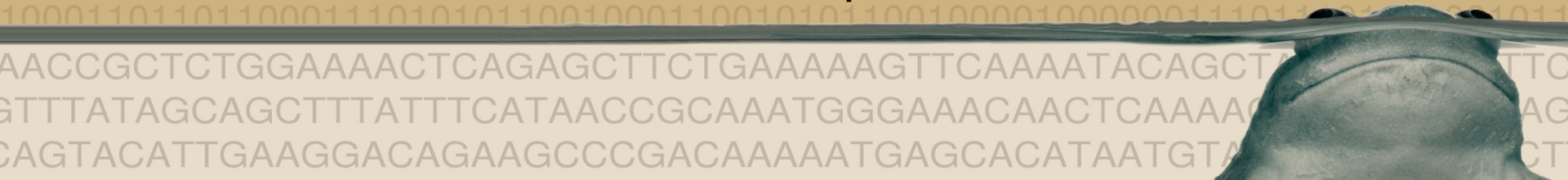


Xenopus functional genomics: Computational analysis of ChIP-seq data

Simon J. van Heeringen

14th International *Xenopus* conference



Support and more info

<http://simonvh.github.com/xenopus2012>

- Contains all presentations
- Links to databases, tools and materials

Mail me: s.vanheeringen@ncmls.ru.nl

Our website:

<http://www.ncmls.nl/gertjanveenstra>

Beyond browsing

- Previous workshop showed a visual approach
 - Good to look at your favorite gene
 - Generate hypothesis
- Usually, a more detailed analysis is necessary
 - Visual inspection can be deceiving
 - Need to know general, genome-wide patterns

Disclaimer

- There are commercial solutions
 - Genomatix
 - CLC bio
 - Avadis
 - And other...
- However, check if they support *Xenopus*!

Disclaimer

- There are commercial solutions
- If you're going to do a lot of analysis, it's worth learning command-line approaches
 - Even if you don't have your own Linux server / computer (“cloud”, Amazon EC2)

Disclaimer

- There are commercial solutions
- If you're going to do a lot of analysis, it's worth learning command-line approaches
- This is an introduction to basic analysis using freely available (web) resources
 - Which can get you quite far

ChIP-seq workflow

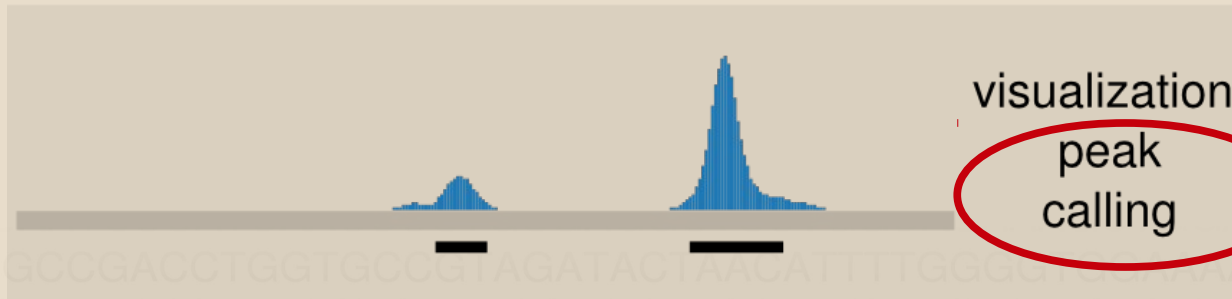
Data type

```
TAACGTGAACCCCTCTATCTTCCTTCACAGATTG
TAGTTTCTCACTTCAAGTTATCCAGCAACCTTGGA
TTTGAACATCATGTTCTGTCATGTTTGGTGCTTG
GACTGCTAATATCCTTATCATTACAAAAGGGTAC
GCTTTTACATTTCGGACCACTTAATAAATGACTAG
TCCTTATCCTATGCTCTTATACCCCATATTACTGC
CAGAACAGGAATGAGGGGTCTCTAAATGGCTGATA
CTGCTAAATGTCAATAACTATAATAGCTATGATT
TGTGGTATTTTATCAAATACATGTTTAAACAAATG
TCCCTATCTTTAAATCCAGTGCCTAAAGAATTG
```

raw
reads



aligned
data



visualization,
peak
calling

Peak calling

- Sounds easy, doesn't it?

Peak calling

- Sounds easy, doesn't it?

