Galaxy for Biologists

A hands-on workshop

Indiana University
19 October 2012

Dave Clements
Emory University

http://galaxyproject.org/

NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT
INDIANA UNIVERSITY

THE CENTER FOR GENOMICS AND BIOINFORMATICS

INDIANA UNIVERSITY

Galaxy

Agenda

Welcome

Basic Analysis with Galaxy

Basic Analysis into Reusable Workflows

NGS Quality Control (time allowing)

Slides on wiki page: http://bit.ly/iugxy

Acknowledgements

Richard LeDuc
William Barnett
Scott Michaels
Radhika Khetani

National Center for Genome Analysis Support (NCGAS)
The Center for Genomics and Bioinformatics
Indiana University

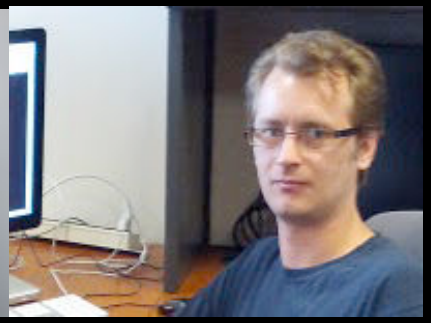
NIH NSF Huck Institute
Penn State University Emory University


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

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

Welcome

Basic Analysis with Galaxy

Basic Analysis into Reusable Workflows

NGS Quality Control (time allowing)

