

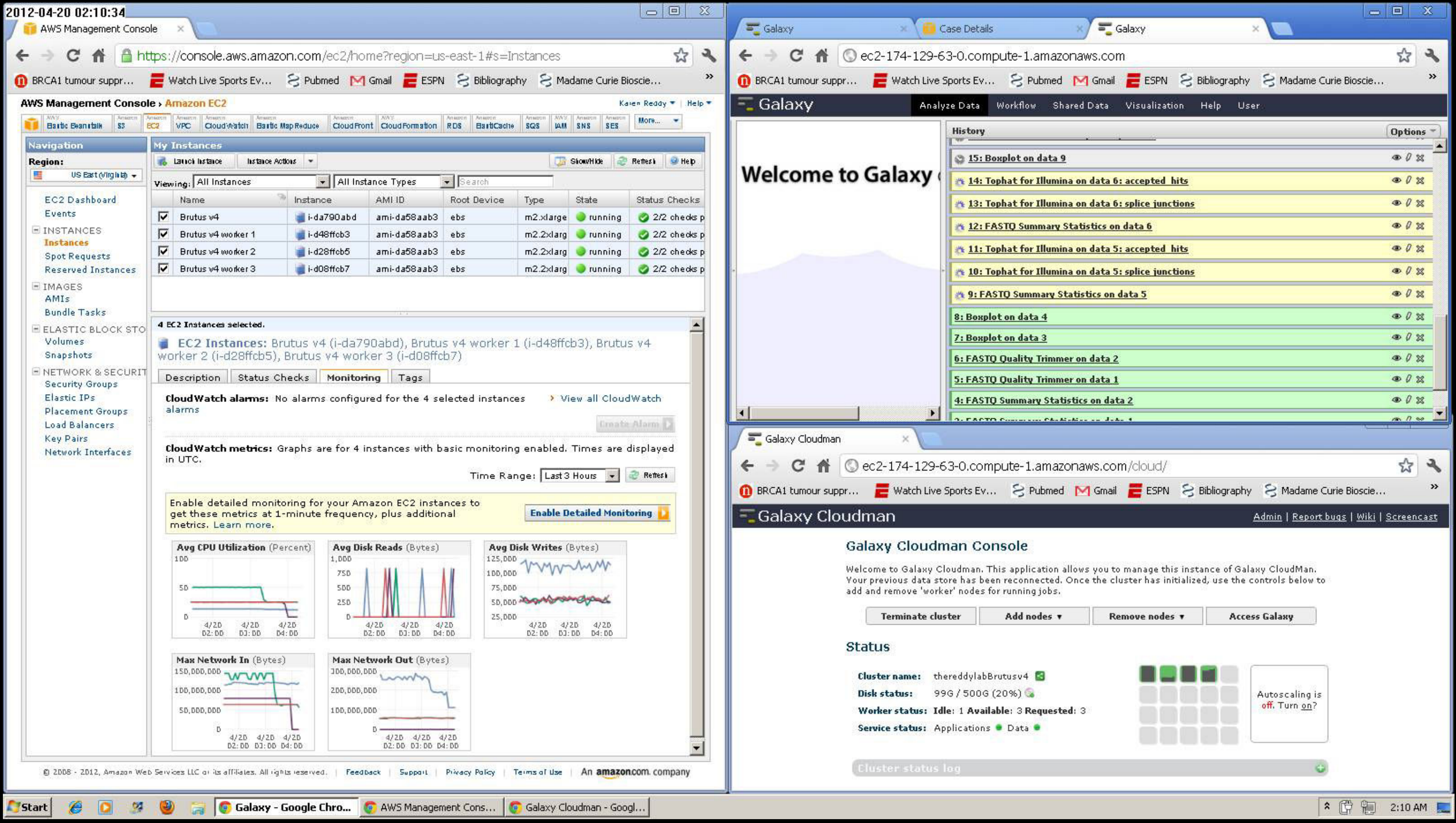
- Institutional issues
  - IT and bioinformaticians not fluent in cloud based computing
  - University does not have clear policies about data on the cloud (especially important for clinical data)

- Not many issues with using—we have been pretty happy…HOWEVER

We have encountered two major issues:

Issue 1 (which led to issue 2)

– Judging capacity planning (http://wiki.g2.bx.psu.edu/CloudMan/CapacityPlanning)



	Name	Instance	AMI ID	Root Device	Type	State	Status Checks
☑	Brutus v4	i-da790abd	ami-da58aab3	ebs	m2.xlarge	● running	✓ 2/2 checks p
☑	Brutus v4 worker 1	i-d48ffcb3	ami-da58aab3	ebs	m2.2xlarg	● running	✓ 2/2 checks p
☑	Brutus v4 worker 2	i-d28ffcb5	ami-da58aab3	ebs	m2.2xlarg	● running	✓ 2/2 checks p
☑	Brutus v4 worker 3	i-d08ffcb7	ami-da58aab3	ebs	m2.2xlarg	● running	✓ 2/2 checks p

4 EC2 Instances selected.

EC2 Instances: Brutus v4 (i-da790abd), Brutus v4 worker 1 (i-d48ffcb3), Brutus v4 worker 2 (i-d28ffcb5), Brutus v4 worker 3 (i-d08ffcb7)

CloudWatch alarms: No alarms configured for the 4 selected instances › View all CloudWatch alarms

CloudWatch metrics: Graphs are for 4 instances with basic monitoring enabled. Times are displayed in UTC.