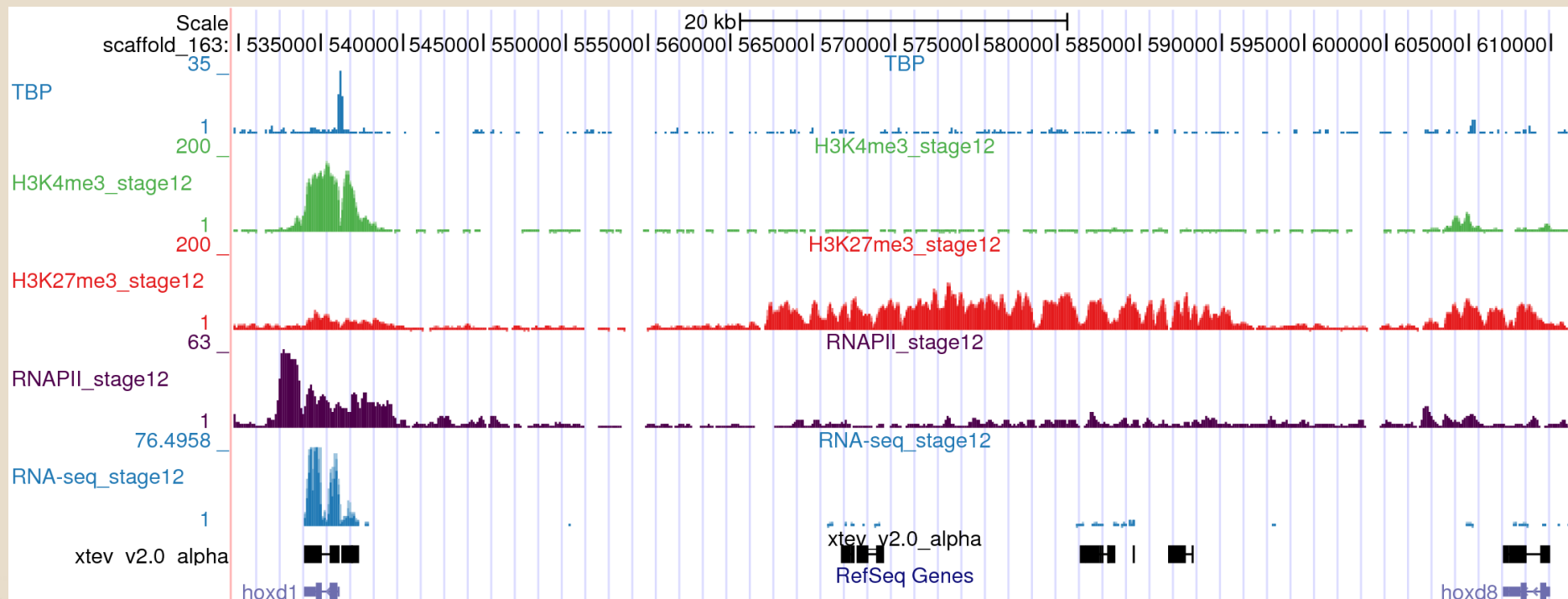
TBF gastrula



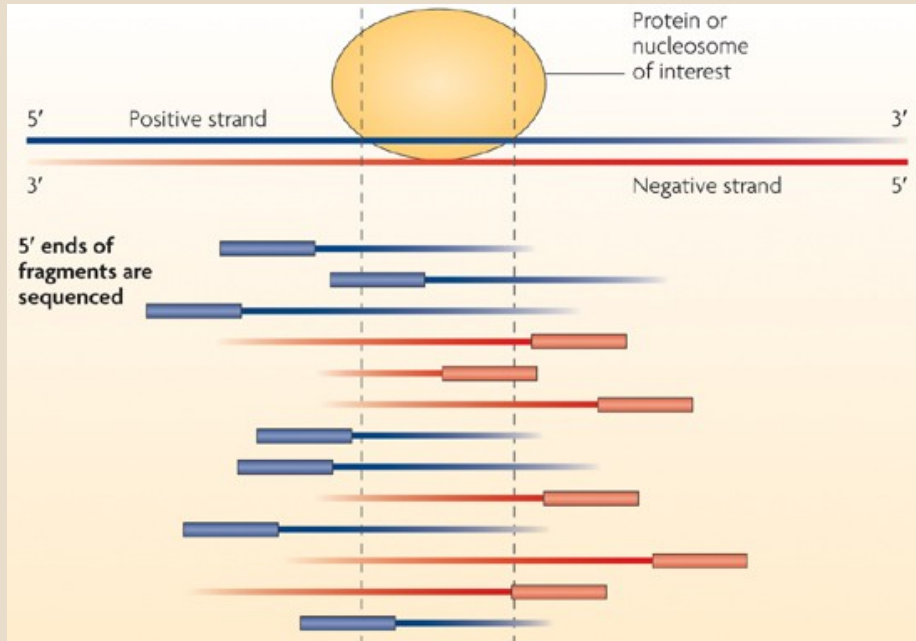
Peak!

Peak calling

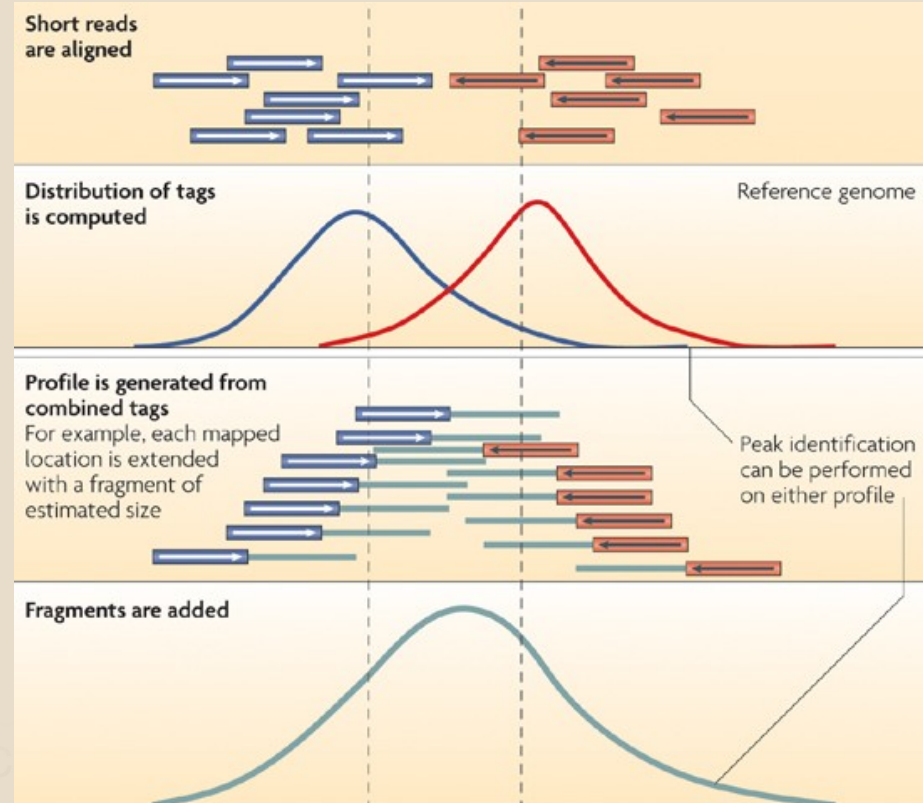
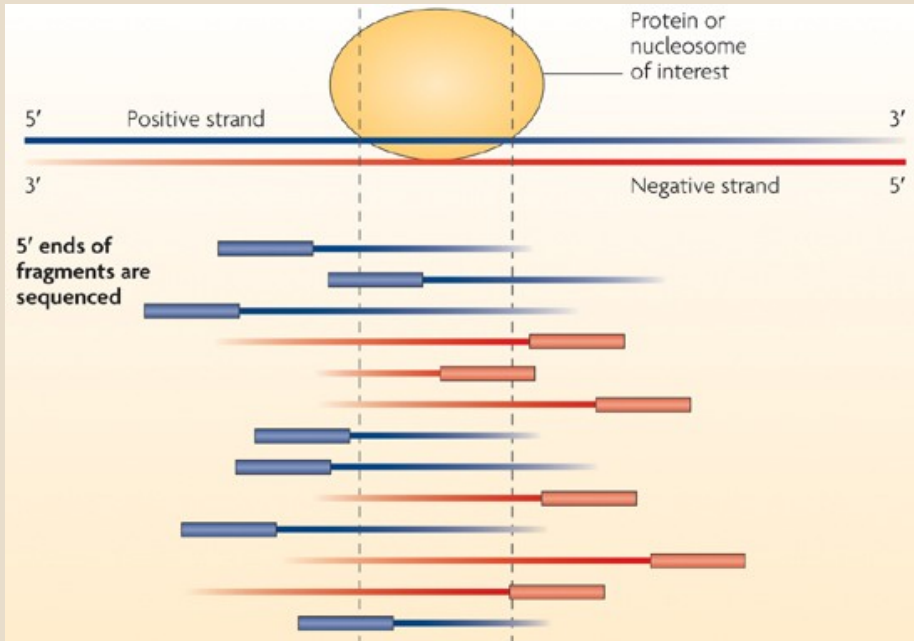
- One peak-caller to call them all?



Strand information



Strand information



Which one to use?

- Many options, I've lost count...
- Many don't support *Xenopus* out-of-the-box
 - “What's 'scaffolds', Precious?”
- My personal (likely biased) advice
 - MACS (widely used)
 - With or without control
 - PeakRanger (modENCODE)
 - With control

• YMMV

Galaxy

- <https://main.g2.bx.psu.edu/>
- Web-based analysis
- Easy-to-use
- Lots of tutorials
- First choice if:
 - you don't have your own analysis server
 - you are unfamiliar with Linux, command-line based tools

Opera

X. tropicalis scaffol...

Galaxy

GEO Accession vie...

