Prediction of transcription factor binding sites from ChIP-Seq data through de novo TFBS motif discovery

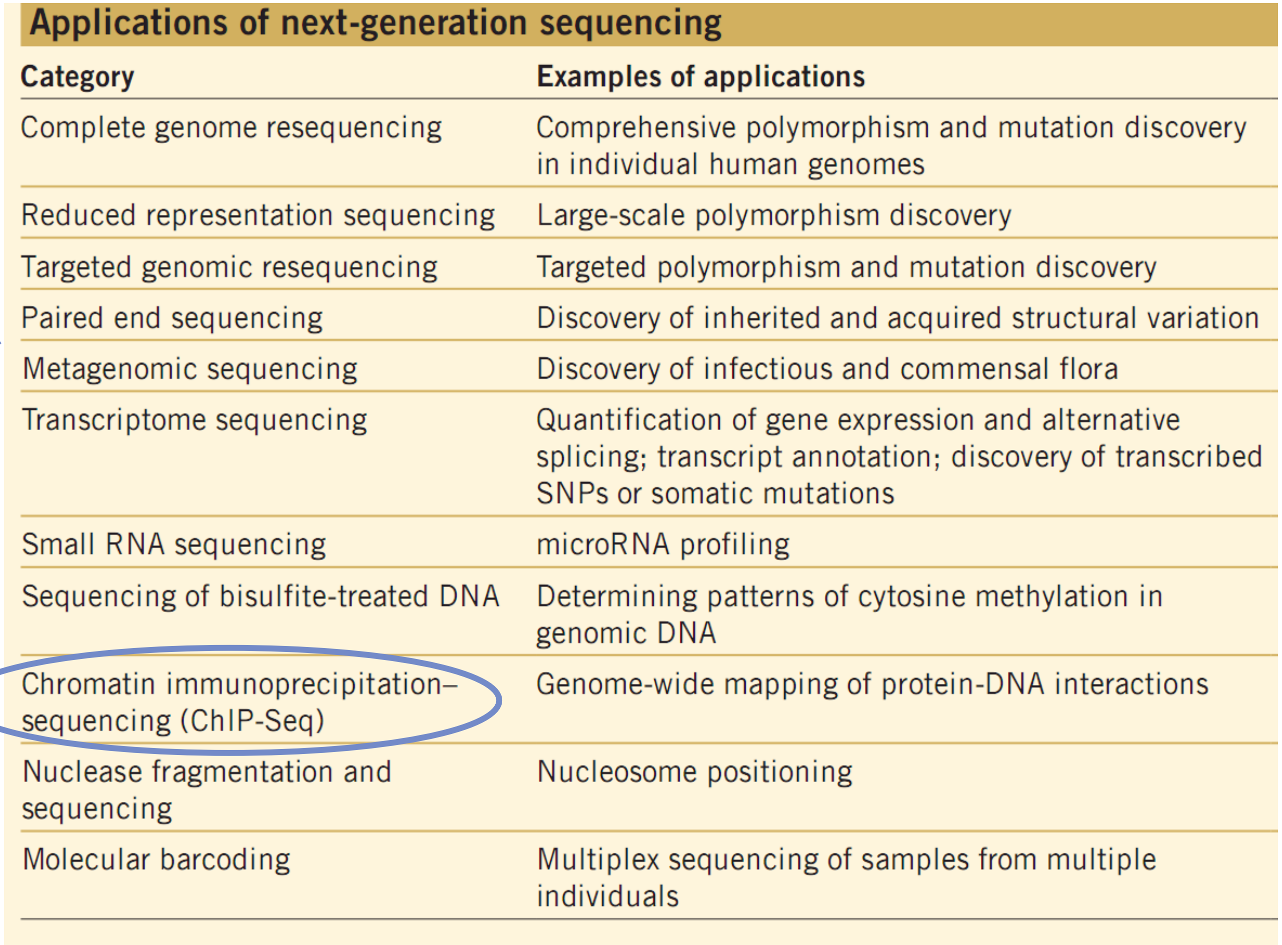
Valentina Boeva, *Institut Curie, Paris*

institutCurie

Inserm
Institut national
de la santé et de la recherche médicale

MINES PARIS

ChIP-Seq – one of the most exciting NGS applications

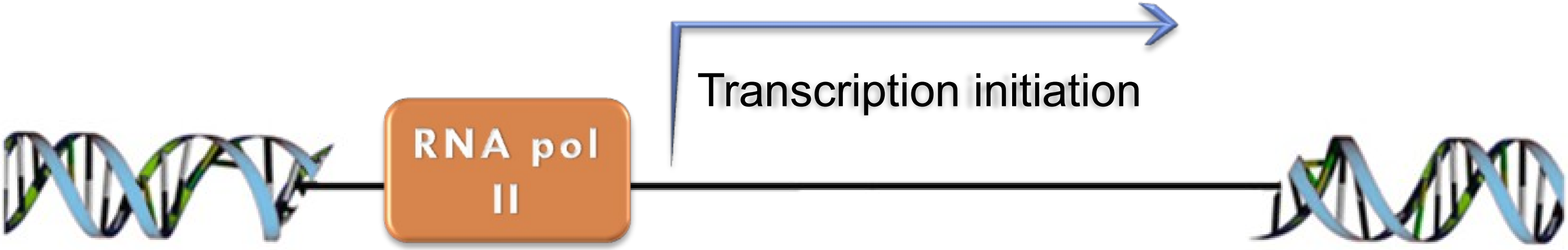


Applications of next-generation sequencing

Category	Examples of applications
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes
Reduced representation sequencing	Large-scale polymorphism discovery
Targeted genomic resequencing	Targeted polymorphism and mutation discovery
Paired end sequencing	Discovery of inherited and acquired structural variation
✓ Metagenomic sequencing	Discovery of infectious and commensal flora
✓ Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations
Small RNA sequencing	microRNA profiling
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions
Nuclease fragmentation and sequencing	Nucleosome positioning
Molecular barcoding	Multiplex sequencing of samples from multiple individuals

