ChIP-seq data analysis with Chipster

Eija Korpelainen
CSC – IT Center for Science, Finland
chipster@csc.fi

What will I learn?

- **Short introduction to ChIP-seq**
- **Analyzing ChIP-seq data**
  - Central concepts
  - Analysis steps
  - File formats

Introduction to ChIP-seq

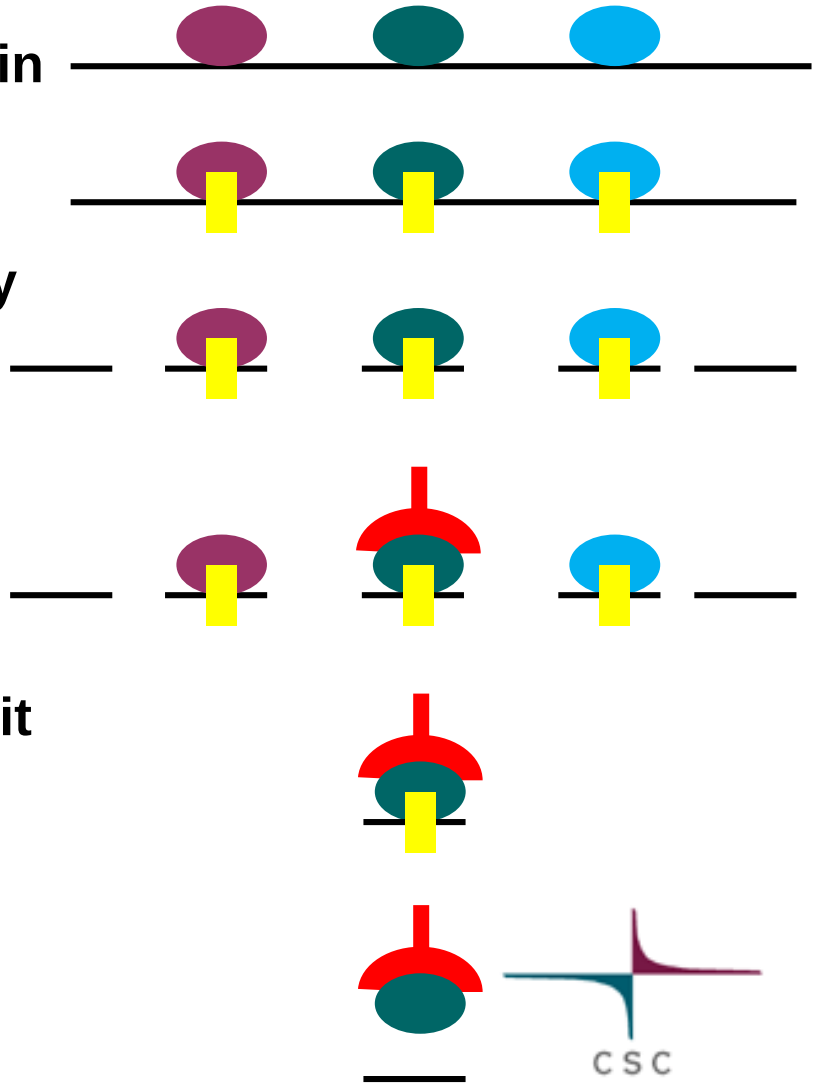
(ChIP-seq = Chromatin immunoprecipitation sequencing)

What can I investigate with ChIP-seq?

- **Locating regulatory elements in genome**
  - Transcription factor binding sites
  - Histone modifications
  - RNA polymerase binding sites
  - Etc etc

What happens in the lab?

- **Proteins are cross-linked to DNA in nucleus**
- **DNA is sheared to small pieces by sonication**
- **Antibodies specific to the protein of interest are added**
- **Antibody is attached to beads so it pulls down (immunoprecipitates) the protein and the piece of DNA.**
- **Cross-link is reversed and the released DNA is sequenced**



ChIP-seq data analysis workflow

- **Quality control, preprocessing**
- **Alignment to reference genome**
- **Manipulation of alignment files**
- **Peak calling**
- **Manipulation of peak files**
- **Visualization of reads and results in genome browser**
- **Retrieval of nearby genes**
- **Pathway analysis**
- **Motif detection**

Aligning reads to genome

Bowtie aligners

- **Fast and memory efficient**
- **Reference genome is indexed to speed up the alignment process**
- **Two version which are very different**
  - Bowtie2
    - Can perform gapped alignment for indels
    - Especially good for longer reads (> 50 bp)
  - Bowtie(1)
    - Can be more sensitive for shorter reads
    - Does not allow gaps

Bowtie(1)

- **Two modes**
  - Limit mismatches across the whole read (v).
  - Limit mismatches only in a user-specified seed region (n). The sum of qualities of all mismatch positions is not allowed to exceed a user-specified number.
- **Bowtie's own default parameters give the first alignment found, even if it is not the best one. Chipster uses the "best" and "strata" options to get the best class alignments**

Best and strata – what do they do?

