bioinformatics.ca

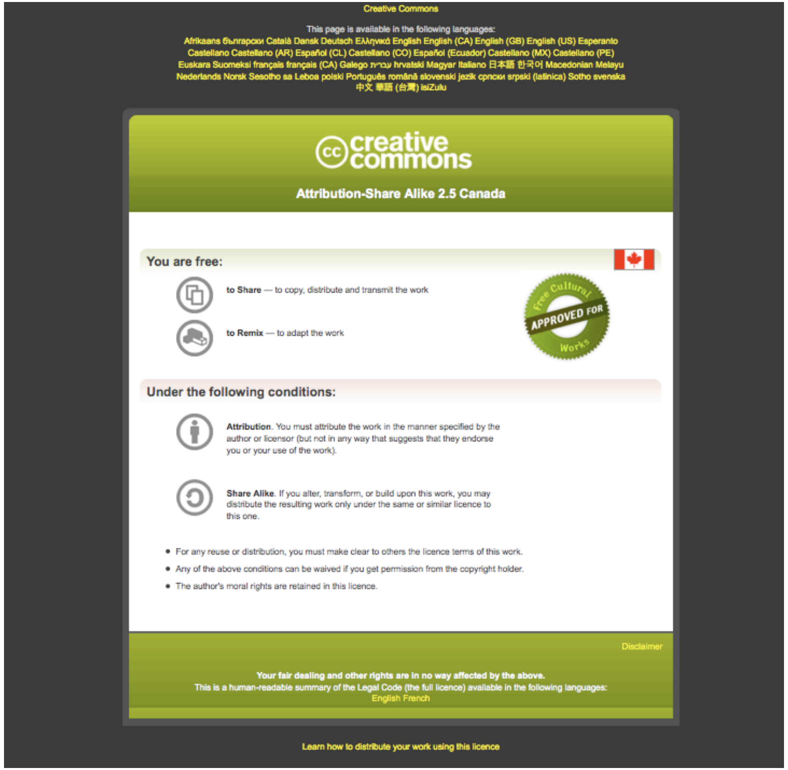
Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io

Supported by

McGill UNIVERSITY



Creative Commons

This page is available in the following languages:
Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska
中文 華語 (台灣) isiZulu

creative commons

Attribution-Share Alike 2.5 Canada

You are free:

to Share — to copy, distribute and transmit the work

to Remix — to adapt the work

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

Disclaimer

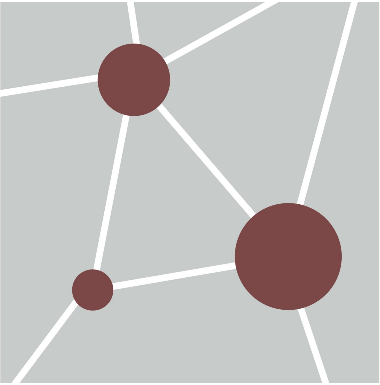
Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French

Learn how to distribute your work using this licence

Module 5 Practical Lab : Pathway Analysis of ChIP_seq Data

bioinformatics.ca

Veronique Voisin
Pathway and Network Analysis of –omics Data
May, 10-12, 2021



Donnelly Centre
for Cellular + Biomolecular Research
UNIVERSITY OF TORONTO

BADER LAB
UNIVERSITY OF TORONTO

Learning Objectives

By the end of this practical lab, you will be able to:

- **Perform pathway analysis of chIP-seq data**
- **Run MEME-chip to detect transcription factor enrichment**

We are going to use the following tools: **GREAT**, **Cytoscape/EnrichmentMap**, **MEME-chIP** and **Cytoscape/ iRegulon** and we will see in examples on how to integrate the analysis of both chIP-seq and RNA-seq data.

Some Tools Available to Analyze ChIP-seq Data

1) GREAT

chIP-seq bed file as input

output →

Pathway analysis results visualized as Cytoscape/EnrichmentMap

2) The MEME Suite Motif-based sequence analysis tools

Transcription factor (TF) motif enrichment in chip-seq peaks

Cytoscape

3) iRegulon

Transcription factor (TF) motif enrichment in gene list (RNAseq)

OR

Find the targets of a transcription factor of interest

ChIP_seq Process

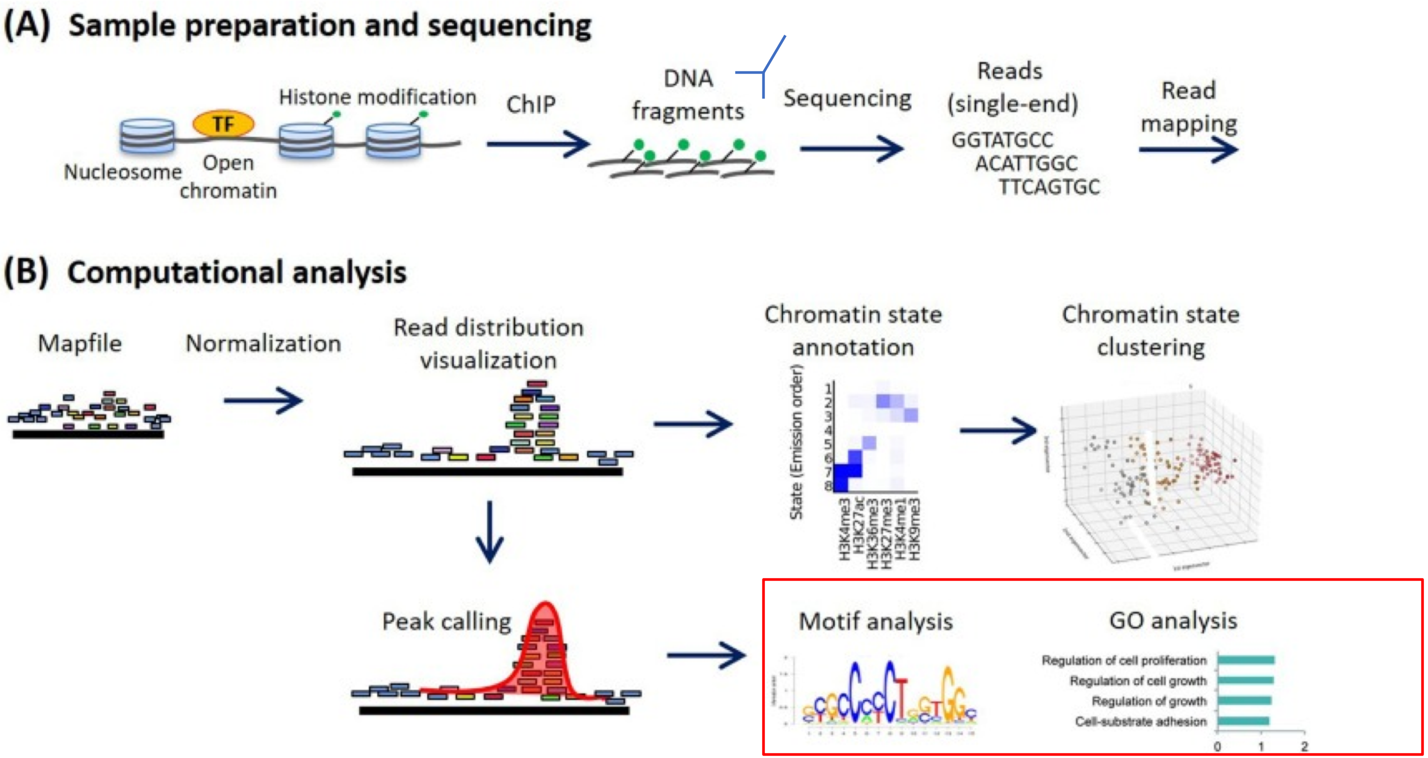
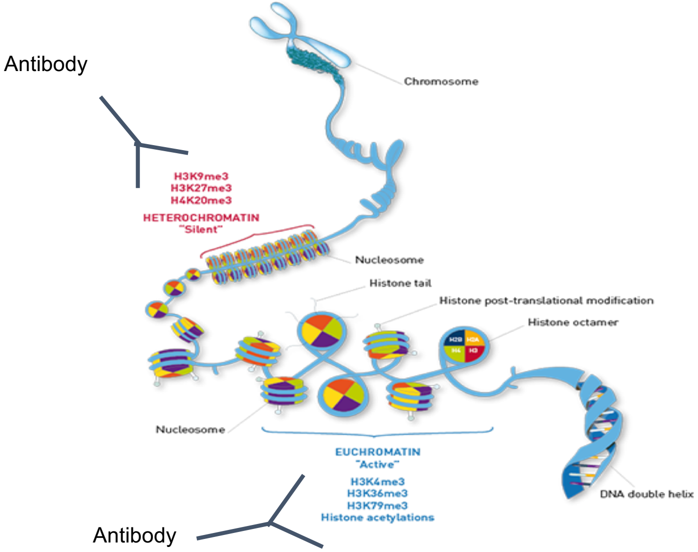