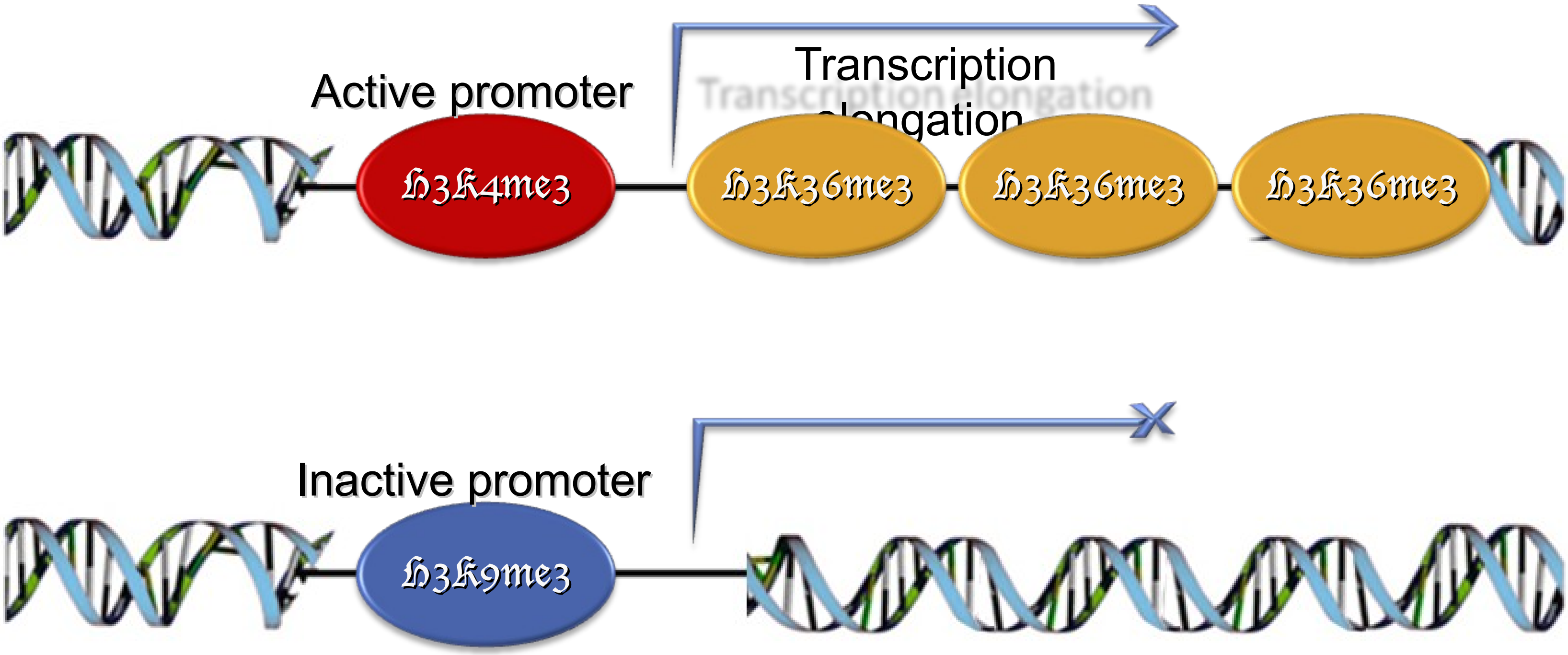
Why is important to find DNA-protein interactions?

- Sites of RNA polymerase



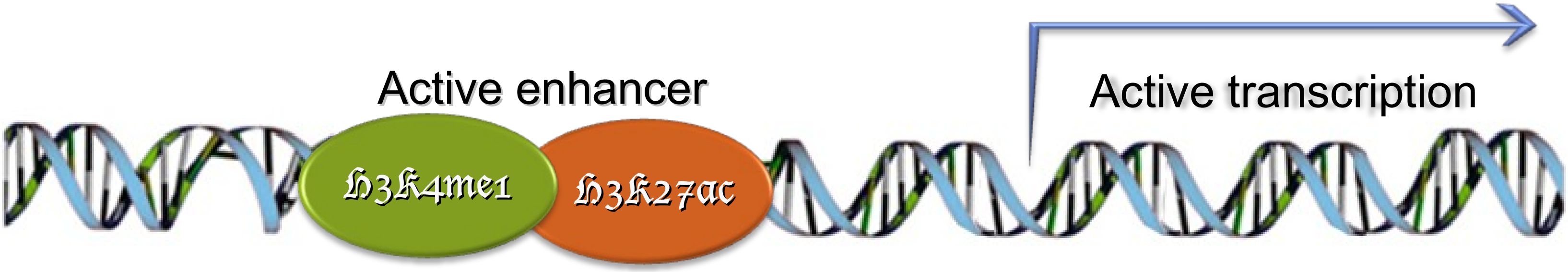
Why is important to find DNA-protein interactions?

- Histone modifications



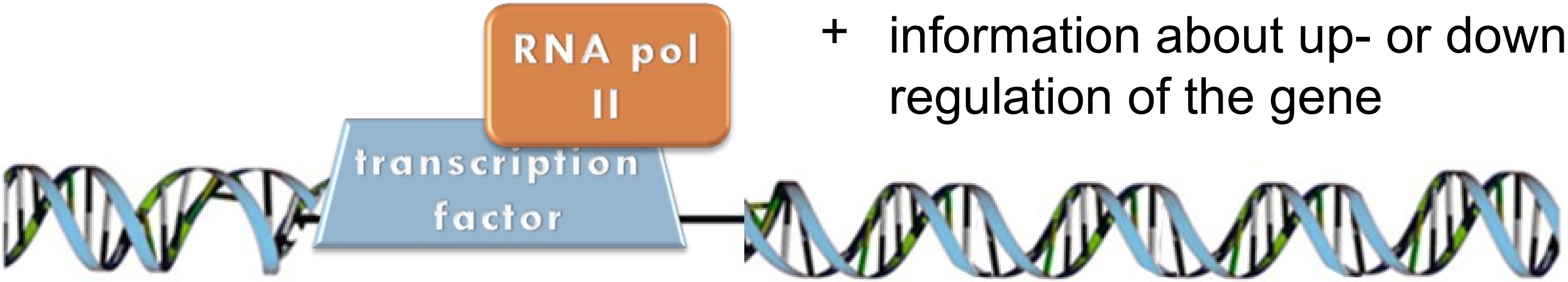
Why is important to find DNA-protein interactions?