- **Best**
  - If a read has several alignments, “-best” forces Bowtie to report them in the best-to-worst order (best is the one with least mismatches).
- **Strata**
  - Forces Bowtie to classify alignments to different categories (stratum) based on the number of mismatches. Only the alignments of the best category are considered.
  - Example: Allowing two mismatches, a read has three alignments:
    - A: 0 mismatches, B: 2 mismatches, C: 2 mismatches
    - A belongs to the best category, B and C form the second category
    - If you ask Bowtie to report reads which have only one best category hit, this read will be reported (even if it has three alignments)

CSC

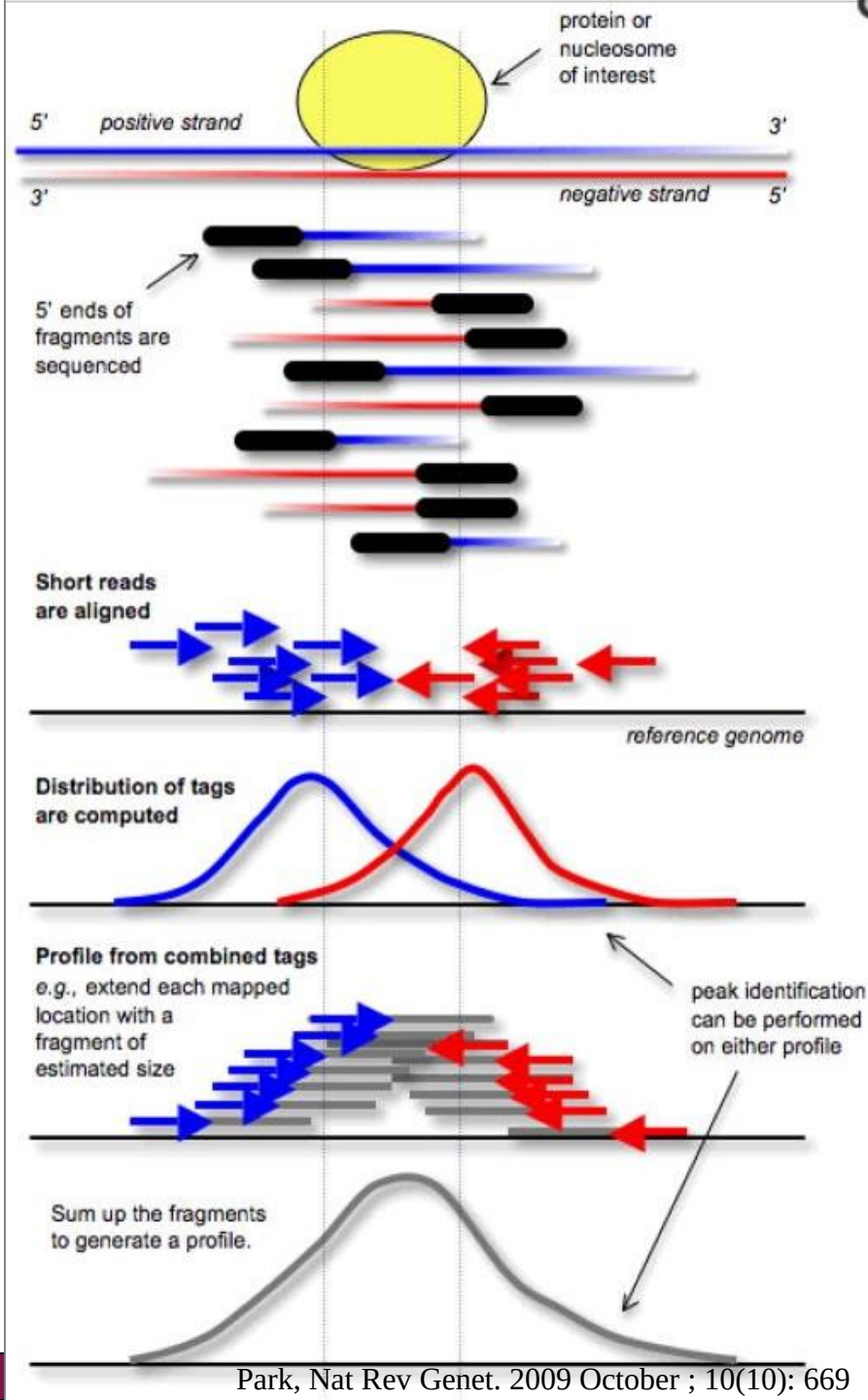
Peak calling

Detecting peaks

- **The goal is to find genomic regions of significant read enrichment**
- **Challenging because ChIP-seq data has regional biases due to**
  - Mappability differences
  - DNA copy number variations
  - Local chromatin structure
  - GC content
- **These background biases can be modelled from a matching control sample**
  - Use input DNA or do IP with a control Ab

How do peaks look?