Image from : https://www.sciencedirect.com/science/article/pii/S1046202320300591

Different Types Of chIP-seq For Which Pathway Analysis May Be Applied

- chIP-seq to **detect histone acetylation** or **histone methylation**



CUT&RUN :

○ **alternative technique**

○ **works for low cell number**

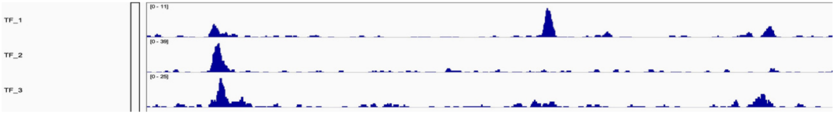
○ **same analysis pipeline as chIP-seq for pathway analysis**

Information To Know Before Starting The Analysis

Narrow or broad peak files?

Transcription Factors:

- narrow peak fikes



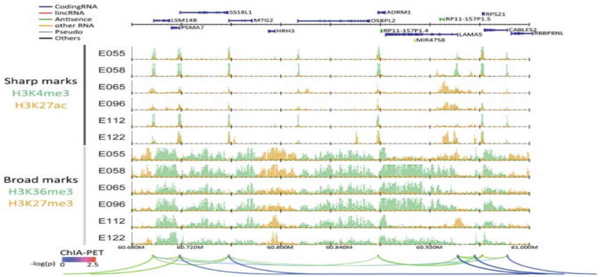
Histone acetylation and methylation:

Narrow Peak file

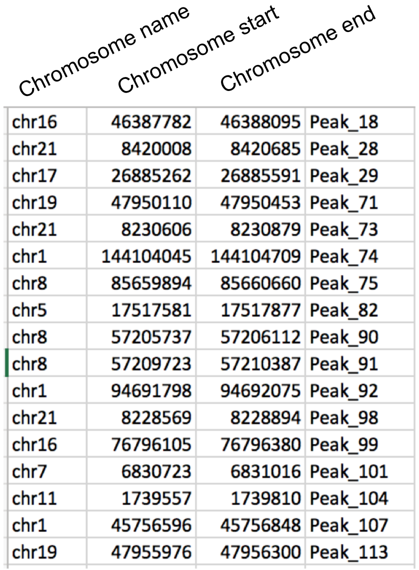
H3K4me3
H3K27ac

Broad Peak file

H3K36me3
H3K27me3



Format of a Bed file



Chromosome name	Chromosome start	Chromosome end	
chr16	46387782	46388095	Peak_18
chr21	8420008	8420685	Peak_28
chr17	26885262	26885591	Peak_29
chr19	47950110	47950453	Peak_71
chr21	8230606	8230879	Peak_73
chr1	144104045	144104709	Peak_74
chr8	85659894	85660660	Peak_75
chr5	17517581	17517877	Peak_82
chr8	57205737	57206112	Peak_90
chr8	57209723	57210387	Peak_91
chr1	94691798	94692075	Peak_92
chr21	8228569	8228894	Peak_98
chr16	76796105	76796380	Peak_99
chr7	6830723	6831016	Peak_101
chr11	1739557	1739810	Peak_104
chr1	45756596	45756848	Peak_107
chr19	47955976	47956300	Peak_113

Genome version

we need to know which genome version was used to align reads:

Human hg18
Human hg19
Human hg38
Mouse mm9
Mouse mm10

Note: increasing number of **biological replicates** increases the specificity of the signal.

How to Select The Peaks For The Pathway Analysis ?