Secure

main.g2.bx.psu.edu

galaxy genome

Galaxy

Analyze DataWorkflowShared DataVisualizationCloudHelpUser

Using 0%

Tools

search tools

Get Data

Send Data

ENCODE Tools

Lift-Over

Text Manipulation

Convert Formats

FASTA manipulation

Filter and Sort

Join, Subtract and Group

Extract Features

Fetch Sequences

Fetch Alignments

Get Genomic Scores

Operate on Genomic Intervals

Statistics

Graph/Display Data

Regional Variation

Multiple regression

Multivariate Analysis

Evolution

Motif Tools

Multiple Alignments

Metagenomic analyses

Phenotype Association

Genome Diversity

EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation

NGS: Mapping

NGS: SAM Tools

NGS: GATK Tools (beta)

NGS: Indel Analysis

NGS: Peak Calling

NGS: RNA Analysis

nobula

.curio.fr

A Galaxy server
for advanced
ChIP-seq analysis

Live Quickies

Q

on:

13

Advanced fastQ
manipulation:

Galactic quickie # 14

454 Mapping:
Single End

Galactic quickie # 15

Uploading Data
using FTP

Galactic quickie # 17

Managing
account histories

Galactic quickie # 19

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or [your own instance](#), you can perform, reproduce, and share complete analyses. The [Galaxy team](#) is a part of [BX at Penn State](#), and the [Biology and Mathematics and Computer Science](#) departments at [Emory University](#). The [Galaxy Project](#) is supported in part by [NSF](#), [NHGRI](#), [The Huck Institutes of the Life Sciences](#), [The Institute for CyberScience at Penn State](#), and [Emory University](#).

Galaxy build: \$Rev 7614:e6444e7a1685\$

galaxyproject

galaxyproject Advanced NGS Course: RNA-seq. @ Amsterdam Medical Centre (AMC) 15 - 17 October. Galaxy & R practicals. [bit.ly/O8TLHc](#) #usegalaxy 2 days ago · reply · retweet · favorite

caseybergman .@nvosa shows us how to use phylotastic tools inside galaxy ([youtu.be/KMME858xOu4](#)) #phylogenetics #bioinformatics #usegalaxy 4 days ago · reply · retweet · favorite

caseybergman .@nvosa shows us how to add a phylotastic service to galaxy ([youtu.be/d-fDngweW-M](#)) #phylogenetics #bioinformatics #usegalaxy 4 days ago · reply · retweet · favorite

more ...

This is a free, public, internet accessible resource. Data transfer and data storage are not encrypted. If there are restrictions on the way your research data can be stored and used, please consult your local institutional review board or the project PI before uploading it to any public site, including this Galaxy server. If you have protected data, large data storage requirements, or short deadlines you are encouraged to setup your own [local Galaxy instance](#) or run [Galaxy on the cloud](#).

History

0 bytes

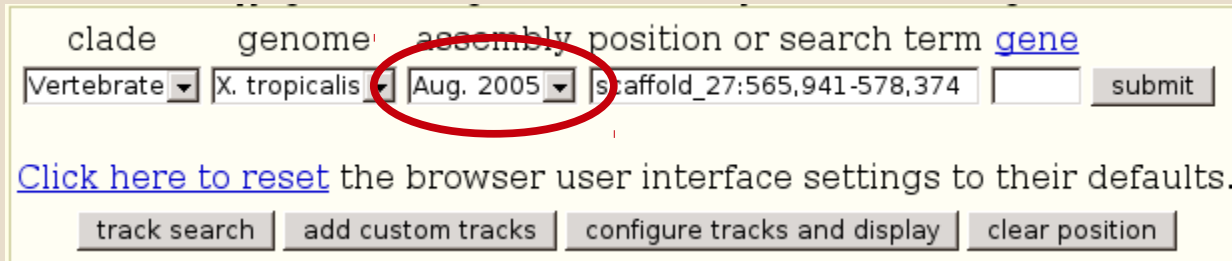
Your history is empty. Click 'Get Data' on the left pane to start

Example

- Peak calling using Galaxy
- TBP ChIP-seq (van Heeringen et al., 2011)
- Download from GEO
- Peak calling with MACS
- Upload peaks to Genome Browser for inspection

Try it out yourself!

- Resources are linked on the tutorial page
- Notice: I'm using **JGI 4.1 / xenTro2** data



The screenshot shows a web interface with the following elements:

- Labels: clade, genome, assembly, position or search term, [gene](#)
- Dropdown menus: Vertebrate, X. tropicalis, Aug. 2005 (circled in red)
- Text input: scaffold_27:565,941-578,374
- Submit button: submit
- Link: [Click here to reset](#) the browser user interface settings to their defaults.
- Buttons: track search, add custom tracks, configure tracks and display, clear position


Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > Accession Display [?] Not logged in | [Login](#) [?]

Scope: Format: Amount: GEO accession:

Sample GSM537039

[Query DataSets for GSM537039](#)

Status	Public on Dec 30, 2010
Title	TBP_ChIPSeq
Sample type	SRA
Source name	whole embryo
Organism	Xenopus (Silurana) tropicalis
Characteristics	development stage: 12 antibody: SL33 against TBP
Growth protocol	Xenopus tropicalis embryos were obtained from a natural mating procedure after human chorion gonadotropin injection, dejellied in 3% cysteine and collected at Nieuwkoop-Faber stage 12
Extracted molecule	genomic DNA
Extraction protocol	Chromatin (400 embryos) was fixed in 1% formaldehyde, sheared using a Branson sonifier, immunoprecipitated using 2 µl of antibody (SL33 against TBP, Ruppert et al. 1996, PMID: 9064287), washed, decrosslinked, DNA was purified and sequenced using a Illumina Genome Analyzer according to the manufacturer.
Library strategy	ChIP-Seq
Library source	genomic
Library selection	ChIP
Instrument model	Illumina Genome Analyzer
Description	Chromatin IP against TBP
Data processing	<p>Reads were mapped to the X.tropicalis genome (version JGI4.1) with ELAND, reads in identical positions were removed and peaks were called using MACS with p-value 1e-7</p> <p>The included bed-files contain xenTro2/JGI4.1 coordinates. These tab-delimited files contain up to 6 columns according to the BED specification.</p> <ol style="list-style-type: none">1) chromosome2) start3) end4) score (see description below)5) not used6) strand <p>The file TBP.bed contains the coordinates of the filtered unique 35-basepair reads, mapped strand-specific to the xenTro2 genome.</p> <p>The TBP_peaks.bed file contains the peaks as determined by MACS using a</p>

using MACS with p-value 1e-7

The included bed-files contain xenTro2/JGI4.1 coordinates. These tab-delimited files contain up to 6 columns according to the BED specification.