Hands On: Basic Analysis

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

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

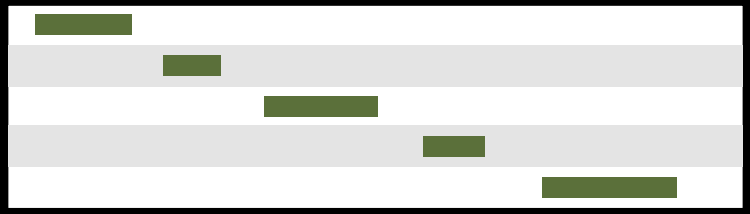
Repetitious Pigs: A Rough Plan

- 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

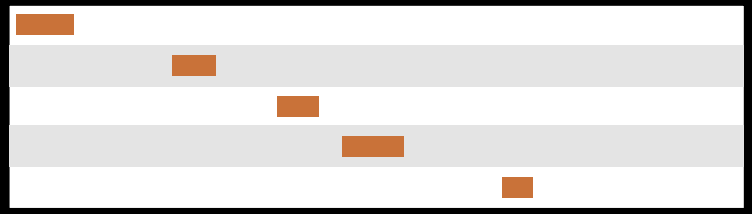
Go to Galaxy

http://bit.ly/IUcrimson

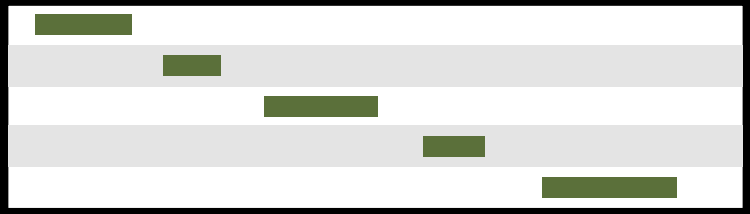
http://bit.ly/IUcream



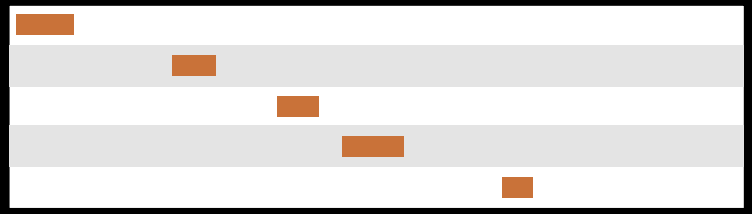
Exons, from UCSC



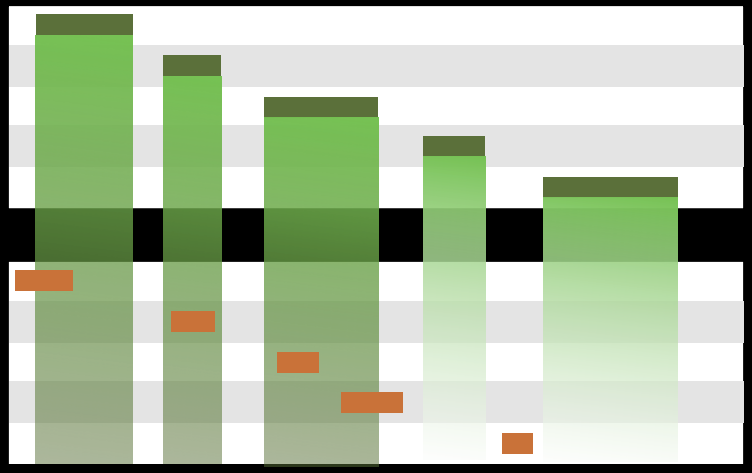