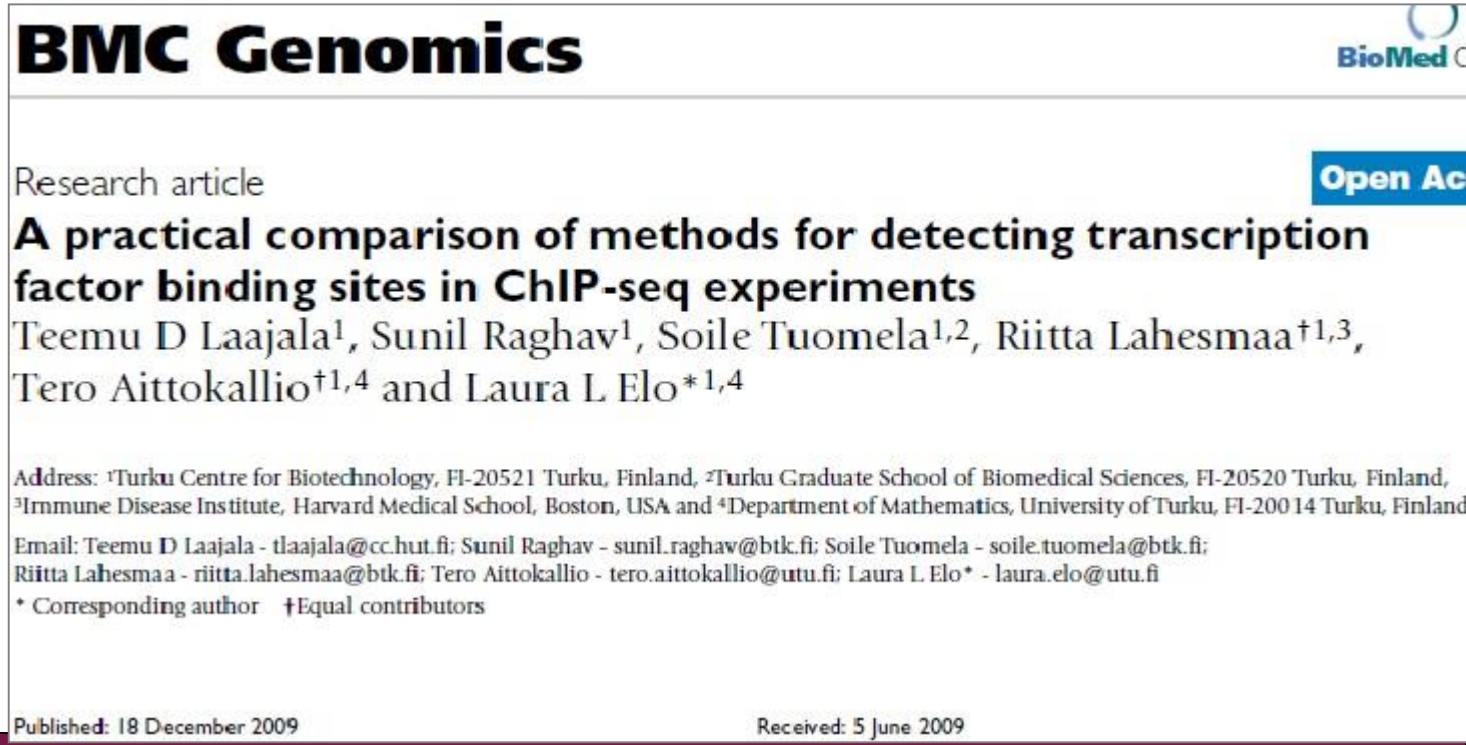
- **Reads are 5’ ends of fragments**
- **Reads form two peaks, one on the forward and one on the reverse strand**
- **The real binding site is in the middle, so the peaks need to shifted towards each other**
- **Note that some histone modifications produce broad peaks with no clear shape**



Park, Nat Rev Genet. 2009 October ; 10(10): 669

Peak callers

- **MACS**
- **MACS 2**
- **PeakSeq**
- **FindPeaks**
- **F-seq**
- **…**



BMC Genomics

BioMed C

Research article

Open Ac

A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments

Teemu D Laajala[1], Sunil Raghav[1], Soile Tuomela[1,2], Riitta Lahesmaa[†1,3], Tero Aittokallio[†1,4] and Laura L Elo[*1,4]

Address: [1]Turku Centre for Biotechnology, FI-20521 Turku, Finland, [2]Turku Graduate School of Biomedical Sciences, FI-20520 Turku, Finland, [3]Immune Disease Institute, Harvard Medical School, Boston, USA and [4]Department of Mathematics, University of Turku, FI-20014 Turku, Finland

Email: Teemu D Laajala - tlaajala@cc.hut.fi; Sunil Raghav - sunil.raghav@btk.fi; Soile Tuomela - soile.tuomela@btk.fi; Riitta Lahesmaa - riitta.lahesmaa@btk.fi; Tero Aittokallio - tero.aittokallio@utu.fi; Laura L Elo* - laura.elo@utu.fi

* Corresponding author †Equal contributors

Published: 18 December 2009 Received: 5 June 2009

MACS

- **MACS = Model-based Analysis of ChIP-seq**
- **Probably the most used peak caller**
- **Currently two versions: MACS (1.4.6) and MACS 2 (2.1.1)**
- **Features:**
  - Models the peak shift size from data
  - Uses a dynamic Poisson distribution to capture local biases
  - Can use control sample to estimate local background
  - Removes duplicate reads which are in excess of what is warranted by the sequencing depth
  - Scales samples linearly to the same number of reads