- 1) chromosome
- 2) start
- 3) end
- 4) score (see description below)
- 5) not used
- 6) strand

The file TBP.bed contains the coordinates of the filtered unique 35-basepair reads, mapped strand-specific to the xenTro2 genome.
The TBP_peaks.bed file contain the peaks as determined by MACS using a p-value threshold of 1e-7. The score column contains the number of reads in the peak as determined by MACS.

Submission date Apr 22, 2010
Last update date Dec 30, 2010
Contact name Gert Jan Veenstra
E-mail(s) g.veenstra@ncmls.ru.nl
Phone +31 24 3610541
Fax +31-24-3610523
URL <http://www.ncmls.nl/gertjanveenstra/home.htm>
Organization name Radboud University
Department Molecular Biology
Street address Geert Grooteplein 28
City Nijmegen
ZIP/Postal code 6525 GA
Country Netherlands

Platform ID [GPL9405](#)
Series (1) [GSE21482](#) Nucleotide composition-linked divergence of vertebrate core promoter architecture

Relations

SRA [SRX019583](#)

Supplementary file	Size	Download	File type/resource
SRX/SRX019/SRX019583		(ftp)	SRA Experiment
GSM537039_TBP.bed.gz	68.9 Mb	(ftp) (http)	BED
GSM537039_TBP_peak.bed.gz	60.7 Kb	(ftp) (http)	BED

Raw data provided as supplementary file
Processed data provided as supplementary file

Opera

X. laevis - UCSC G... X

Galaxy

GEO Accession vie... X

Web

main.g2.bx.psu.edu

Search with Google

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

 table browser

EBI SRA

 ENA SRA

BioMart

 Central server

GrameneMart

Flymine

modENCODE fly

modENCODE modMine

Ratmine

YeastMine

modENCODE worm

WormBase

EuPathDB

EncodeDB

EpiGRAPH

Send Data

ENCODE Tools

Lift-Over

Text Manipulation

Convert Formats

FASTA manipulation

Filter and Sort

Join, Subtract and Group

Extract Features

Fetch Sequences

Fetch Alignments

Get Genomic Scores

Upload File (version 1.1.3)

File Format:

Auto-detect

Which format? See help below

File:

"/home/simon/tmp/GSN

Choose...

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

File	Size	Date
<div>Please <u>create</u> or <u>log in to</u> a Galaxy account to view files uploaded via FTP.</div>		
<div>This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at main.g2.bx.psu.edu using your Galaxy credentials (email address and password).</div>		

Convert spaces to tabs:

☐ Yes

Use this option if you are entering intervals by hand.

Genome:

opicalis Aug. 2005 (JGI 4.1/xenTro2) (xenTro2)

Execute

Auto-detect

The system will attempt to detect Axt, Fasta, Fastqsolexa, Gff, Gff3, Html, Lav, Maf, Tabular, Wiggle, Bed and Interval (Bed with headers) formats. If your file is not detected properly as one of the known formats, it most likely means that it has some format problems (e.g., different number of columns on different rows). You can still coerce the system to set your data to the format you think it should be. You can also upload compressed files, which will automatically be decompressed.

Ab1

A binary sequence file in 'ab1' format with a '.ab1' file extension. You must manually select this 'File Format' when uploading the file.

History

0 bytes

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

Opera

X. laevis - UCSC G... X

Galaxy

GEO Accession vie... X

Web

main.g2.bx.psu.edu

Elements: 65/65

Search with Google

Galaxy

Analyze Data

Workflow

Shared Data

Visualization

Cloud

Help

User

Using 0%

Tools

Statuses

Graph/Display Data

Regional Variation

Multiple regression

Multivariate Analysis

Evolution

Motif Tools

Multiple Alignments

Metagenomic analyses

Phenotype Association

Genome Diversity

EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation

NGS: Mapping

NGS: SAM Tools

NGS: GATK Tools (beta)

NGS: 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

NGS: Picard (beta)

BEDTools

snpEff

RGENETICS

SNP/WGA: Data: Filters

SNP/WGA: QC: LD: Plots

SNP/WGA: Statistical Models

MACS (version 1.0.1)

Experiment Name:

MACS in Galaxy

Paired End Sequencing:

Single End

ChIP-Seq Tag File:

2: TBP_scaffold_1.bed

ChIP-Seq Control File:

Selection is Optional

Effective genome size:

1500000000

default: 2.7e+9

Tag size:

35

Band width:

300

Pvalue cutoff for peak detection:

1e-05

default: 1e-5

Select the regions with MFOLD high-confidence enrichment ratio against background to build model:

32

Parse xls files into into distinct interval files:

Save shifted raw tag count at every bp into a wiggle file:

Do not create wig file (faster)

Use fixed background lambda as local lambda for every peak region:

up to 9X more time consuming

3 levels of regions around the peak region to calculate the maximum lambda as local lambda:

1000,5000,10000

Build Model:

Build the shifting model

Diagnosis report:

Generate diagnosis report (faster)

History

0 bytes

2: TBP_scaffold_1.bed

Opera

Manage Cust...

Galaxy

GEO Accessio...

X. tropicalis s...

Veenstra Lab...

Postvak IN - ...

simonvh/xen...

Xenopus2012

←

→

↺

🔑

Web

main.g2.bx.psu.edu

★

Search with Google

🔍

Galaxy

Analyze Data

Workflow

Shared Data

Visualization

Cloud

Help

User

Using 0%

Tools

⚙️

⌵

Statuses

Graph/Display Data

Regional Variation

Multiple regression

Multivariate Analysis

Evolution

Motif Tools

Multiple Alignments

Metagenomic analyses

Phenotype Association

Genome Diversity

EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation

NGS: Mapping

NGS: SAM Tools

NGS: GATK Tools (beta)

NGS: 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

NGS: Picard (beta)

BEDTools

snpEff

RGENETICS

SNP/WGA: Data: Filters

SNP/WGA: QC: LD: Plots

SNP/WGA: Statistical Models

⏪

⏩

📄

☁️

🔄

📧

✓

The following job has been successfully added to the queue:

3: MACS on data 2 (peaks: bed)

4: MACS on data 2 (html report)

You can check the status of queued jobs and view the resulting data by refreshing the **History** pane.