- Histone modifications



Why is important to find DNA-protein interactions?

- Transcription factors (TFs) involved in regulation of cell growth, DNA repair and cell death pathways

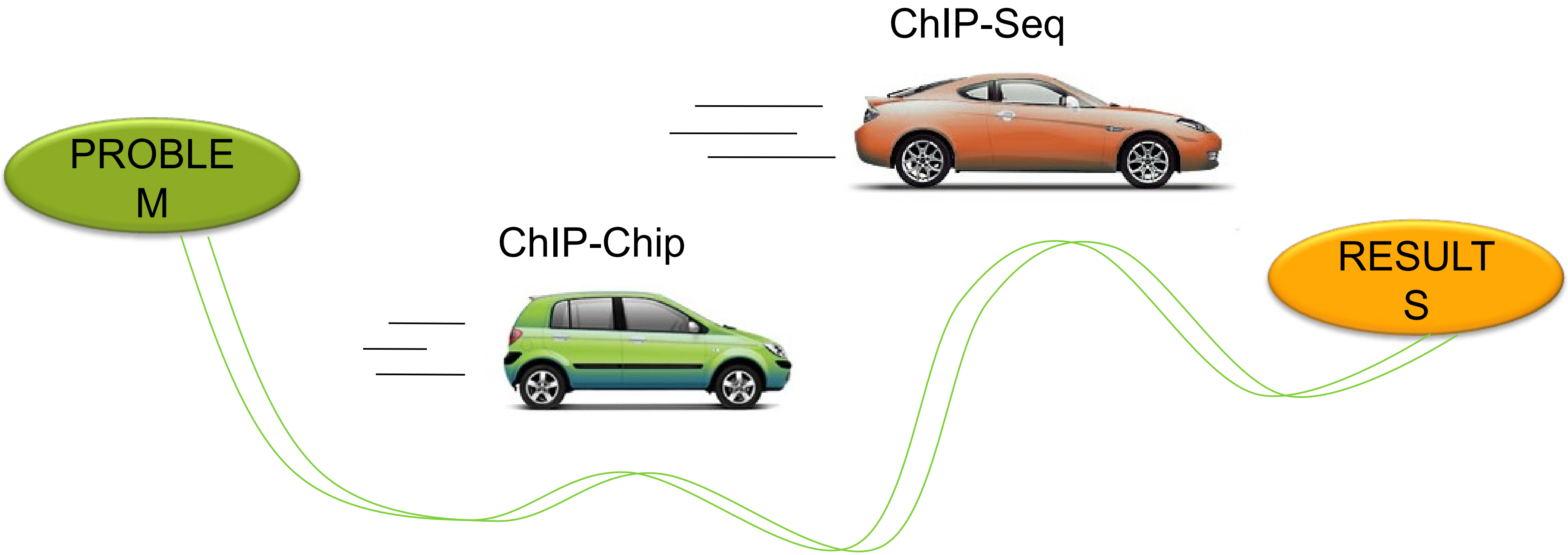


+ information about up- or down regulation of the gene

➡ Direct targets!

