Cistrome Project

An Integrative Platform to Analyze ChIP-chip/seq Data

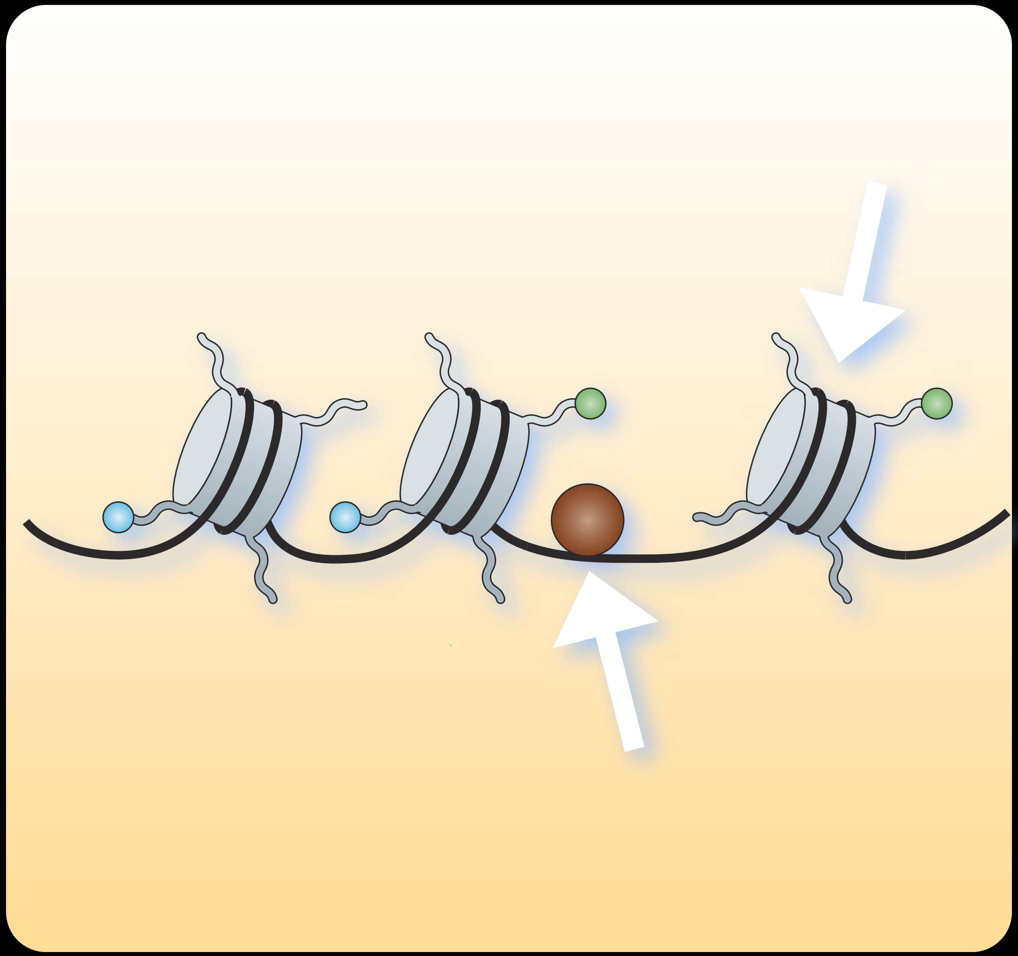
Tao Liu

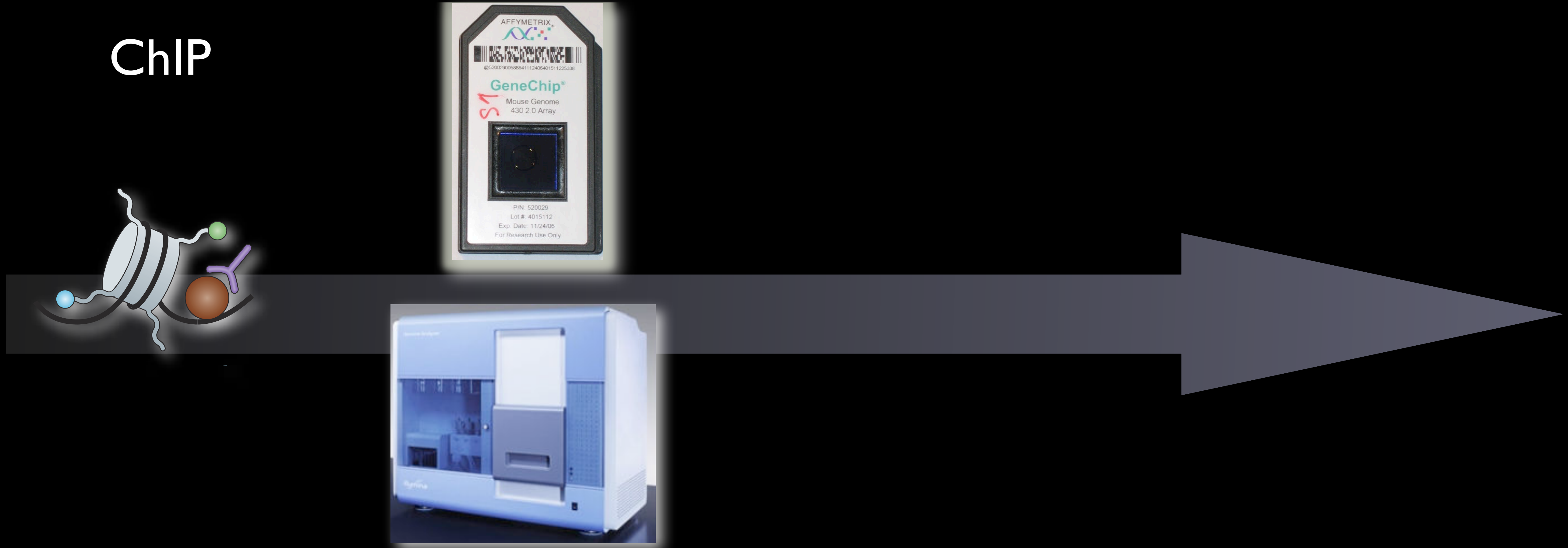
X. Shirley Liu lab

Department of Biostatistics and Computational Biology
Dana-Farber Cancer Institute
Harvard School of Public Health

The Cistrome

- cis-acting targets of a trans-acting factor on a genome scale.
- Cistrome defines the genomic locations where the regulatory protein-DNA interactions occur, such as TF binding, epigenetic modification and nucleosome occupancy.



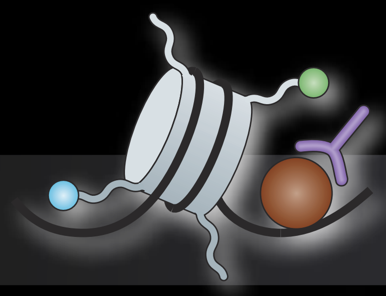


ChIP

Array

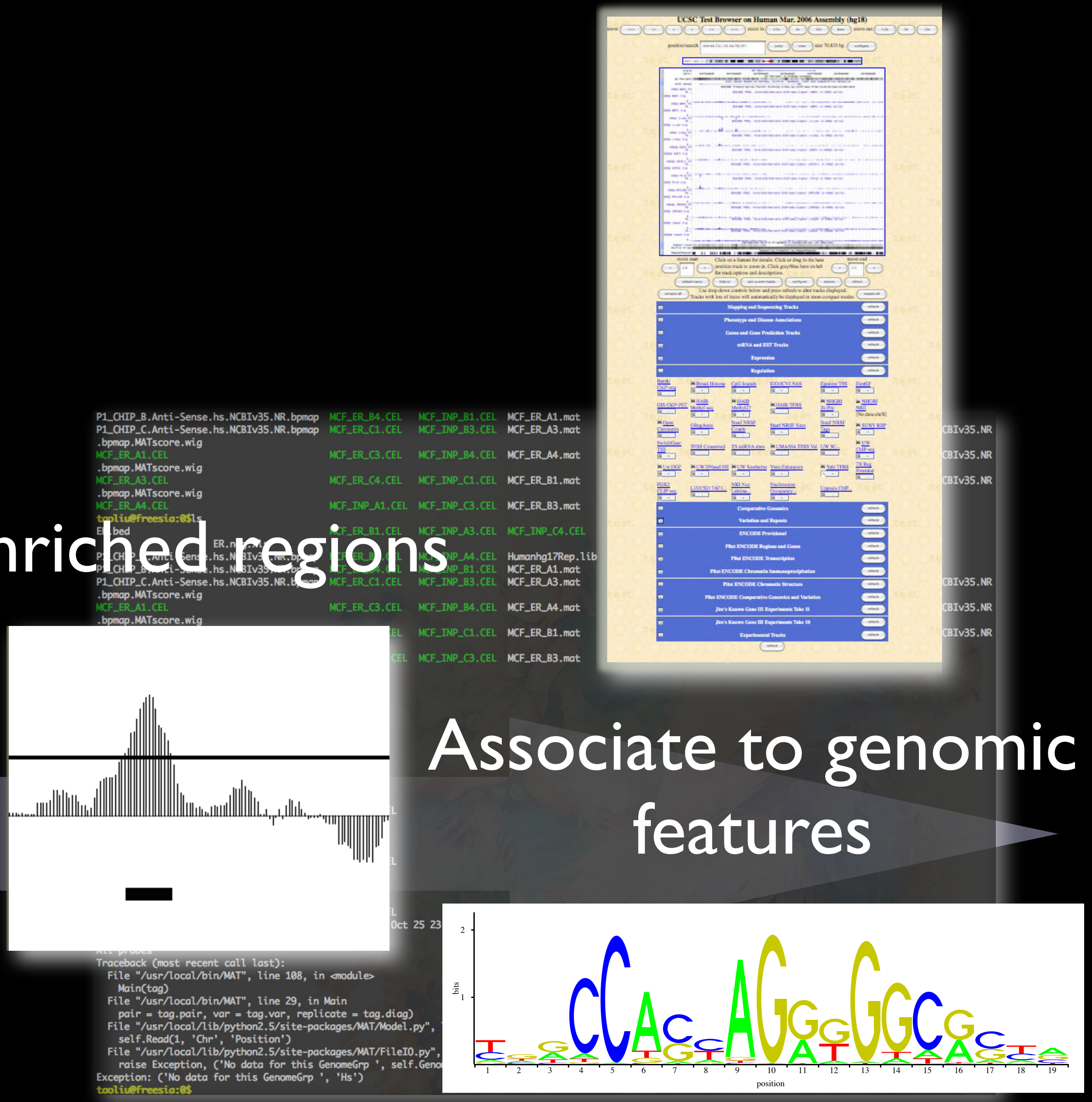
Next
Generation
Sequencing

ChIP



Array

Next Generation Sequencing



```
P1_CHIP_B.Anti-Sense.hs.NCBIv35.NR.bpmap  MCF_ER_B4.CEL   MCF_INP_B1.CEL  MCF_ER_A1.mat
P1_CHIP_C.Anti-Sense.hs.NCBIv35.NR.bpmap  MCF_ER_C1.CEL   MCF_INP_B3.CEL  MCF_ER_A3.mat
.bpmap.MATscore.wig
MCF_ER_A1.CEL                             MCF_ER_C3.CEL   MCF_INP_B4.CEL  MCF_ER_A4.mat
.bpmap.MATscore.wig
MCF_ER_A3.CEL                             MCF_ER_C4.CEL   MCF_INP_C1.CEL  MCF_ER_B1.mat
.bpmap.MATscore.wig
MCF_ER_A4.CEL                             MCF_INP_A1.CEL  MCF_INP_C3.CEL  MCF_ER_B3.mat
taoliu@freesia:0$ls
```

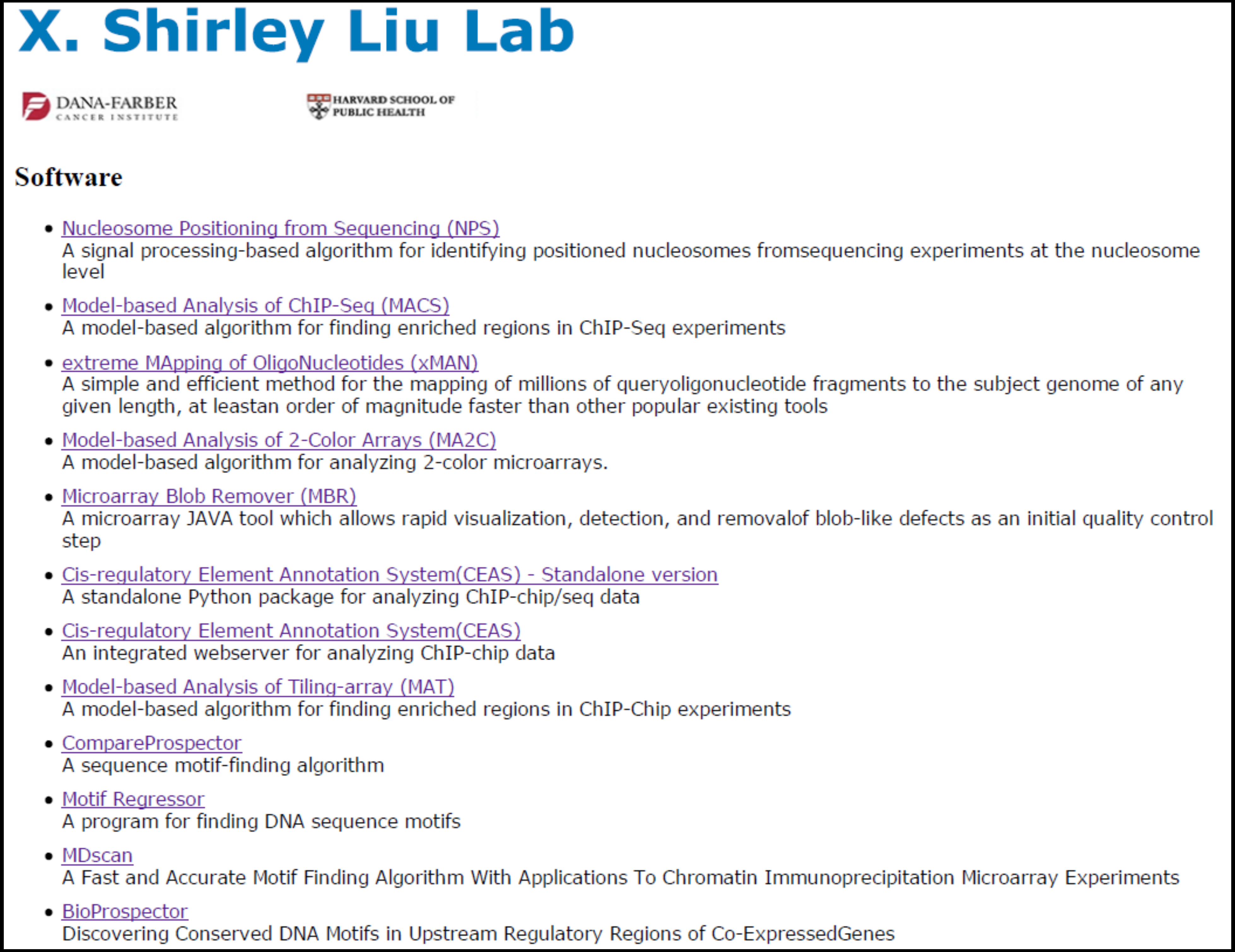
```
Traceback (most recent call last):
  File "/usr/local/bin/MAT", line 108, in <module>
    Main(tag)
  File "/usr/local/bin/MAT", line 29, in Main
    pair = tag.pair, var = tag.var, replicate = tag.diag)
  File "/usr/local/lib/python2.5/site-packages/MAT/Model.py",
    self.Read(1, 'Chr', 'Position')
  File "/usr/local/lib/python2.5/site-packages/MAT/FileIO.py",
    raise Exception, ('No data for this GenomeGrp ', self.Geno
Exception: ('No data for this GenomeGrp ', 'Hs')
taoliu@freesia:0$
```

Enriched regions

Associate to genomic features

Motif discovery

And more...



X. Shirley Liu Lab

DANA-FARBER CANCER INSTITUTE

HARVARD SCHOOL OF PUBLIC HEALTH

Software

- Nucleosome Positioning from Sequencing (NPS)
  A signal processing-based algorithm for identifying positioned nucleosomes fromsequencing experiments at the nucleosome level
- Model-based Analysis of ChIP-Seq (MACS)
  A model-based algorithm for finding enriched regions in ChIP-Seq experiments
- extreme MApping of OligoNucleotides (xMAN)
  A simple and efficient method for the mapping of millions of queryoligonucleotide fragments to the subject genome of any given length, at leastan order of magnitude faster than other popular existing tools
- Model-based Analysis of 2-Color Arrays (MA2C)
  A model-based algorithm for analyzing 2-color microarrays.
- Microarray Blob Remover (MBR)
  A microarray JAVA tool which allows rapid visualization, detection, and removalof blob-like defects as an initial quality control step
- Cis-regulatory Element Annotation System(CEAS) - Standalone version
  A standalone Python package for analyzing ChIP-chip/seq data
- Cis-regulatory Element Annotation System(CEAS)
  An integrated webserver for analyzing ChIP-chip data
- Model-based Analysis of Tiling-array (MAT)
  A model-based algorithm for finding enriched regions in ChIP-Chip experiments
- CompareProspector
  A sequence motif-finding algorithm
- Motif Regressor
  A program for finding DNA sequence motifs
- MDscan
  A Fast and Accurate Motif Finding Algorithm With Applications To Chromatin Immunoprecipitation Microarray Experiments
- BioProspector
  Discovering Conserved DNA Motifs in Upstream Regulatory Regions of Co-ExpressedGenes

A series of our in-house tools

Challenges

- User experience
- Sufficient computation power and storage
- Data availability

Challenges

- User experience
  - Web applications to ease the downstream data analysis for both experimental and computational biologists -- Galaxy
- Sufficient computation power and storage
  - Deploy it in the Cloud
- Data availability
  - Collect public available genome scale ChIP-chip/seq data.

cistrome.org

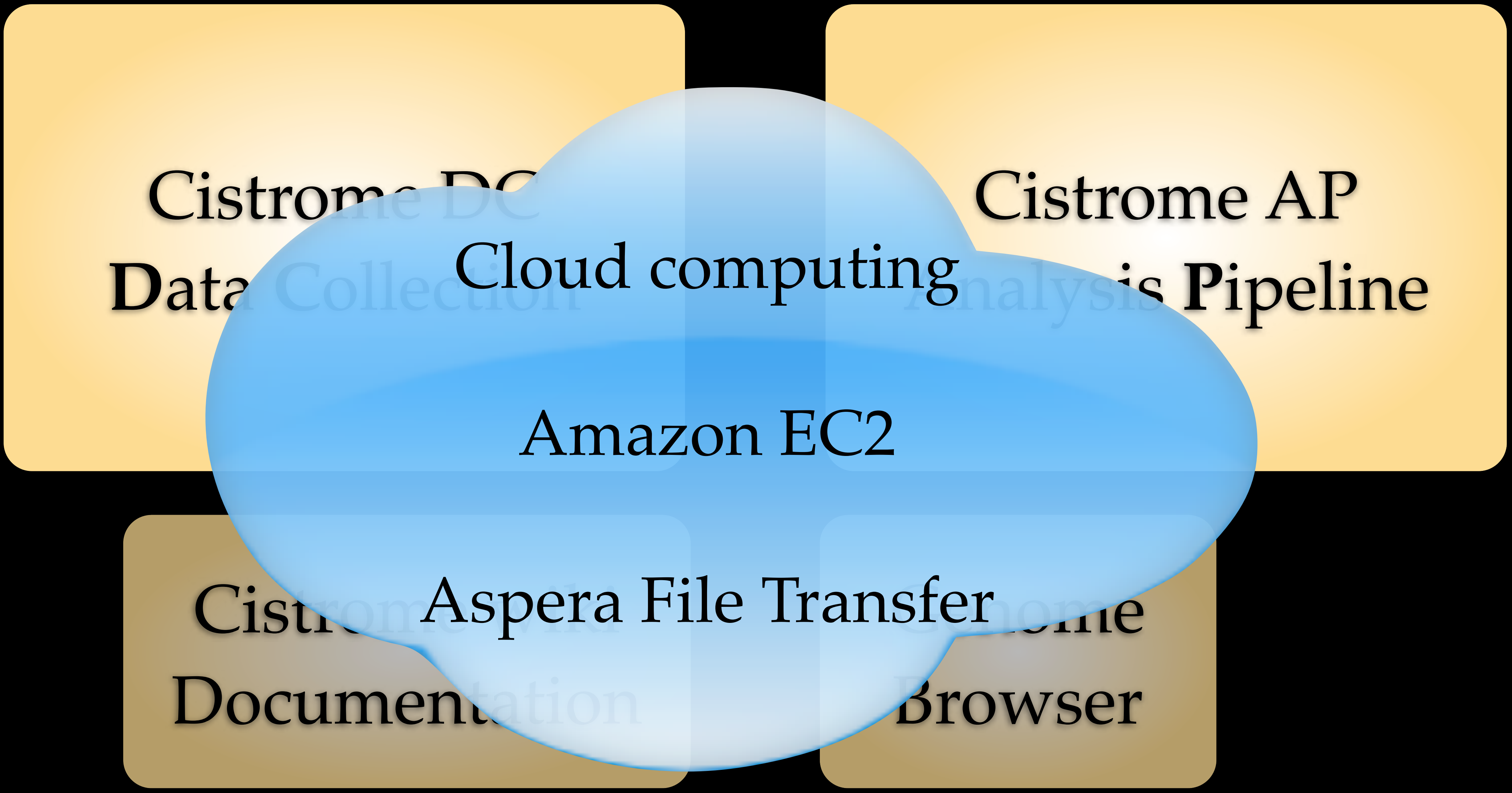
Cistrome DC
Data **C**ollection

Cistrome AP
Analysis **P**ipeline

Cistrome wiki
Documentation

Genome
Browser

cistrome.org



Cistrome DC
Data Collection

Cistrome AP
Analysis **Pipeline**

Cistrome Wiki
Documentation

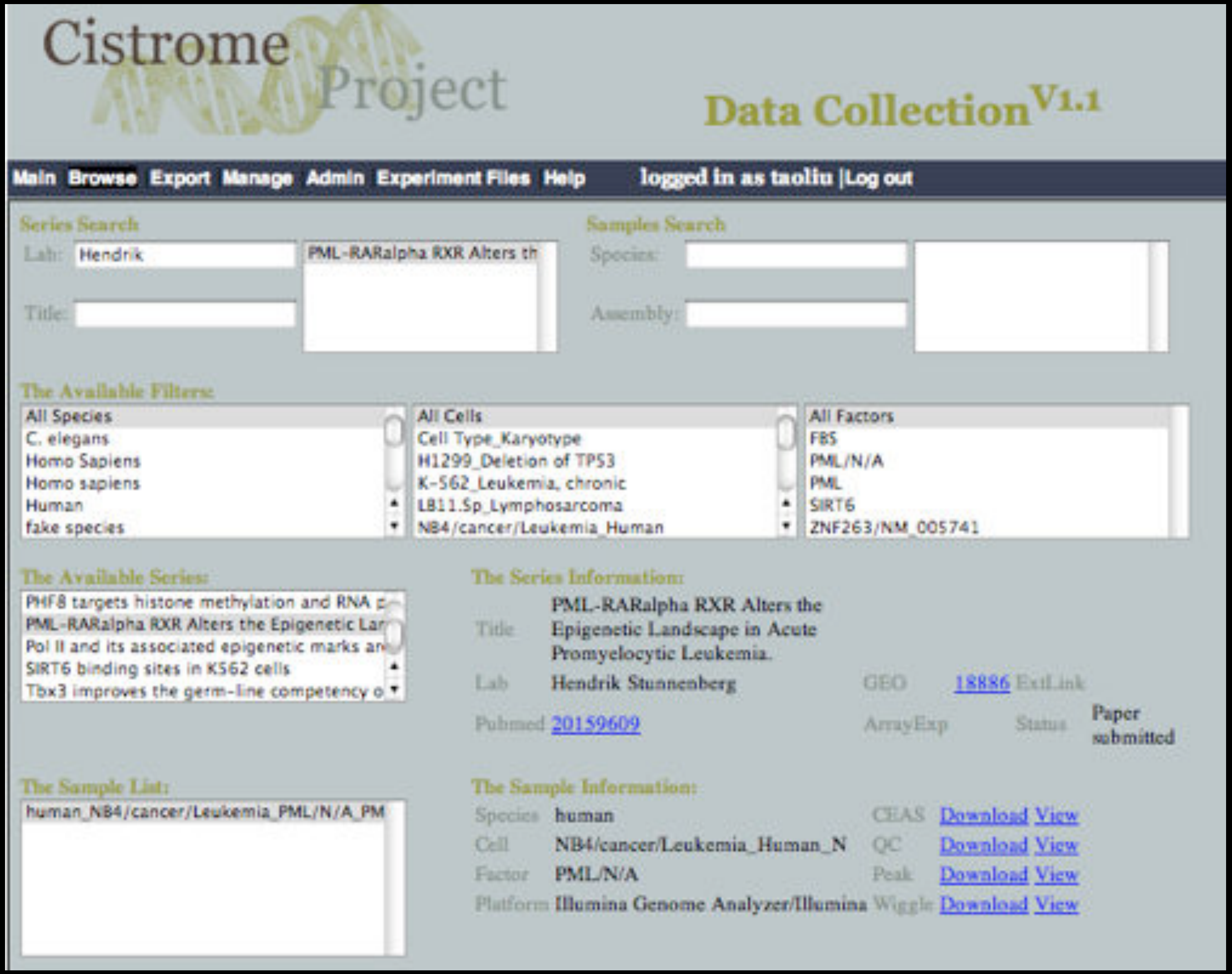
Genome
Browser

Cloud computing

Amazon EC2

Aspera File Transfer

Cistrome DC



Series Search

Lab: Hendrik

Title:

PML-RARalpha RXR Alters th

Samples Search

Species:

Assembly:

The Available Filters:

All Species
C. elegans
Homo Sapiens
Homo sapiens
Human
fake species

All Cells
Cell Type_Karyotype
H1299_Deletion of TP53
K-562_Leukemia, chronic
LB11.Sp_Lymphosarcoma
NB4/cancer/Leukemia_Human

All Factors
FBS
PML/N/A
PML
SIRT6
ZNF263/NM_005741

The Available Series:

PHF8 targets histone methylation and RNA p
PML-RARalpha RXR Alters the Epigenetic Lar
Pol II and its associated epigenetic marks ar
SIRT6 binding sites in K562 cells
Tbx3 improves the germ-line competency o

The Series Information:

Title: PML-RARalpha RXR Alters the Epigenetic Landscape in Acute Promyelocytic Leukemia.
Lab: Hendrik Stunnenberg
GEO: 18886 ExtLink
Pubmed: 20159609
ArrayExp:
Status: Paper submitted

The Sample List:

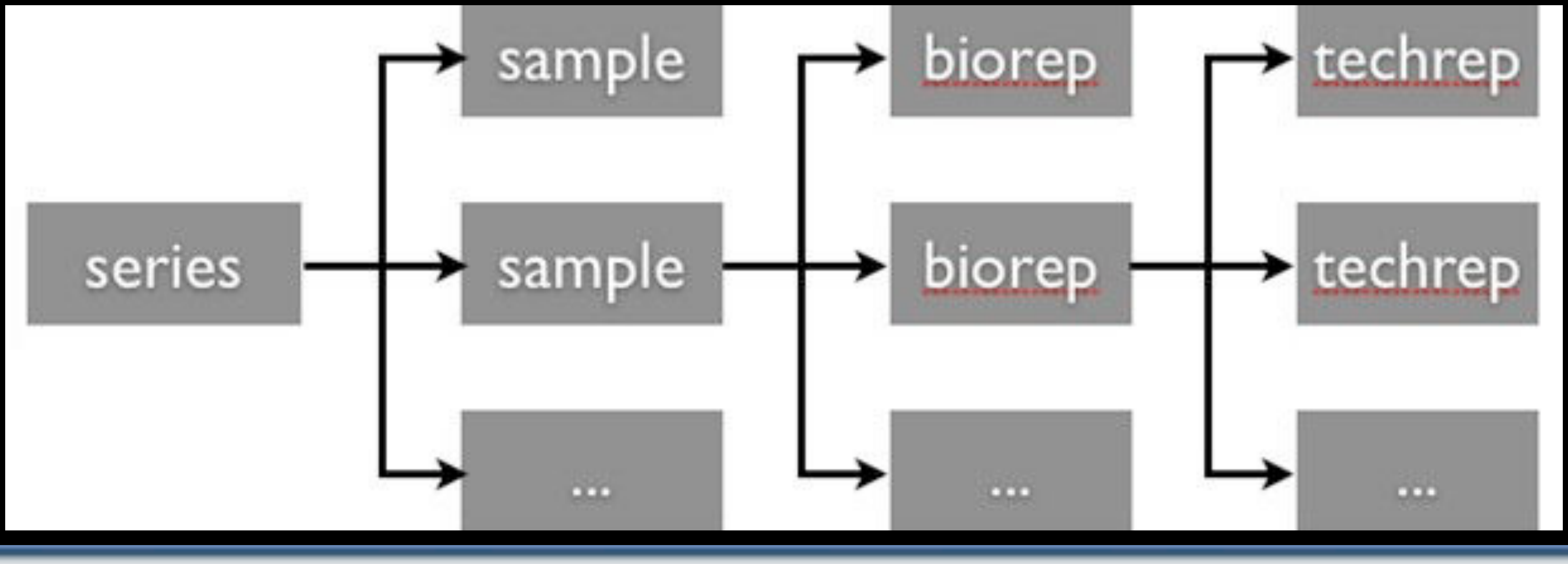
human_NB4/cancer/Leukemia_PML/N/A_PM

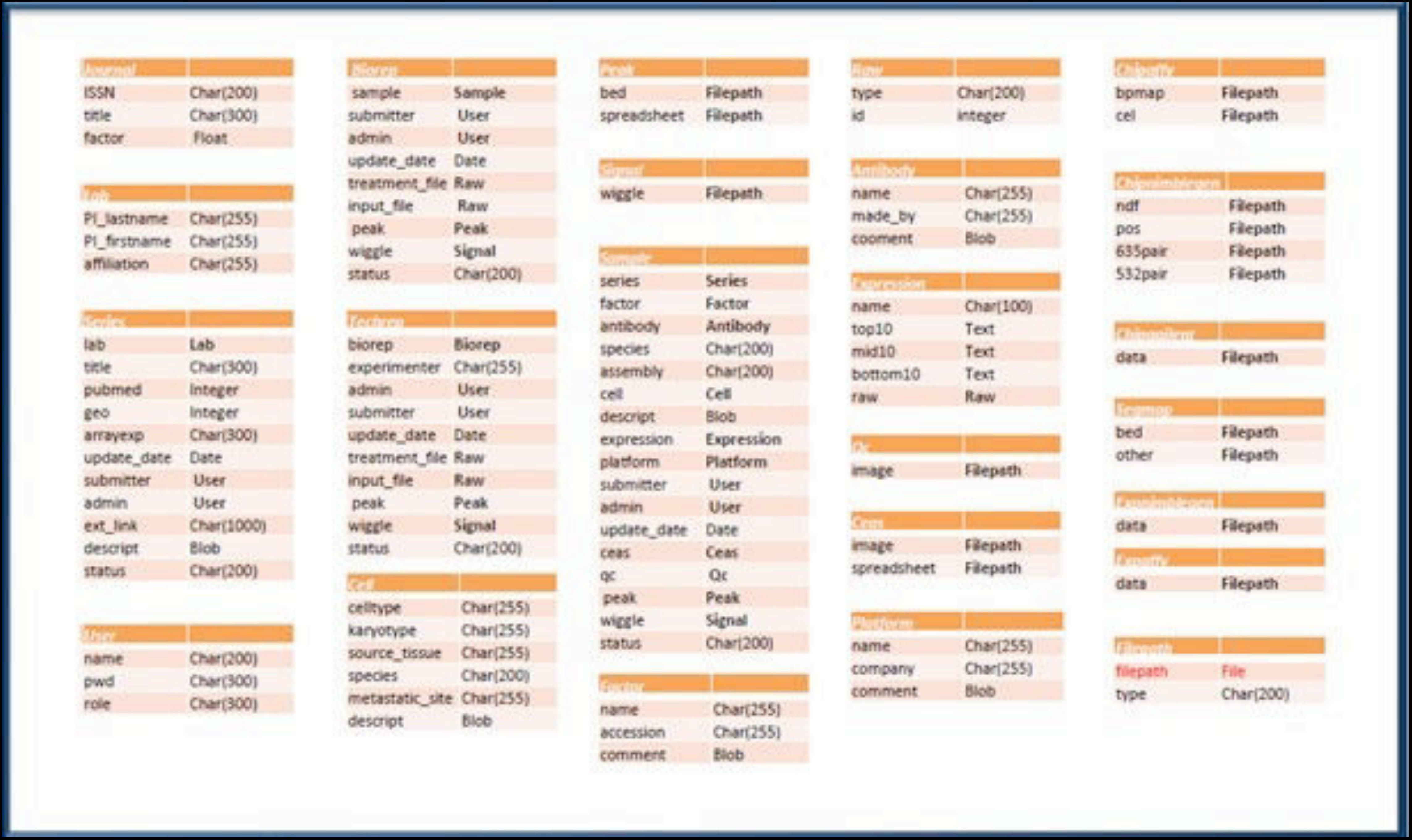
The Sample Information:

Species	human	CEAS	Download View
Cell	NB4/cancer/Leukemia_Human_N	QC	Download View
Factor	PML/N/A	Peak	Download View
Platform	Illumina Genome Analyzer/Illumina	Wiggle	Download View

- Published Data collected from GEO/ArrayExpress/author's website.
- Functions for browsing and searching.
- Collaborating project unpublished data, protected by password.

Javascript
Ajax
XML
HTML5





Journal	
ISSN	Char(200)
title	Char(300)
factor	Float

Lab	
PI_lastname	Char(255)
PI_firstname	Char(255)
affiliation	Char(255)

Series	
lab	Lab
title	Char(300)
pubmed	Integer
geo	Integer
arrayexp	Char(300)
update_date	Date
submitter	User
admin	User
ext_link	Char(1000)
descript	Blob
status	Char(200)

User	
name	Char(200)
pwd	Char(300)
role	Char(300)

Biorep	
sample	Sample
submitter	User
admin	User
update_date	Date
treatment_file	Raw
input_file	Raw
peak	Peak
wiggle	Signal
status	Char(200)

Techrep	
biorep	Biorep
experimenter	Char(255)
admin	User
submitter	User
update_date	Date
treatment_file	Raw
input_file	Raw
peak	Peak
wiggle	Signal
status	Char(200)

Cell	
celltype	Char(255)
karyotype	Char(255)
source_tissue	Char(255)
species	Char(200)
metastatic_site	Char(255)
descript	Blob

Peak	
bed	Filepath
spreadsheet	Filepath

Signal	
wiggle	Filepath

Sample	
series	Series
factor	Factor
antibody	Antibody
species	Char(200)
assembly	Char(200)
cell	Cell
descript	Blob
expression	Expression
platform	Platform
submitter	User
admin	User
update_date	Date
ceas	Ceas
qc	Qc
peak	Peak
wiggle	Signal
status	Char(200)

Factor	
name	Char(255)
accession	Char(255)
comment	Blob

Raw	
type	Char(200)
id	integer

Antibody	
name	Char(255)
made_by	Char(255)
cooment	Blob

Expression	
name	Char(100)
top10	Text
mid10	Text
bottom10	Text
raw	Raw

Qc	
image	Filepath

Ceas	
image	Filepath
spreadsheet	Filepath

Platform	
name	Char(255)
company	Char(255)
comment	Blob

Chipaffy	
bpmap	Filepath
cel	Filepath

Chipnimblegen	
ndf	Filepath
pos	Filepath
635pair	Filepath
532pair	Filepath

Chipagilent	
data	Filepath

Seqmap	
bed	Filepath
other	Filepath

Expnimblegen	
data	Filepath

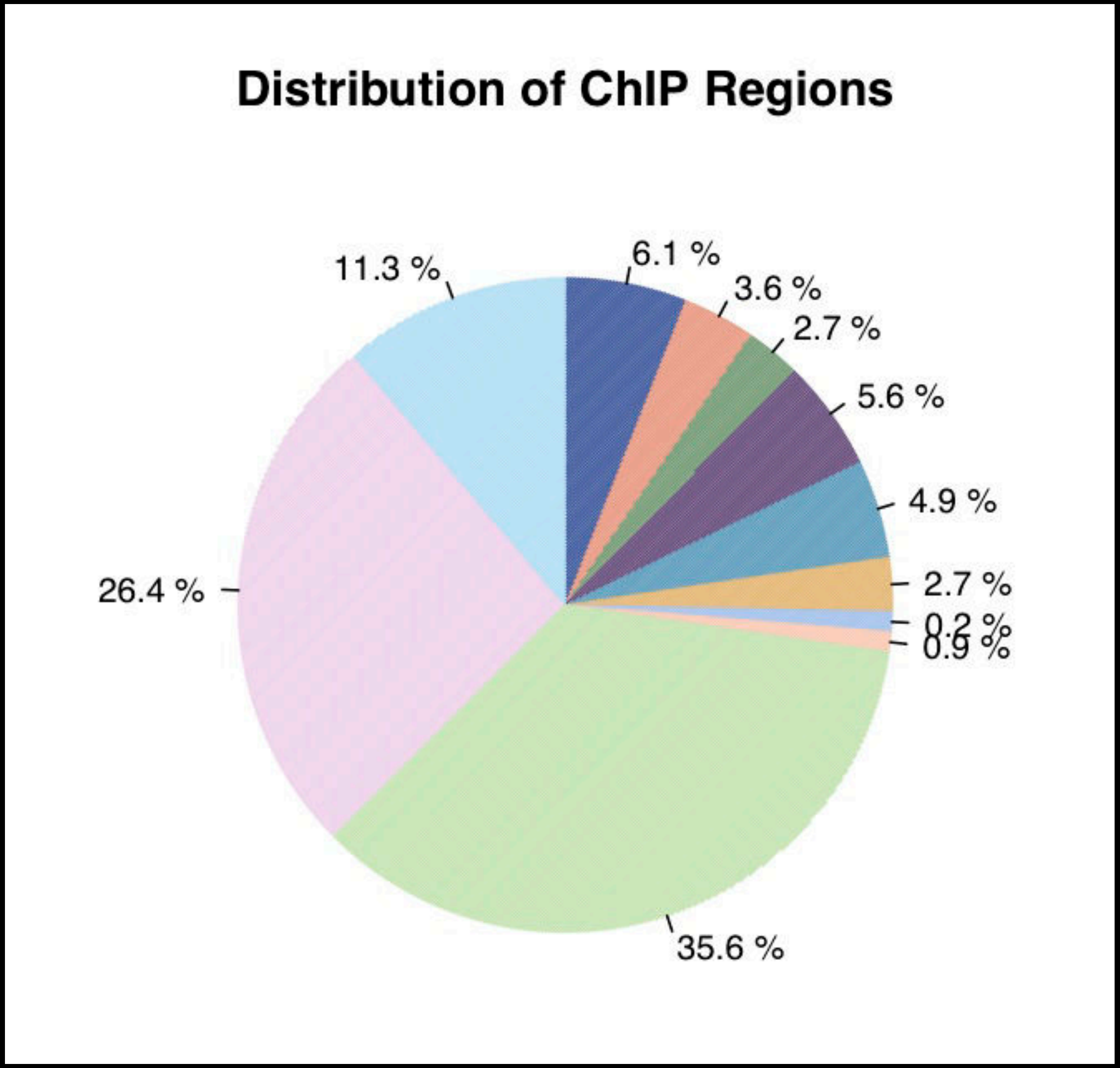
Expaffy	
data	Filepath

Filepath	
filepath	File
type	Char(200)

Cistrome AP using Galaxy

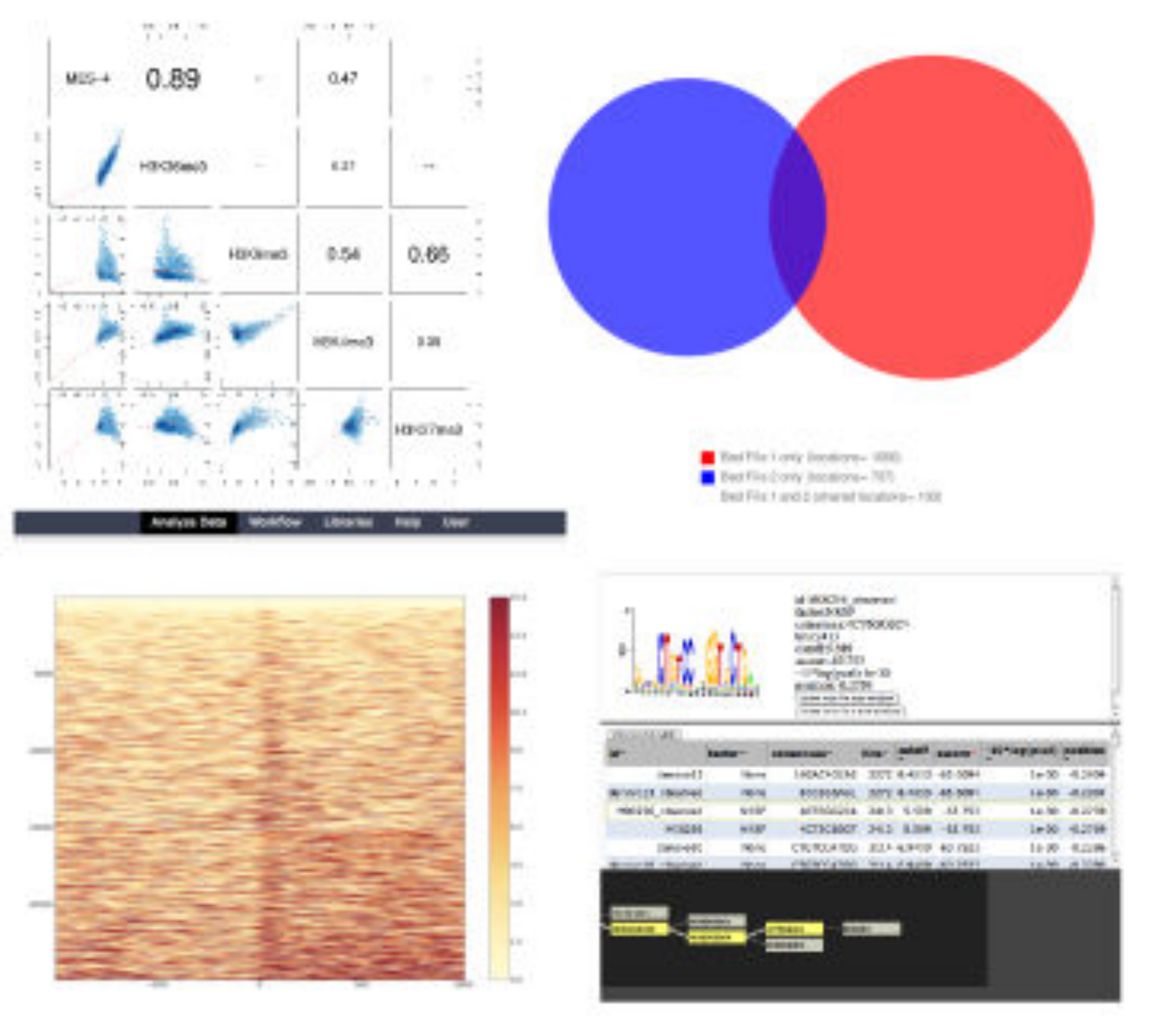
- Peak calling -- MAT, MA2C and MACS
- Annotation and Visualization -- CEAS (sitepro, ceas, and gca) + Conservation plot + Heatmap
- Motif analysis – motif discovery and motif scan
- Correlation -- e.g. many wiggle files of score profile
- Expression and GO
- Liftover and wiggle file standardization
- Import from DC and upload raw data
- Export published dataset to public DC
- Other Galaxy Low-Level Operations

Cistrome AP using Galaxy



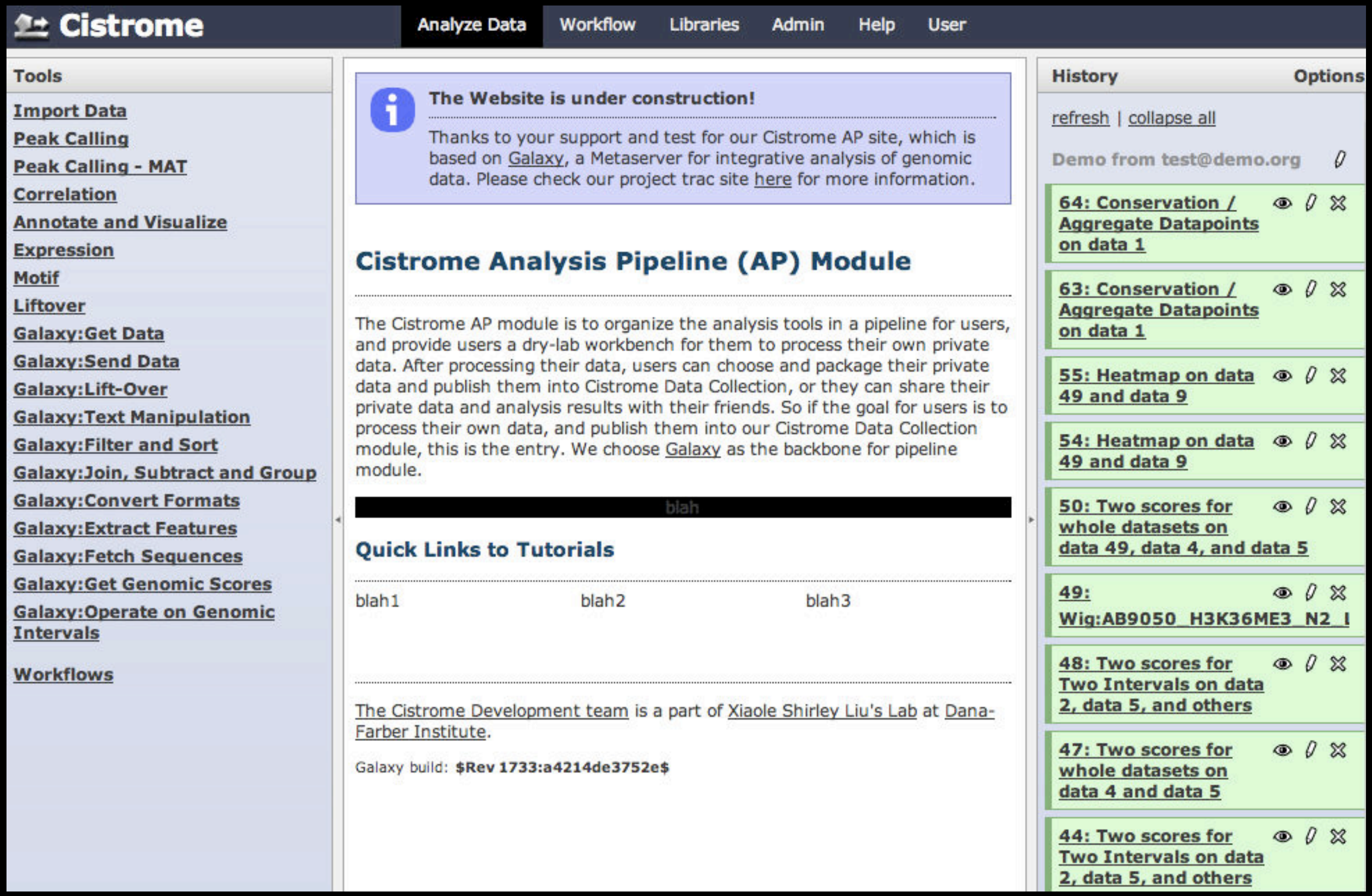
- Peak calling -- MAT, MA2C and MACS
- Annotation and Visualization -- CEAS (sitepro, ceas, and gca) + Conservation plot + Heatmap
- Motif analysis – motif discovery and motif scan
- Correlation -- e.g. many wiggle files of score profile
- Expression and GO
- Liftover and wiggle file standardization
- Import from DC and upload raw data
- Export published dataset to public DC
- Other Galaxy Low-Level Operations

Cistrome AP using Galaxy



- Peak calling -- MAT, MA2C and MACS
- Annotation and Visualization -- CEAS (sitepro, ceas, and gca) + Conservation plot + Heatmap
- Motif analysis – motif discovery and motif scan
- Correlation -- e.g. many wiggle files of score profile
- Expression and GO
- Liftover and wiggle file standardization
- Import from DC and upload raw data
- Export published dataset to public DC
- Other Galaxy Low-Level Operations

Cistrome AP using Galaxy



The Website is under construction!