When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

History

⚙️

2.0 MB

4: MACS on data 2 (html report)

3: MACS on data 2 (peaks: bed)

52 regions, 1 comments

format: bed, database: xenTro2

display at UCSC [main](#)

view in [GeneTrack](#)

display in IGB [Local](#) [Web](#)

display at Ensembl [Current](#)

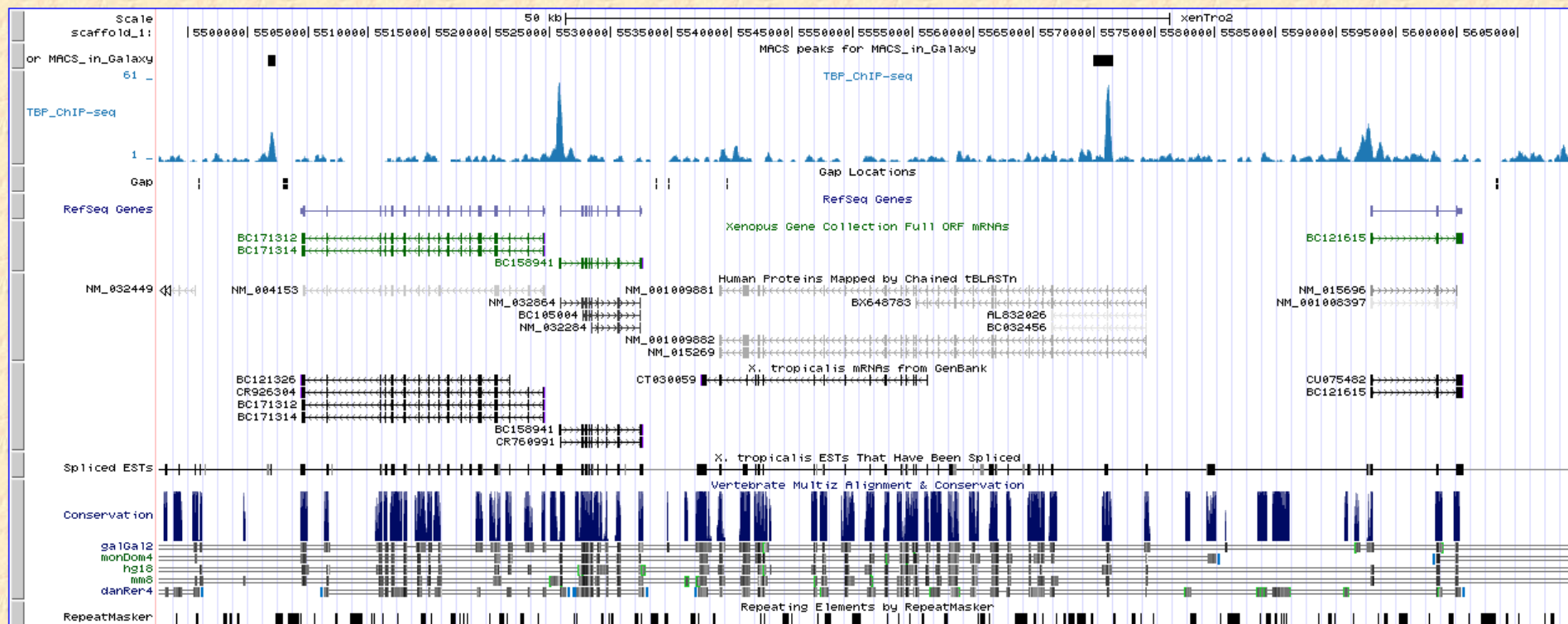
1. Chrom	2. Start	3. End	4. Name	5
track name="MACS peaks for MACS_in_Galaxy"				
scaffold_1	135	425	MACS_peak_1	57.64
scaffold_1	90495	92200	MACS_peak_2	99.93
scaffold_1	46425	46467	MACS_peak_3	64.97
scaffold_1	49413	49572	MACS_peak_4	109.53
scaffold_1	55904	56086	MACS_peak_5	243.24

2: TBP scaffold 1.bed

UCSC Genome Browser on X. tropicalis Aug. 2005 Assembly (xenTro2)

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search scaffold_1:5,492,633-5,609,93 [gene](#) jump clear size 117,300 bp. configure



move start < 2.0 > 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. move end < 2.0 >

track search default tracks default order hide all manage custom tracks configure reverse resize refresh

collapse all

Use drop-down controls below and press refresh to alter tracks displayed.

Tracks with lots of items will automatically be displayed in more compact modes.

expand all



Custom Tracks

refresh

[MACS peaks for](#) [TBP_ChIP-seq](#)
[MACS_in_Galaxy](#)