CSC

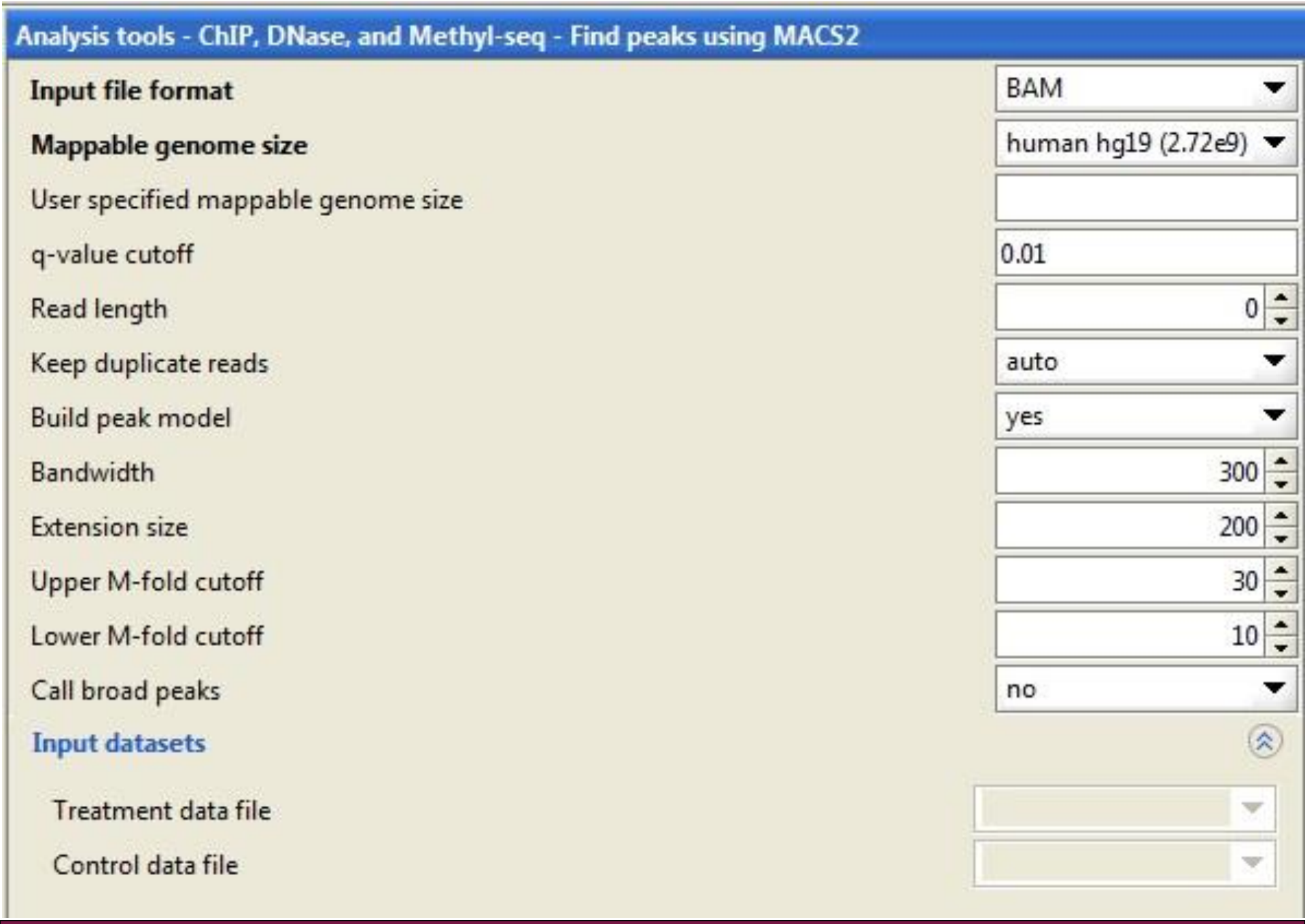
Peak calling – how does it work?

- **Find genomic regions which have more reads than expected**
- **Look at the enriched regions and calculate the distance d between the forward and reverse peaks**
- **Extend reads towards the center to length d**
- **Slide a window of size 2*d across the genome to select candidate peaks**
- **Calculate p-value by using a local lambda**
  - Is the peak higher than expected by chance? What is the probability to find a peak higher than x?
- **Calculate q-value using Benjamini-Hochberg correction**

CSC

Steps performed by MACS

- **Remove duplicates**
- **Slide a window across the genome to find enriched regions which have m-fold more reads than lambda (expected read number)**
  - Window size = 2 * sonication size (called bandwidth)
  - By default 10 < m < 30
  - lambda = (read length * number of reads) / mappable genome size
- **Take 1000 regions to calculate the distance (d) between the forward and reverse peaks**
- **Extend reads towards the center to length d (shifts the peaks)**
- **Scale the samples to the same read number**
- **Slide a window of size 2*d across the genome to select candidate peaks**
- **Calculate p-value by using a <u>local</u> lambda**
  - Use the control sample to estimate the local lambda
  - Is the peak higher than expected by chance? What is the probability to find a peak higher than x?
- **Calculate q-value using Benjamini-Hochberg correction**


Analysis tools - ChIP, DNase, and Methyl-seq - Find peaks using MACS2
Input file format
BAM
Mappable genome size
human hg19 (2.72e9)
User specified mappable genome size
q-value cutoff
0.01
Read length
0
Keep duplicate reads
auto
Build peak model
yes
Bandwidth
300
Extension size
200
Upper M-fold cutoff
30
Lower M-fold cutoff
10
Call broad peaks
no
Input datasets
Treatment data file
Control data file