+ Motif finding ➡ Possible cofactors

ChIP-Seq is more precise than ChIP-on-Chip



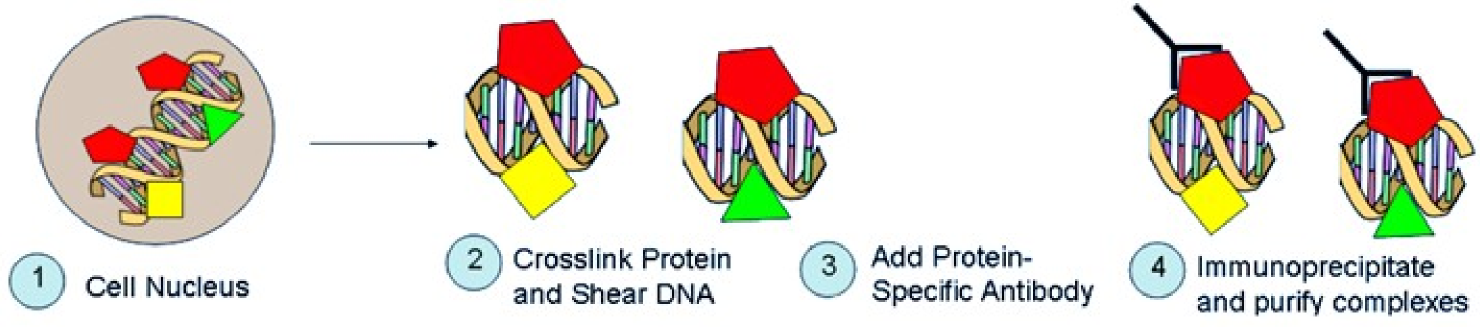
PROBLE
M

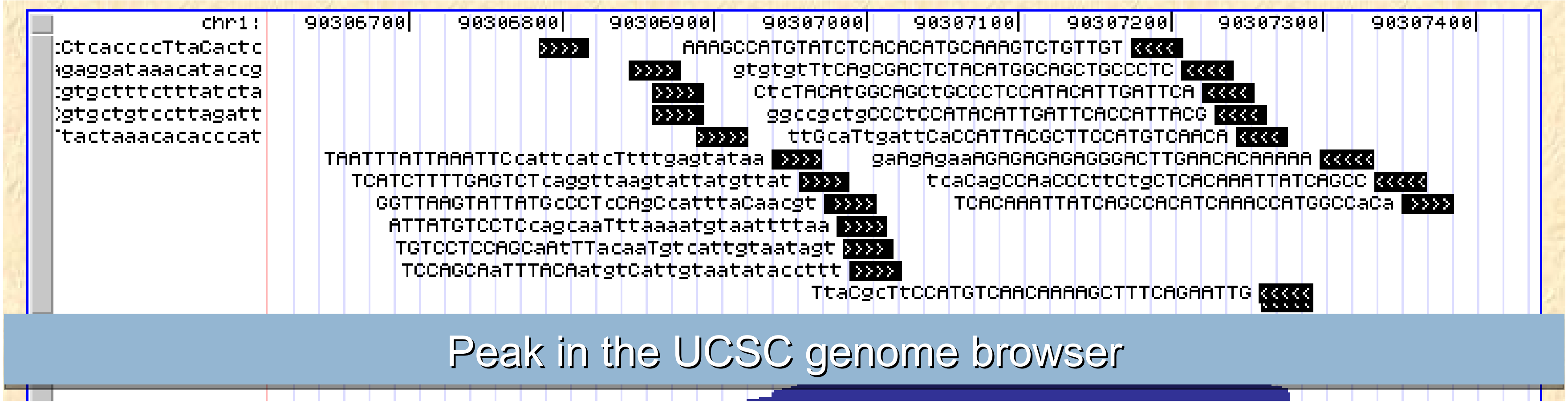
ChIP-Chip

ChIP-Seq

RESULT
S

Mains steps of ChIP-Seq technique


1 Cell Nucleus
2 Crosslink Protein and Shear DNA
3 Add Protein-Specific Antibody
4 Immunoprecipitate and purify complexes



```
                    chr1:    90306700|    90306800|    90306900|    90307000|    90307100|    90307200|    90307300|    90307400|
:CtcaccccTtaCactc                               >>>>         AAAGCCATGTATCTCACACATGCAAAGTCTGTTGT <<<<
agaggataaacataccg                                    >>>>        gtgtgtTtCAgCGACTCTACATGGCAGCTGCCCTC <<<<
:gtgctttctttatcta                                     >>>>        CtcTACAtGGCAGCtGCCCTCCATACATTGATTCA <<<<
:gtgctgtccttagatt                                     >>>>         ggccgctgCCCtCCATACATTGATTCACCATTACG <<<<
'tactaaacacacccat                                        >>>>>       ttGcaTtgattCaCCATTACGCTTCCATGTCAACA <<<<
             TAATTTATTAAATTCcattcatcTtttgagtataa >>>>       gaAgAgaaAGAGAGAGAGGGACTTGAACACAAAAA <<<<<
                TCATCTTTTGAGTCTcaggttaagtattatgttat >>>>          tcaCagCCAaCCCttCtgCTCACAAATTATCAGCC <<<<<
                  GGTTAAGTATTATGcCCTcCAgCcatttaCaacgt >>>>          TCACAAATTATCAGCCACATCAAACCATGGCCaCa >>>>
                   ATTATGTCCTCcagcaaTttaaaatgtaattttaa >>>>
                    TGTCCTCCAGCaAtTTacaaTgtcattgtaatagt >>>>
                     TCCAGCAaTTTACAatgtCattgtaatatacttt >>>>
                                              TtaCgcTtCCATGTCAACAAAAGCTTTCAGAATTG <<<<<
```

Peak in the UCSC genome browser

There is > a dozen of tools to detect read clusters

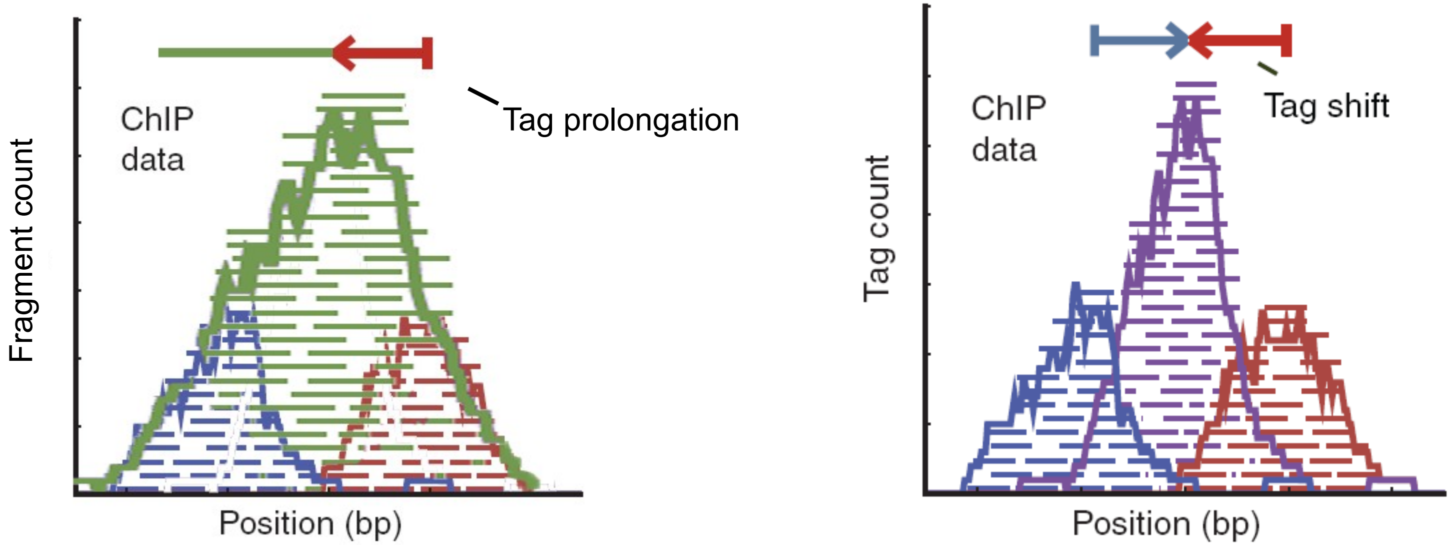
- FindPeaks
- QuEST
- CisGenome
- GLITR
- F-Seq
- SICER
- PeakSeq
- Spp
- Useq
- SiSSRs
- MACS
- ERANGE

Most of tools translate read clusters into peaks

- FindPeaks
- QuEST
- CisGenome
- GLITR
- F-Seq
- SICER
- PeakSeq
- Spp
- Useq
- SiSSRs
- MACS
- ERANGE