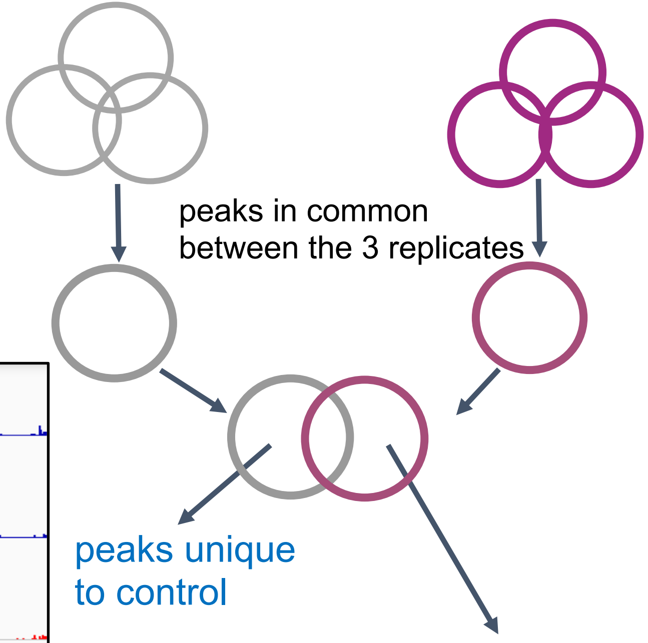
BEDtools Combine Peak Calls

MACS2 bed file

Compare Peak Calls ***DiffBind***

control

treated

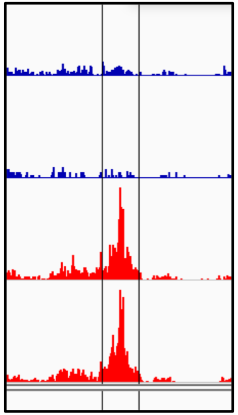


peaks in common
between the 3 replicates

peaks unique
to control

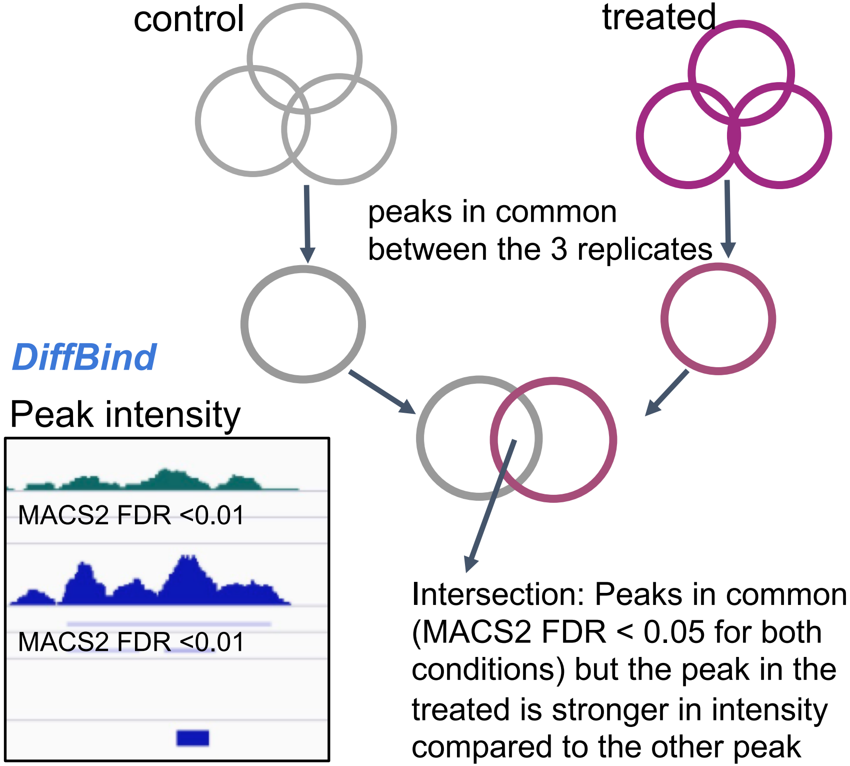
peaks unique
to treated

Peak
presence



control

treated



peaks in common
between the 3 replicates

DiffBind

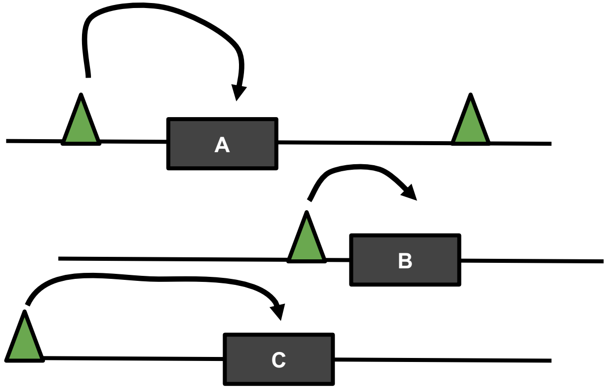
Peak intensity

MACS2 FDR <0.01

MACS2 FDR <0.01

Intersection: Peaks in common
(MACS2 FDR < 0.05 for both
conditions) but the peak in the
treated is stronger in intensity
compared to the other peak
(Diffbind FDR <0.05)

How To Perform Pathway Analysis On ChIPseq Data?



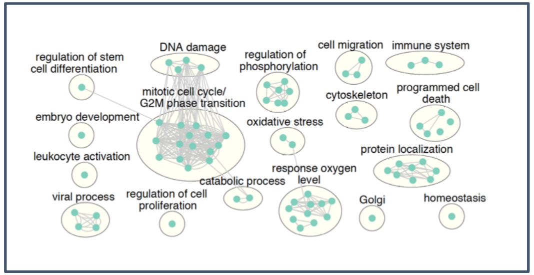
Peaks
(bed file)