Thanks to your support and test for our Cistrome AP site, which is based on Galaxy, a Metaserver for integrative analysis of genomic data. Please check our project trac site here for more information.

Cistrome Analysis Pipeline (AP) Module

The Cistrome AP module is to organize the analysis tools in a pipeline for users, and provide users a dry-lab workbench for them to process their own private data. After processing their data, users can choose and package their private data and publish them into Cistrome Data Collection, or they can share their private data and analysis results with their friends. So if the goal for users is to process their own data, and publish them into our Cistrome Data Collection module, this is the entry. We choose Galaxy as the backbone for pipeline module.

blah

Quick Links to Tutorials

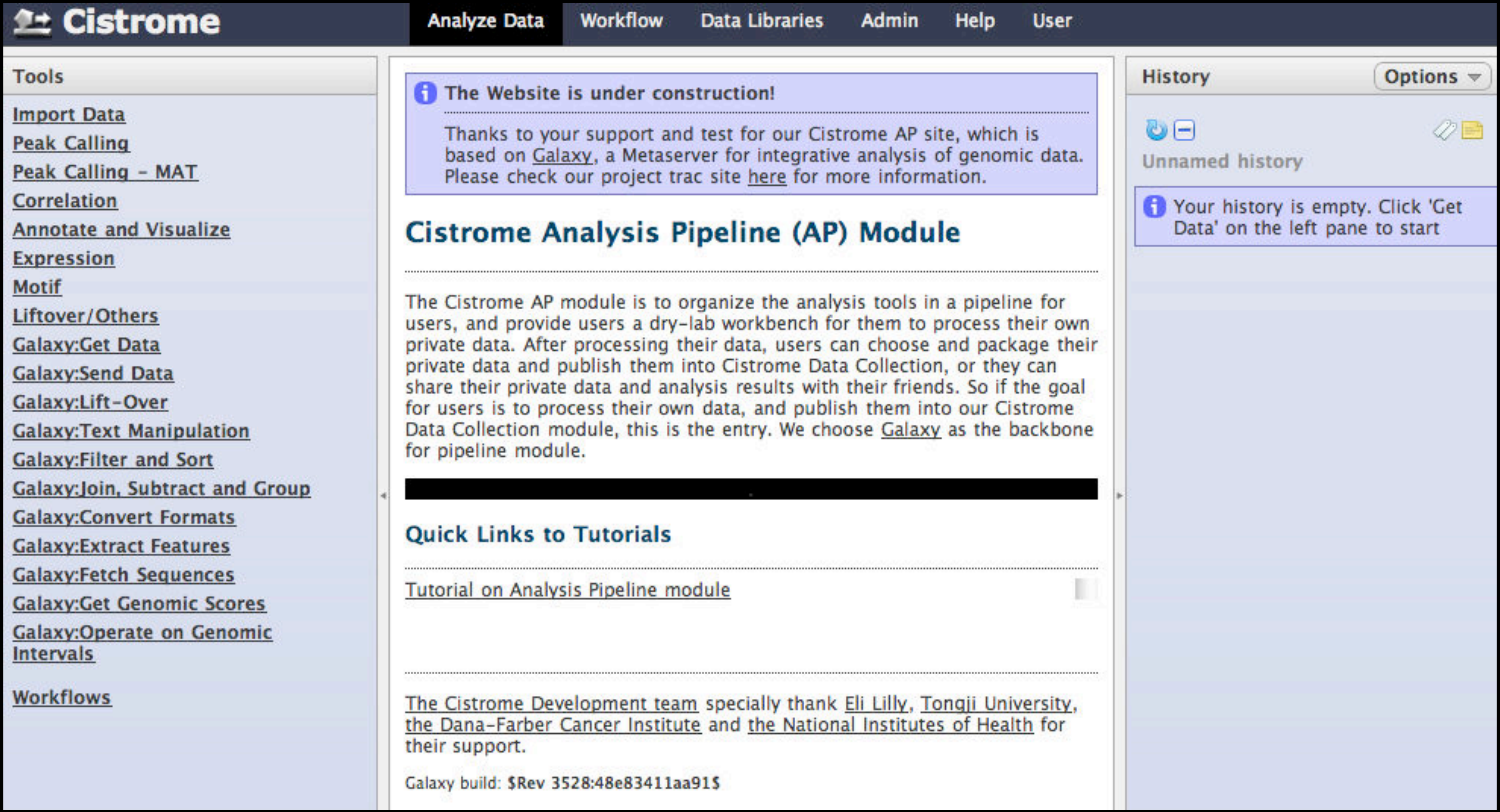
blah1 blah2 blah3

The Cistrome Development team is a part of Xiaole Shirley Liu's Lab at Dana-Farber Institute.

Galaxy build: **$Rev 1733:a4214de3752e$**

- Peak calling -- MAT, MA2C and MACS
- Annotation and Visualization -- CEAS (sitepro, ceas, and gca) + Conservation plot + Heatmap
- Motif analysis – motif discovery and motif scan
- Correlation -- e.g. many wiggle files of score profile
- Expression and GO
- Liftover and wiggle file standardization
- Import from DC and upload raw data
- Export published dataset to public DC
- Other Galaxy Low-Level Operations

Cistrome AP



The Website is under construction!

Thanks to your support and test for our Cistrome AP site, which is based on Galaxy, a Metaserver for integrative analysis of genomic data. Please check our project trac site here for more information.

Cistrome Analysis Pipeline (AP) Module

The Cistrome AP module is to organize the analysis tools in a pipeline for users, and provide users a dry-lab workbench for them to process their own private data. After processing their data, users can choose and package their private data and publish them into Cistrome Data Collection, or they can share their private data and analysis results with their friends. So if the goal for users is to process their own data, and publish them into our Cistrome Data Collection module, this is the entry. We choose Galaxy as the backbone for pipeline module.

Quick Links to Tutorials

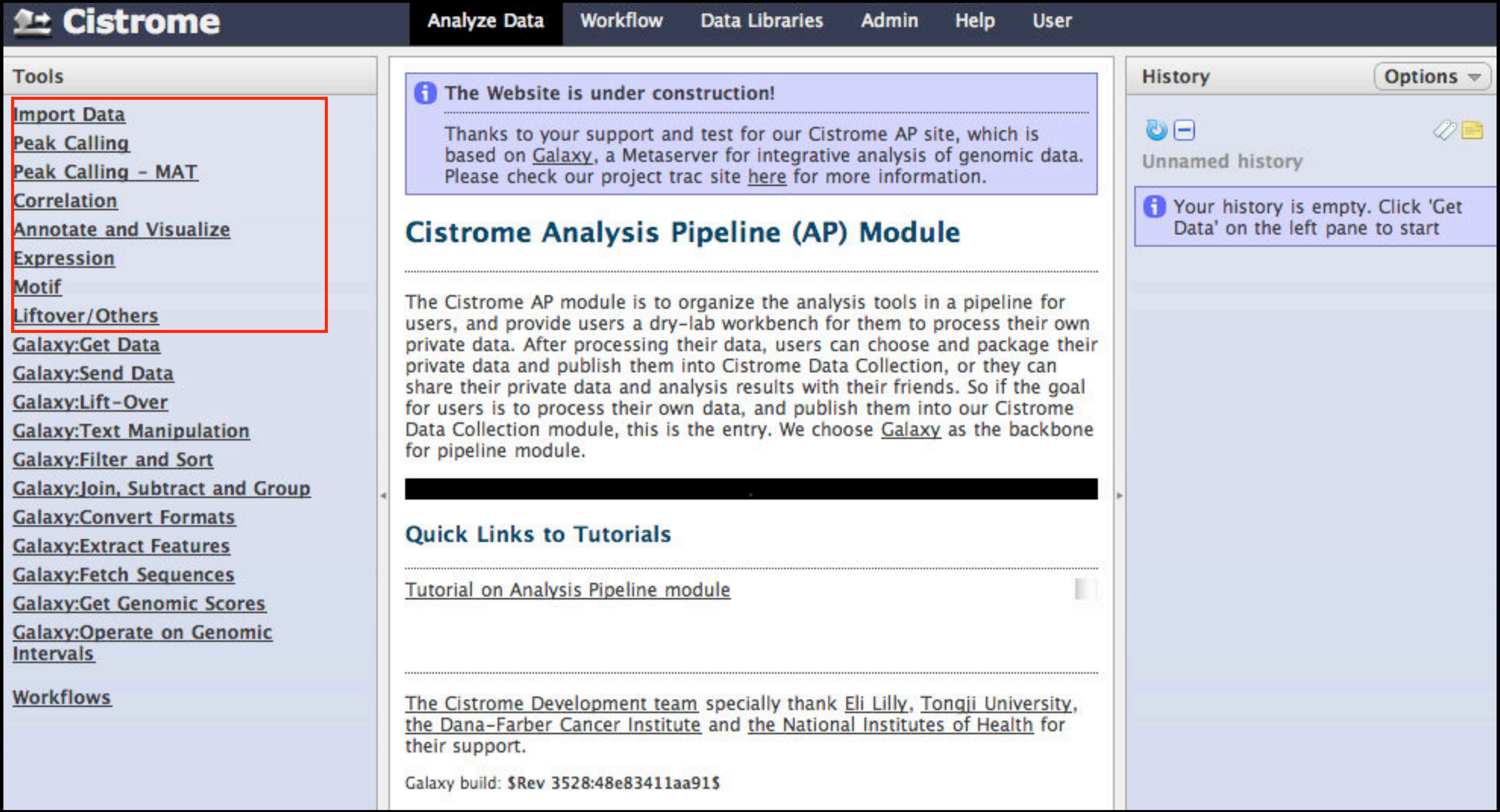
Tutorial on Analysis Pipeline module

The Cistrome Development team specially thank Eli Lilly, Tongji University, the Dana-Farber Cancer Institute and the National Institutes of Health for their support.

Galaxy build: $Rev 3528:48e83411aa91$

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

Cistrome AP



The Website is under construction!

Thanks to your support and test for our Cistrome AP site, which is based on Galaxy, a Metaserver for integrative analysis of genomic data. Please check our project trac site here for more information.

Cistrome Analysis Pipeline (AP) Module

The Cistrome AP module is to organize the analysis tools in a pipeline for users, and provide users a dry-lab workbench for them to process their own private data. After processing their data, users can choose and package their private data and publish them into Cistrome Data Collection, or they can share their private data and analysis results with their friends. So if the goal for users is to process their own data, and publish them into our Cistrome Data Collection module, this is the entry. We choose Galaxy as the backbone for pipeline module.

Quick Links to Tutorials

Tutorial on Analysis Pipeline module

The Cistrome Development team specially thank Eli Lilly, Tongji University, the Dana-Farber Cancer Institute and the National Institutes of Health for their support.

Galaxy build: $Rev 3528:48e83411aa91$

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

Cistrome documentation
&
local genome browser

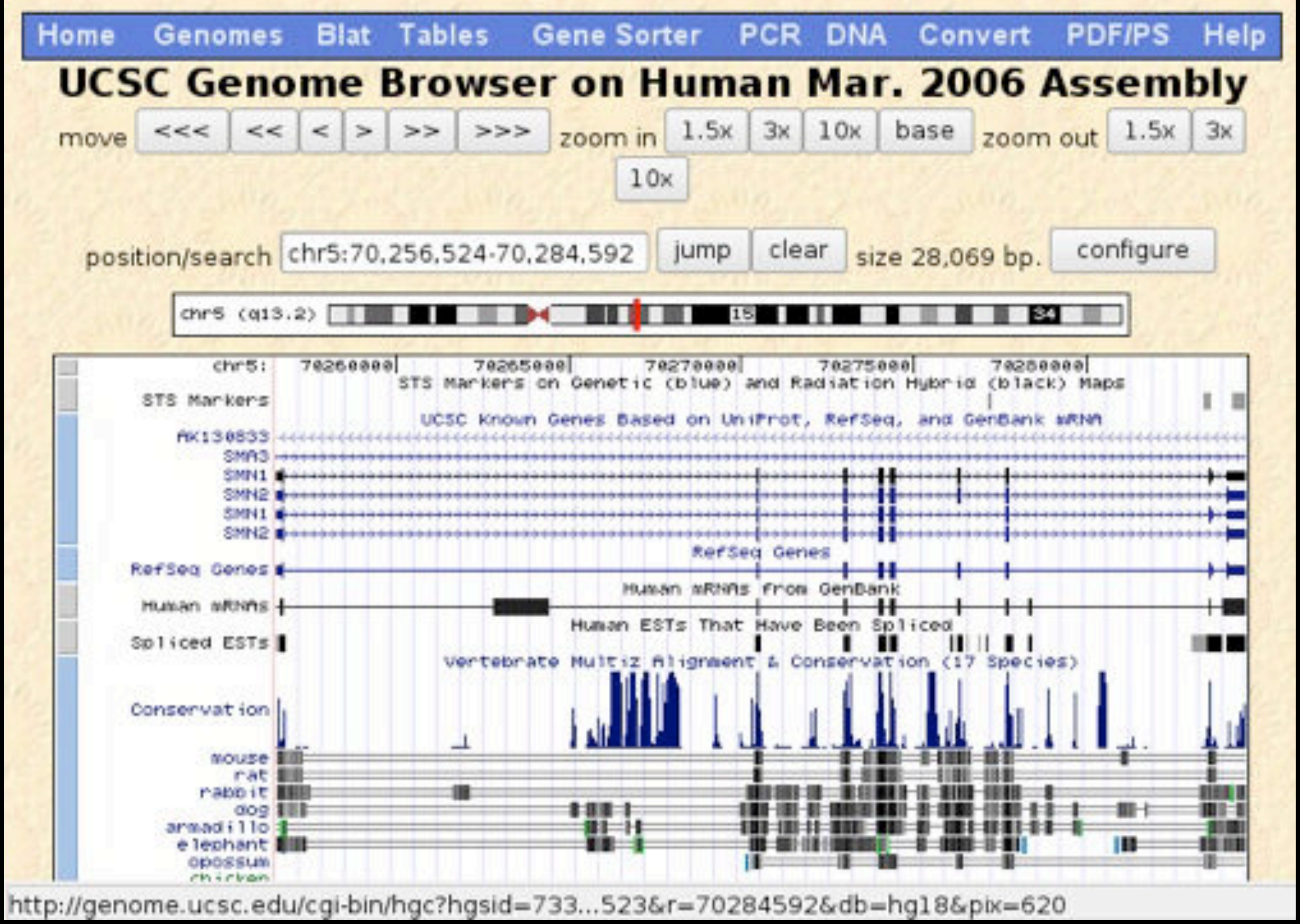
Tutorial For Analysis Pipeline

1. Summary

The Cistrome Analysis Pipeline module provides users a flexible dry-lab workbench for general analysis tools for ChIP-chip and ChIP-seq experiments. The whole module is built on Galaxy, a powerful tools integration platform for both biologists and bioinformatics. It contains the following categories for tools:

1. Import and Upload raw data; Export to DC;
2. Peak calling -- MAT, MA2C and MACS;
3. Correlation -- two wiggle files or two wiggle files with two bed files;
4. CEAS (sitepro, ceas, and gca) + Conservation + Heatmap;
5. Motif analysis;
6. Expression and GO;
7. Other Galaxy Low-Level Operations.

The following image is a typical homepage for AP module. If you are familiar with Galaxy, they are similar. The top menu provides links for analysis tools page(Analyze Data), workflows (Workflow), data importation from DC (Libraries), help and account management (User). The left panel provides a menu for analysis tools (**AP tools**), the middle panel is for parameters form and views for results, and the right panel lists your data and results (**AP data**) stored in the system. For more detail about the interface, please check Galaxy. Since this website is mainly based on galaxy, we will focus on the differences between our module and Galaxy including our special tools for Cistrome study, the interface to the Cistrome Data Collection module.



UCSC Genome Browser on Human Mar. 2006 Assembly

position/search chr5:70,256,524-70,284,592 size 28,069 bp.

http://genome.ucsc.edu/cgi-bin/hgc?hgsid=733...523&r=70284592&db=hg18&pix=620

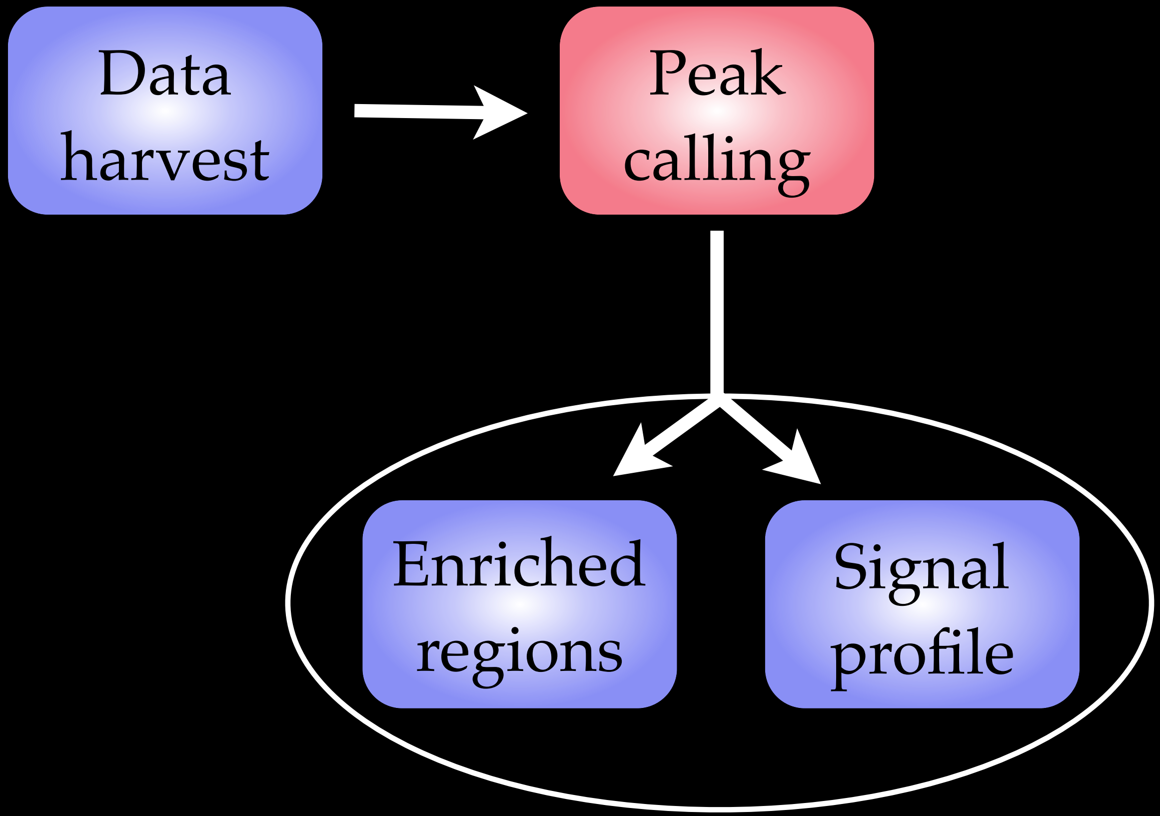
trac Integrated SCM & Project Management

