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
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
- 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 available
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

Greil, Moorman and van Steensel 2006
DNA Adenine Methylase Identification DamID

1. Create Dam-Protein X Fusion Vectors
2. Virally Transduce Cells and Allow for Protein Expression
3. Prepare Genomic DNA
4. DpnI Digest gDNA (Cuts methylated GATCs)
5. Ligate on Adaptors
6. DpnII Digest DNA (Cuts unmethylated GATCs)
7. PCR Amplification
8. Analyze DNA by:
   - q-PCR
   - Microarrays
Randomisation:
End Repair and ligate

Sonication

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

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

*Merge 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 fragment 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](https://main.g2.bx.psu.edu/u/jeremy/p/galaxy-rna-seq-analysis-exercise)
On-line tutorials, guides and workflows invaluable
**PROTOCOL**

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

Cole Trapnell$^{1,2}$, Adam Roberts$^3$, Loyal Goff$^{1,3}$, Geon Perteas$^4$, Daehwan Kim$^{1,2}$, David R. Kelley$^5$, Harold Pimentel$^6$, Steven L. Salzberg$^7$, John L. Rinn$^8,9$ & Lior Pachter$^{8,9}$

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

Published online 1 March 2012. doi:10.1038/aspect.2012.015
Visualization-UCSC
Visualization-UCSC
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)

We use “Brutus” configuration now: head=High Memory XL
Worker(s)=High Memory 2XL
Issue 2: Keys, keys and more keys

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, IncRNA, 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!!!!!!!