↓

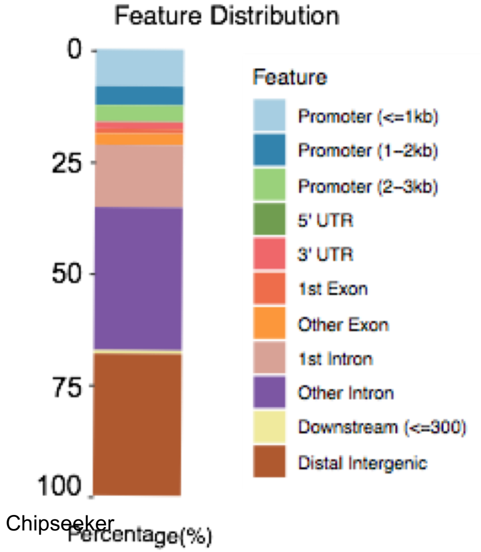
genes

↓

pathways

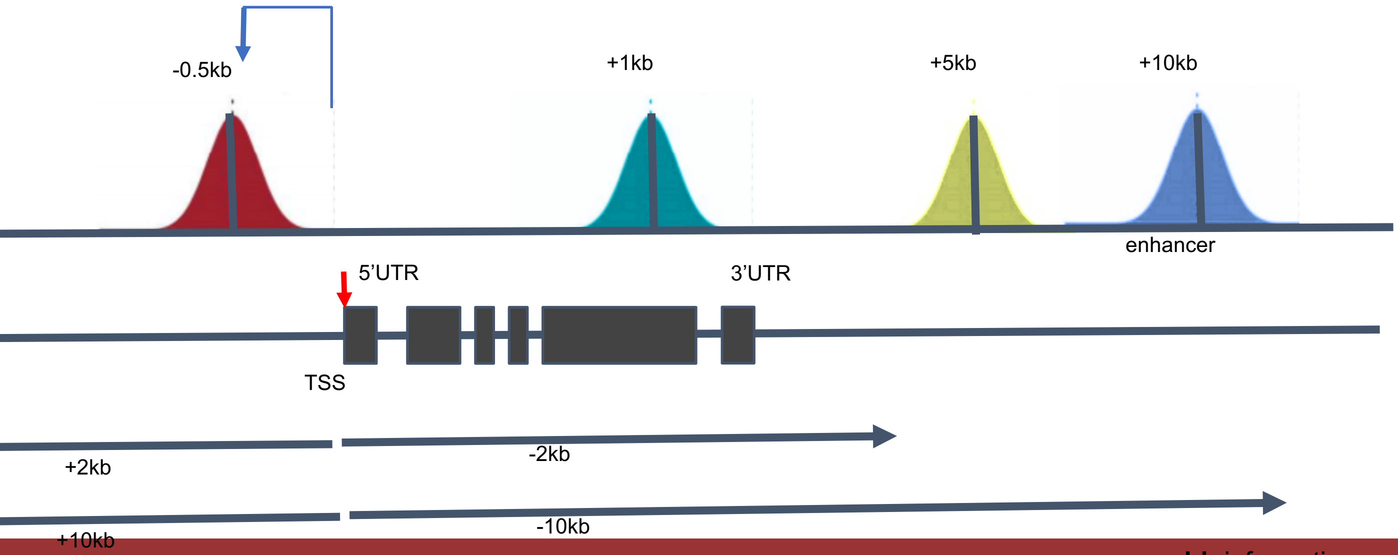


From Peaks to Genes… and then to Pathways



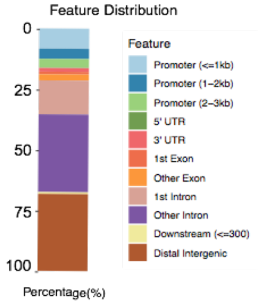
Chipseeker

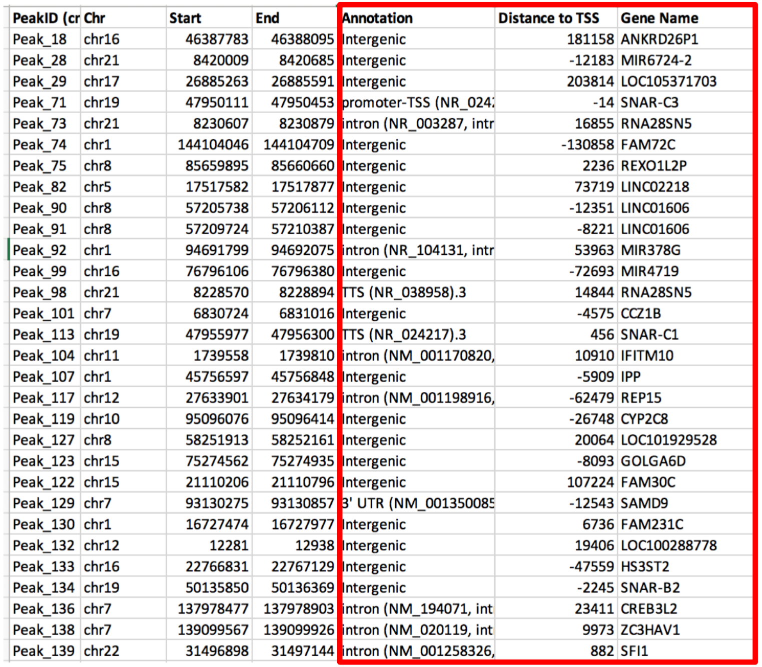
- Feature distribution: promoter, exonic, intronic, intergenic.
- Pathway analysis can be done only if we associate peaks to genes
- Rules are usually defined depending on the distance starting from the TSS (transcription start site of genes) to the middle/summit of peaks
- Proximal rule
- Distal rule
- How to choose a rule?



How to annotate chIP-seq peaks for pathway analysis?

ChIPseeker (R package)





PeakID (cr	Chr	Start	End	Annotation	Distance to TSS	Gene Name
Peak_18	chr16	46387783	46388095	Intergenic	181158	ANKRD26P1
Peak_28	chr21	8420009	8420685	Intergenic	-12183	MIR6724-2
Peak_29	chr17	26885263	26885591	Intergenic	203814	LOC105371703
Peak_71	chr19	47950111	47950453	promoter-TSS (NR_024:	-14	SNAR-C3
Peak_73	chr21	8230607	8230879	intron (NR_003287, intr	16855	RNA28SN5
Peak_74	chr1	144104046	144104709	Intergenic	-130858	FAM72C
Peak_75	chr8	85659895	85660660	Intergenic	2236	REXO1L2P
Peak_82	chr5	17517582	17517877	Intergenic	73719	LINC02218
Peak_90	chr8	57205738	57206112	Intergenic	-12351	LINC01606
Peak_91	chr8	57209724	57210387	Intergenic	-8221	LINC01606
Peak_92	chr1	94691799	94692075	intron (NR_104131, intr	53963	MIR378G
Peak_99	chr16	76796106	76796380	Intergenic	-72693	MIR4719
Peak_98	chr21	8228570	8228894	TTS (NR_038958).3	14844	RNA28SN5
Peak_101	chr7	6830724	6831016	Intergenic	-4575	CCZ1B
Peak_113	chr19	47955977	47956300	TTS (NR_024217).3	456	SNAR-C1
Peak_104	chr11	1739558	1739810	intron (NM_001170820,	10910	IFITM10
Peak_107	chr1	45756597	45756848	Intergenic	-5909	IPP
Peak_117	chr12	27633901	27634179	intron (NM_001198916,	-62479	REP15
Peak_119	chr10	95096076	95096414	Intergenic	-26748	CYP2C8
Peak_127	chr8	58251913	58252161	Intergenic	20064	LOC101929528
Peak_123	chr15	75274562	75274935	Intergenic	-8093	GOLGA6D
Peak_122	chr15	21110206	21110796	Intergenic	107224	FAM30C
Peak_129	chr7	93130275	93130857	3' UTR (NM_00135008!	-12543	SAMD9
Peak_130	chr1	16727474	16727977	Intergenic	6736	FAM231C
Peak_132	chr12	12281	12938	Intergenic	19406	LOC100288778
Peak_133	chr16	22766831	22767129	Intergenic	-47559	HS3ST2
Peak_134	chr19	50135850	50136369	Intergenic	-2245	SNAR-B2
Peak_136	chr7	137978477	137978903	intron (NM_194071, int	23411	CREB3L2
Peak_138	chr7	139099567	139099926	intron (NM_020119, int	9973	ZC3HAV1
Peak_139	chr22	31496898	31497144	intron (NM_001258326,	882	SFI1

Options:

- select peaks closer to TSS for a proximal analysis.
- Select promoter region only
- Or keep all

Remove duplicate gene names

Galaxy Training!