Repeats, from UCSC



Exons, from UCSC



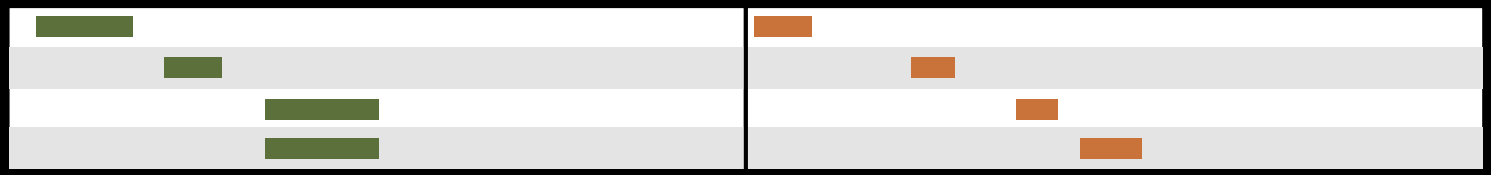
Repeats, from UCSC

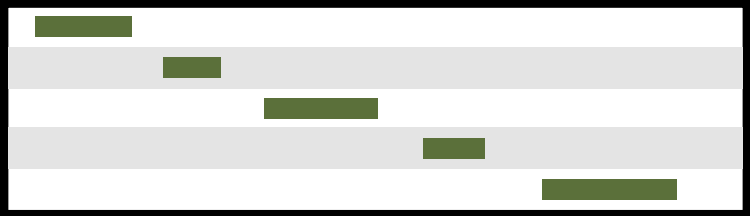


Exons, from UCSC

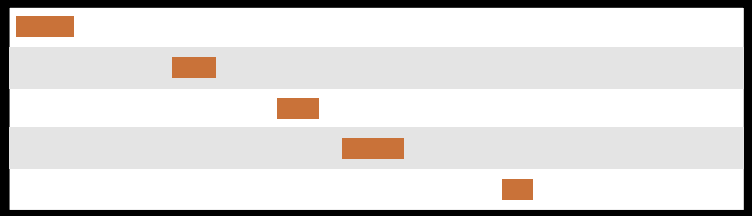
Repeats, from UCSC

Overlap pairings

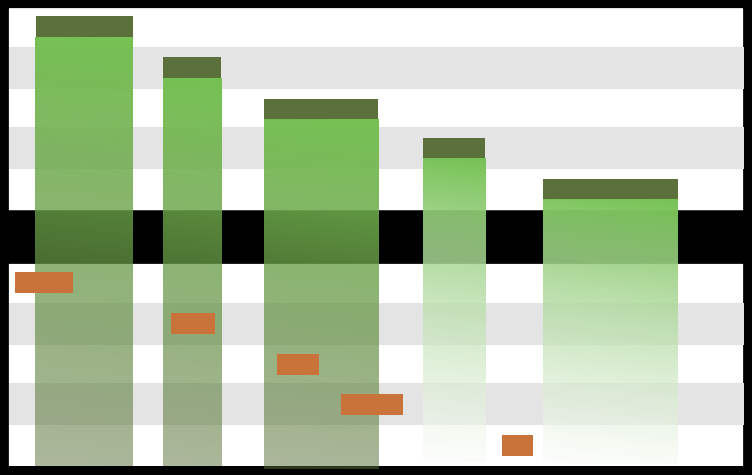




Exons, from UCSC



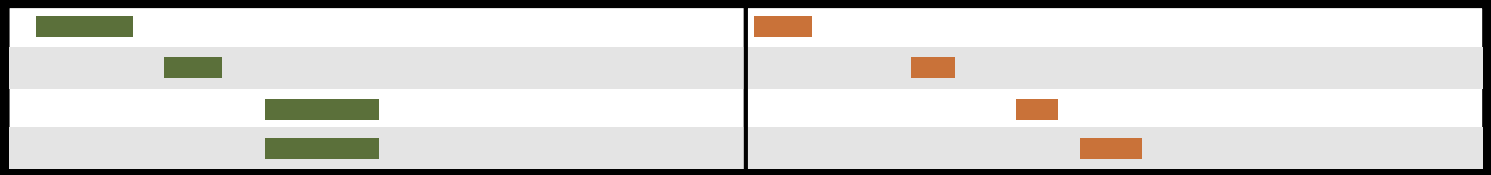
Repeats, from UCSC



Exons, from UCSC

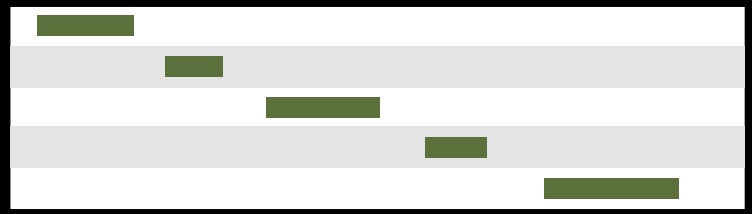
Repeats, from UCSC

Overlap pairings