SUBVERSION

Cistrome Work flow

Data harvest

Cistrome Work flow



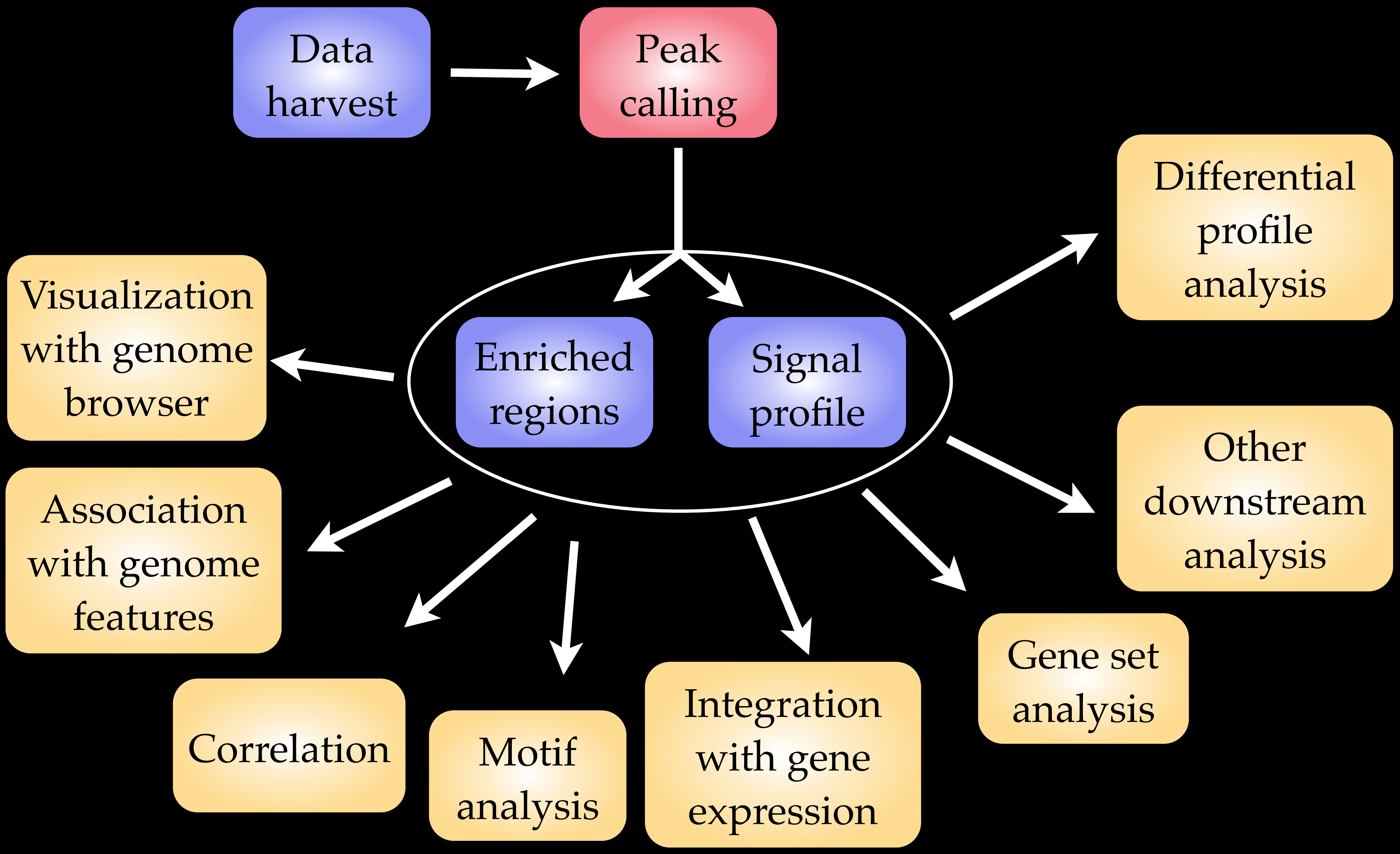
Data harvest

Peak calling

Enriched regions

Signal profile

Cistrome Work flow



Data harvest

Peak calling

Enriched regions

Signal profile

Visualization with genome browser

Association with genome features

Correlation

Motif analysis

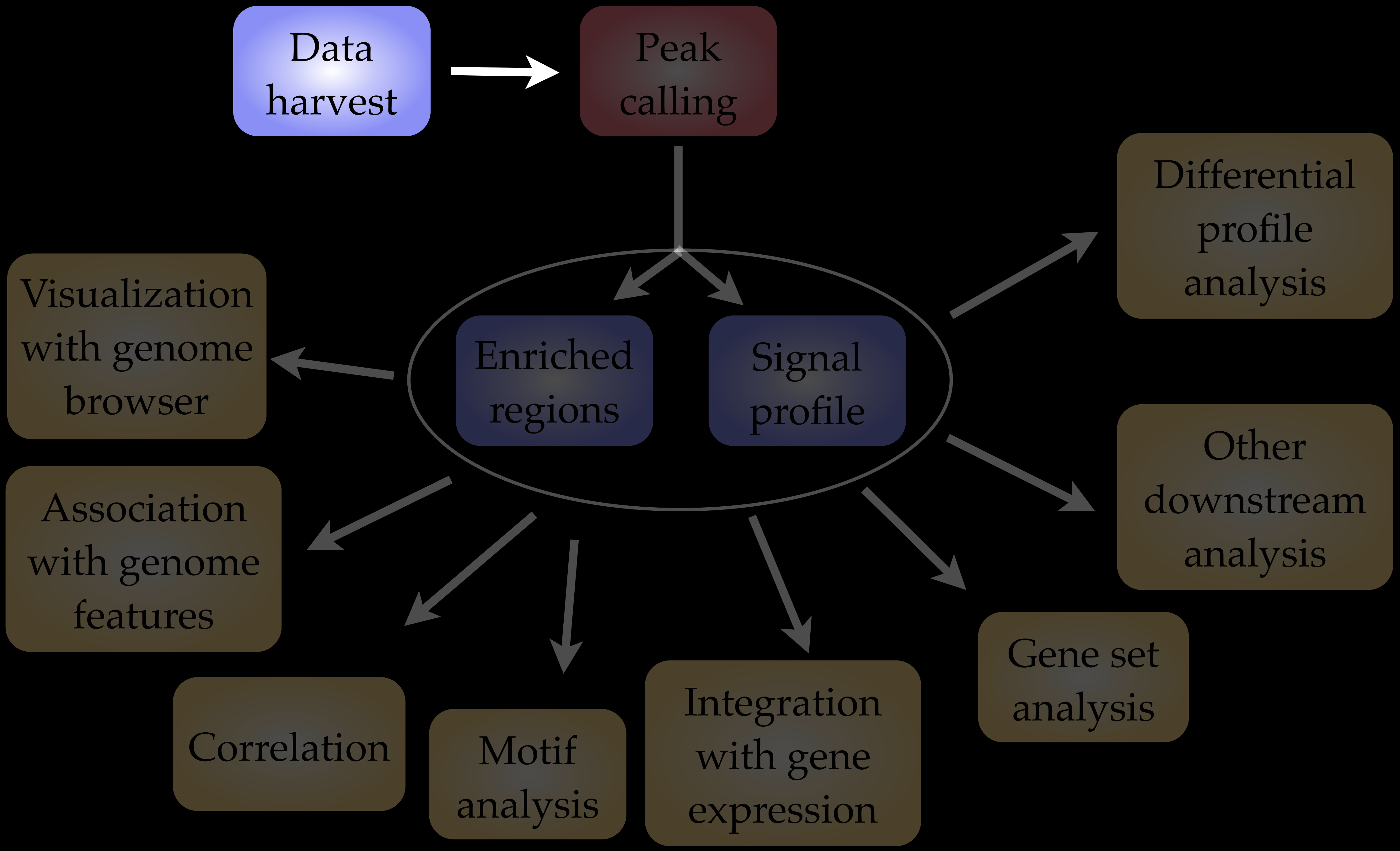
Integration with gene expression

Gene set analysis

Other downstream analysis

Differential profile analysis

Cistrome Work flow



Data harvest

Peak calling

Enriched regions

Signal profile

Visualization with genome browser

Association with genome features

Correlation

Motif analysis

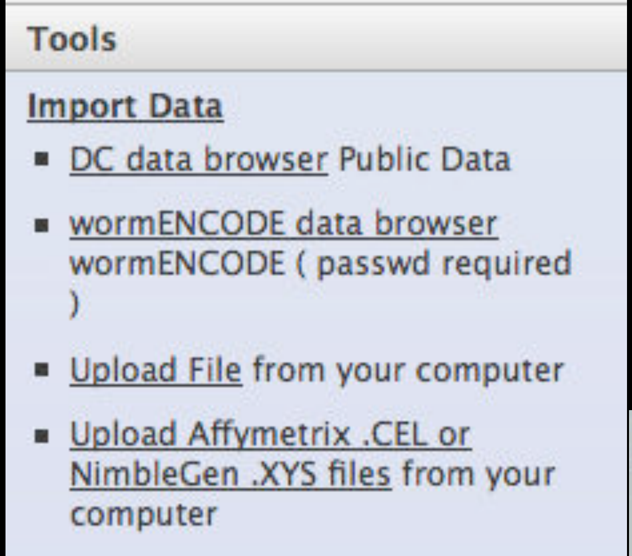
Integration with gene expression

Gene set analysis

Other downstream analysis

Differential profile analysis

Import data to AP



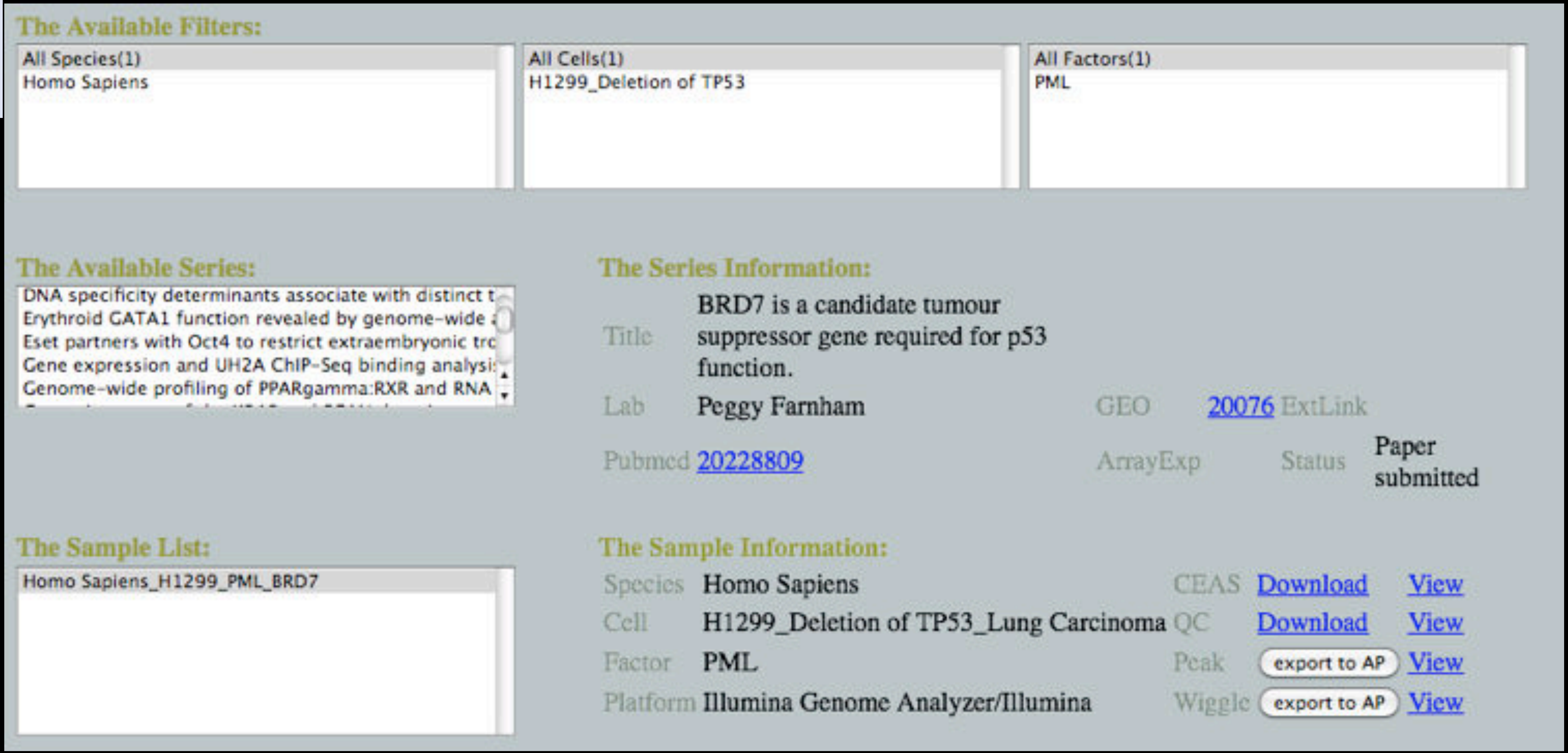
Tools

Import Data

- DC data browser Public Data
- wormENCODE data browser wormENCODE (passwd required)
- Upload File from your computer
- Upload Affymetrix .CEL or NimbleGen .XYS files from your computer

>600 Public ChIP-seq datasets to be added

~500 modENCODE ChIP-chip/seq



The Available Filters:

All Species(1)	All Cells(1)	All Factors(1)
Homo Sapiens	H1299_Deletion of TP53	PML

The Available Series:

DNA specificity determinants associate with distinct t
Erythroid GATA1 function revealed by genome-wide
Eset partners with Oct4 to restrict extraembryonic tr
Gene expression and UH2A ChIP-Seq binding analysi
Genome-wide profiling of PPARgamma:RXR and RNA

The Series Information:

Title	BRD7 is a candidate tumour suppressor gene required for p53 function.				
Lab	Peggy Farnham	GEO	20076 ExtLink		
Pubmed	20228809	ArrayExp		Status	Paper submitted

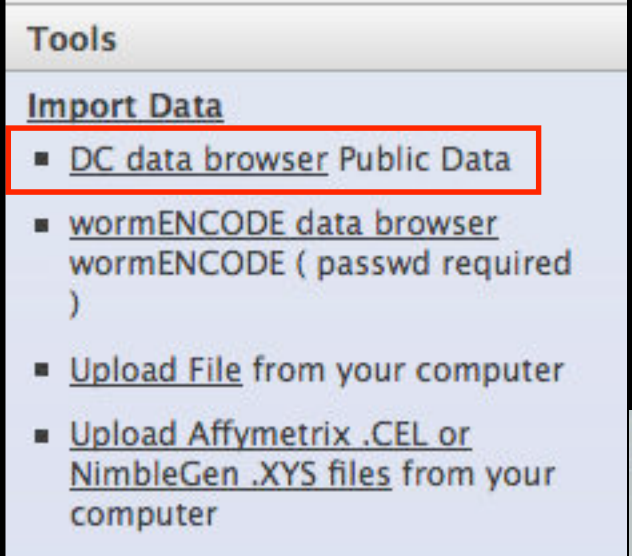
The Sample List:

Homo Sapiens_H1299_PML_BRD7

The Sample Information:

Species	Homo Sapiens	CEAS	Download	View
Cell	H1299_Deletion of TP53_Lung Carcinoma	QC	Download	View
Factor	PML	Peak	export to AP	View
Platform	Illumina Genome Analyzer/Illumina	Wiggle	export to AP	View

Import data to AP

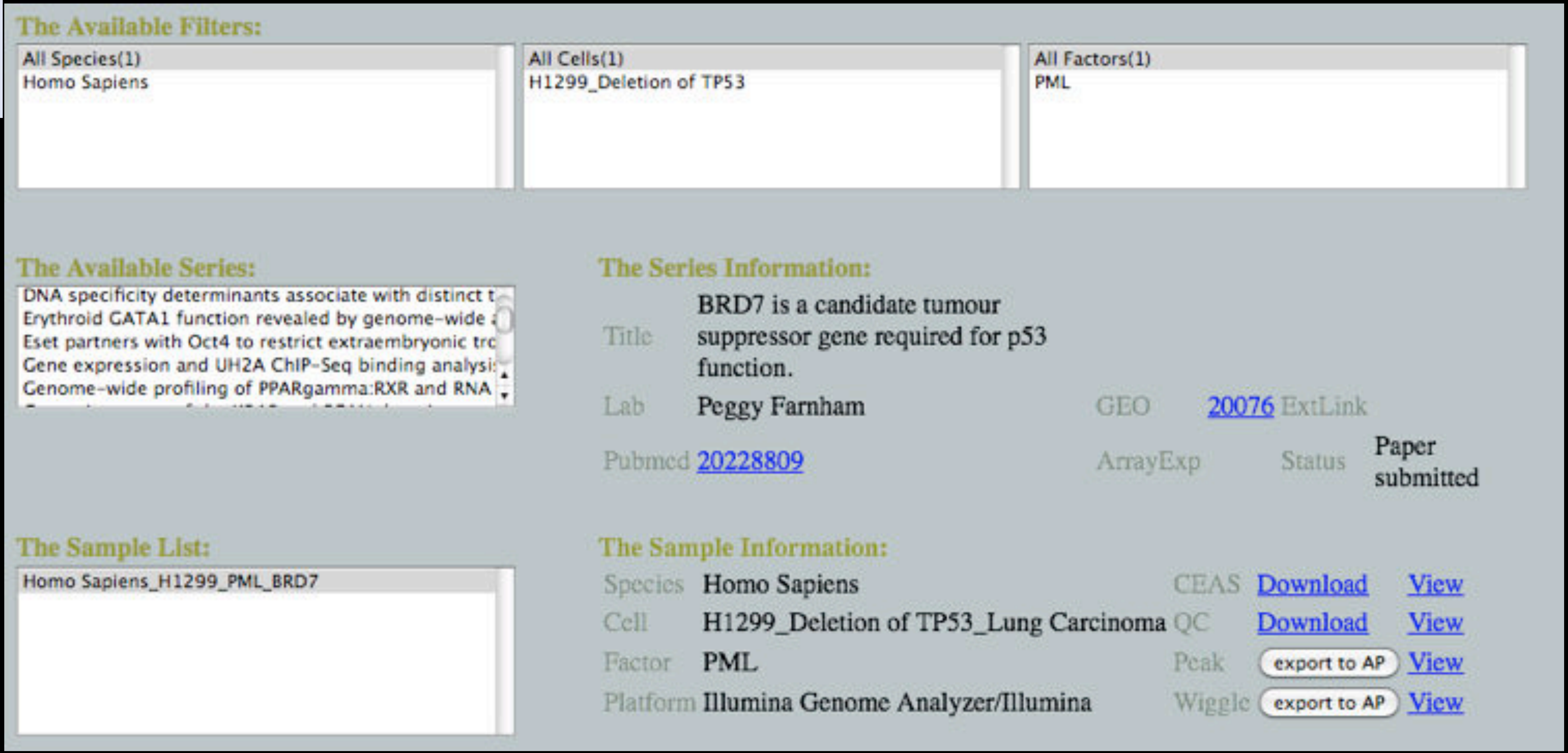


Tools

Import Data

- DC data browser Public Data
- wormENCODE data browser wormENCODE (passwd required)
- Upload File from your computer
- Upload Affymetrix .CEL or NimbleGen .XYS files from your computer

>600 Public ChIP-seq datasets to be added
~500 modENCODE ChIP-chip/seq



The Available Filters:

All Species(1)	All Cells(1)	All Factors(1)
Homo Sapiens	H1299_Deletion of TP53	PML

The Available Series:

DNA specificity determinants associate with distinct t
Erythroid GATA1 function revealed by genome-wide a
Eset partners with Oct4 to restrict extraembryonic trc
Gene expression and UH2A ChIP-Seq binding analysi:
Genome-wide profiling of PPARgamma:RXR and RNA

The Series Information:

Title: BRD7 is a candidate tumour suppressor gene required for p53 function.
Lab: Peggy Farnham
GEO: 20076 ExtLink
Pubmed: 20228809
ArrayExp
Status: Paper submitted

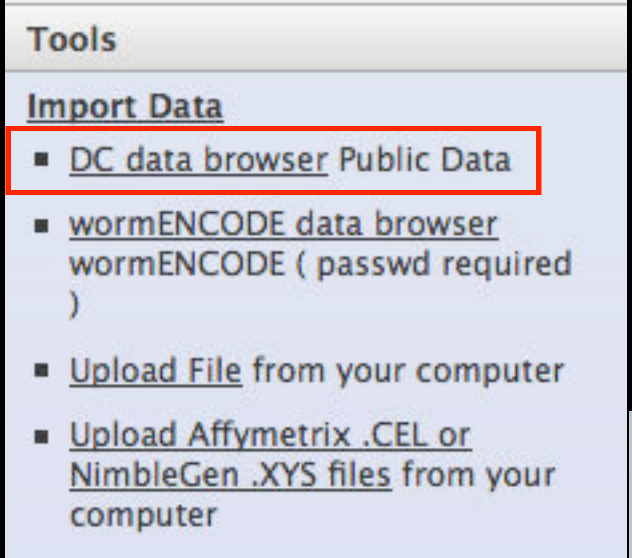
The Sample List:

Homo Sapiens_H1299_PML_BRD7

The Sample Information:

Species: Homo Sapiens
Cell: H1299_Deletion of TP53_Lung Carcinoma
Factor: PML
Platform: Illumina Genome Analyzer/Illumina

Import data to AP



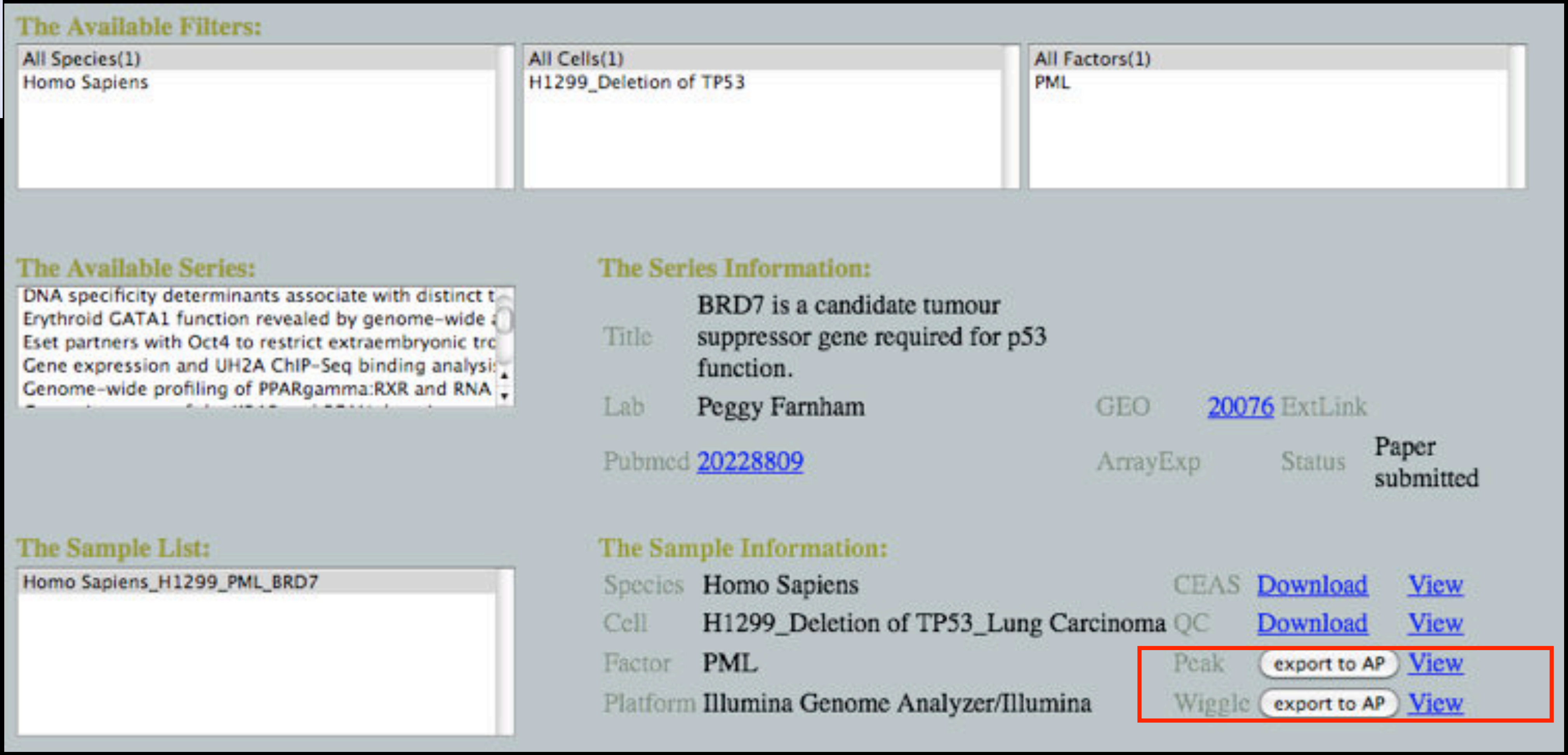
Tools

Import Data

- DC data browser Public Data
- wormENCODE data browser wormENCODE (passwd required)
- Upload File from your computer
- Upload Affymetrix .CEL or NimbleGen .XYS files from your computer

>600 Public ChIP-seq datasets to be added

~500 modENCODE ChIP-chip/seq



The Available Filters:

All Species(1)	All Cells(1)	All Factors(1)
Homo Sapiens	H1299_Deletion of TP53	PML

The Available Series:

DNA specificity determinants associate with distinct t
Erythroid GATA1 function revealed by genome-wide a
Eset partners with Oct4 to restrict extraembryonic trc
Gene expression and UH2A ChIP-Seq binding analysi:
Genome-wide profiling of PPARgamma:RXR and RNA

The Series Information:

Title: BRD7 is a candidate tumour suppressor gene required for p53 function.

