ChIP-seq: Peak Calling (answers to questions)

Stéphanie Le Gras
(slegras@igbmc.fr)

Exercise: peak calling

- 1.
  - Search for "macs2 callpeak" in the search field (tool panel)
  - Click on the name of the tool
  - Set parameters:
    - ChIP-Seq Treatment File: mitf.bam
    - ChIP-Seq Control File: ctrl.bam
    - Effective genome size: Human
    - Outputs: select Peaks as tabular file, summits, Summary page (html), Plot in PDF
    - Click on 

Exercise: peak calling

- 2. Macs2 callpeak generates 5 datasets:
  - List of the peaks (tabular format)

List of arguments used to run Macs2

Peaks

	A	B	C	D	E	F	G	H	I	J
1	# This file is generated by MACS version 2.1.0.20151222									
2	# Command line: callpeak --name MACS2 -t /galaxy13/files/052/dataset_52866.dat -c /galaxy22/files/052/dataset_52865.dat --f									
3	# ARGUMENTS LIST:									
4	# name = MACS2									
5	# format = BAM									
6	# ChIP-seq file = ['/galaxy13/files/052/dataset_52866.dat']									
7	# control file = ['/galaxy22/files/052/dataset_52865.dat']									
8	# effective genome size = 2.45e+09									
9	# band width = 300									
10	# model fold = [5, 50]									
11	# qvalue cutoff = 5.00e-02									
12	# Larger dataset will be scaled towards smaller dataset.									
13	# Range for calculating regional lambda is: 1000 bps and 10000 bps									
14	# Broad region calling is off									
15	# tag size is determined as 54 bps									
16	# total tags in treatment: 23124393									
17	# tags after filtering in treatment: 6223075									
18	# maximum duplicate tags at the same position in treatment = 1									
19	# Redundant rate in treatment: 0.73									
20	# total tags in control: 19949607									
21	# tags after filtering in control: 4798380									
22	# maximum duplicate tags at the same position in control = 1									
23	# Redundant rate in control: 0.76									
24	# d = 75									
25	# alternative fragment length(s) may be 75 bps									
26	chr	start	end	length	abs_summit	pileup	-log10(pvalue)	fold_enrichment	-log10(qvalue)	name
27	chr1	980686	980816	131	980745	8.48	10.38277	7.29361	6.46786	MACS2_peak_1
28	chr1	983821	983925	105	983877	6.94	9.11038	6.77148	5.34984	MACS2_peak_2
29	chr1	1031344	1031475	132	1031406	6.17	6.82634	5.21345	3.25879	MACS2_peak_3
30	chr1	1079424	1079564	141	1079490	12.34	18.30659	10.88735	13.88358	MACS2_peak_4
31	chr1	1304817	1304958	142	1304891	13.11	20.10101	11.51679	15.56374	MACS2_peak_5

Exercise: peak calling

- 2. Macs2 callpeak generates 5 datasets:
  - List of the peaks (tabular format)

26	chr	start	end	length	abs_summit	pileup	-log10(pvalue)	fold_enrichment	-log10(qvalue)	name
27	chr1	980686	980816	131	980745	8.48	10.38277	7.29361	6.46786	MACS2_peak_1
28	chr1	983821	983925	105	983877	6.94	9.11038	6.77148	5.34984	MACS2_peak_2
29	chr1	1031344	1031475	132	1031406	6.17	6.82634	5.21345	3.25879	MACS2_peak_3
30	chr1	1079424	1079564	141	1079490	12.34	18.30659	10.88735	13.88358	MACS2_peak_4
31	chr1	1304817	1304958	142	1304891	13.11	20.10101	11.51679	15.56374	MACS2_peak_5

  - chr: chromosome name
  - start: start position of peak
  - end: end position of peak
  - length: length of peak region
  - abs_summit: absolute peak summit position
  - pileup: pileup height at peak summit
  - -log10(pvalue): -log10(pvalue) for the peak summit (e.g. pvalue =1e-10, then this value should be 10)
  - fold_enrichment: fold enrichment for this peak summit against random Poisson distribution with local lambda
  - -log10(qvalue): -log10(qvalue) at peak summit
  - name: peak name