MACS 2 result table

- **abs_summit = peak summit position**
- **pileup = hight at the peak summit**
- **neglog10pvalue = -log p-value for the summit**
- **fold enrichment = summit / local lambda**
- **neglog10qvalue = -log q-value for the summit**
  - FDR, Benjamini-Hochberg corrected p-value

chr	start	end	length	abs_summit	pileup	neglog10pvalue	fold_enrichment	neglog10qvalue
chr1	74	456	383	189	61	14.31	3.26	11.42
chr1	18657	18870	214	18799	28	6.12	2.9	3.6
chr1	22911	23232	322	23049	43	28.81	10.28	25.64
chr1	129968	130287	320	130147	53	40.56	13.32	37.27
chr1	133767	134100	334	133924	39	31.1	14.17	27.9
chr1	310247	310471	225	310383	65	36.54	8.23	33.29
chr1	313500	313922	423	313681	97	105.9	30.64	102.2
chr1	413910	414211	302	414063	50	43.13	16.71	39.81
chr1	655465	655786	322	655605	51	38.29	12.93	35.02
chr1	658826	659213	388	659061	54	41.7	13.71	38.4

MACS 2 log file

```
INFO  @ Mon, 15 Sep 2014 06:49:59:
# ARGUMENTS LIST:
# name = macs2
# format = BAM
# ChIP-seq file = treatment.bam
# control file = control.bam
# effective genome size = 2.47e+08
# band width = 300
# model fold = 10,30
# qvalue cutoff = 1.00e-02
# Larger dataset will be scaled towards smaller dataset.
# Range for calculating regional lambda is: 1000 bps and 10000 bps
# Broad region calling is off

INFO  @ Mon, 15 Sep 2014 06:49:59: #1 read tag files...
INFO  @ Mon, 15 Sep 2014 06:49:59: #1 read treatment tags...
INFO  @ Mon, 15 Sep 2014 06:49:59: tag size: 27
INFO  @ Mon, 15 Sep 2014 06:50:09:  1000000
INFO  @ Mon, 15 Sep 2014 06:50:19:  2000000
INFO  @ Mon, 15 Sep 2014 06:50:29:  3000000
INFO  @ Mon, 15 Sep 2014 06:50:29: #1.2 read input tags...
INFO  @ Mon, 15 Sep 2014 06:50:39:  1000000
INFO  @ Mon, 15 Sep 2014 06:50:48:  2000000
INFO  @ Mon, 15 Sep 2014 06:50:53: #1 tag size is determined as 27 bps
INFO  @ Mon, 15 Sep 2014 06:50:53: #1 tag size = 27
INFO  @ Mon, 15 Sep 2014 06:50:53: #1  total tags in treatment: 3014699
INFO  @ Mon, 15 Sep 2014 06:50:53: #1 calculate max duplicate tags in single position based on binomal distribution...
INFO  @ Mon, 15 Sep 2014 06:50:53: #1  max_dup_tags based on binomal = 2
INFO  @ Mon, 15 Sep 2014 06:50:53: #1 filter out redundant tags at the same location and the same strand by allowing at most 2 tag(s)
INFO  @ Mon, 15 Sep 2014 06:50:54: #1  tags after filtering in treatment: 2942747
INFO  @ Mon, 15 Sep 2014 06:50:54: #1  Redundant rate of treatment: 0.02
INFO  @ Mon, 15 Sep 2014 06:50:54: #1  total tags in control: 2442445
INFO  @ Mon, 15 Sep 2014 06:50:54: #1  for control, calculate max duplicate tags in single position based on binomal distribution...
INFO  @ Mon, 15 Sep 2014 06:50:54: #1  max_dup_tags based on binomal = 2
INFO  @ Mon, 15 Sep 2014 06:50:54: #1 filter out redundant tags at the same location and the same strand by allowing at most 2 tag(s)
INFO  @ Mon, 15 Sep 2014 06:50:54: #1  tags after filtering in control: 2305411
INFO  @ Mon, 15 Sep 2014 06:50:54: #1  Redundant rate of control: 0.06
INFO  @ Mon, 15 Sep 2014 06:50:54: #1 finished!
INFO  @ Mon, 15 Sep 2014 06:50:54: #2 Build Peak Model...
INFO  @ Mon, 15 Sep 2014 06:50:56: #2 number of paired peaks: 1311
INFO  @ Mon, 15 Sep 2014 06:50:58: #2 finished!
INFO  @ Mon, 15 Sep 2014 06:50:58: #2 predicted fragment length is 188 bps
INFO  @ Mon, 15 Sep 2014 06:50:58: #2.2 Generate R script for model : macs2_model.r
INFO  @ Mon, 15 Sep 2014 06:50:58: #3 Call peaks...
INFO  @ Mon, 15 Sep 2014 06:50:58: #3 pileup treatment data by extending tags towards 3' to 188 length
INFO  @ Mon, 15 Sep 2014 06:51:10: #3 calculate d local lambda for control data
INFO  @ Mon, 15 Sep 2014 06:51:20: #3 calculate small local lambda for control data
INFO  @ Mon, 15 Sep 2014 06:51:44: #3 calculate large local lambda for control data
INFO  @ Mon, 15 Sep 2014 06:52:16: #3 Build score track ...
INFO  @ Mon, 15 Sep 2014 06:54:02: #3 Calculate qvalues ...
INFO  @ Mon, 15 Sep 2014 06:55:38: #3 Saving p-value to q-value table ...
INFO  @ Mon, 15 Sep 2014 06:55:38: #3 Assign qvalues ...
INFO  @ Mon, 15 Sep 2014 06:56:17: #3 Call peaks with given -log10qvalue cutoff: 2.00 ...
INFO  @ Mon, 15 Sep 2014 06:57:03: #4 Write output xls file... macs2_peaks.xls
```