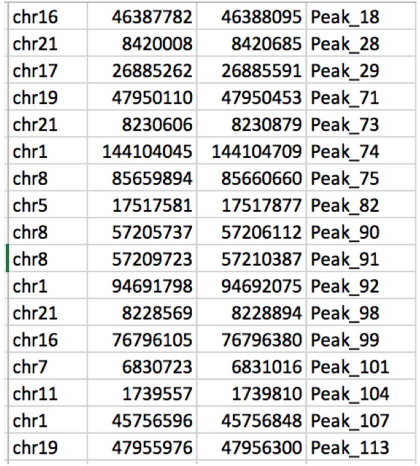
Use of Galaxy to annotate your peaks? (not tested)

https://training.galaxyproject.org/training-material/topics/introduction/tutorials/galaxy-intro-peaks2genes/tutorial.html

GREAT Predicts Functions Of Cis-regulatory Regions.

http://great.stanford.edu/public/html/

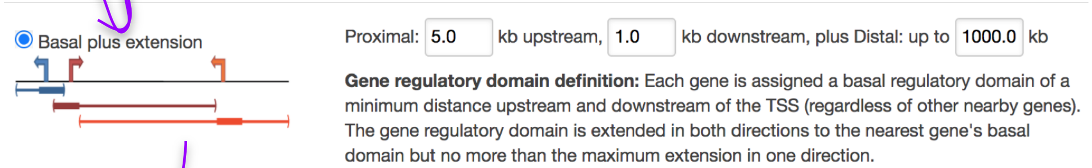
Input: bed file (selected peaks)



chr16	46387782	46388095	Peak_18
chr21	8420008	8420685	Peak_28
chr17	26885262	26885591	Peak_29
chr19	47950110	47950453	Peak_71
chr21	8230606	8230879	Peak_73
chr1	144104045	144104709	Peak_74
chr8	85659894	85660660	Peak_75
chr5	17517581	17517877	Peak_82
chr8	57205737	57206112	Peak_90
chr8	57209723	57210387	Peak_91
chr1	94691798	94692075	Peak_92
chr21	8228569	8228894	Peak_98
chr16	76796105	76796380	Peak_99
chr7	6830723	6831016	Peak_101
chr11	1739557	1739810	Peak_104
chr1	45756596	45756848	Peak_107
chr19	47955976	47956300	Peak_113

Peak file (# of peaks) can be larged

Step1: Find genes near peaks: define the rule

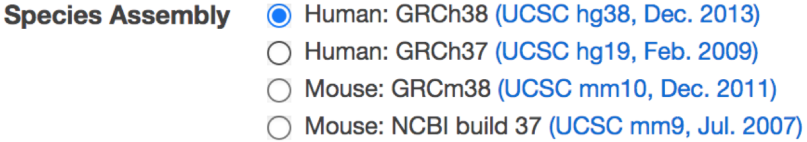


● Basal plus extension

Proximal: 5.0 kb upstream, 1.0 kb downstream, plus Distal: up to 1000.0 kb

Gene regulatory domain definition: Each gene is assigned a basal regulatory domain of a minimum distance upstream and downstream of the TSS (regardless of other nearby genes). The gene regulatory domain is extended in both directions to the nearest gene's basal domain but no more than the maximum extension in one direction.

Step2: Pathway enrichment analysis



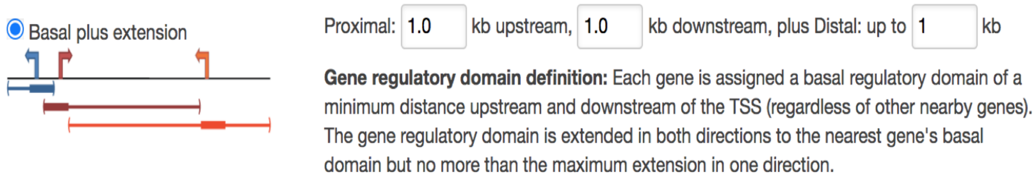
Species Assembly

● Human: GRCh38 (UCSC hg38, Dec. 2013)
○ Human: GRCh37 (UCSC hg19, Feb. 2009)
○ Mouse: GRCm38 (UCSC mm10, Dec. 2011)
○ Mouse: NCBI build 37 (UCSC mm9, Jul. 2007)

Tip: If you have genomic regions defined for a different species or assembly from the ones we currently support, you can use the UCSC LiftOver utility to convert to a supported assembly

Rules To Associate Peaks And Genes

PROXIMAL RULE



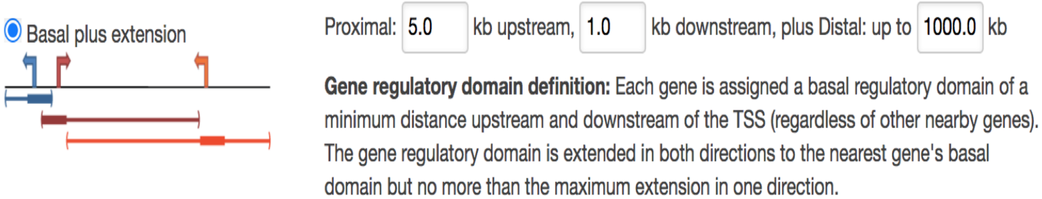
● Basal plus extension

Proximal: 1.0 kb upstream, 1.0 kb downstream, plus Distal: up to 1 kb

Gene regulatory domain definition: Each gene is assigned a basal regulatory domain of a minimum distance upstream and downstream of the TSS (regardless of other nearby genes). The gene regulatory domain is extended in both directions to the nearest gene's basal domain but no more than the maximum extension in one direction.

- **Proximal rules** reduce the problem to a size of a gene list (count how many genes with a peak is contained in a tested pathway). We can use any tools that are using a gene list and we can use the **Fisher's exact test**.
- But associating only proximal peaks loses a lot of information.