Exercise: peak calling

- List of the peaks (Narrowpeak format)

1	2	3	4	5	6	7	8	9	10
chr1	980685	980816	MACS2_peak_1	64	.	7.29361	10.38277	6.46786	59
chr1	983820	983925	MACS2_peak_2	53	.	6.77148	9.11038	5.34984	56
chr1	1031343	1031475	MACS2_peak_3	32	.	5.21345	6.82634	3.25879	62
chr1	1079423	1079564	MACS2_peak_4	138	.	10.88735	18.30659	13.88358	66
chr1	1304816	1304958	MACS2_peak_5	155	.	11.51679	20.10101	15.56374	74
chr1	1441082	1441181	MACS2_peak_6	124	.	10.25923	16.71260	12.40068	71

1. chr
2. Start of peak
3. End of peak
4. Peak name
5. Integer score for display
7. fold-change
8. -log10pvalue
9. -log10qvalue
10. Relative summit position to peak start

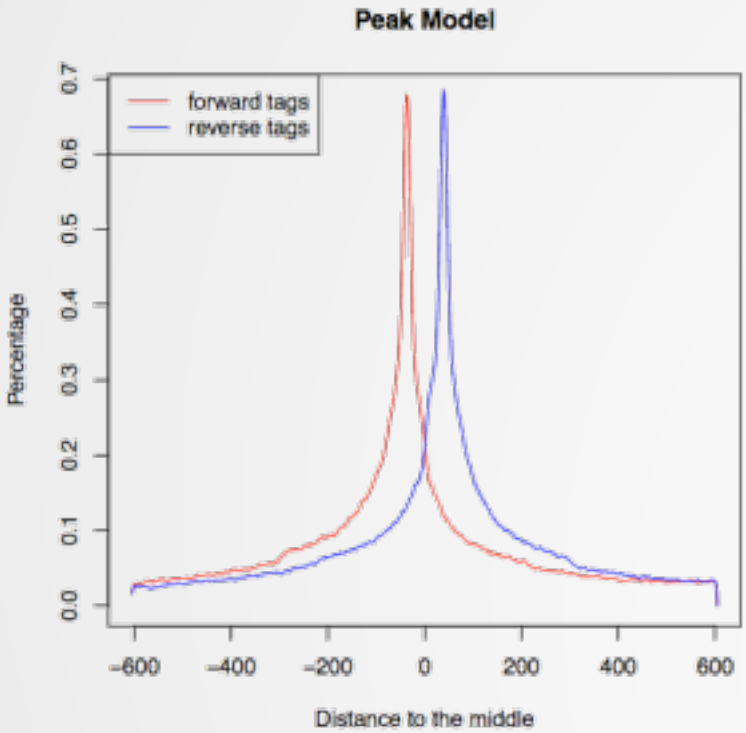
Exercise: peak calling

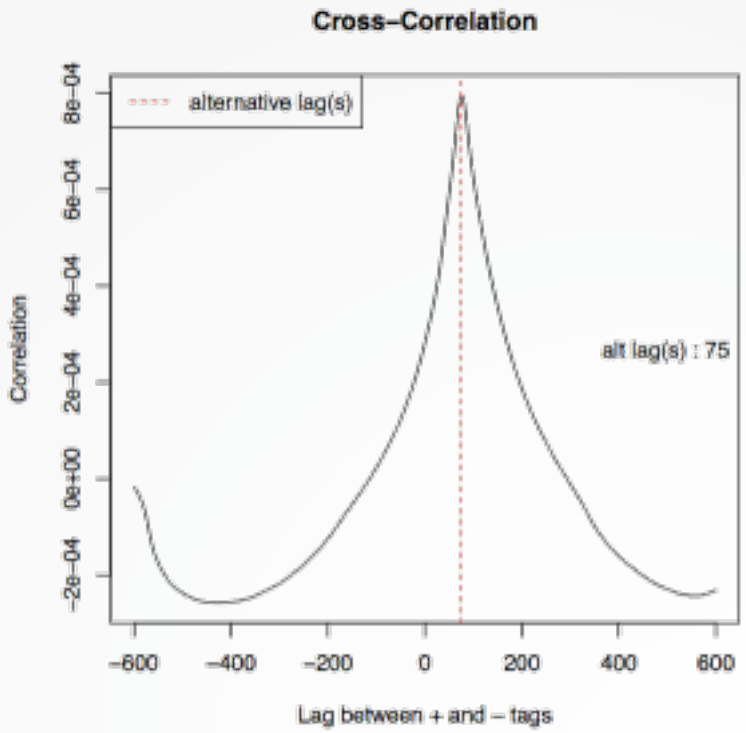
- List of the peak summits (BED): contains the peak summit location for each peak.