Clusters → Peaks

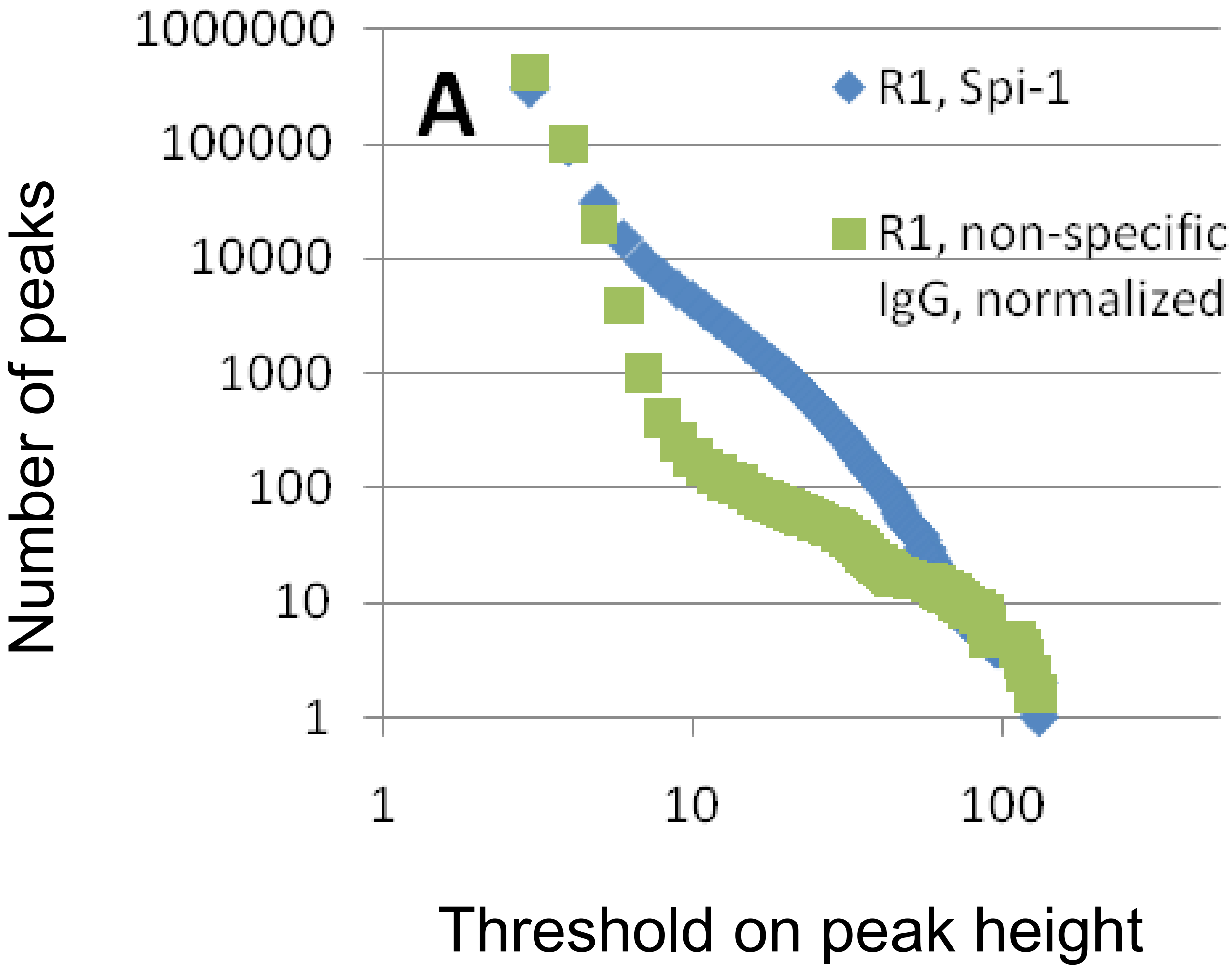
two ways



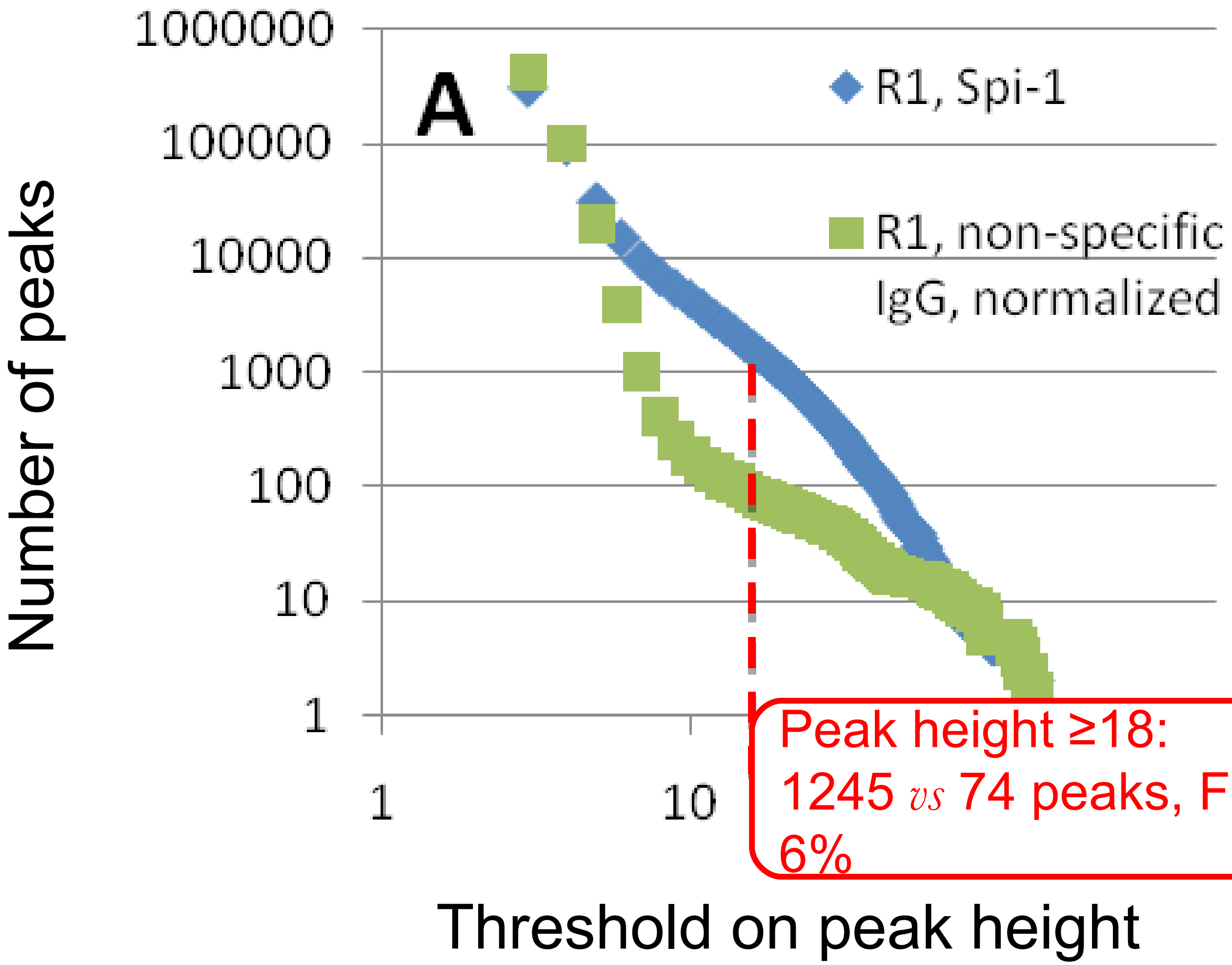
Adopted from S. Pepke et al., 2009 Nat Methods