DISTAL RULE



● Basal plus extension

Proximal: 5.0 kb upstream, 1.0 kb downstream, plus Distal: up to 1000.0 kb

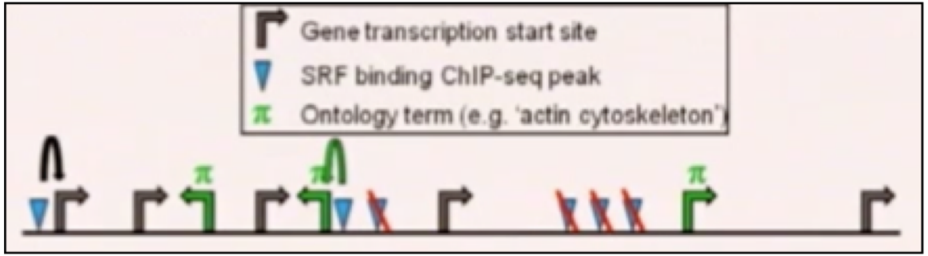
Gene regulatory domain definition: Each gene is assigned a basal regulatory domain of a minimum distance upstream and downstream of the TSS (regardless of other nearby genes). The gene regulatory domain is extended in both directions to the nearest gene's basal domain but no more than the maximum extension in one direction.

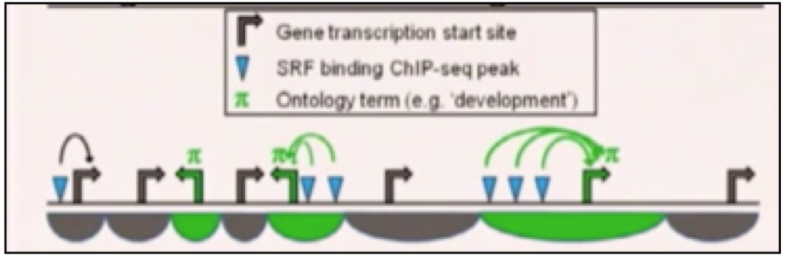
- **Associating distal peaks** to genes but applying the Fisher's exact test can lead to spurious enrichment results (it biases the results toward pathways enriched in genes located in the genome to desert regions like developmental pathways).
- The way GREAT is doing to correct for bias is: 1) define genomic regions that contains peaks associated with genes 2) for a tested pathway, count how many of the peaks land with the genomic regions associated with the tested pathway compared to genomics regions with peaks not associated with the tested pathway. It is using a **binomial test**.

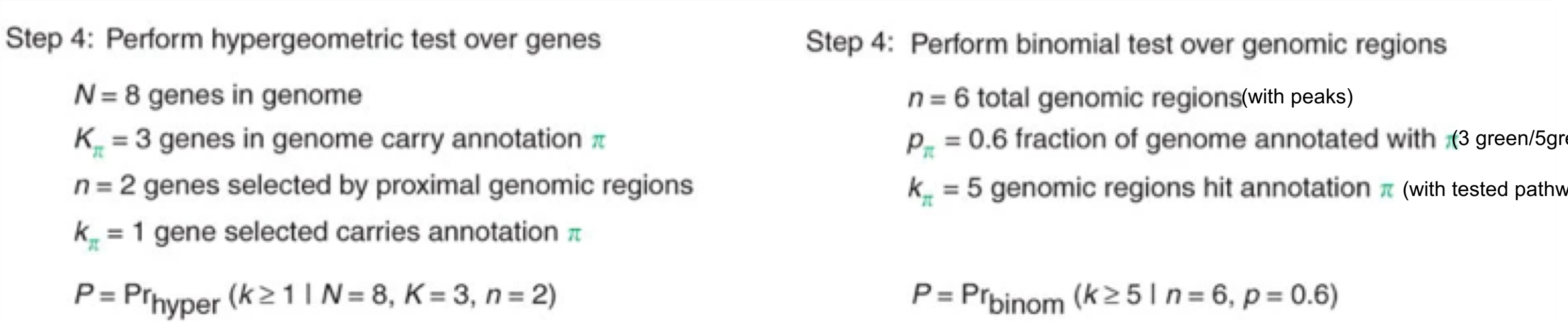
GREAT Statistics Fisher’s Exact Test Versus Binomial Test

Proximal: Hypergeometric test over genes

Distal: Binomial test over genomic regions


Gene transcription start site
SRF binding ChIP-seq peak
Ontology term (e.g. ‘actin cytoskeleton’)
π
π
π


Gene transcription start site
SRF binding ChIP-seq peak
Ontology term (e.g. ‘development’)
π
π
π



Step 4: Perform hypergeometric test over genes

$N = 8$ genes in genome

$K_\pi = 3$ genes in genome carry annotation π

$n = 2$ genes selected by proximal genomic regions

$k_\pi = 1$ gene selected carries annotation π

$P = \Pr_{\text{hyper}} (k \geq 1 \mid N = 8, K = 3, n = 2)$

Step 4: Perform binomial test over genomic regions

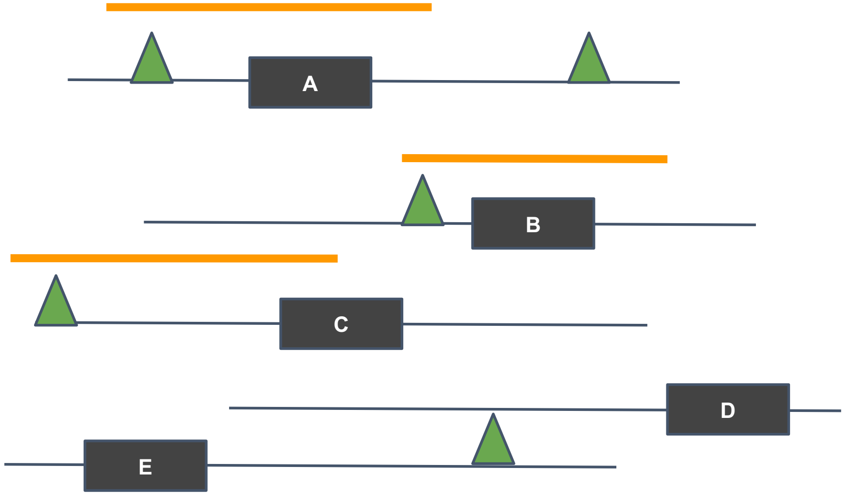
$n = 6$ total genomic regions(with peaks)

$p_\pi = 0.6$ fraction of genome annotated with π(3 green/5grey)

$k_\pi = 5$ genomic regions hit annotation π (with tested pathway)