dense

full

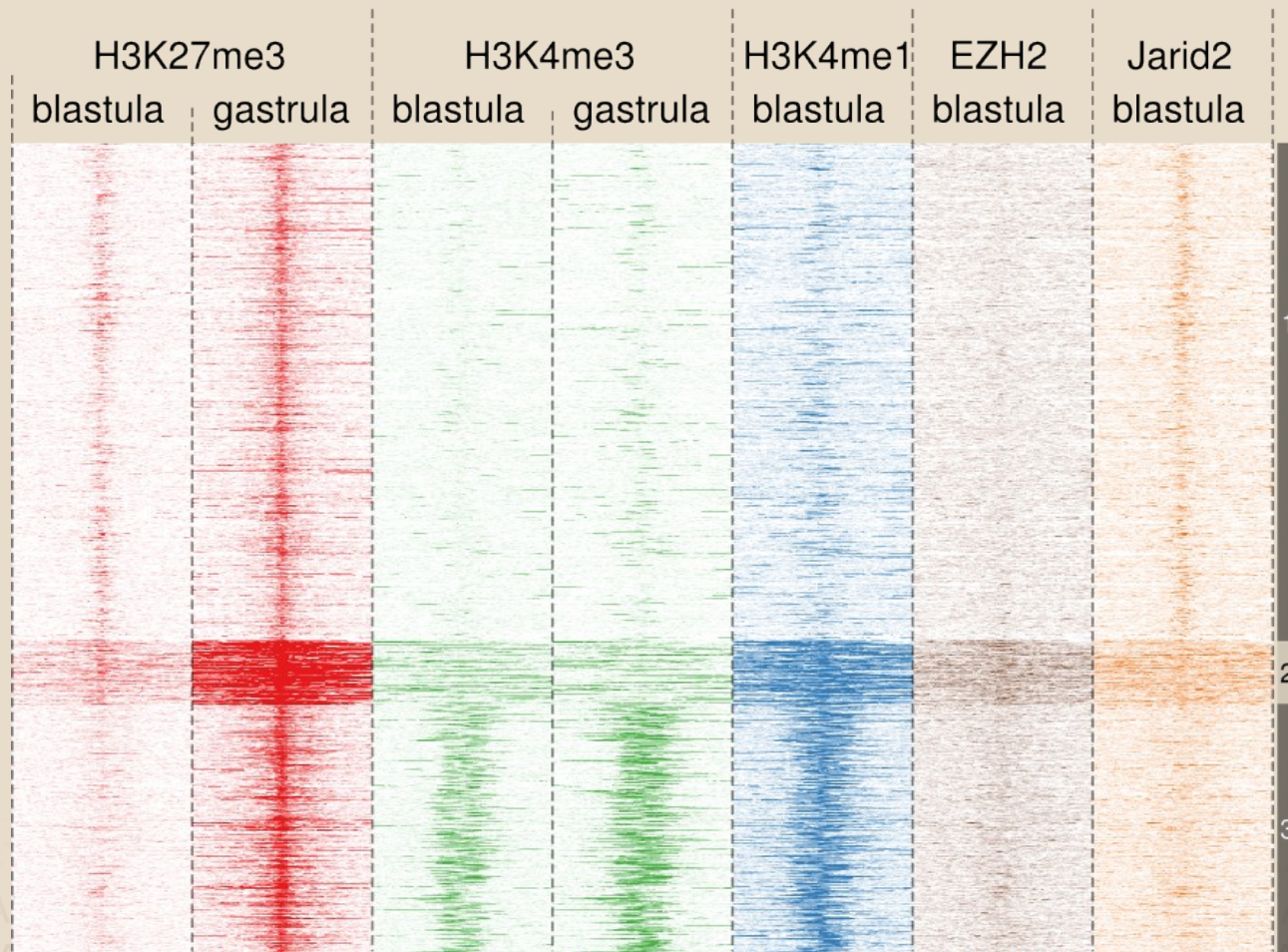


Mapping and Sequencing Tracks

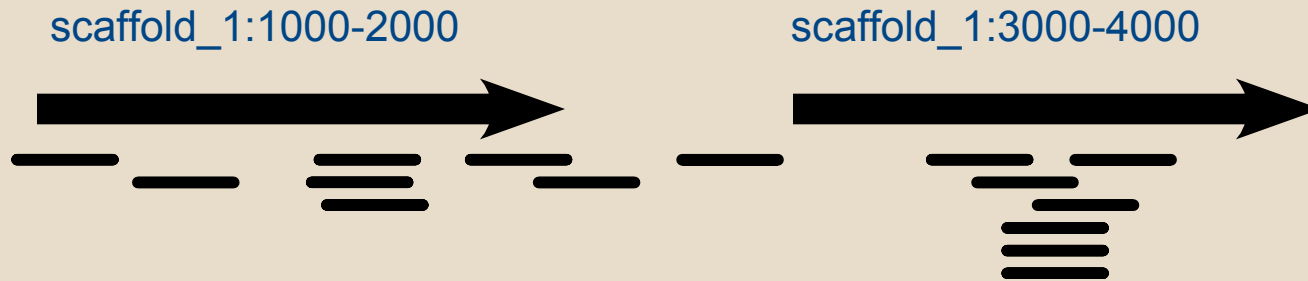
refresh

Heatmaps

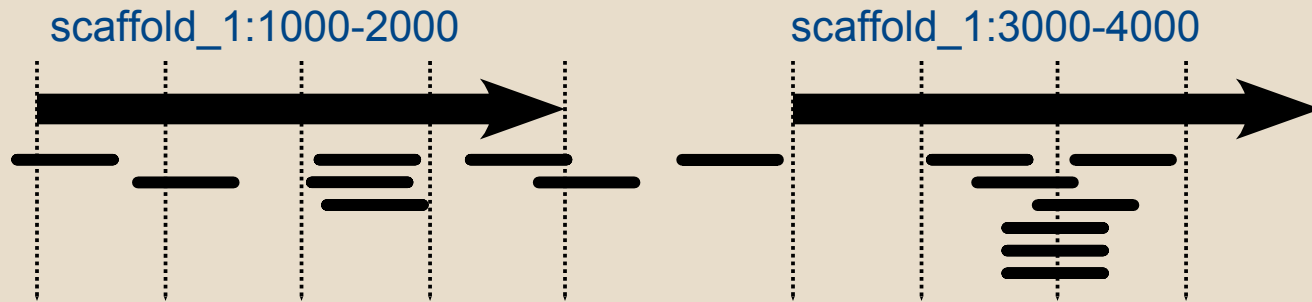
H3K27me3 peaks in blastula and/or gastrula
(5,652 peaks)



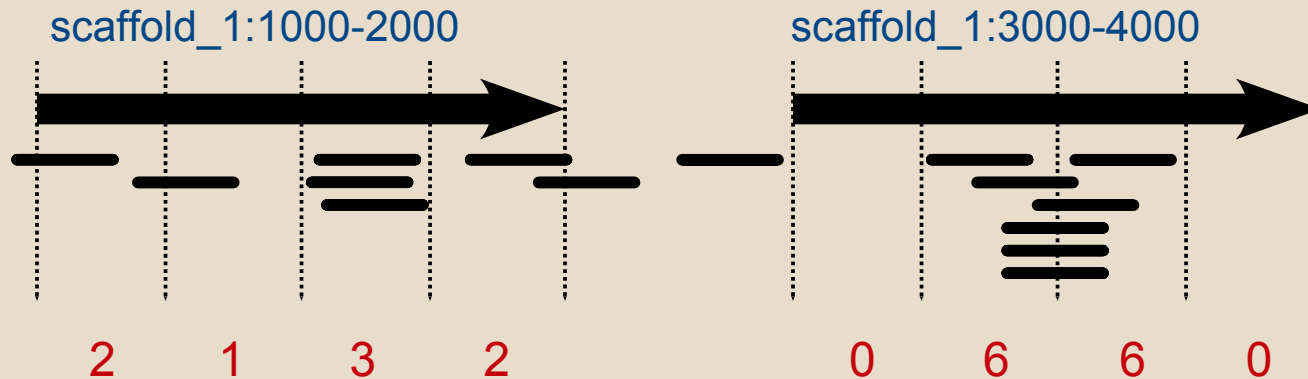
Heatmaps



Heatmaps



Heatmaps



Heatmap data:

scaffold_1	1000	2000	2	1	3	2
scaffold_1	3000	4000	0	6	6	0

Heatmaps (from easy to “difficult”)

- seqMINER
<http://bips.u-strasbg.fr/seqminer/tiki-index.php>
 - Graphical Interface
 - Easy to use
- fluff <http://github.com/simonvh/fluff>
 - Command-line
 - Colorful plots
- R
 - Powerful language for statistical computing
 - Many plotting functions
 - Heatmaps using gplots library

seqMINER

- Java program
 - Runs on Windows, Mac OS X, Linux
 - Easy to use
 - Easy to play around with settings
 - Runs on modest hardware
 - However: more tracks = more memory needed
 - Export is somewhat iffy
 - it gets confused by scaffolds

Simple example

- Goal: profile of H3K4me3 and H3K27me3 around TSS of genes
- Download BED data from GEO (GSE14025)
- Create the TSS of Xtev genes using UCSC Table Browser

Xtev genes

- Based on Xenbase, JGI, Ensembl, Refseq genes and EST data (Akkers, 2009)
- Available at <http://www.ncmls.nl/gertjanveenstra>
- Upload track as custom track in UCSC Genome Browser
- Then, use the Table Browser

clade: genome: assembly:

group: Custom Tracks **track:** Xtev_v1.0 [manage custom tracks](#)

table:	ct_Xtewl0_8432	describe table schema
--------	----------------	-----------------------

region: ☒ genome ☐ position scaffold_163:520000-6900 lookup define regions

identifiers (names/accessions): paste list | upload list

filter: create

intersection: create

correlation:

output format: BED - browser extensible data Send output to ☐ [Galaxy](#) ☐ [GREAT](#)

output file: (leave blank to keep output in browser)

file type returned: ☒ plain text ☐ gzip compressed

get output | summary/statistics

To reset **all** user cart settings (including custom tracks), [click here](#).