Lab: Peggy Farnham

GEO: 20076 ExtLink

Pubmed: 20228809

ArrayExp

Status: Paper submitted

The Sample List:

Homo Sapiens_H1299_PML_BRD7

The Sample Information:

Species: Homo Sapiens

Cell: H1299_Deletion of TP53_Lung Carcinoma

Factor: PML

Platform: Illumina Genome Analyzer/Illumina

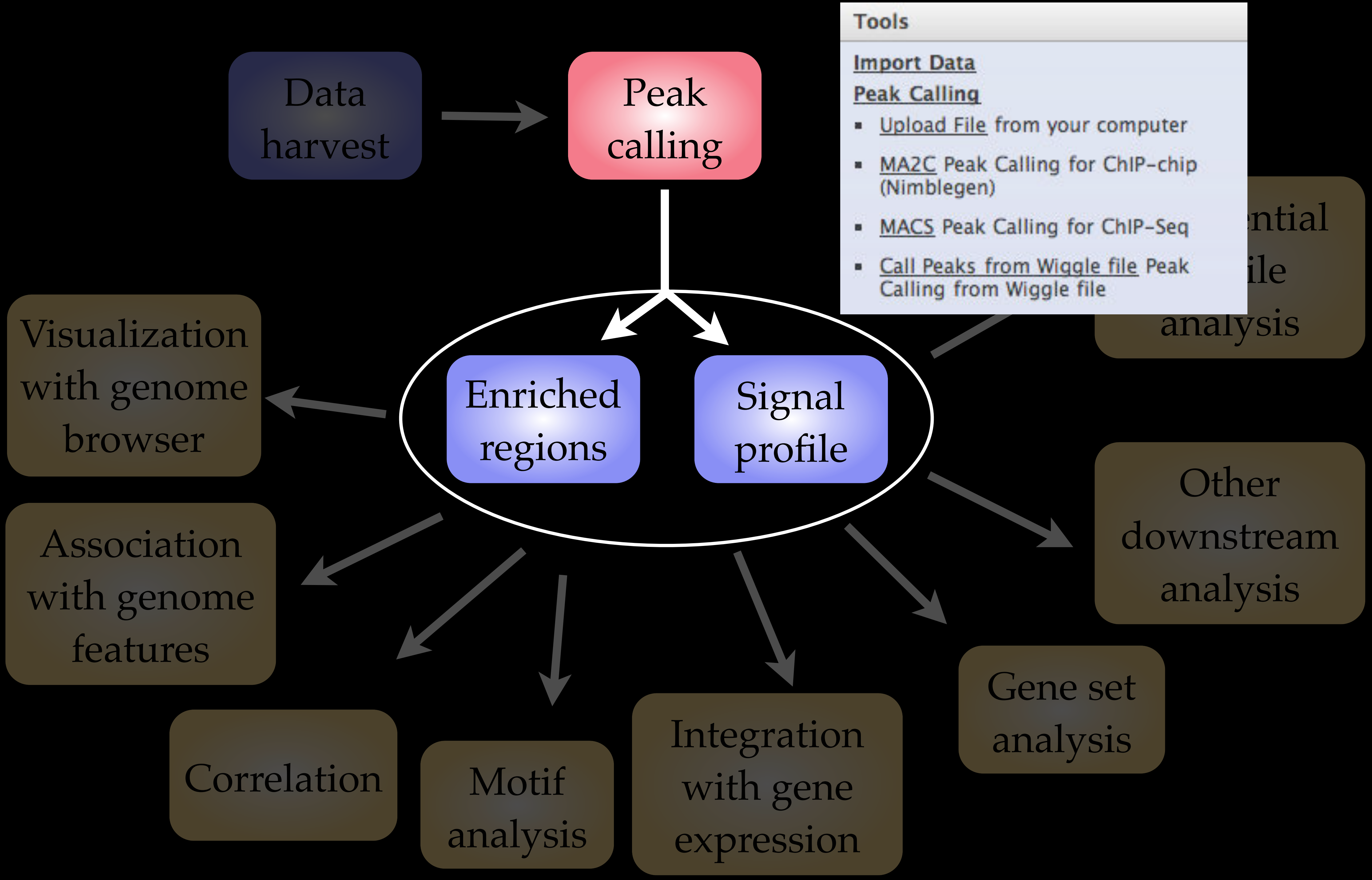
CEAS: Download View

QC: Download View

Peak: export to AP View

Wiggle: export to AP View

Cistrome Work flow



Import Data

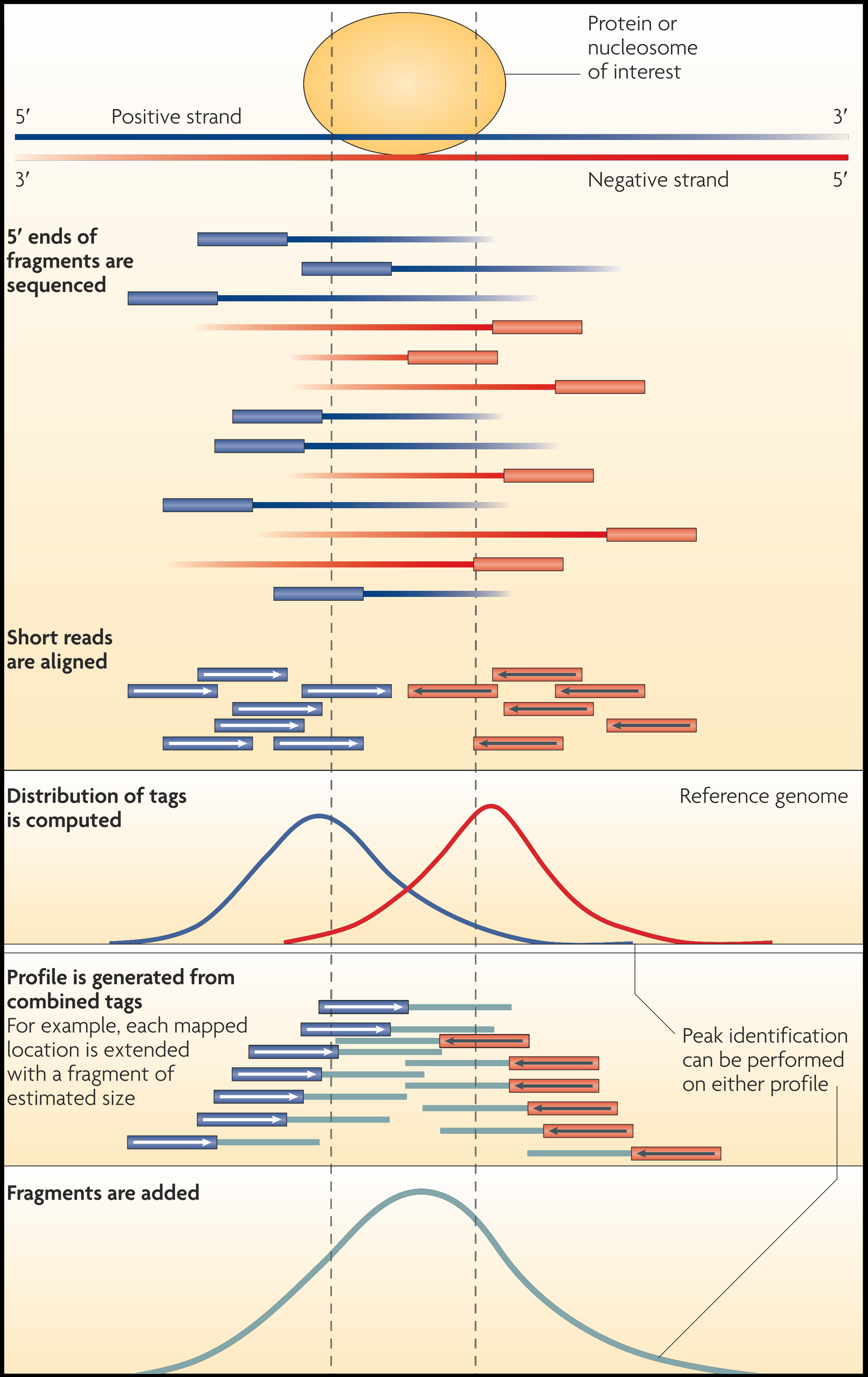
Peak Calling

- Upload File from your computer
- MA2C Peak Calling for ChIP-chip (Nimblegen)
- MACS Peak Calling for ChIP-Seq
- Call Peaks from Wiggle file Peak Calling from Wiggle file

MACS

Model based Analysis of ChIP-Seq

- Tags need to be shifted and extended with a estimated fragment size.
- Model the local bias caused by open chromatin effects, or sequence mappability.



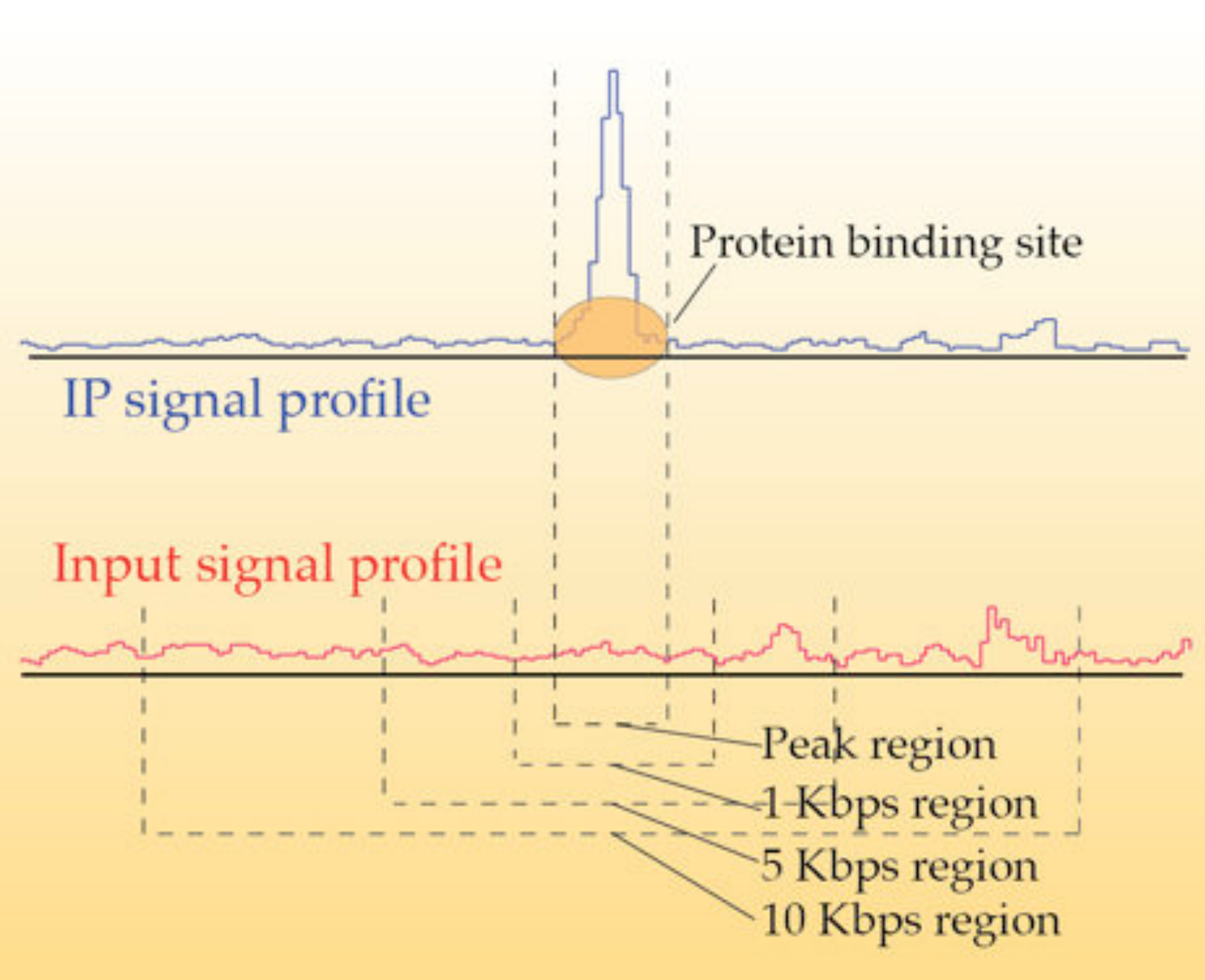
Nat Rev Genet (2009) vol. 10 (10) pp. 669-80

MACS

Model based Analysis of **C**hIP-**S**eq

- Tags need to be shifted and extended with a estimated fragment size.

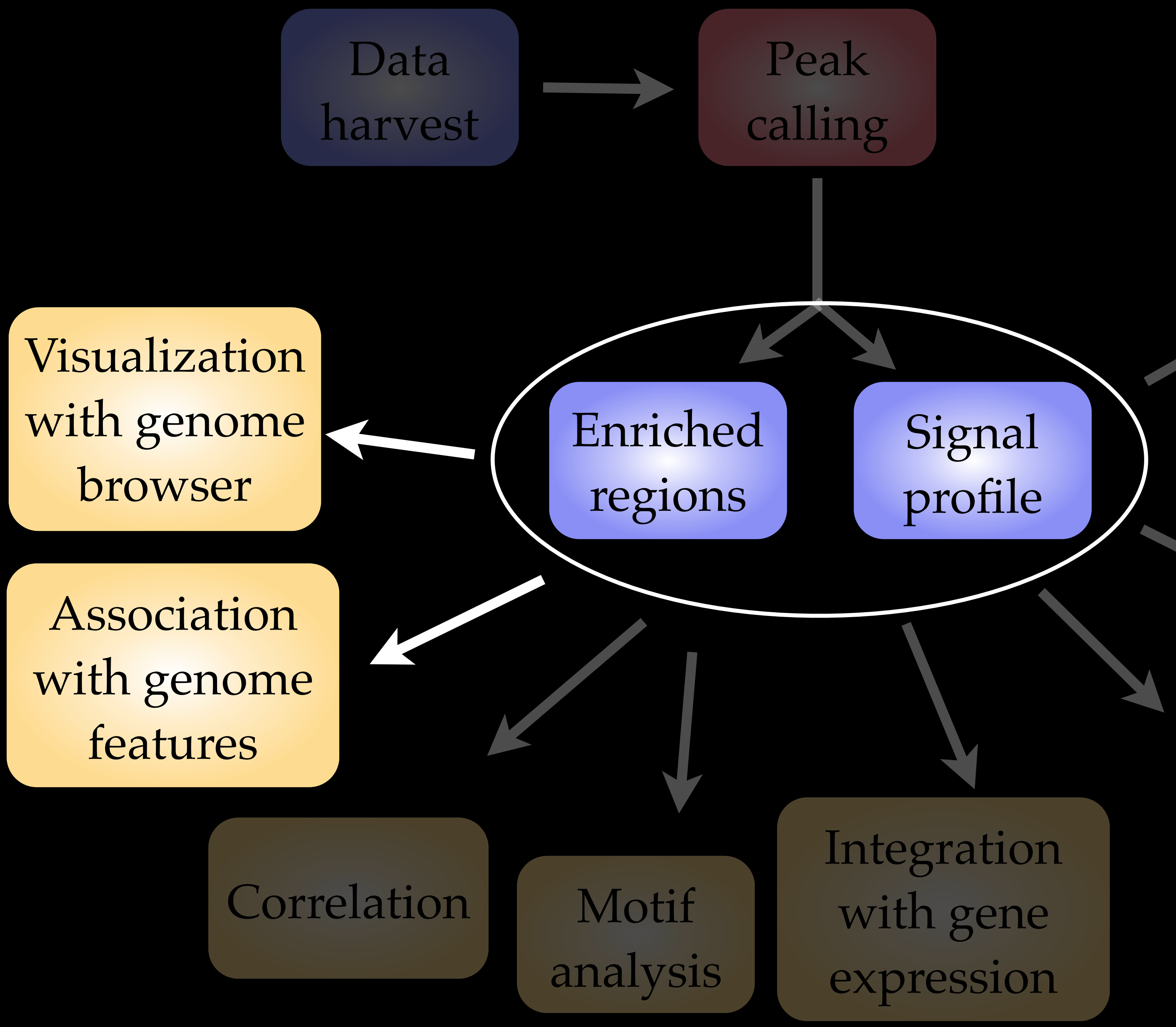
- Model the local bias caused by open chromatin effects, or sequence mappability.

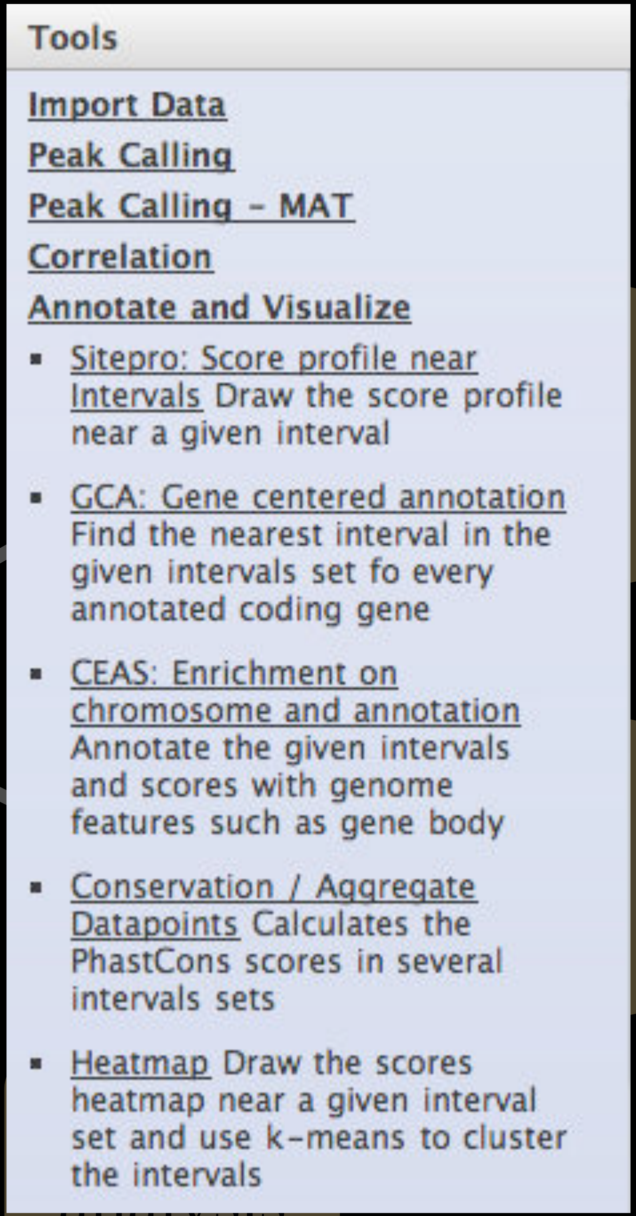


$$\lambda_{local} = \max(\lambda_{BG}, [\lambda_{peak}, \lambda_{1k},]\lambda_{5k}, \lambda_{10k})$$

$$pvalue(\lambda; k) = 1 - e^{-\lambda} \sum_{i=0}^{k-1} \frac{\lambda^i}{i!}$$

Cistrome Work flow




Tools

Import Data

Peak Calling

Peak Calling - MAT

Correlation

Annotate and Visualize

- Sitepro: Score profile near Intervals Draw the score profile near a given interval
- GCA: Gene centered annotation Find the nearest interval in the given intervals set fo every annotated coding gene
- CEAS: Enrichment on chromosome and annotation Annotate the given intervals and scores with genome features such as gene body
- Conservation / Aggregate Datapoints Calculates the PhastCons scores in several intervals sets
- Heatmap Draw the scores heatmap near a given interval set and use k-means to cluster the intervals

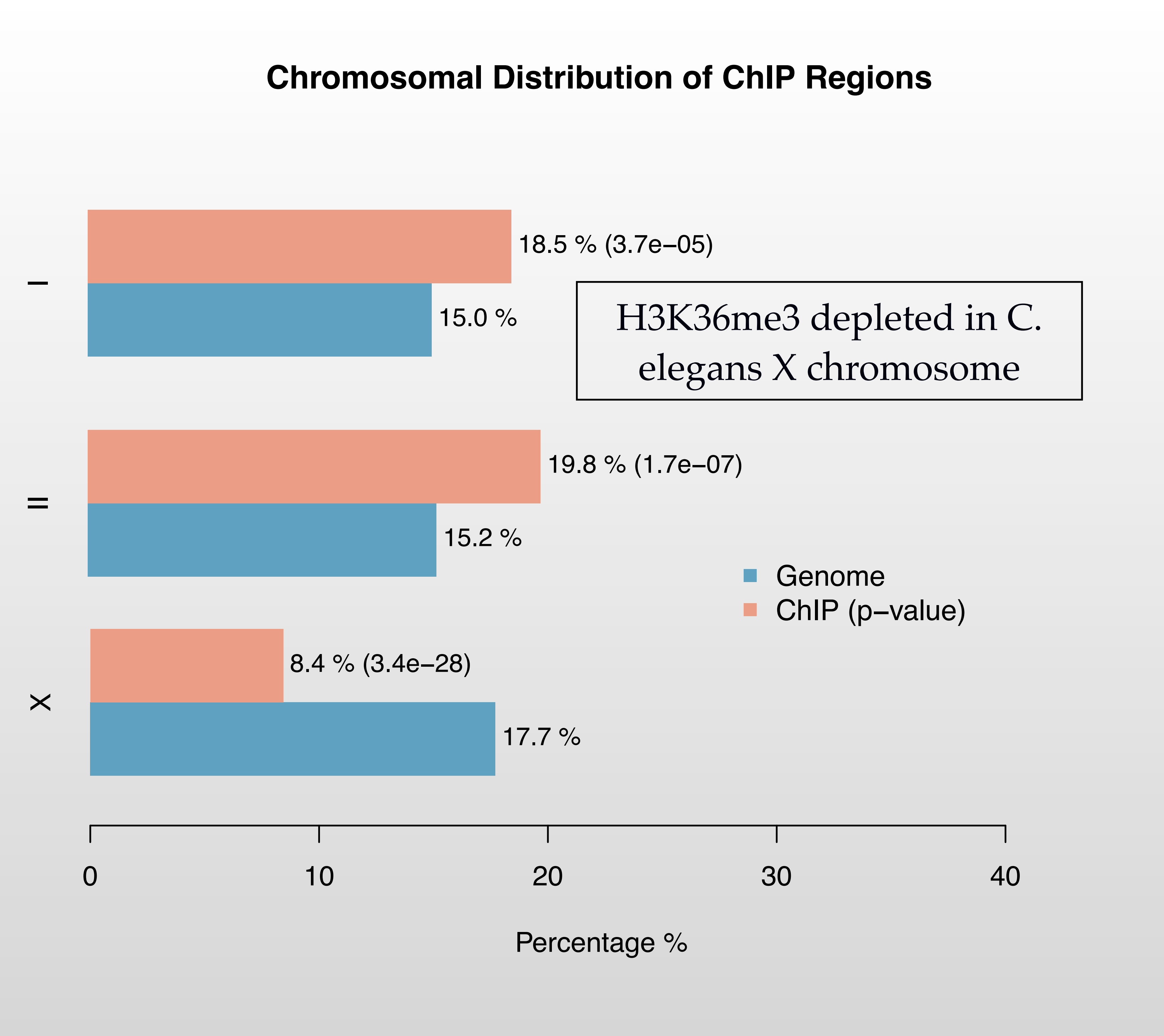
CEAS

Cis-regulatory **E**lement **A**nnotation **S**ystem

- A easy solution to generate graphic report and statistics
- Published in *Bioinformatics* (2009) vol. 25 (19) pp. 2605-6
- Coded in Python

ChIP region annotation

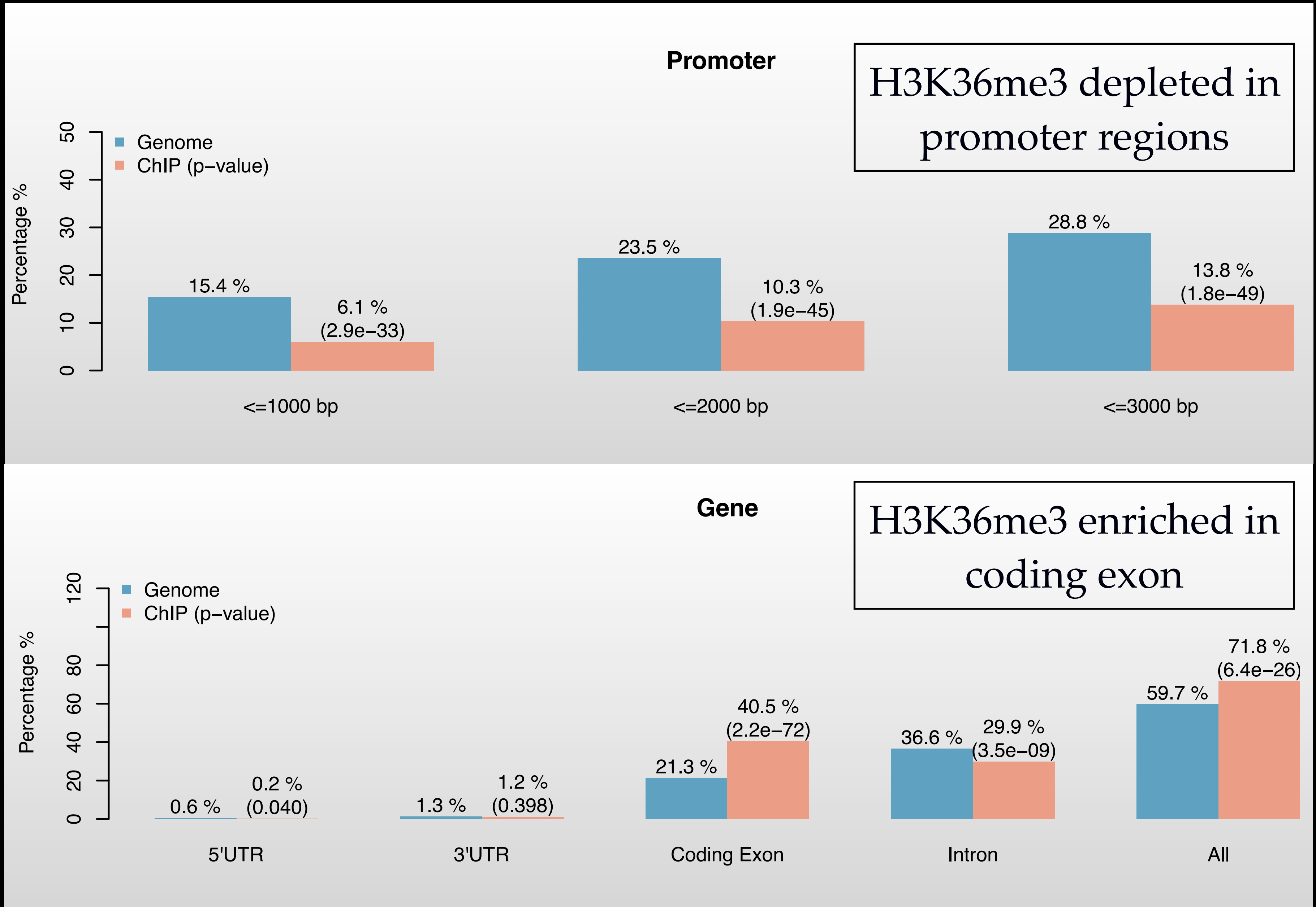
- Chromosomal distribution of ChIP enriched regions
- Feature distribution
- Signal profiles