+ some statistics to eliminate ‘false’ peaks

Threshold selection based on peak height

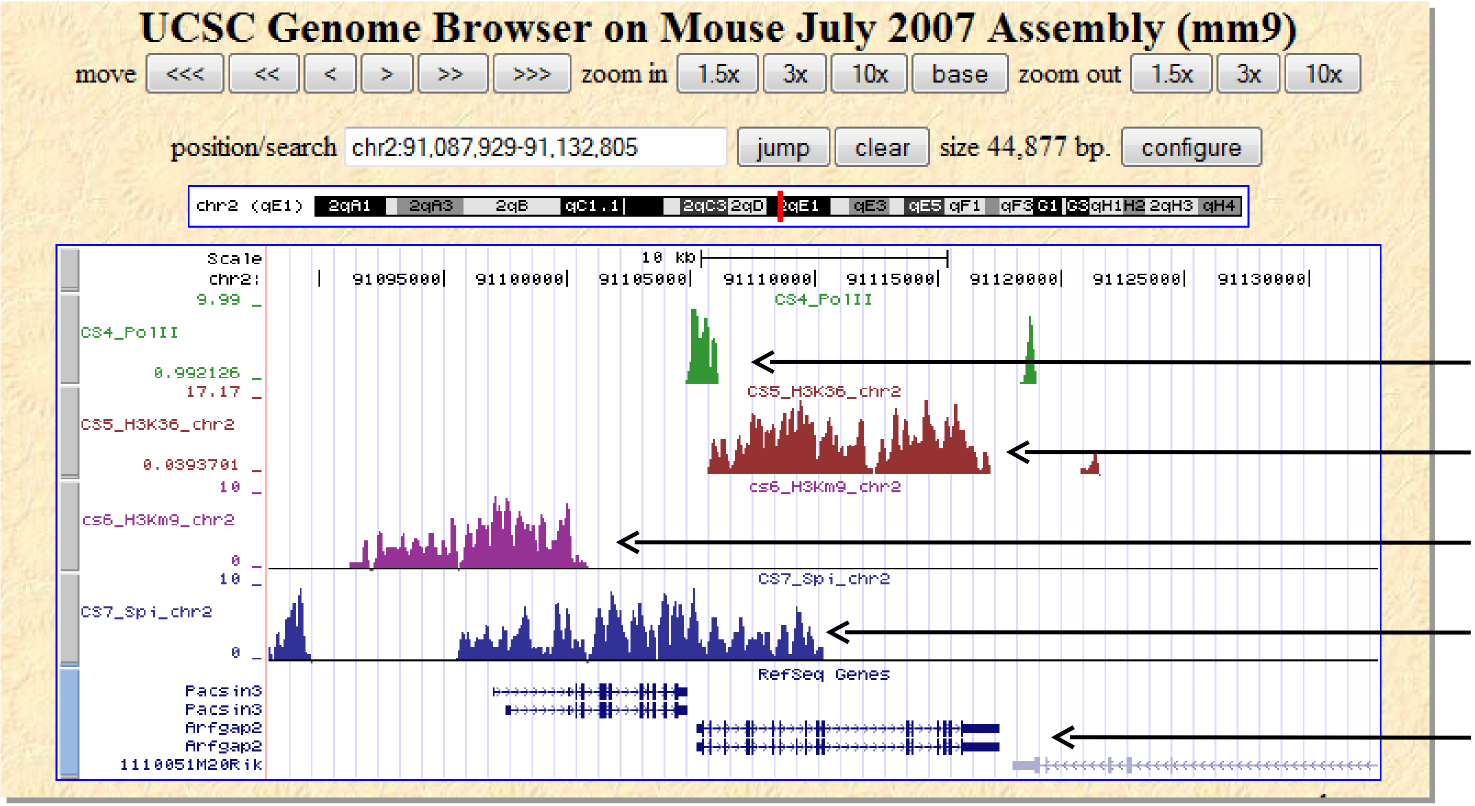


Threshold selection based on peak height



Peak height ≥18: 1245 *vs* 74 peaks, FDR = 6%

Visualization in the UCSC Genome Browser



UCSC Genome Browser on Mouse July 2007 Assembly (mm9)

position/search chr2:91,087,929-91,132,805

size 44,877 bp.