$P = \Pr_{\text{binom}} (k \geq 5 \mid n = 6, p = 0.6)$

Example : Integration of chIPseq and ATACseq



Legend:

■ gene

▲ Chip seq peaks found in treated condition

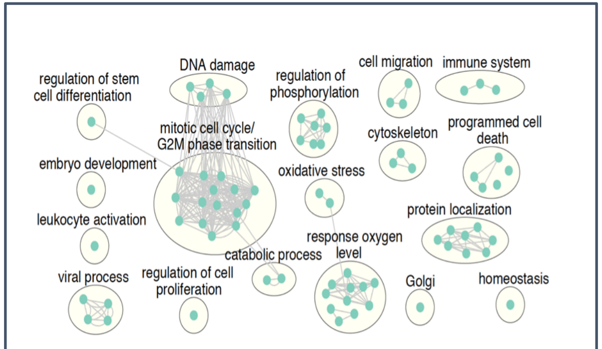
▬ Open chromatin region specific fortreated condition (ATAC-seq)

chIP-seq peaks ↓ GREAT ↓ Export of the GO BP table

ATAC-seq peaks ↓ GREAT ↓ Export of the GO BP table

↓

Cytoscape / EnrichmentMap

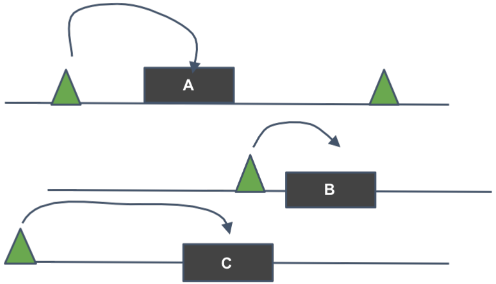


enrichment map FDR 0.05

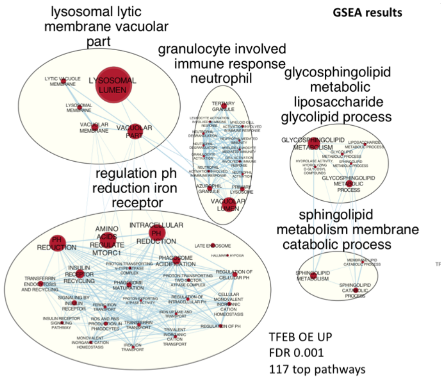
MEME-ChIP is a web-based tool for analyzing motifs in large sequence data sets. It can analyze peak regions identified by ChIP-seq

chip-seq data :

- overexpression of a specific transcription factor called TFEB

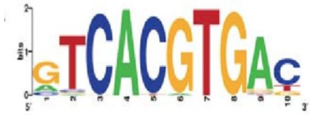


.bed file
GREAT/EnrichmentMap



TFEB OE UP
FDR 0.001
117 top pathways
Node size proportional to NES

MEME-chip:
find overenrichment of known DNA motif in chipseq sequences



TFEB is the first known motif found significantly enriched in If yes, we have proved that TFEB
Is binding and regulating the expression of the lysosomal genes in our model system.

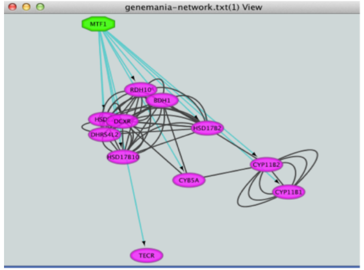
iRegulon: detects TF that co-regulate a gene list (RNAseq) ---> help us to link chIP-seq and RNA-seq results

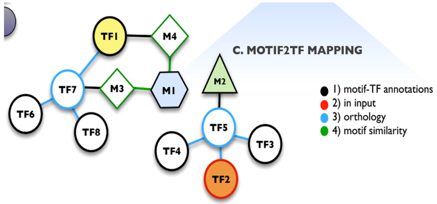
iREgulon (Cytoscape app, bulk RNAseq , gene list)

iRegulon detects the TF, the targets and the motifs/tracks from a set of genes.

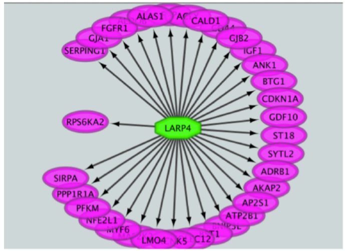
Look at **pySCENIC for single cell RNAseq!**

1.Find predicted transcription factor regulating genes in my gene list





2.Find predicted targets of a transcription factor of interest

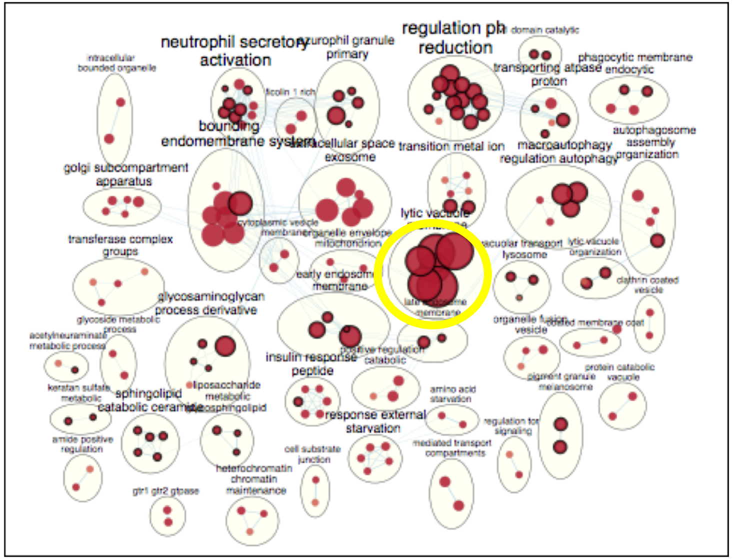


iREgulon: Find Predicted Transcription Factor Regulating an Enriched Pathway

RNAseq data :

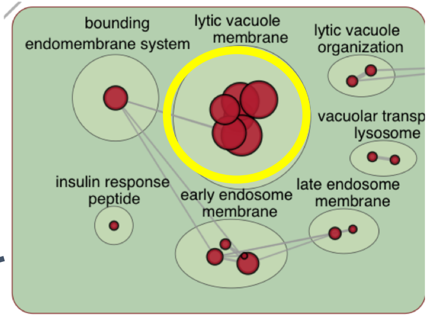
- overexpression of a specific transcription factor(TF) called TFEB
- upregulated genes are the TF targets + secondary events