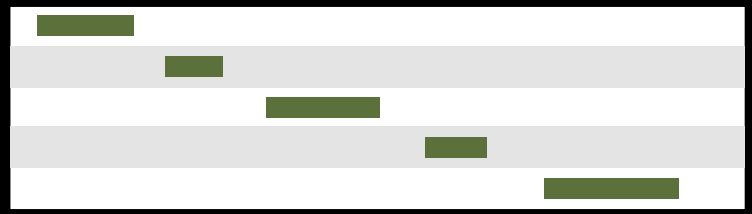
Exon overlap counts

Exon overlap counts



Exons, from UCSC

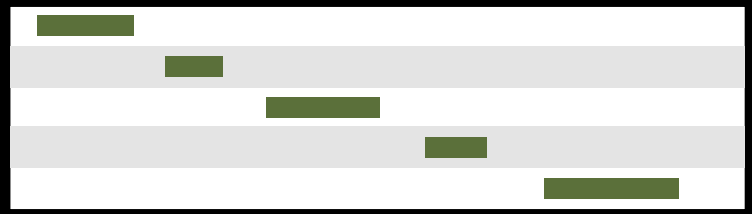
Exon overlap counts



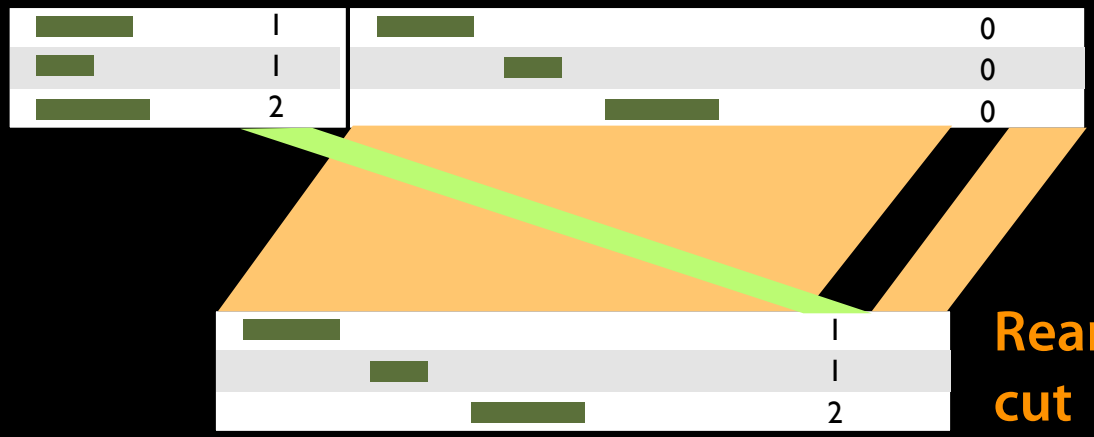
Exons, from UCSC

Join on exon name

Exon overlap counts



Exons, from UCSC



Join on exon name

Rearrange columns w/ cut

Agenda

Welcome

Basic Analysis with Galaxy

Basic Analysis into Reusable Workflows

NGS Quality Control (time allowing)

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.