Enable detailed monitoring for your Amazon EC2 instances to get these metrics at 1-minute frequency, plus additional metrics. Learn more.

History

- 15: Boxplot on data 9
- 14: Tophat for Illumina on data 6: accepted_hits
- 13: Tophat for Illumina on data 6: splice junctions
- 12: FASTQ Summary Statistics on data 6
- 11: Tophat for Illumina on data 5: accepted_hits
- 10: Tophat for Illumina on data 5: splice junctions
- 9: FASTQ Summary Statistics on data 5
- 8: Boxplot on data 4
- 7: Boxplot on data 3
- 6: FASTQ Quality Trimmer on data 2
- 5: FASTQ Quality Trimmer on data 1
- 4: FASTQ Summary Statistics on data 2

Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application allows you to manage this instance of Galaxy CloudMan. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to add and remove 'worker' nodes for running jobs.

Status

Cluster name: thereddylabBrutusv4
Disk status: 99G / 500G (20%)
Worker status: Idle: 1 Available: 3 Requested: 3
Service status: Applications ● Data ●

We use "Brutus" configuration now: head=High Memory XL
Worker(s)=High Memory 2XL

Issue 2: Keys, keys and more keys

Secret key

Security Key

Private Key

Public Key

Key pair

Access Key

Connecting EBS to EC2 drives

Keys, keys and more keys

What key(s) do we need for what?

A bit of difficulty in biologists communicating with technical support (*which key??*)

Lessons learned and words of advice

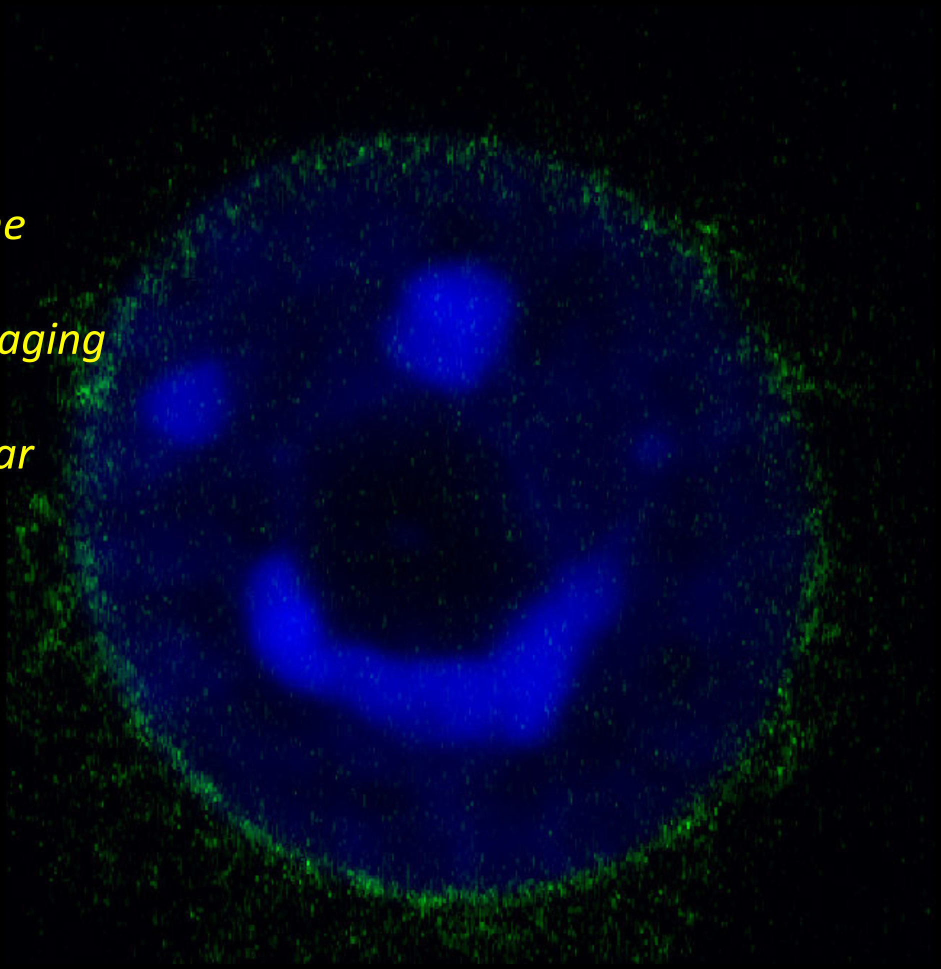
- Better communication between tech support and the end user would make life easier
- Use the tutorial/wiki (http://wiki.g2.bx.psu.edu/CloudMan)
- Capacity planning. Go big.
  - We like using at least Hi-Mem Double XL workers
- Keep small head node up and running (EC2) and shunt all jobs to workers—this will avoid mounting issues and data loss.

Acknowledgements:

Mo Heydarian- RNA –seq, lncRNA, proteome

Xianrong "Jose" Wong-nuclear structure in aging

Teresa Romeo—DamID mapping and nuclear structure in development



Agilent (tiled OLIDS-chromosome paints)

Alice Yamada and team

GALAXY TEAM AND COMMUNITY!!!!!!!