1. chr	2. Start of peak	3. End of peak	4. Peak name	5. -log10pvalue
1	2	3	4	5
chr1	980744	980745	MACS2_peak_1	6.46786
chr1	983876	983877	MACS2_peak_2	5.34984
chr1	1031405	1031406	MACS2_peak_3	3.25879
chr1	1079489	1079490	MACS2_peak_4	13.88358
chr1	1304890	1304891	MACS2_peak_5	15.56374
chr1	1441153	1441154	MACS2_peak_6	12.40068

Exercise: peak calling

- PDF images about the model based on your data

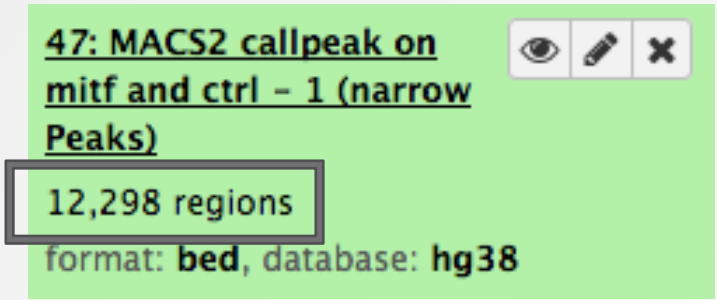




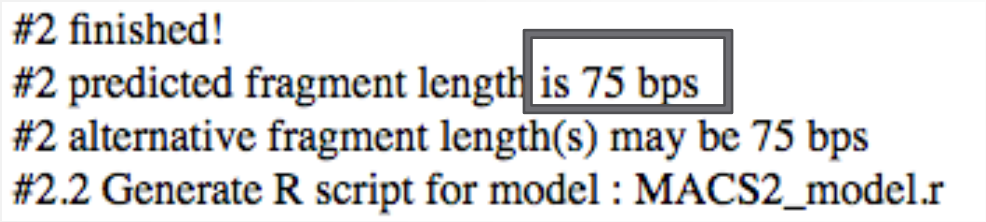
- Log of MACS - output during Macs2 run (HTML)

Exercise: peak calling

- There is 12,298 peaks



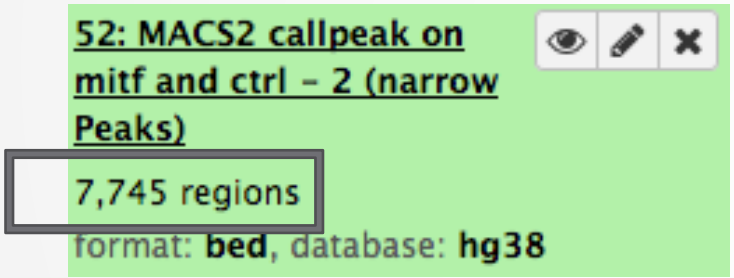
- 3. Look at the HTML dataset



- The d value estimated by MACS seems a bit small. Let's try to re-run MACS with the expected fragment size : 200

Exercise: peak calling

- 4.
  - Click on the name of one of the datasets generated by Macs2.
  - Click on ↻ to display Macs2 form with the same parameters as for the previous run of Macs2
  - In Build Model, select Do not build the shifting model (--nomodel)
  - Enter 100 in the text box "The arbitrary extension size in bp"
  - Click on ✔ Execute
- 5.
  - 7,745 peaks are now found



  - NOTE: the graphs (showing the d values estimate) are no longer generated