RNApolII

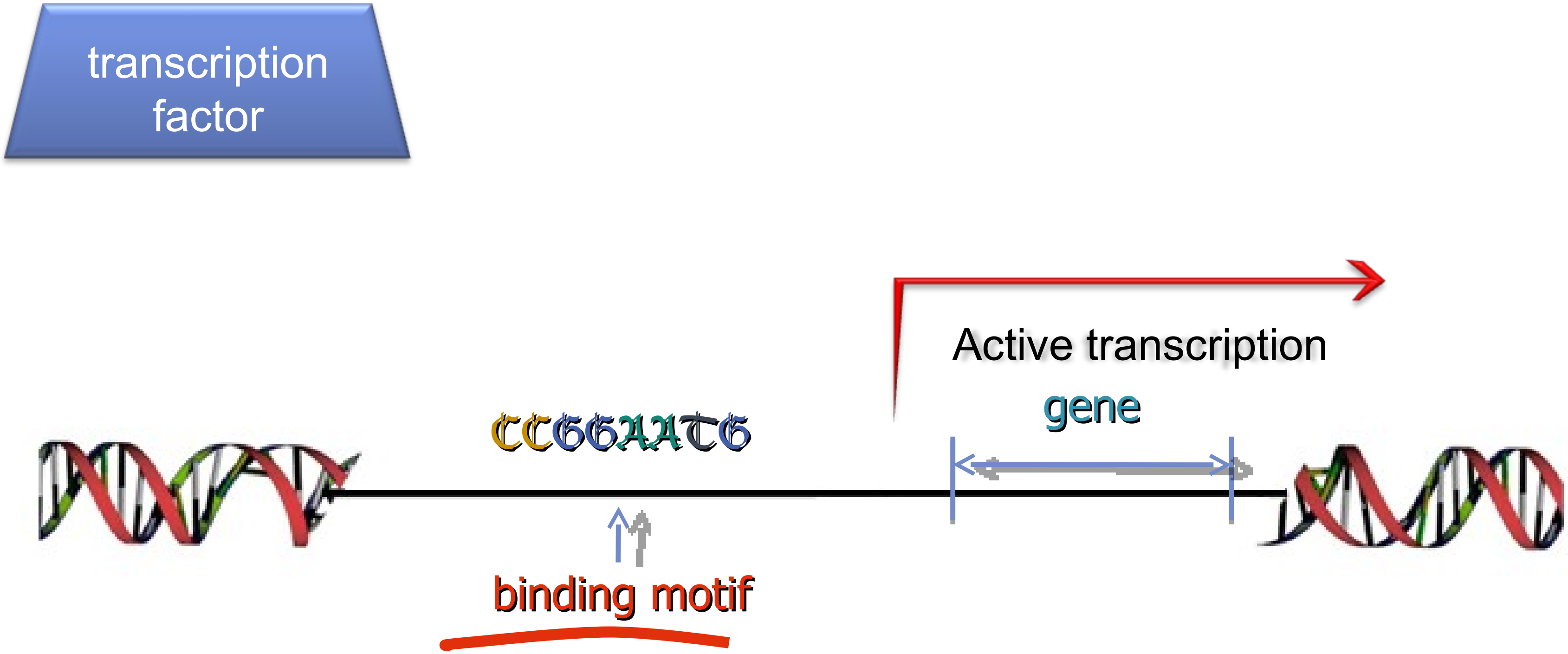
H3K36me3

H3K9me3

Spi-1 sites

Genes

Transcription factors: binding site usually contains a binding motif



transcription factor

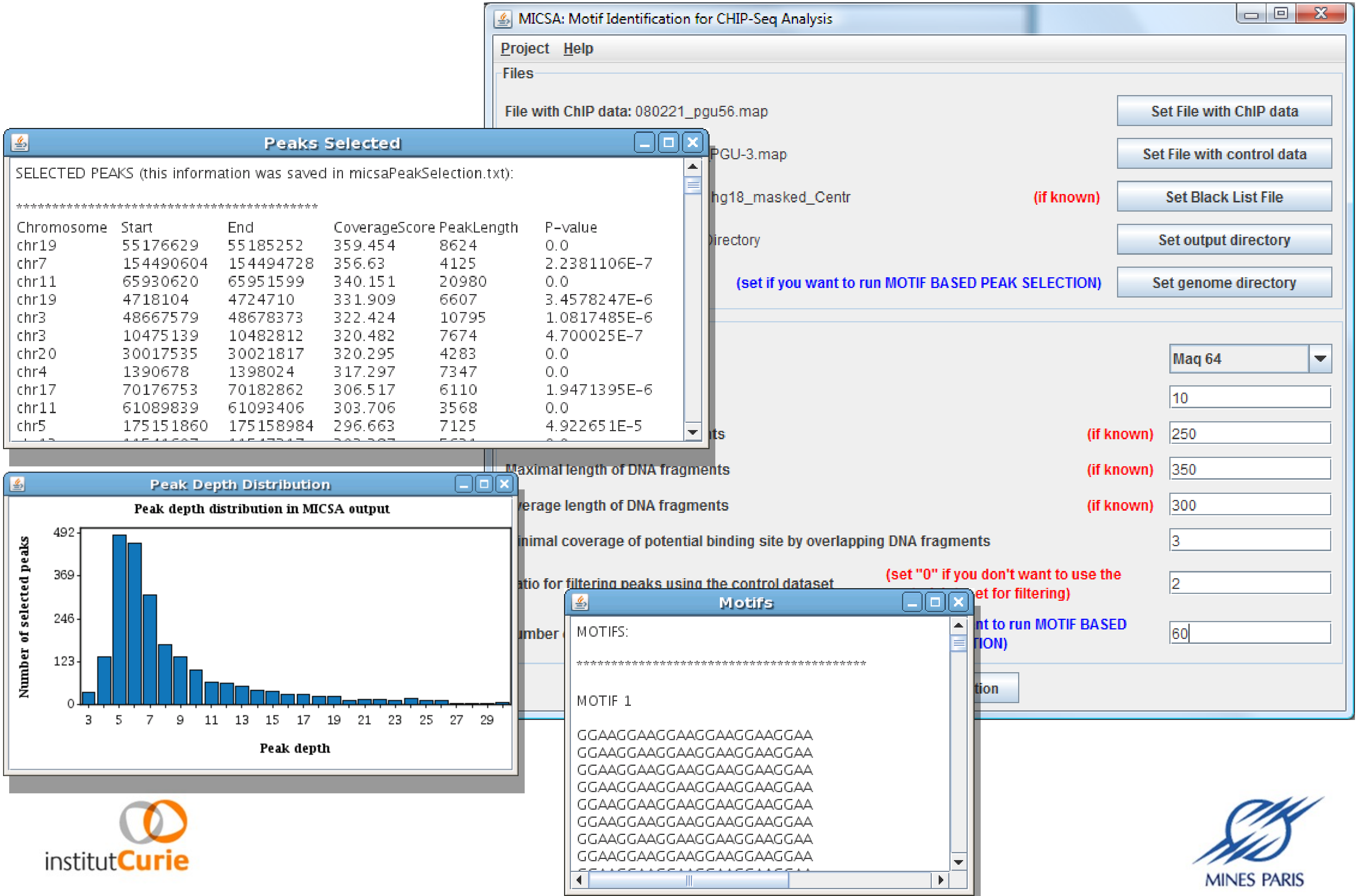
Active transcription

gene

CCGGAATG

binding motif

MICSA: a package for analysis of ChIP-Seq data for transcription factors



MICSA: Motif Identification for CHIP-Seq Analysis

File with ChIP data: 080221_pgu56.map

PGU-3.map

hg18_masked_Centr (if known)

irectory

(set if you want to run MOTIF BASED PEAK SELECTION)