H3K36me3 depleted in C. elegans X chromosome

ChIP region annotation

- Chromosomal distribution of ChIP enriched regions
- Feature distribution
- Signal profiles

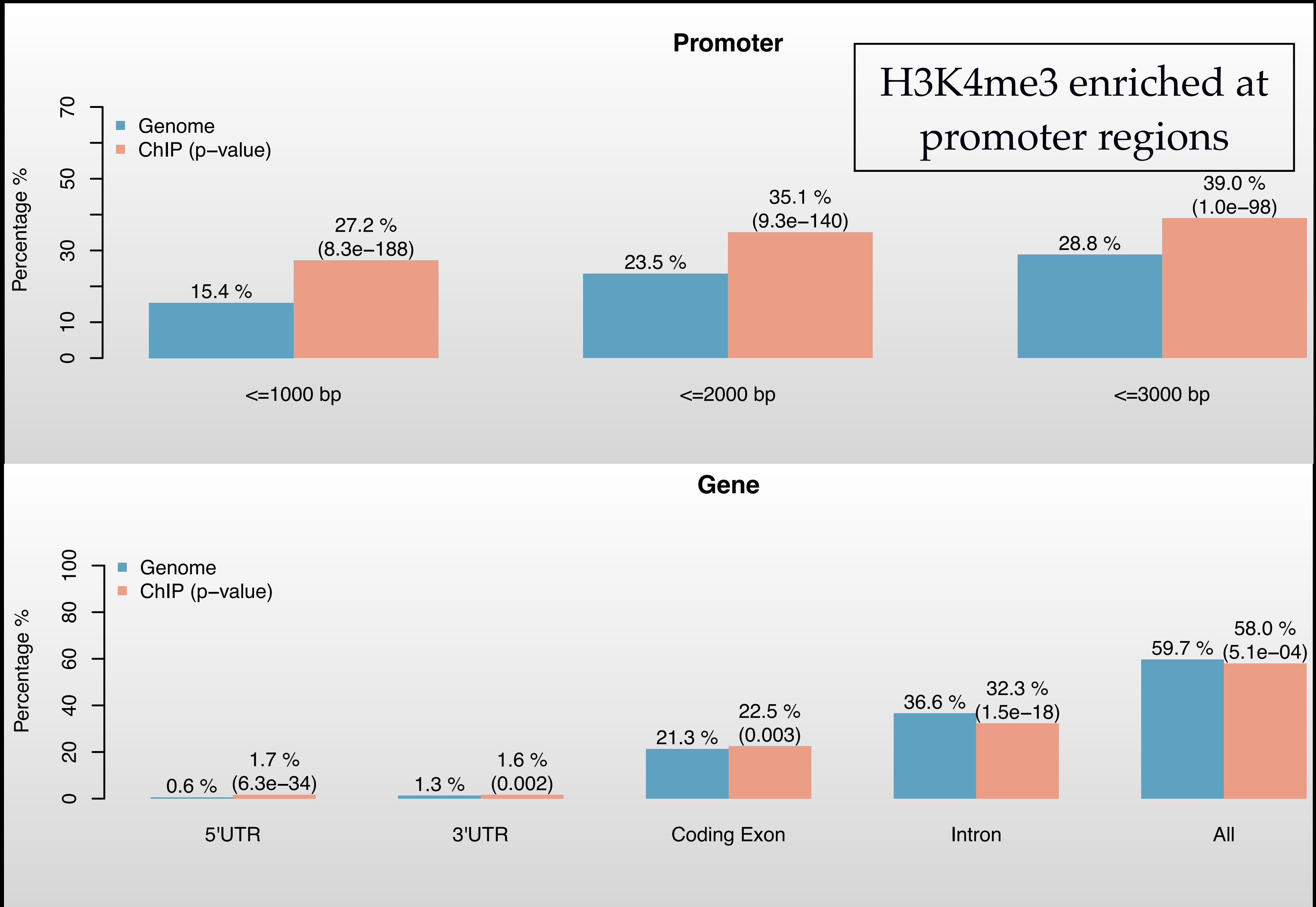


H3K36me3 depleted in promoter regions

H3K36me3 enriched in coding exon

ChIP region annotation

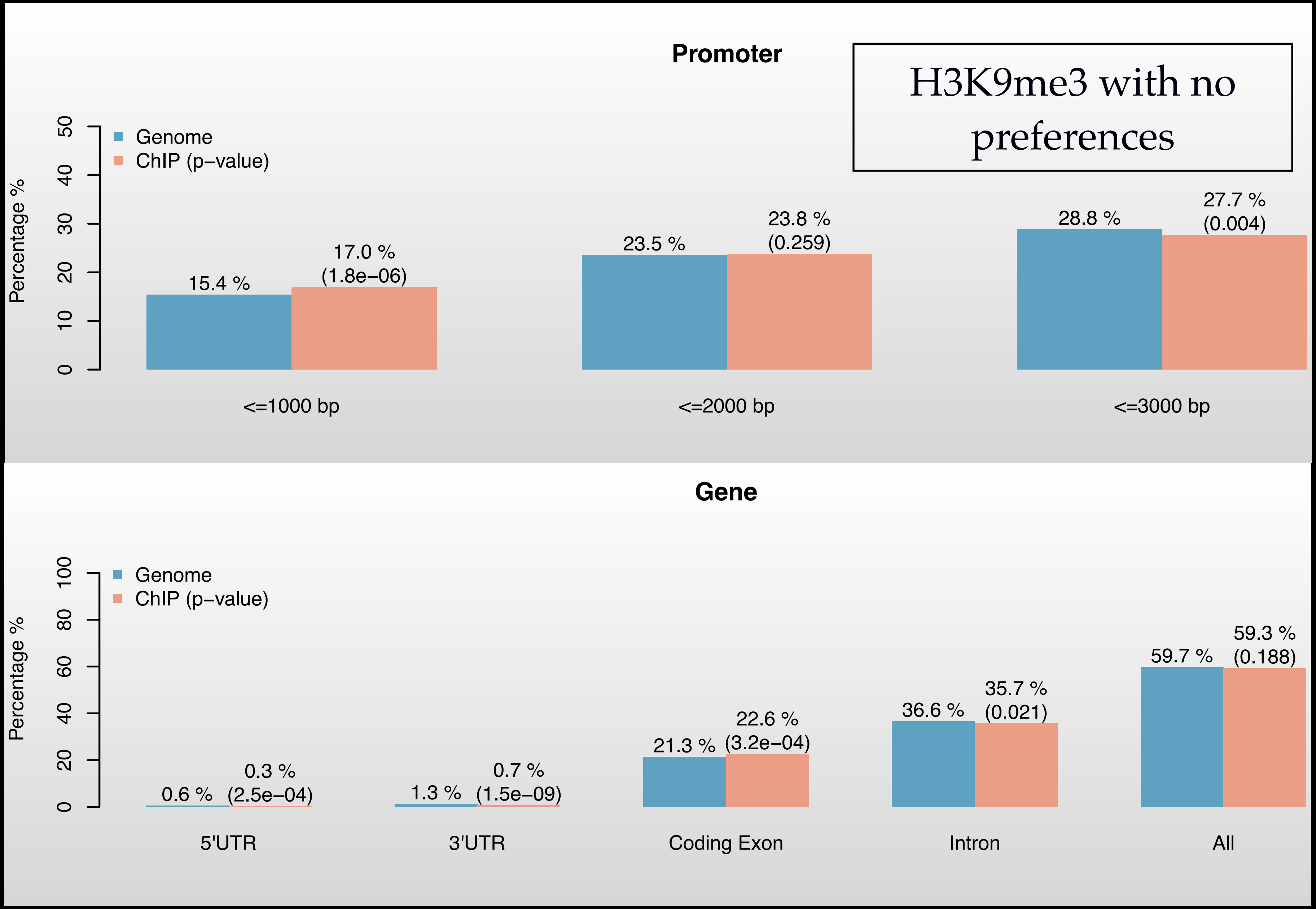
- Chromosomal distribution of ChIP enriched regions
- Feature distribution
- Signal profiles



H3K4me3 enriched at promoter regions

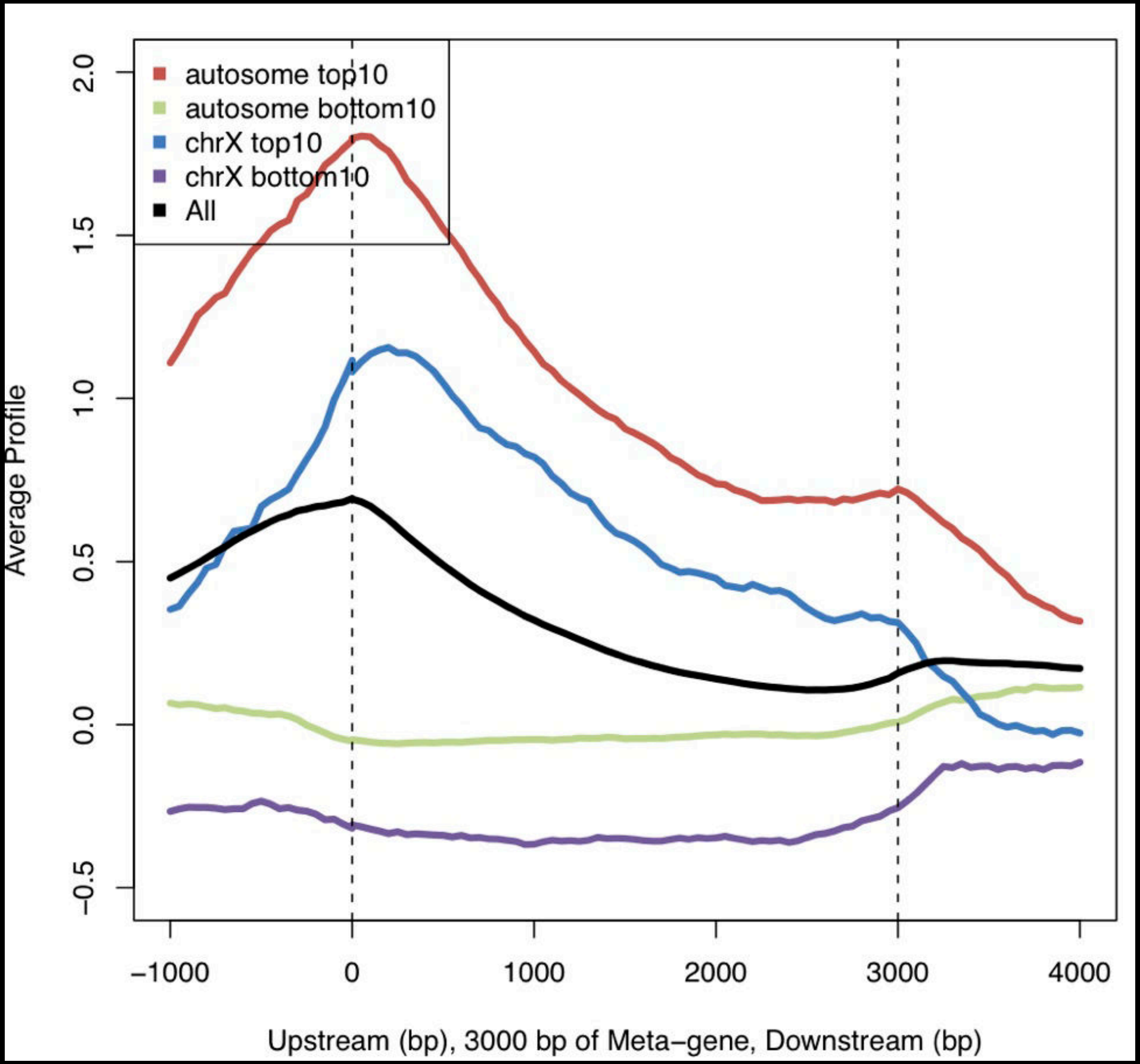
ChIP region annotation

- Chromosomal distribution of ChIP enriched regions
- Feature distribution
- Signal profiles


H3K9me3 with no preferences

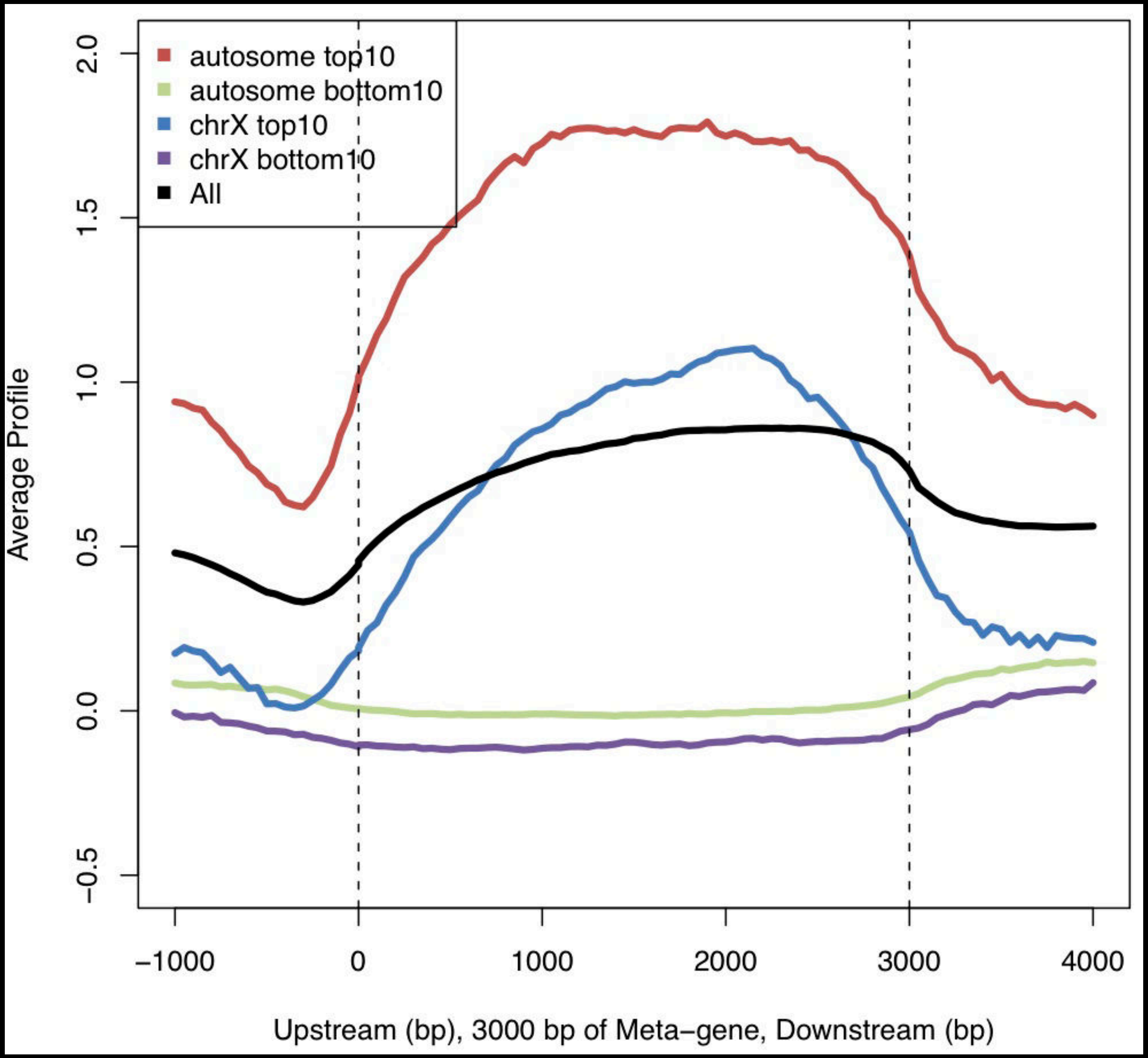
Signal annotation

- Chromosomal distribution of ChIP enriched regions
- Feature distribution
- Signal profiles



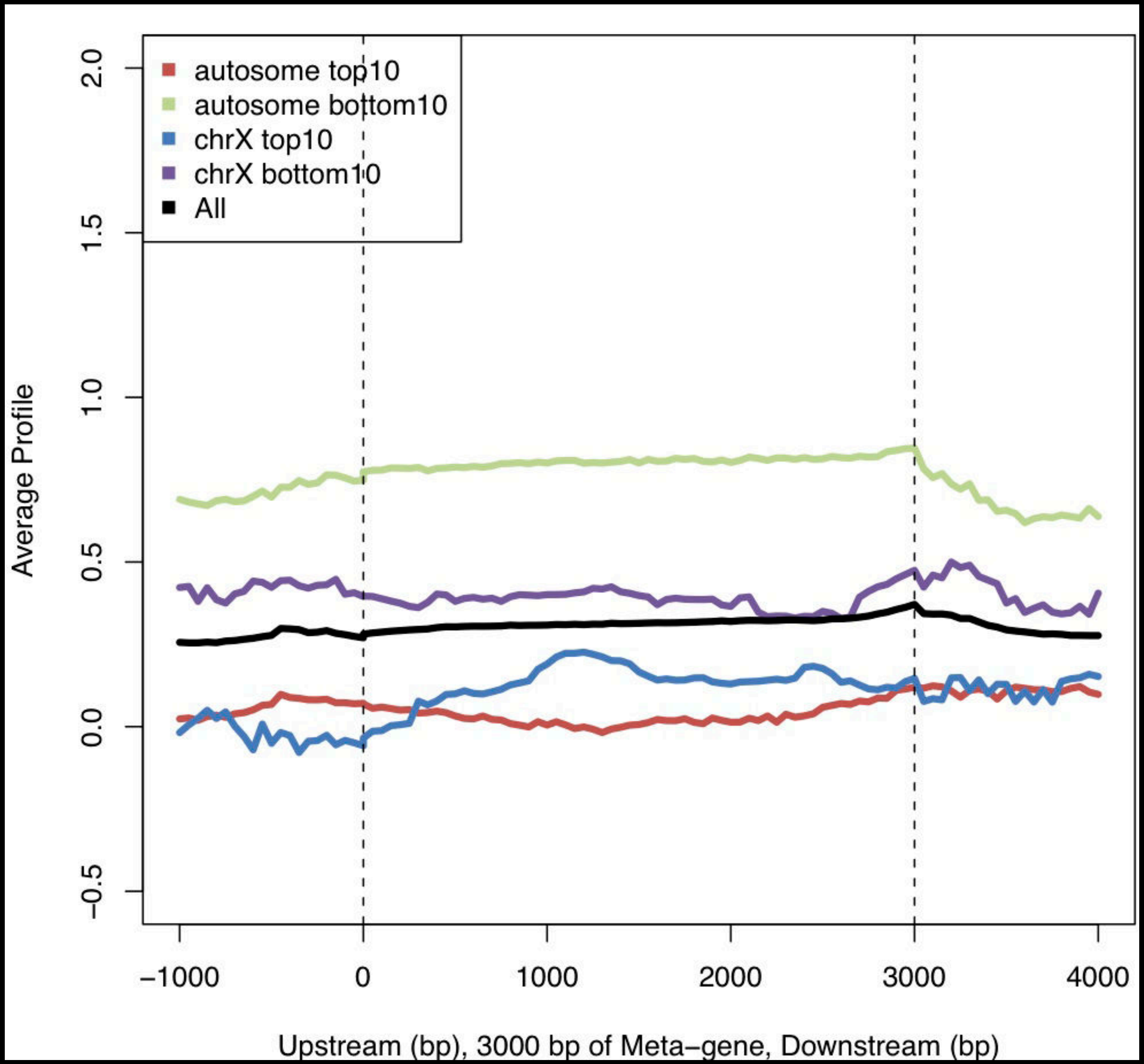
Signal annotation

- Chromosomal distribution of ChIP enriched regions
- Feature distribution
- Signal profiles

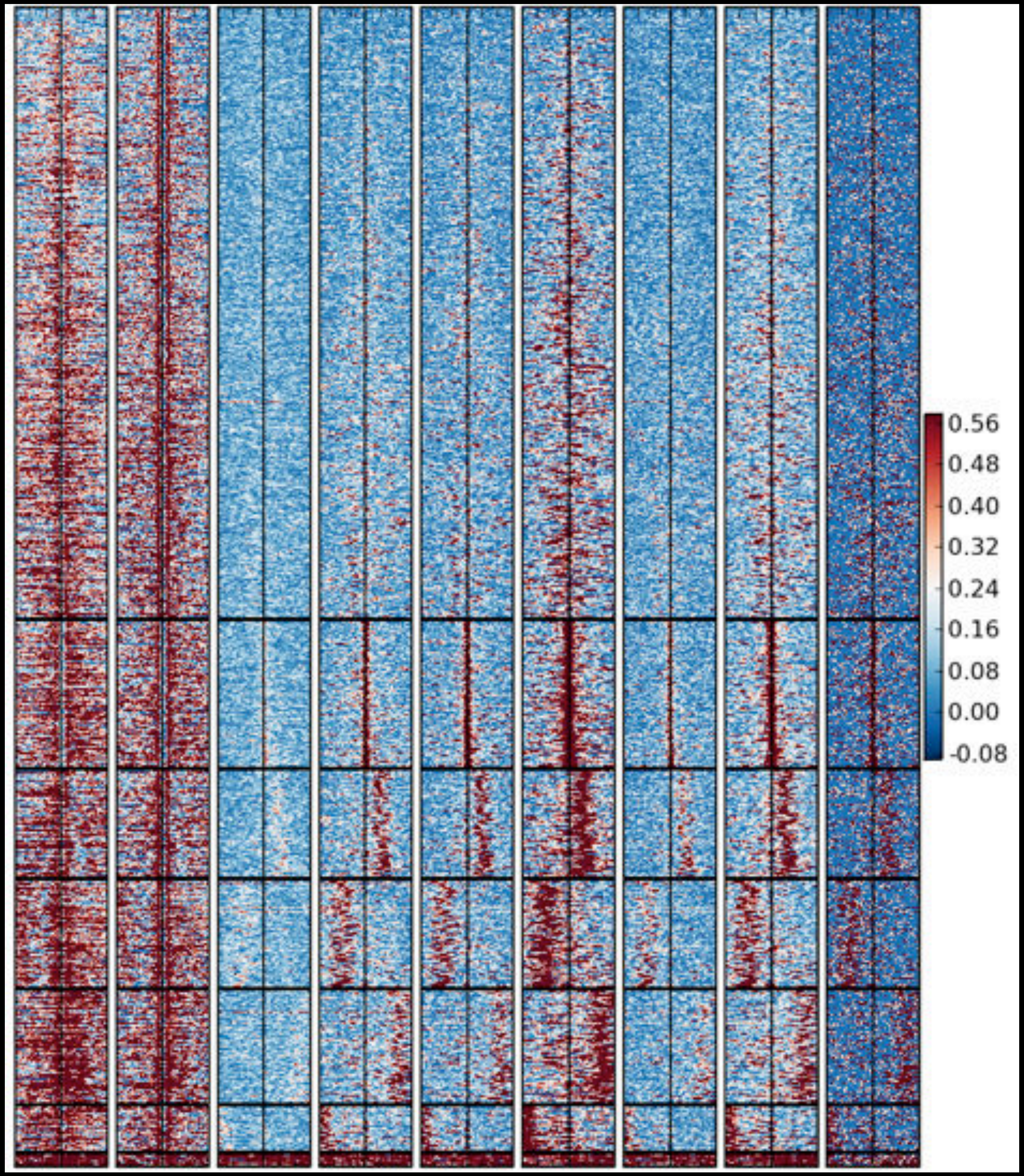


Signal annotation

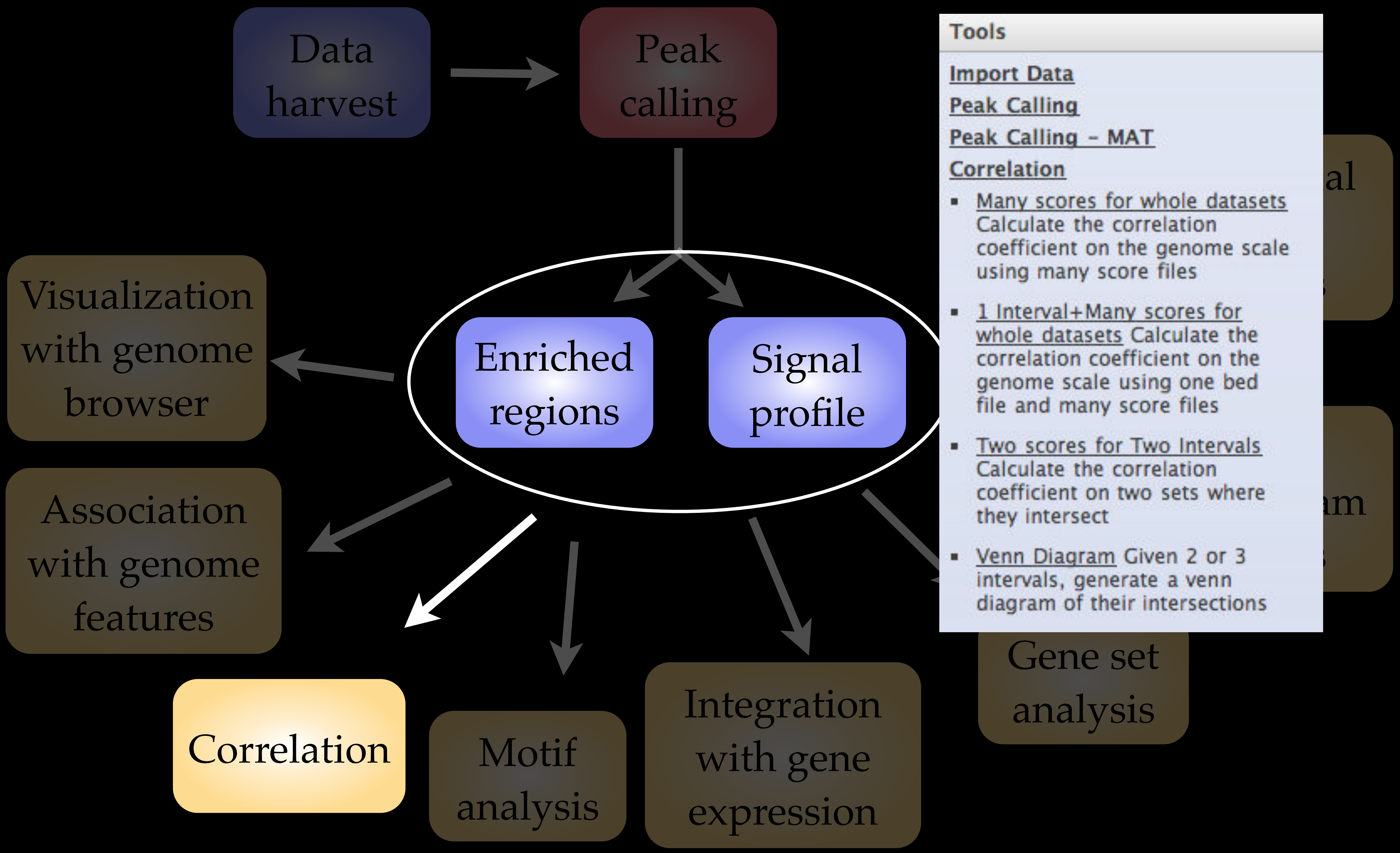
- Chromosomal distribution of ChIP enriched regions
- Feature distribution
- Signal profiles