CSC

BED

- **5 obligatory columns: chr, start, end, name, score (-log pvalue)**
- **0-based, like BAM**

column0	column1	column2	column3	column4
chr1	120	488	MACS_peak_1	8.76
chr1	18736	19053	MACS_peak_2	4.6
chr1	22915	23281	MACS_peak_3	24.09
chr1	129952	130322	MACS_peak_4	31.48
chr1	133748	134105	MACS_peak_5	22.06
chr1	310255	310563	MACS_peak_6	12.11

CSC

Narrow peak format (BED6+4)

- **0 = chr, 1 = start, 2 = end, 3 = name**
- **4 = score for display**
- **6 = fold change**
- **7 = -log10 pvalue**
- **8 = -log10 qvalue**
- **9 = summit position relative to peak start**

column0	column1	column2	column3	column4	column5	column6	column7	column8	column9
chr1	73	456	MACS_peak_1	114	.	3.26	14.31	11.42	115
chr1	18656	18870	MACS_peak_2	36	.	2.90	6.12	3.60	142
chr1	22910	23232	MACS_peak_3	256	.	10.28	28.81	25.64	138
chr1	129967	130287	MACS_peak_4	372	.	13.32	40.56	37.27	179
chr1	133766	134100	MACS_peak_5	279	.	14.17	31.10	27.90	157
chr1	310246	310471	MACS_peak_6	332	.	8.23	36.54	33.29	136
chr1	313499	313922	MACS_peak_7	1022	.	30.64	105.90	102.20	181
chr1	413909	414211	MACS_peak_8	398	.	16.71	43.13	39.81	153

Matching sets of genomic regions

Why?

Useful for many questions, for example:

- **Are the peaks located close to genes?**
- **What genes are closest upstream to the peaks?**
- **Do the peaks overlap with transcription start sites?**
- **Do two lists of peaks overlap?**
- **Which peaks are specific to the first list?**
- **….**

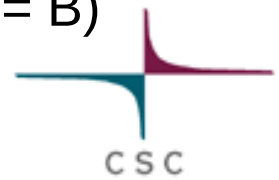
Software packages for region matching

- **BEDTools**
  - Supports BED, GTF, VCF, BAM
  - Rich functionality

- **Chipster's own region matching tools**
  - Support BED
  - Tolerant for chromosome naming (chr1 vs 1)

Intersect BED (BEDTools)

- **Looks for overlapping regions between two BED/GFF/VCF files.**
  - One of the files can also be BAM.
  - Option for strand-awareness
- **Reporting options**
  - Only the overlapping region
  - Original region in file A or B
  - Region in A so that the overlapping part is removed
  - Remove the portion of a region that is overlapped by another region
- **The B file is loaded to memory**
  - So the smaller one should be B (e.g. BAM = A, BED = B)



Closest BED (BEDTools)

- **Looks for overlapping regions between two BED/GFF/VCF files and if no overlap is found, the closest region is reported.**
- **Reports a region in A region followed by its closest region in B.**
- **Option for strand-awareness**
- **E.g. What is the nearest gene to this peak?**

Window BED (BEDTools)

- **Looks for overlapping regions between two BED/GFF/VCF files after adding a given number of bases upstream and downstream of regions in A.**
  - One of the files can be BAM
- **Reports the regions in A which overlap with regions in B.**
- **Option for strand-awareness**

Dataset for peak calling exercises

- **Data from Rozowsky *et al.* Nature Biotechnology 27, 66 - 75 (2009)**
- **Interferon gamma stimulated HeLa S3 cells**
  - ChIP with STAT1 antibody
  - Control sample: Input DNA
- **For the interest of time we use only reads which map to chr 1**