RNASeq : GSEA + EnrichmentMap



neutrophil secretory activation
neutrophil granule primary
regulation ph reduction
domain catalytic
intracellular bounded organelle
ficolin 1 rich
transporting atpase proton
phagocytic membrane endocytic
bounding endomembrane system
extracellular space exosome
transition metal ion
macroautophagy regulation autophagy
autophagosome assembly organization
golgi subcompartment apparatus
cytoplasmic vesicle membrane
organelle envelope mitochondrion
lytic vacuole membrane
vacuolar transport lysosome
lytic vacuole organization
transferase complex groups
early endosome membrane
clathrin coated vesicle
glycosaminoglycan process derivative
late endosome membrane
glycoside metabolic process
acetylneuraminate metabolic process
organelle fusion vesicle
coated membrane coat
insulin response peptide
positive regulation catabolic
keratan sulfate metabolic
sphingolipid catabolic ceramide
liposaccharide metabolic sphingolipid
pigment granule melanosome
protein catabolic vacuole
amino acid starvation
response external starvation
regulation tor signaling
amide positive regulation
cell substrate junction
mediated transport compartments
heterochromatin chromatin maintenance
gtr1 gtr2 gtpase

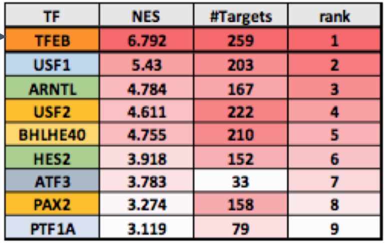
Genes up-regulated (FDR 0.05):
Gene list → imported as a network
in Cytoscape → iRegulon



LYSOSOME RELATED FUNCTIONS

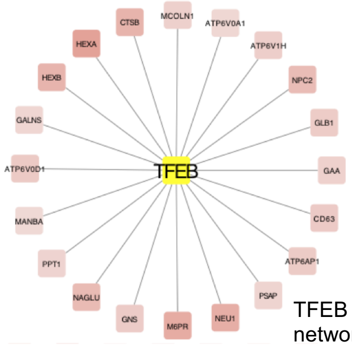
Gene list →
g:Profiler /
Enrichment Map

iRegulon results:



TF	NES	#Targets	rank
TFEB	6.792	259	1
USF1	5.43	203	2
ARNTL	4.784	167	3
USF2	4.611	222	4
BHLHE40	4.755	210	5
HES2	3.918	152	6
ATF3	3.783	33	7
PAX2	3.274	158	8
PTF1A	3.119	79	9

iRegulon: option 1 “Predict regulators and targets”



TFEB regulatory network

CBW Epigenomics Workshop: Learn How To Align Your Reads And Call The Peaks Using MACS2

Epigenomics Analysis

3 days: September 13 - September 15, 2021

Online

Module 1: Introduction to ChIP Sequencing and Analysis

Module 2: ChIP-Seq Alignment, Peak Calling, and Visualization

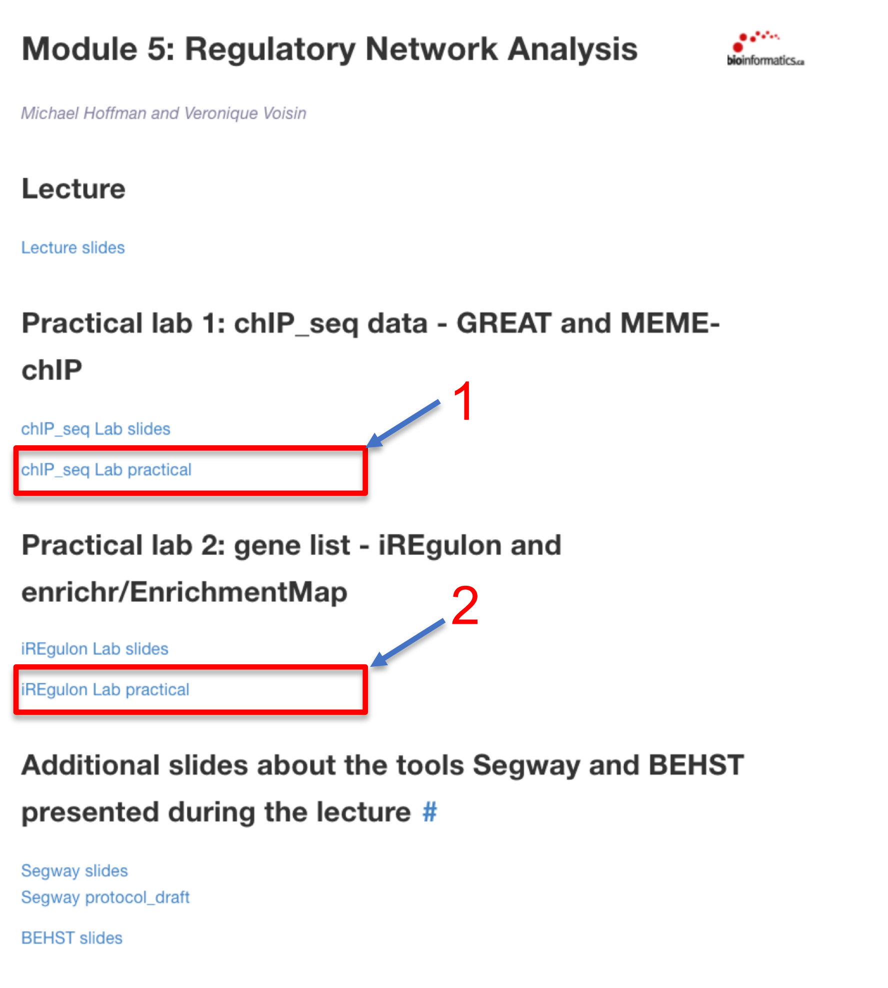
Module 3: Introduction to WGBS and Analysis

Module 4: Downstream Analysis and Integrative Tools

References

https://www.bioconductor.org/help/course-materials/2016/CSAMA/lab-5-chipseq/Epigenetics.html

https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html



Module 5: Regulatory Network Analysis

Michael Hoffman and Veronique Voisin

Lecture

Lecture slides

Practical lab 1: chIP_seq data - GREAT and MEME-chIP

chIP_seq Lab slides

chIP_seq Lab practical

Practical lab 2: gene list - iREgulon and enrichr/EnrichmentMap

iREgulon Lab slides

iREgulon Lab practical

Additional slides about the tools Segway and BEHST presented during the lecture

Segway slides
Segway protocol_draft

BEHST slides

We are on a Coffee Break & Networking Session

Workshop Sponsors:

Canadian Centre for Computational Genomics

HPC4Health

OICR Ontario Institute for Cancer Research

MiCM McGill initiative in Computational Medicine