Heatmap and clustering



Cistrome Work flow



Data harvest

Peak calling

Enriched regions

Signal profile

Visualization with genome browser

Association with genome features

Correlation

Motif analysis

Integration with gene expression

Gene set analysis

al

am

Tools

[Import Data]

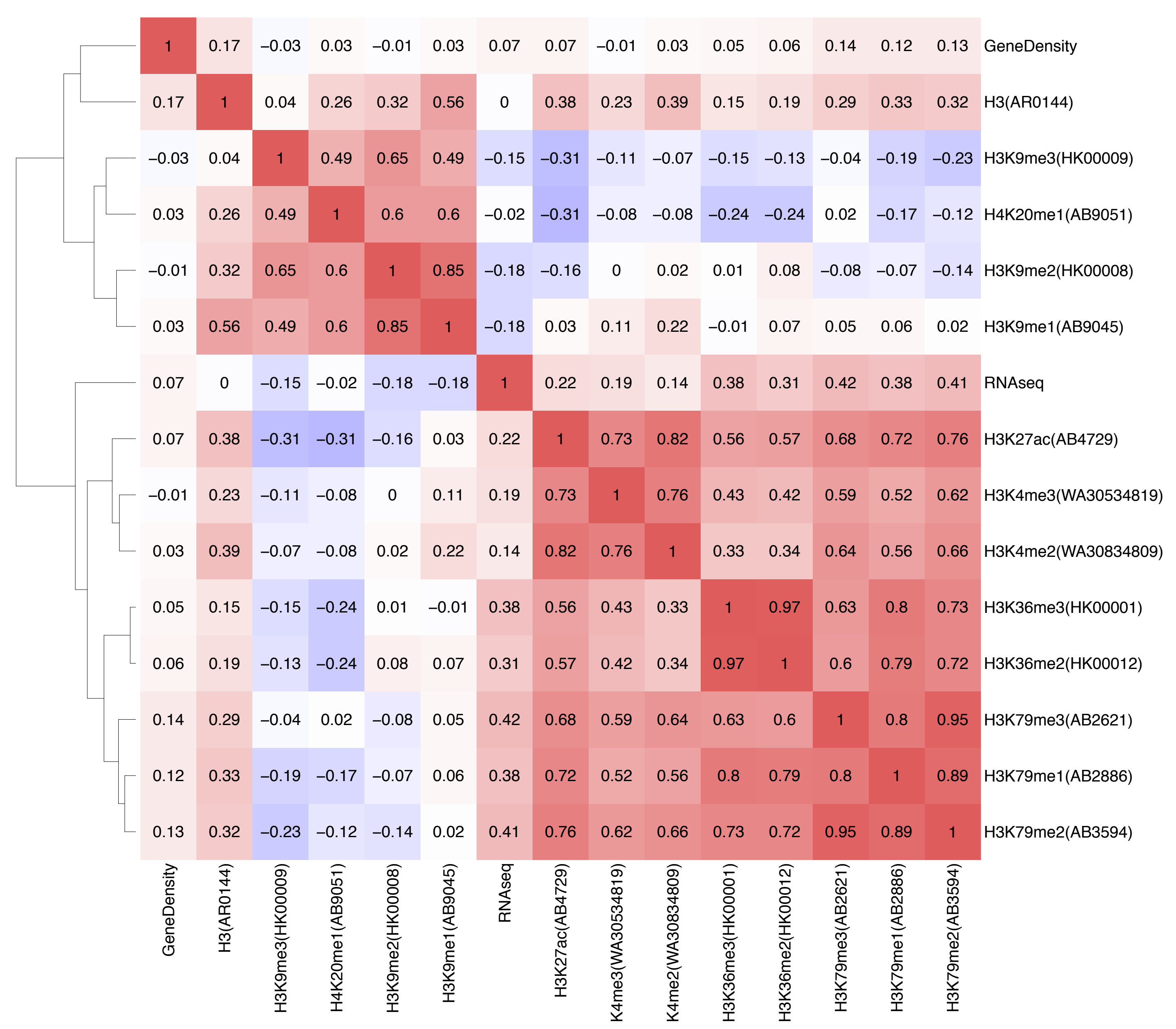
[Peak Calling]

[Peak Calling - MAT]

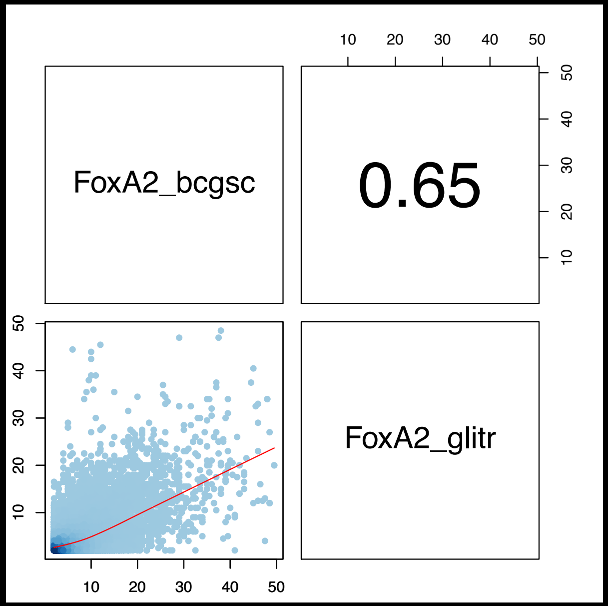
[Correlation]

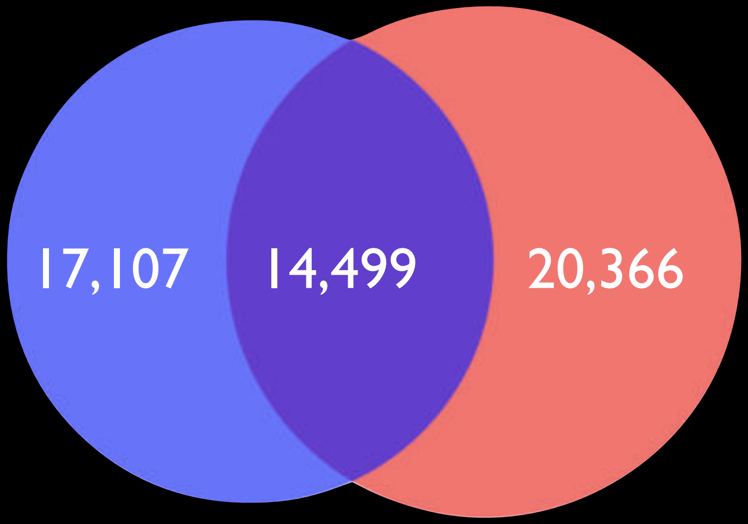
- [Many scores for whole datasets] Calculate the correlation coefficient on the genome scale using many score files
- [1 Interval+Many scores for whole datasets] Calculate the correlation coefficient on the genome scale using one bed file and many score files
- [Two scores for Two Intervals] Calculate the correlation coefficient on two sets where they intersect
- [Venn Diagram] Given 2 or 3 intervals, generate a venn diagram of their intersections

Correlation

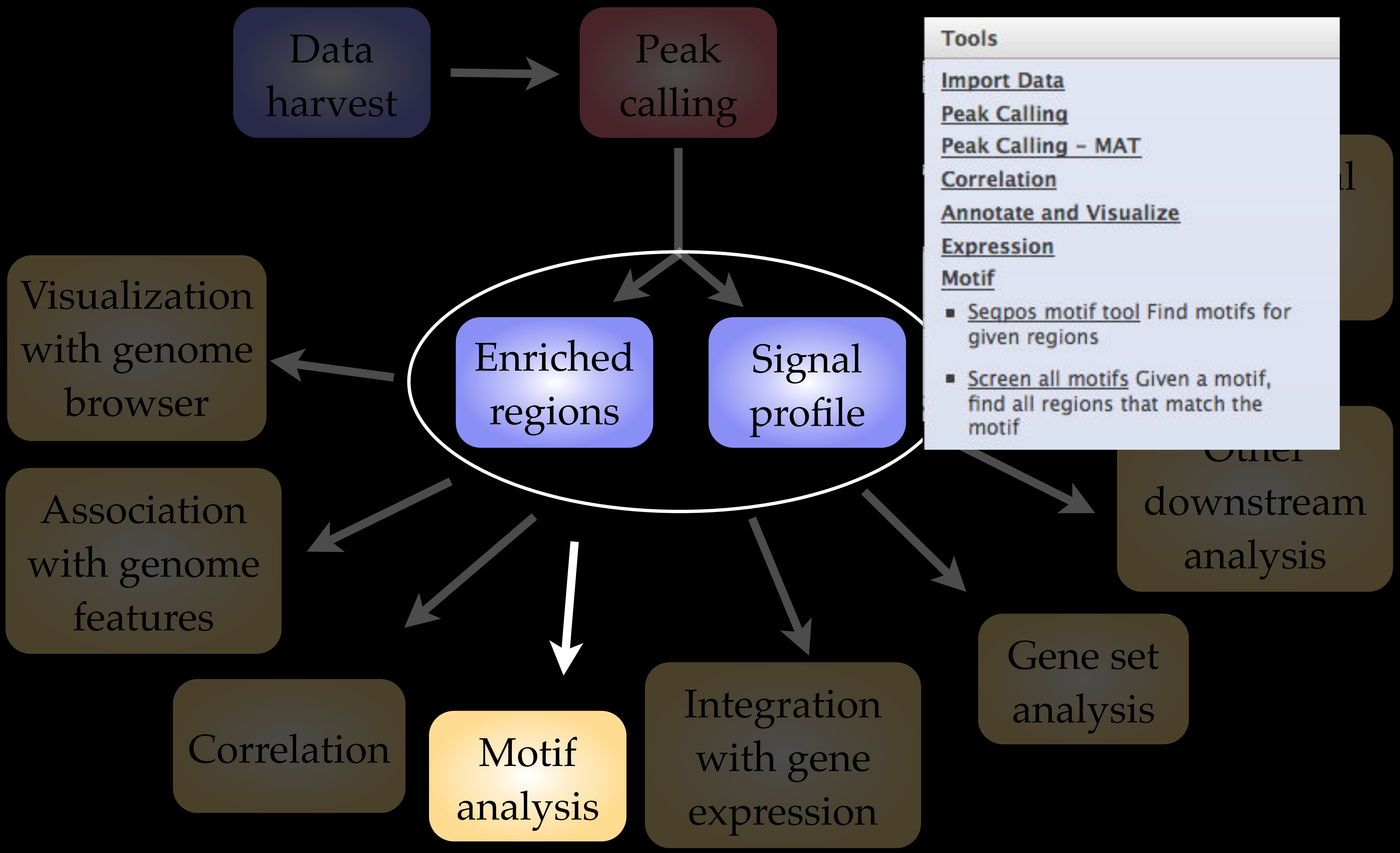


	GeneDensity	H3(AR0144)	H3K9me3(HK00009)	H4K20me1(AB9051)	H3K9me2(HK00008)	H3K9me1(AB9045)	RNAseq	H3K27ac(AB4729)	K4me3(WA30534819)	K4me2(WA30834809)	H3K36me3(HK00001)	H3K36me2(HK00012)	H3K79me3(AB2621)	H3K79me1(AB2886)	H3K79me2(AB3594)
GeneDensity	1	0.17	−0.03	0.03	−0.01	0.03	0.07	0.07	−0.01	0.03	0.05	0.06	0.14	0.12	0.13
H3(AR0144)	0.17	1	0.04	0.26	0.32	0.56	0	0.38	0.23	0.39	0.15	0.19	0.29	0.33	0.32
H3K9me3(HK00009)	−0.03	0.04	1	0.49	0.65	0.49	−0.15	−0.31	−0.11	−0.07	−0.15	−0.13	−0.04	−0.19	−0.23
H4K20me1(AB9051)	0.03	0.26	0.49	1	0.6	0.6	−0.02	−0.31	−0.08	−0.08	−0.24	−0.24	0.02	−0.17	−0.12
H3K9me2(HK00008)	−0.01	0.32	0.65	0.6	1	0.85	−0.18	−0.16	0	0.02	0.01	0.08	−0.08	−0.07	−0.14
H3K9me1(AB9045)	0.03	0.56	0.49	0.6	0.85	1	−0.18	0.03	0.11	0.22	−0.01	0.07	0.05	0.06	0.02
RNAseq	0.07	0	−0.15	−0.02	−0.18	−0.18	1	0.22	0.19	0.14	0.38	0.31	0.42	0.38	0.41
H3K27ac(AB4729)	0.07	0.38	−0.31	−0.31	−0.16	0.03	0.22	1	0.73	0.82	0.56	0.57	0.68	0.72	0.76
H3K4me3(WA30534819)	−0.01	0.23	−0.11	−0.08	0	0.11	0.19	0.73	1	0.76	0.43	0.42	0.59	0.52	0.62
H3K4me2(WA30834809)	0.03	0.39	−0.07	−0.08	0.02	0.22	0.14	0.82	0.76	1	0.33	0.34	0.64	0.56	0.66
H3K36me3(HK00001)	0.05	0.15	−0.15	−0.24	0.01	−0.01	0.38	0.56	0.43	0.33	1	0.97	0.63	0.8	0.73
H3K36me2(HK00012)	0.06	0.19	−0.13	−0.24	0.08	0.07	0.31	0.57	0.42	0.34	0.97	1	0.6	0.79	0.72
H3K79me3(AB2621)	0.14	0.29	−0.04	0.02	−0.08	0.05	0.42	0.68	0.59	0.64	0.63	0.6	1	0.8	0.95
H3K79me1(AB2886)	0.12	0.33	−0.19	−0.17	−0.07	0.06	0.38	0.72	0.52	0.56	0.8	0.79	0.8	1	0.89
H3K79me2(AB3594)	0.13	0.32	−0.23	−0.12	−0.14	0.02	0.41	0.76	0.62	0.66	0.73	0.72	0.95	0.89	1





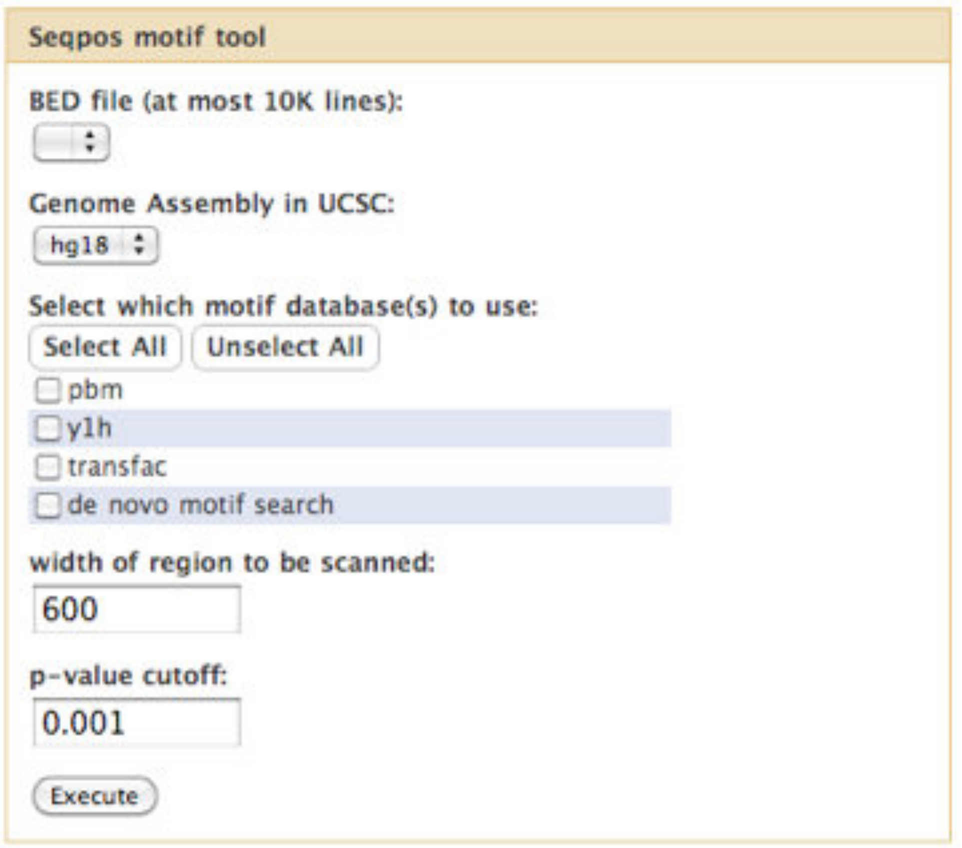
Cistrome Work flow



Tools

- Import Data
- Peak Calling
- Peak Calling - MAT
- Correlation
- Annotate and Visualize
- Expression
- Motif
  - Seqpos motif tool Find motifs for given regions
  - Screen all motifs Given a motif, find all regions that match the motif

Motif Analysis
SeqPos



Seqpos motif tool

BED file (at most 10K lines):

Genome Assembly in UCSC: hg18

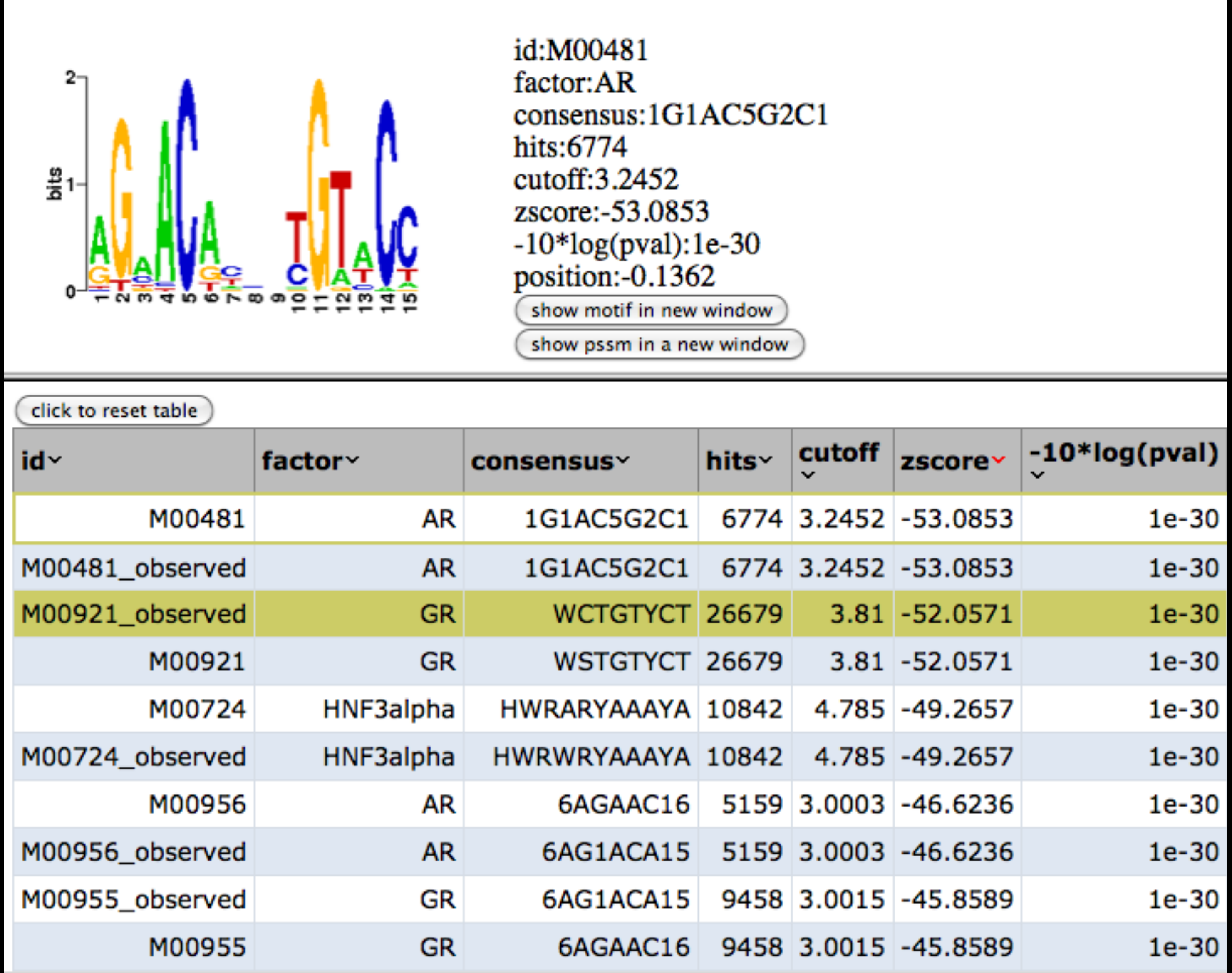
Select which motif database(s) to use:

- ☐ pbm
- ☐ y1h
- ☐ transfac
- ☐ de novo motif search

width of region to be scanned: 600

p-value cutoff: 0.001

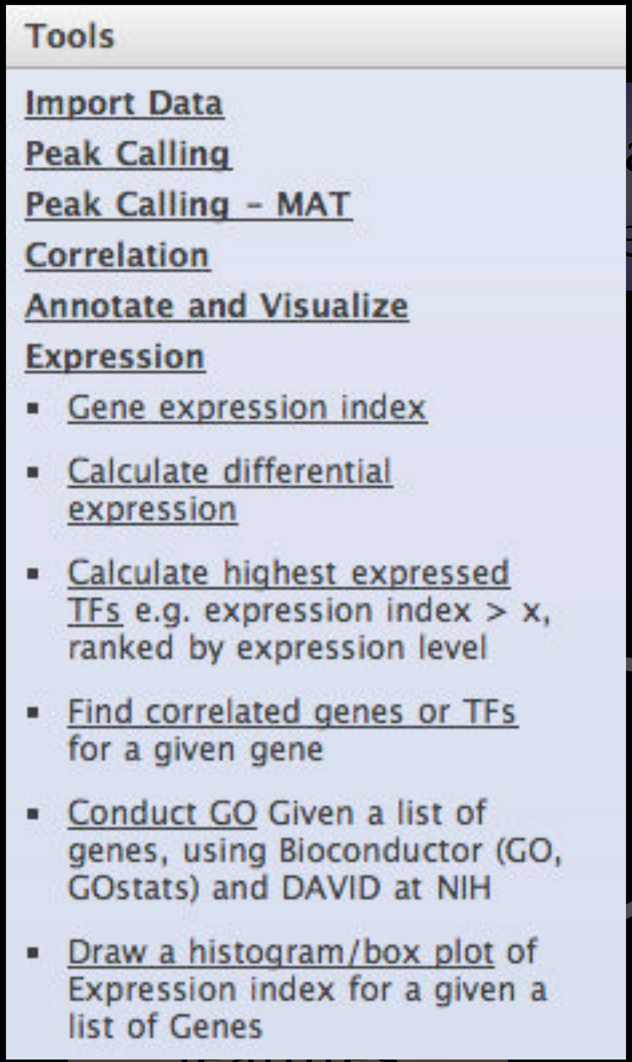
- Search the sequence motifs around peak center positions
- Search in known motif databases
- Perform de novo motif discovery based on MDscan algorithm
- Cluster similar motifs



id:M00481
factor:AR
consensus:1G1AC5G2C1
hits:6774
cutoff:3.2452
zscore:-53.0853
-10*log(pval):1e-30
position:-0.1362

id	factor	consensus	hits	cutoff	zscore	-10*log(pval)
M00481	AR	1G1AC5G2C1	6774	3.2452	-53.0853	1e-30
M00481_observed	AR	1G1AC5G2C1	6774	3.2452	-53.0853	1e-30
M00921_observed	GR	WCTGTYCT	26679	3.81	-52.0571	1e-30
M00921	GR	WSTGTYCT	26679	3.81	-52.0571	1e-30
M00724	HNF3alpha	HWRARYAAAYA	10842	4.785	-49.2657	1e-30
M00724_observed	HNF3alpha	HWRWRYAAAYA	10842	4.785	-49.2657	1e-30
M00956	AR	6AGAAC16	5159	3.0003	-46.6236	1e-30
M00956_observed	AR	6AG1ACA15	5159	3.0003	-46.6236	1e-30
M00955_observed	GR	6AG1ACA15	9458	3.0015	-45.8589	1e-30
M00955	GR	6AGAAC16	9458	3.0015	-45.8589	1e-30

Cistrome Work flow



Tools

Import Data

Peak Calling

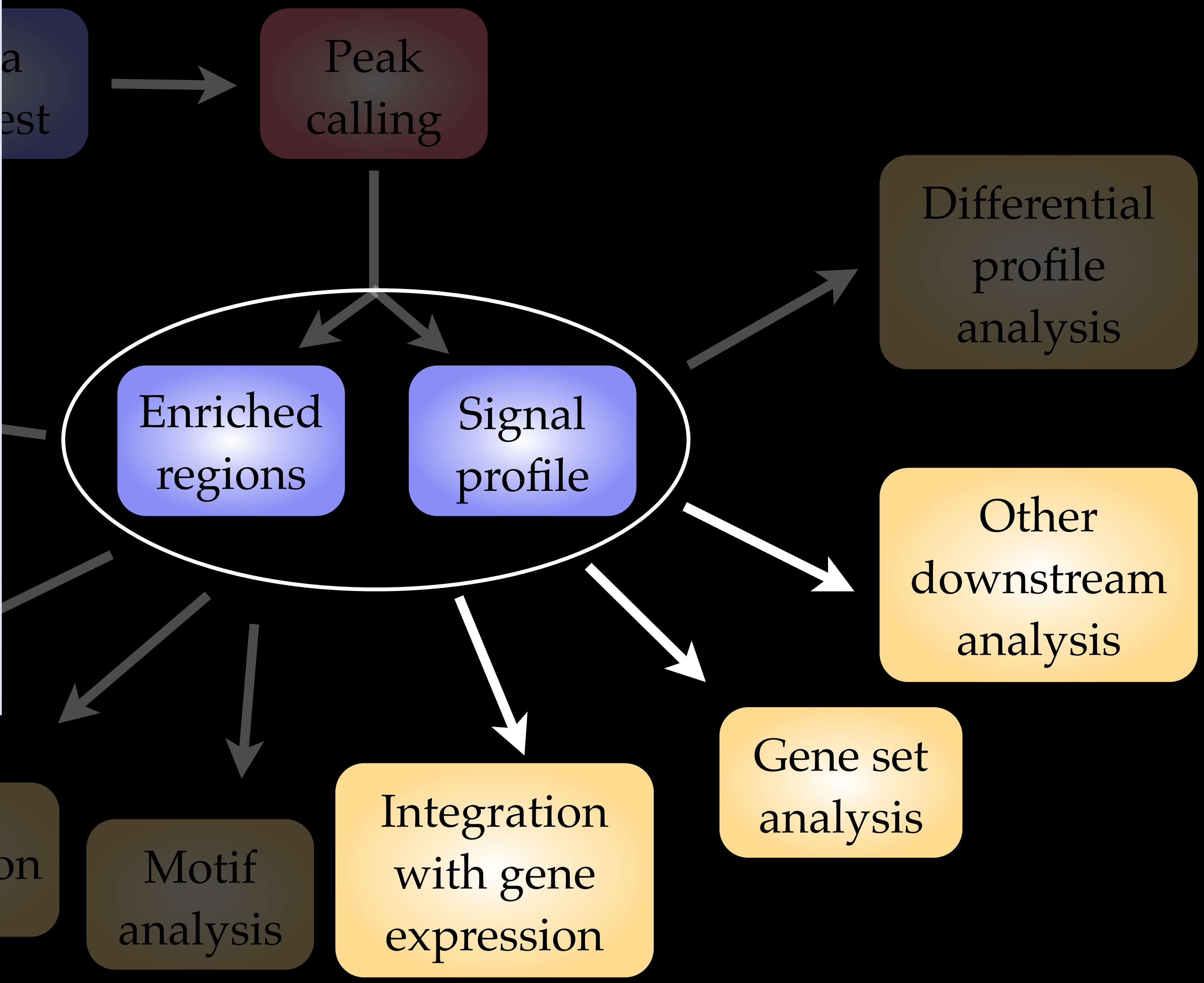
Peak Calling - MAT

Correlation

Annotate and Visualize

Expression

- Gene expression index
- Calculate differential expression
- Calculate highest expressed TFs e.g. expression index > x, ranked by expression level
- Find correlated genes or TFs for a given gene
- Conduct GO Given a list of genes, using Bioconductor (GO, GOstats) and DAVID at NIH
- Draw a histogram/box plot of Expression index for a given a list of Genes



a
est

Peak calling

Enriched regions

Signal profile

Differential profile analysis

Other downstream analysis

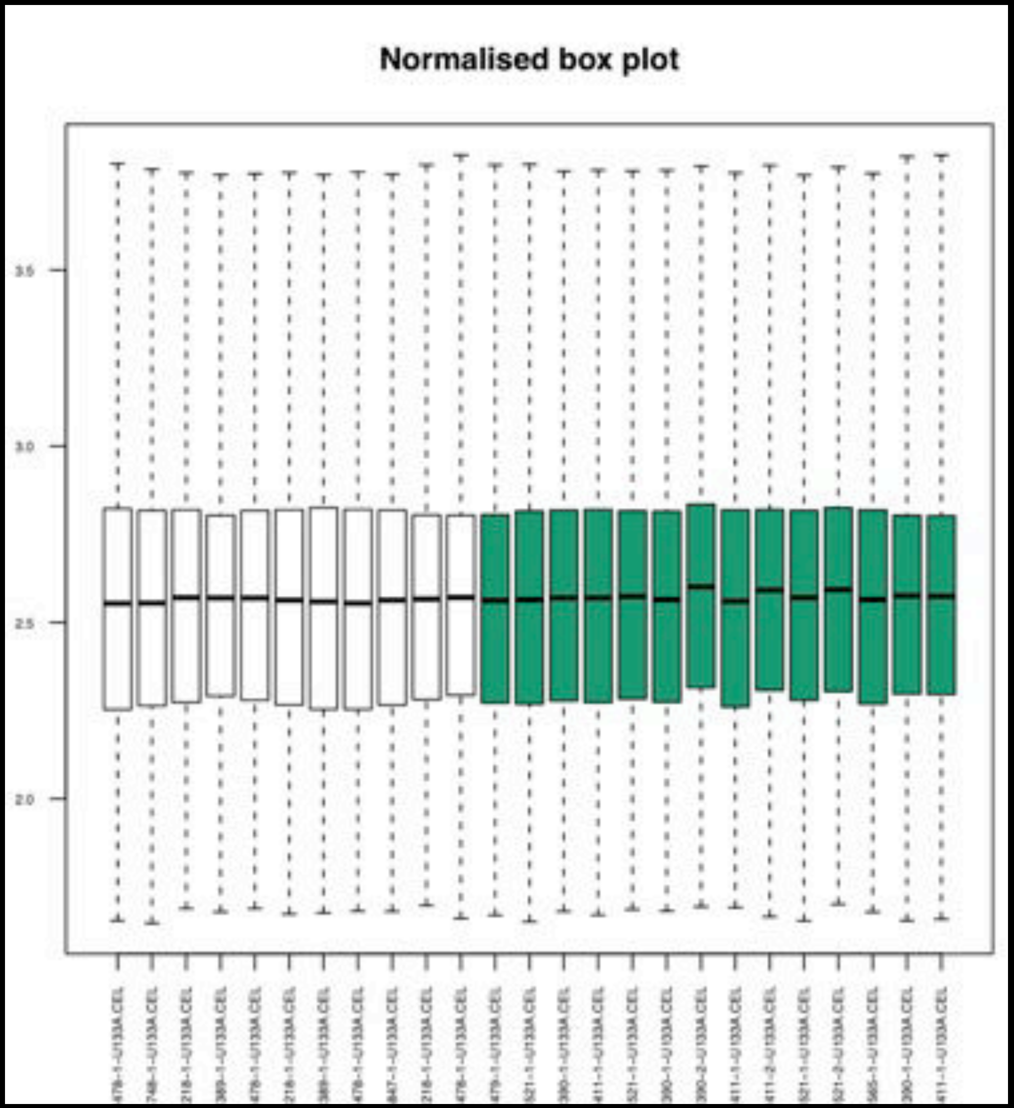
Gene set analysis

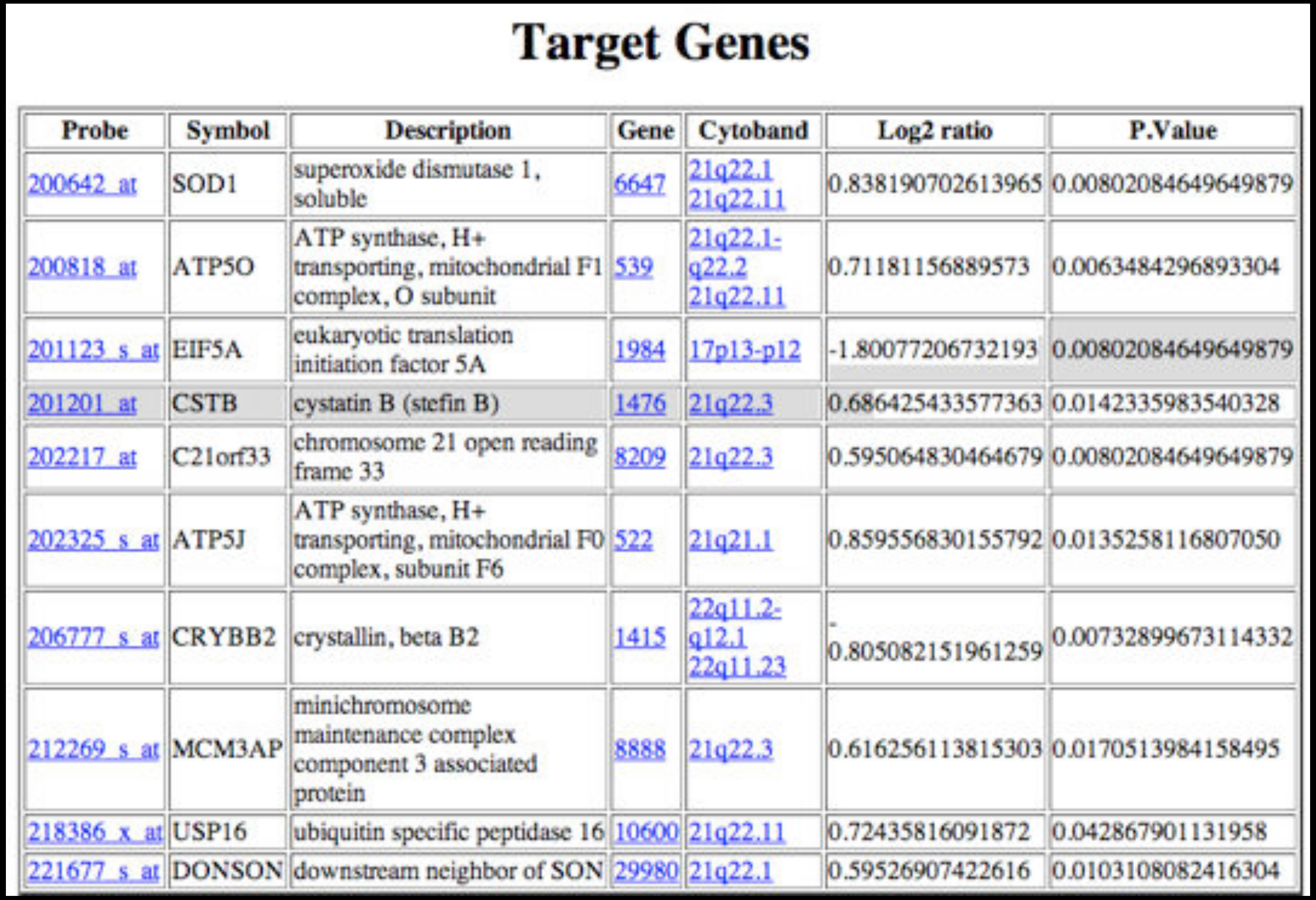
Integration with gene expression

Motif analysis

Correlation

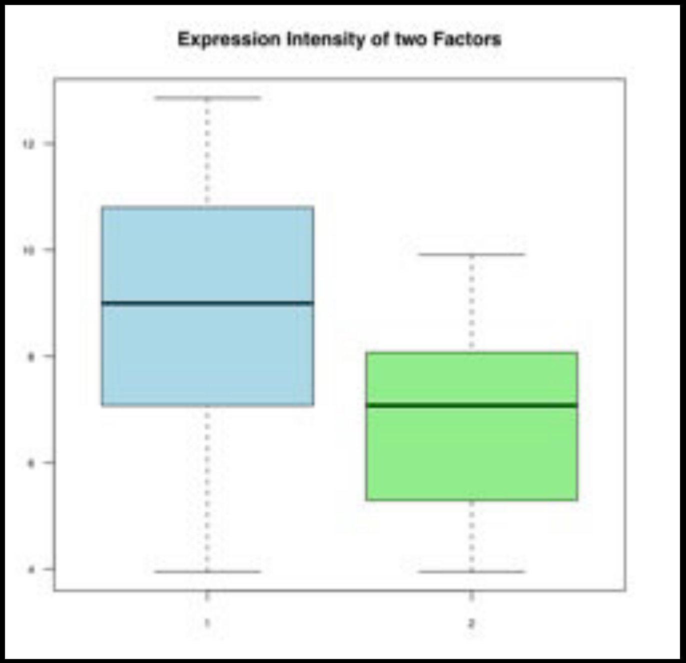
features



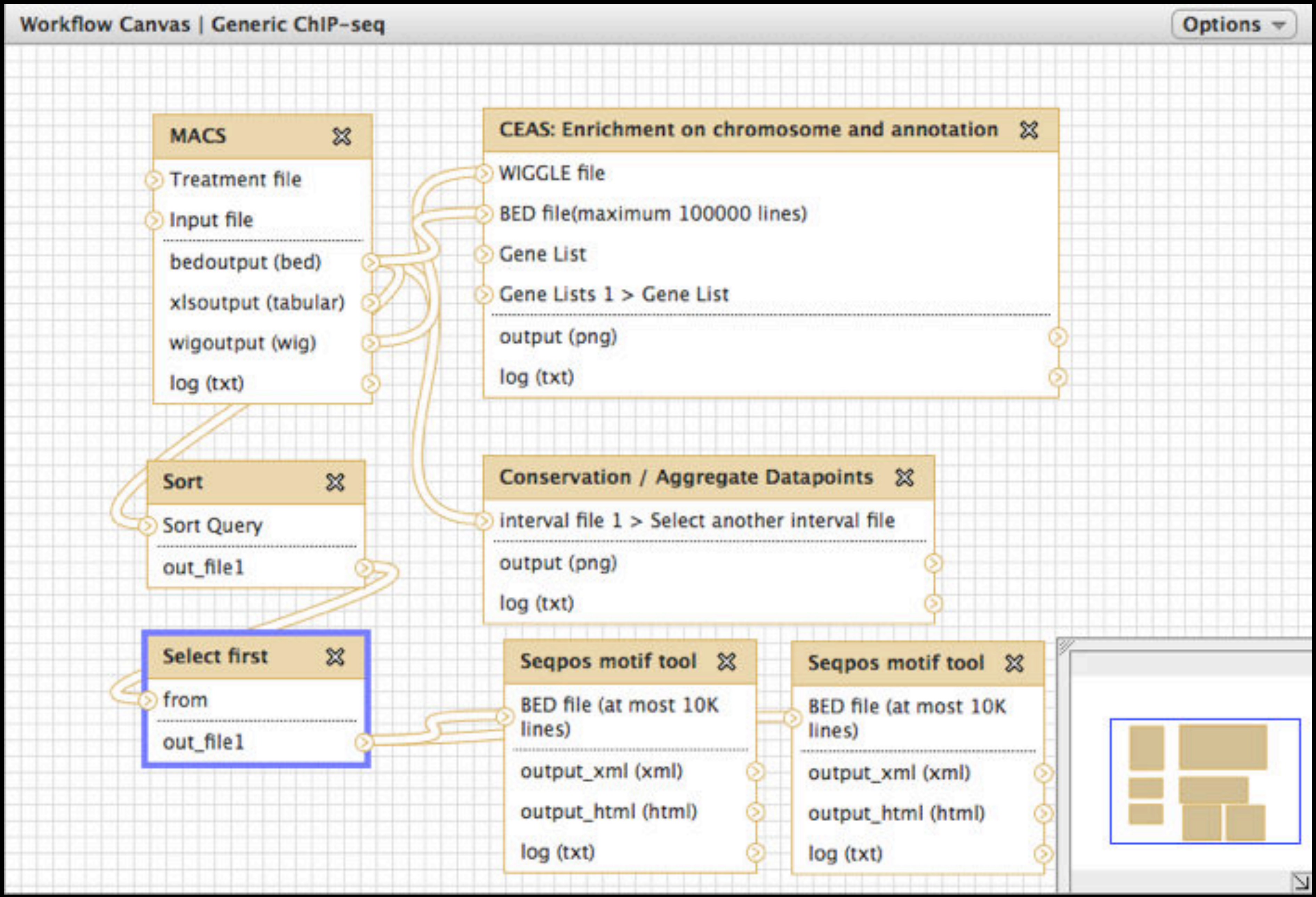


Target Genes

Probe	Symbol	Description	Gene	Cytoband	Log2 ratio	P.Value
200642_at	SOD1	superoxide dismutase 1, soluble	6647	21q22.1 21q22.11	0.838190702613965	0.00802084649649879
200818_at	ATP5O	ATP synthase, H+ transporting, mitochondrial F1 complex, O subunit	539	21q22.1-q22.2 21q22.11	0.71181156889573	0.0063484296893304
201123_s_at	EIF5A	eukaryotic translation initiation factor 5A	1984	17p13-p12	-1.80077206732193	0.00802084649649879
201201_at	CSTB	cystatin B (stefin B)	1476	21q22.3	0.686425433577363	0.0142335983540328
202217_at	C21orf33	chromosome 21 open reading frame 33	8209	21q22.3	0.595064830464679	0.00802084649649879
202325_s_at	ATP5J	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit F6	522	21q21.1	0.859556830155792	0.0135258116807050
206777_s_at	CRYBB2	crystallin, beta B2	1415	22q11.2-q12.1 22q11.23	-0.805082151961259	0.00732899673114332
212269_s_at	MCM3AP	minichromosome maintenance complex component 3 associated protein	8888	21q22.3	0.616256113815303	0.0170513984158495
218386_x_at	USP16	ubiquitin specific peptidase 16	10600	21q22.11	0.72435816091872	0.042867901131958
221677_s_at	DONSON	downstream neighbor of SON	29980	21q22.1	0.59526907422616	0.0103108082416304



Easy button


Workflow Canvas | Generic ChIP-seq
Options
MACS
Treatment file
Input file
bedoutput (bed)
xlsoutput (tabular)
wigoutput (wig)
log (txt)
CEAS: Enrichment on chromosome and annotation
WIGGLE file
BED file(maximum 100000 lines)
Gene List
Gene Lists 1 > Gene List
output (png)
log (txt)
Sort
Sort Query
out_file1
Conservation / Aggregate Datapoints
interval file 1 > Select another interval file
output (png)
log (txt)
Select first
from
out_file1
Seqpos motif tool
BED file (at most 10K lines)
output_xml (xml)
output_html (html)
log (txt)
Seqpos motif tool
BED file (at most 10K lines)
output_xml (xml)
output_html (html)
log (txt)

Future work

- More public ChIP-chip/seq datasets, precompiled in Cistrome DC.
- Correlation map of all avaiable data in DC.
- Integration. e.g. Galaxy->DC
- Follow Galaxy.

Acknowledgement

- Cistrome Authors
  - DFCI Team:
    - X. Shirley Liu, Myles Brown, Tao Liu, Cliff Meyer, Len Taing, Jian Ma, Meng Zhou, Ying Lei, Hyunjin Gene Shin, Yong Zhang
  - Eli Lilly Team:
    - Yaron Turpaz, Jorge Cesar Andrade Ortiz, Teck Kwong Bernett Lee, Swee Seong Wong, Kar Joon Chew, Kevin Yeung Chung Pak, Michael Poidinger, Patrick Hartman
- Acknowledgement:
  - Yi Wang, Scott Taing, Jacqueline Wentz , Lingling Shen, Josiah Altschuler, Zeynep Coban, Wenbo Wang, Mayako Michino, Meng Zhou, Jeremy Wu, Hansen He, Yiwen Chen, Utz Pape, Bo Jiang, Xiangfeng Wang
  - Edmond Joseph Breen, Shan Ho Tan
  - Wei Li (Baylor), Jun Song (UCSF), Mathieu Lupien (Dartmouth)
  - Jason Lieb (UNC), Susan Strome (UCSC), Julie Ahringer (Cambridge), Mitchelle Lazar (UPENN), Christopher Glass (UCSD)