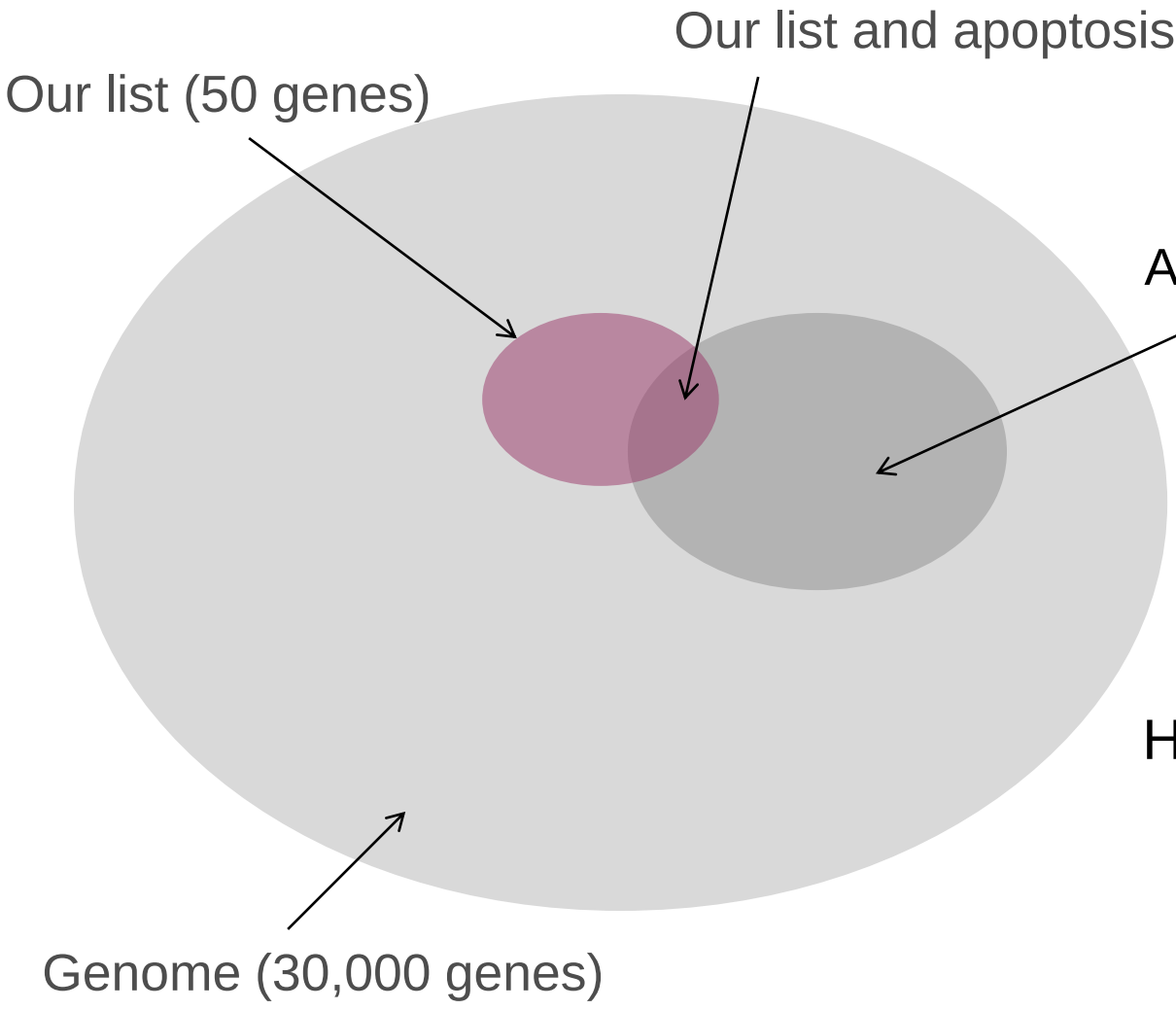
Pathway analysis

Pathway analysis – what and why?

- **ChIP-seq peaks can have thousands of neighboring genes. We want to understand what is the biological meaning of this list**
  - What pathways are regulated by the transcription factor?
- **Typically accomplished by enrichment analysis**
  - Check which known pathways the genes belong to
  - Are there enriched pathways (which have more members in this list that would be expected by chance)?
- **Pathway databases in Chipster**
  - Gene ontology (GO)
  - ConsensusPathDB

CSC

Pathway enrichment analysis:



$$H_0 : \frac{\frac{10}{50}}{\frac{200}{30000}} = 30 >> 1$$

ConsensusPathDB

- **One-stop shop: Integrates pathway information from 32 databases covering**
  - biochemical pathways
  - protein-protein, genetic, metabolic, signaling, gene regulatory and drug-target interactions

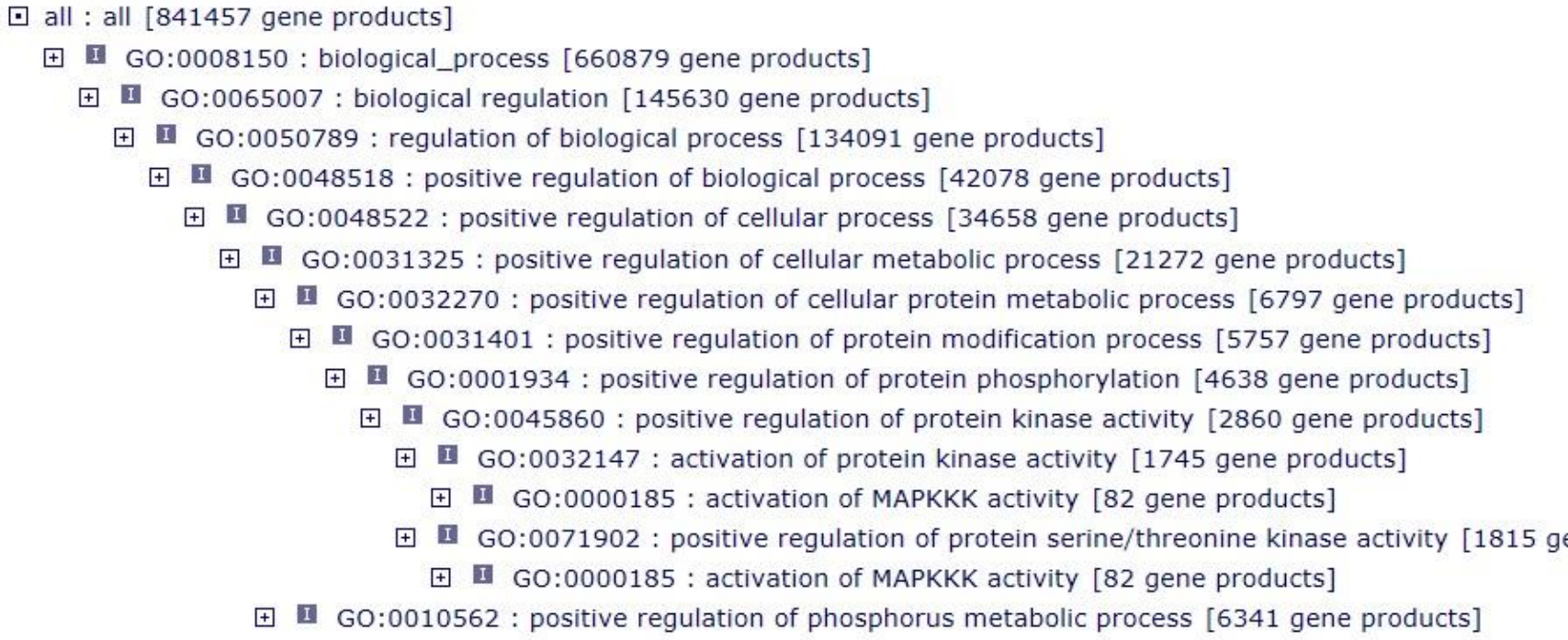
- **Developed by Ralf Herwig's group at the Max-Planck Institute in Berlin**