This section provides brief line-by-line descriptions of the Table Browser controls. For more information on using this program, see the [Table Browser User's Guide](#).

- ~~tracks: Selects the annotation track data to work with. This list displays all tracks belonging to the group specified in the group list~~



Output ct_Xtev10_8432 as BED

☐ Include [custom track](#) header:

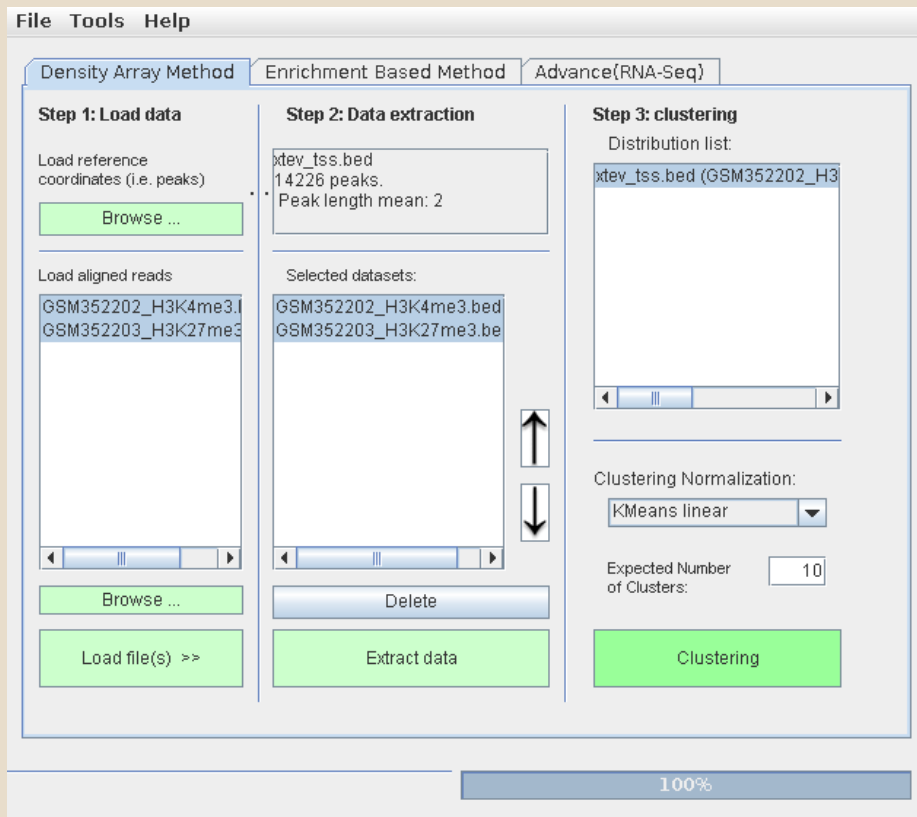
name=
description=
visibility=
url=

Create one BED record per:

- ☐ Whole Gene
- ☒ Upstream by bases
- ☐ Exons plus bases at each end
- ☐ Introns plus bases at each end
- ☐ 5' UTR Exons
- ☐ Coding Exons
- ☐ 3' UTR Exons
- ☐ Downstream by bases