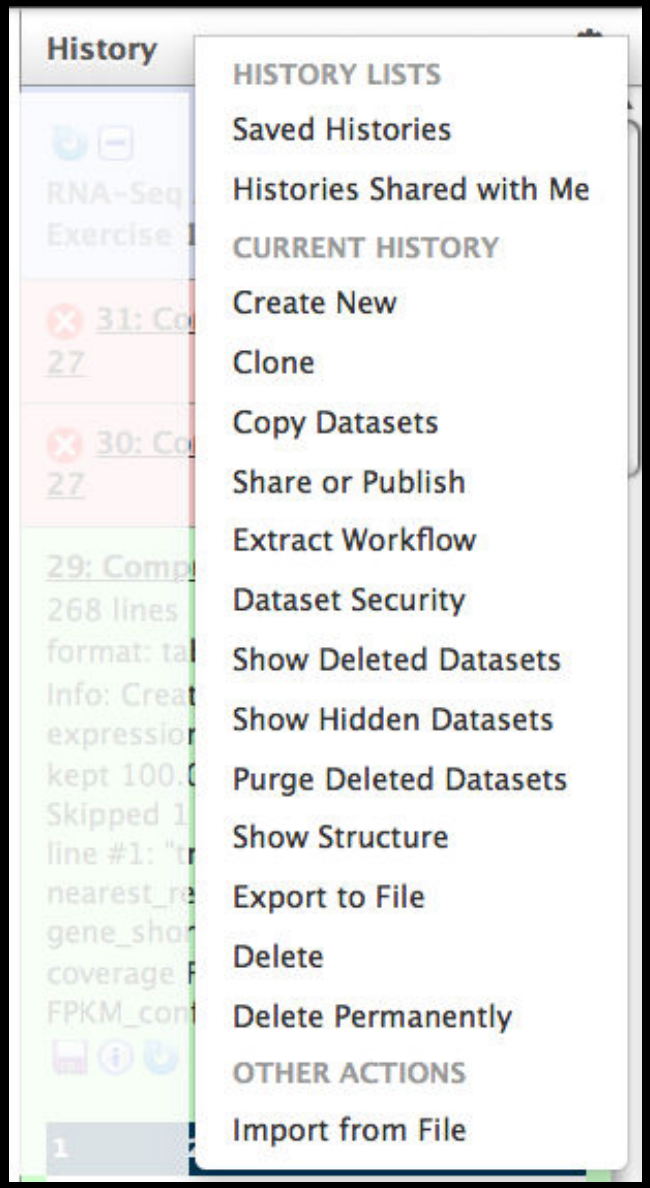
Resume any previous history

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

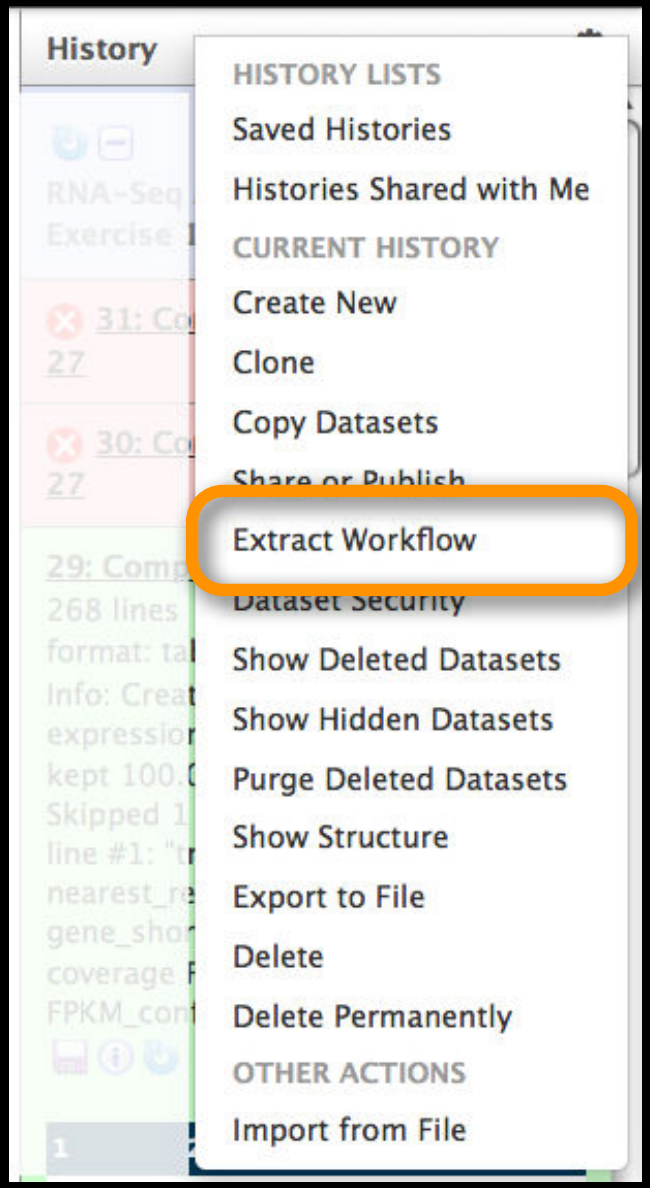
Guided: rerun with same inputs

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



Agenda

Welcome

Basic Analysis with Galaxy

Basic Analysis into Reusable Workflows

NGS Quality Control (time allowing)

FASTQ Format

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!*

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

S - Sanger        Phred+33,  93 values  (0, 93) (0 to 60 expected in raw reads)
I - Illumina 1.3 Phred+64,  62 values  (0, 62) (0 to 40 expected in raw reads)
X - Solexa        Solexa+64, 67 values (-5, 62) (-5 to 40 expected in raw reads)
```

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

We'll do the early steps of a ChIP-Seq Exercise:

(but this also applies to lots of other NGS data)

- 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.
- Identify zinc-finger CTCF transcription factor tags in mouse
- All datasets are FASTQ

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 → Data Libraries
    - ChIP-Seq Datasets
      - Import all

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: Still interested?

See the rest of the slides at the end of the talk.

- 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)

Agenda

Welcome
Basic Analysis with Galaxy
Basic Analysis into Reusable Workflows
NGS Quality Control (time allowing)

The Galaxy Needs *You!*

http://galaxyproject.org/wiki/GalaxyIsHiring

Galaxy URLs to Remember

http://galaxyproject.org

http://usegalaxy.org

http://getgalaxy.org

Workshop Feedback

Please help.

http://bit.ly/IUFeedbackG4B

Thanks

http://bit.ly/IUFeedbackG4B

http://galaxyproject.org/GCC2013

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?*

ChIP-Seq Exercise: The Rest of the 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.

Keywor

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.

Advice on many NGS tools in Galaxy