- **ConsensusPathDB over-representation analysis tool is integrated in Chipster**
  - runs on the MPI server in Berlin

CSC

Gene Ontology (GO)

- **Controlled vocabulary of terms for describing gene product characteristics**
- **3 ontologies**
  - Biological process
  - Molecular function
  - Cellular component
- **Hierarchical structure**



Motif detection

Motif detection – what and why?

- **Goal is to find sequence motifs that occur in all detected peaks**
- **Useful for many questions, such as:**
  - What sequence does the transcription factor bind to?
  - How well did the experiment work?
    - Does every peak contain the (known) binding site?
    - Is the site located close to the summit of the peak?

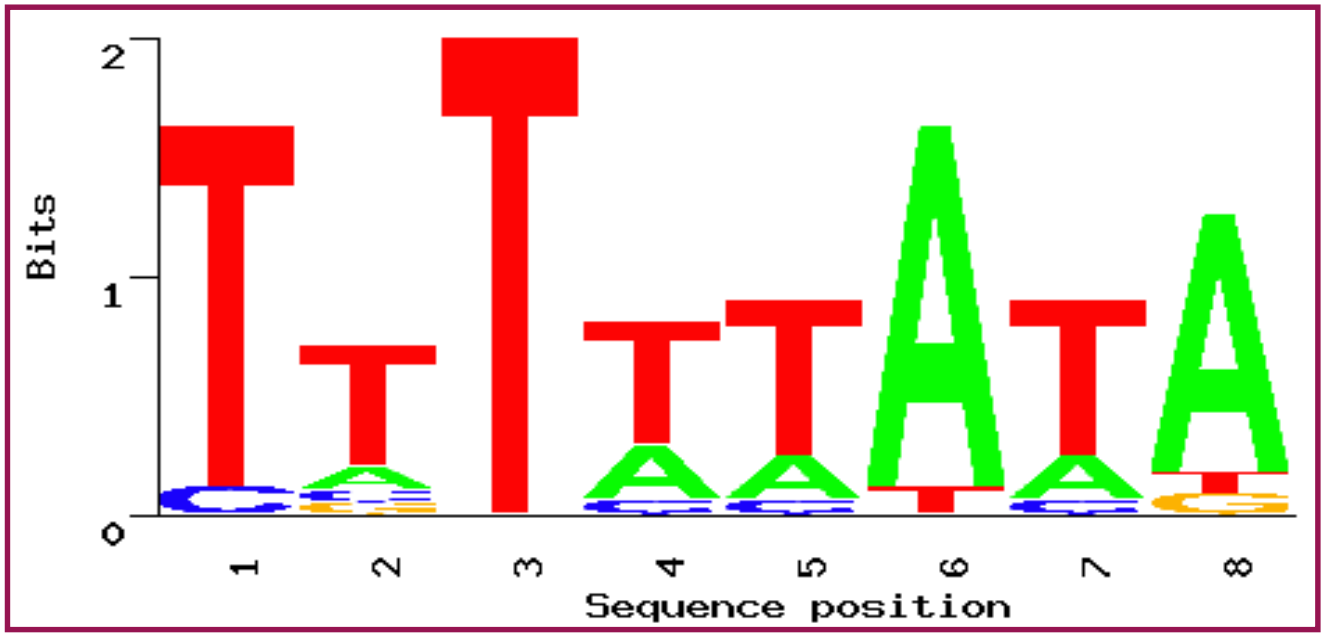
Different ways to represent motifs:

Consensus, matrix and sequence logo

TTTTTATA
TTTAAATA
TTTTTAAA
CTTTTATA
TTTATATA
TTTAAAAA
TATTTATA
TTTTTACA
TTTTAAAA
TATATATA
TTTTTTTT
TCTTTATA
TTTCCATG
TGTTTATA

TTTTTATA

Pos	1	2	3	4	5	6	7	8
A	0	2	0	4	3	13	3	12
C	1	1	0	1	1	0	1	0
G	0	1	0	0	0	0	0	1
T	13	10	14	9	10	1	10	1



Motif detection tools in Chipster

- **Find motifs with GADEM and match to JASPAR**
  - Uses the rGADEM package to detect motifs
    - GADEM = Genetic Algorithm guided formation of spaced Dyads coupled with EM for Motif identification
  - Uses the motIV package to match the detected motifs to the JASPAR database of known motifs
- **Dimont**
  - Extracts sequences from the peaks and detects motifs in them
  - Can scan new sequences with the discovered motif