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

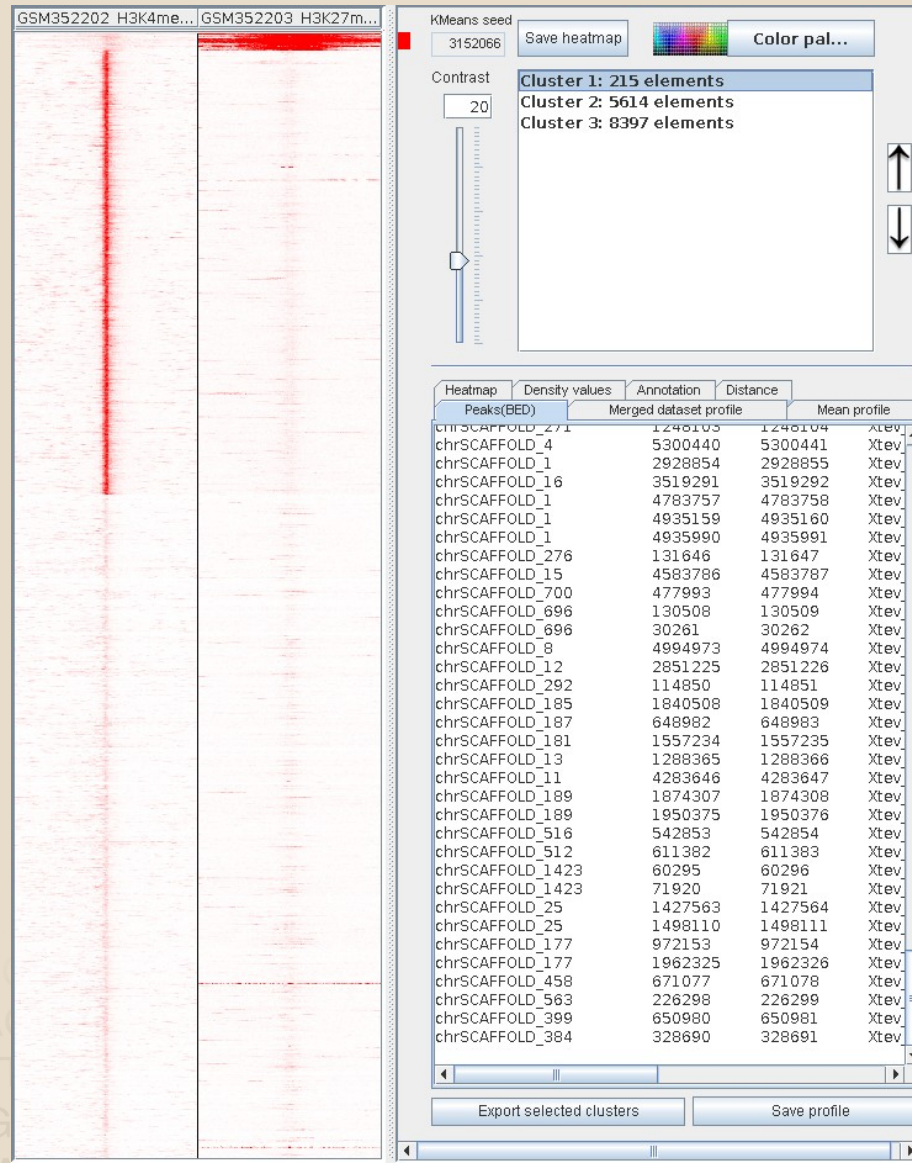
seqMINER



- load reference
- load reads (BED/BAM)
- extract data
- cluster

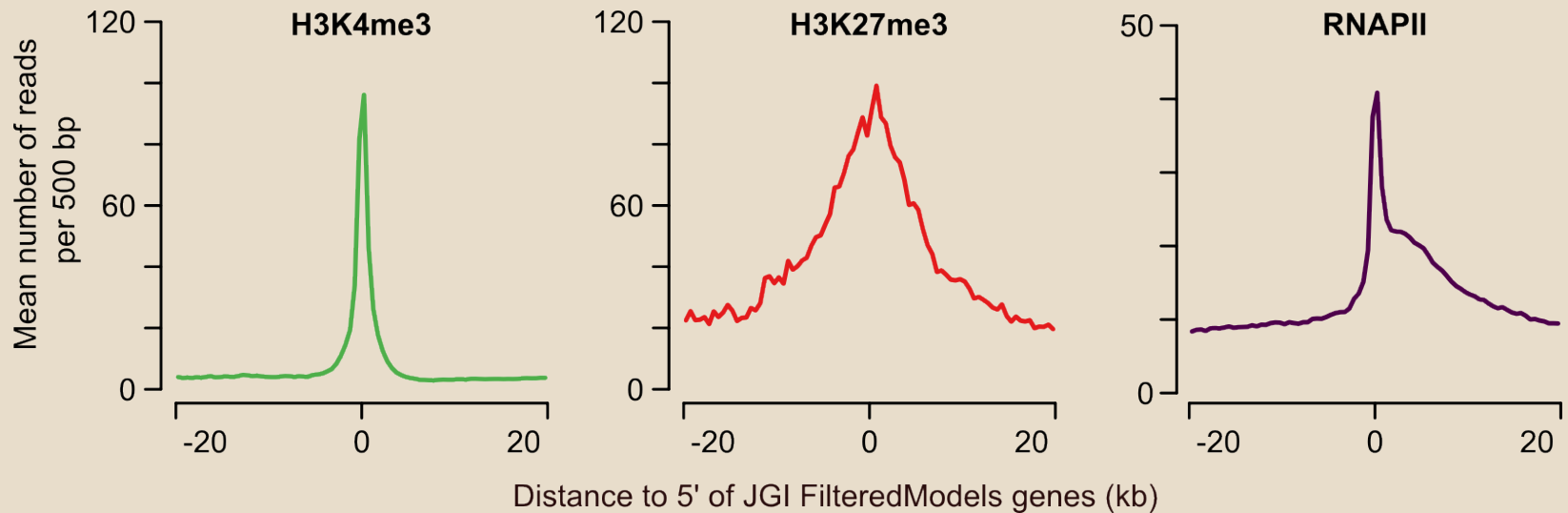
Tip: select window size under Tools

Result



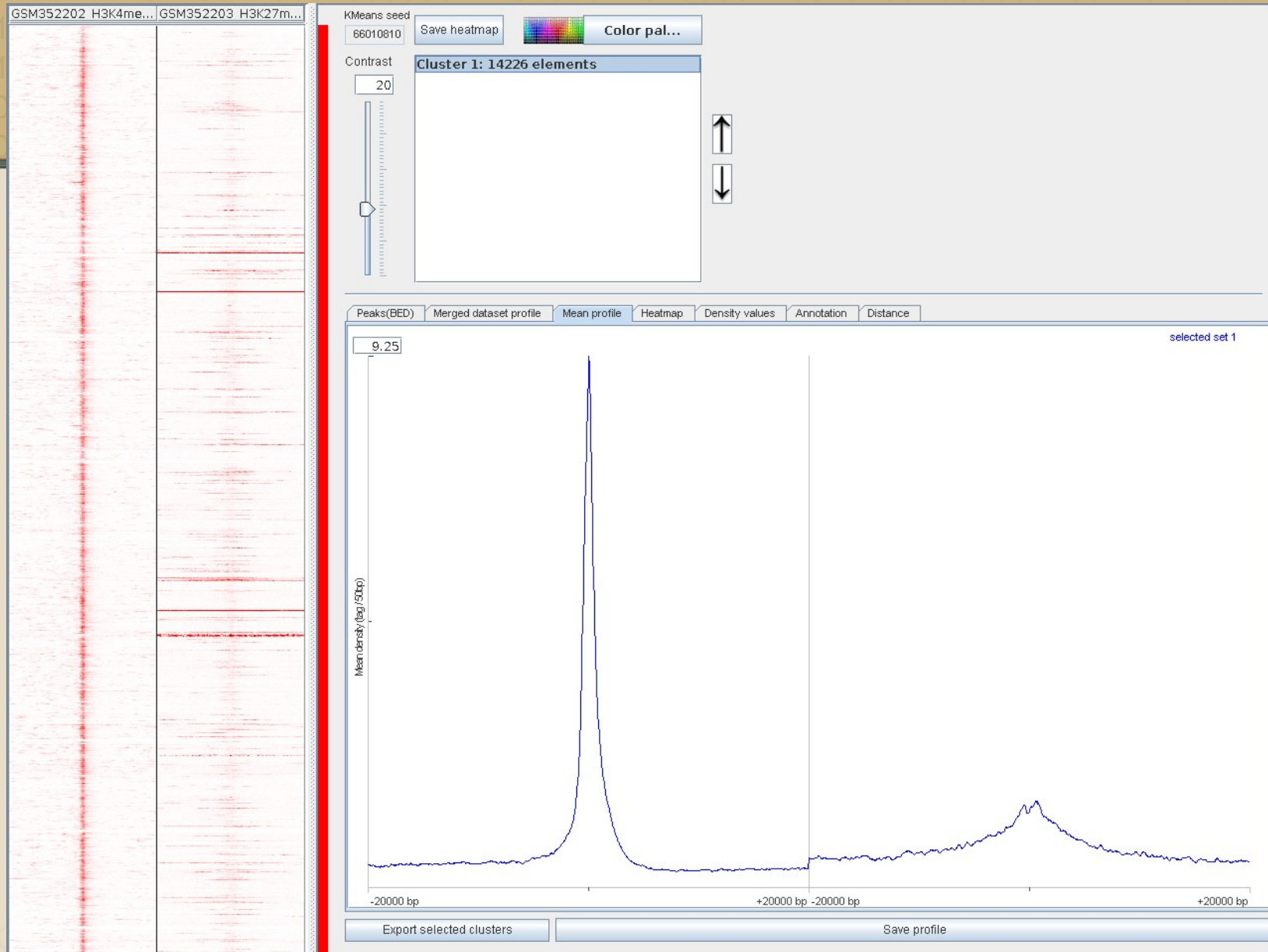
- Play around with number of clusters
- Vary contrast to see how that looks

Profiles



Creating profiles

- Many ways
- seqMINER can do it for each cluster
 - To generate a profile of the whole dataset, just use 1 cluster ;-)





NCMLS
Radboud University, Nijmegen

Robert Akkers
Ozren Bogdanovic
Ulrike Jacobi
Gert Jan Veenstra

MRC

National Institute
for Medical Research

Mike Gilchrist
Ilya Patrushev



Netherlands Organisation for Scientific Research

Radboud University Nijmegen



Thank you
for your
attention





Feel free to share or re-use these slides!
Creative Commons Attribution 3.0
Unported License, except where noted.

Xenopus laevis images (c) Shutterstock