Maq 64

10

ts (if known) 250

Maximal length of DNA fragments (if known) 350

verage length of DNA fragments (if known) 300

inimal coverage of potential binding site by overlapping DNA fragments 3

atio for filtering peaks using the control dataset (set "0" if you don't want to use the et for filtering) 2

umber nt to run MOTIF BASED TION) 60

Peaks Selected

```
SELECTED PEAKS (this information was saved in micsaPeakSelection.txt):

********************************************

Chromosome  Start      End        CoverageScore PeakLength  P-value
chr19       55176629   55185252   359.454       8624        0.0
chr7        154490604  154494728  356.63        4125        2.2381106E-7
chr11       65930620   65951599   340.151       20980       0.0
chr19       4718104    4724710    331.909       6607        3.4578247E-6
chr3        48667579   48678373   322.424       10795       1.0817485E-6
chr3        10475139   10482812   320.482       7674        4.700025E-7
chr20       30017535   30021817   320.295       4283        0.0
chr4        1390678    1398024    317.297       7347        0.0
chr17       70176753   70182862   306.517       6110        1.9471395E-6
chr11       61089839   61093406   303.706       3568        0.0
chr5        175151860  175158984  296.663       7125        4.922651E-5
```

Peak Depth Distribution

Peak depth distribution in MICSA output

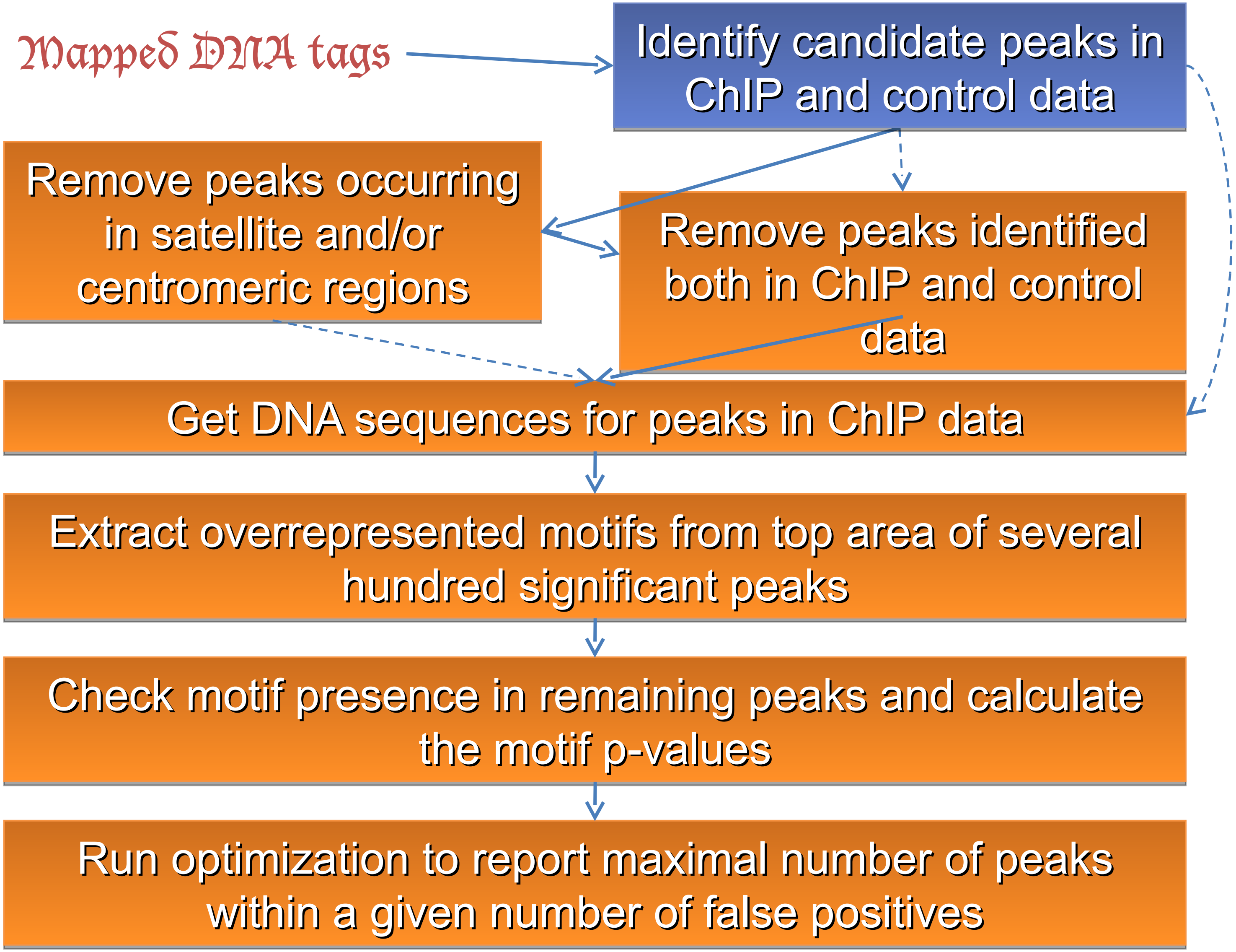
Motifs

```
MOTIFS:

********************************************

MOTIF 1

GGAAGGAAGGAAGGAAGGAAGGAA
GGAAGGAAGGAAGGAAGGAAGGAA
GGAAGGAAGGAAGGAAGGAAGGAA
GGAAGGAAGGAAGGAAGGAAGGAA
GGAAGGAAGGAAGGAAGGAAGGAA
GGAAGGAAGGAAGGAAGGAAGGAA
GGAAGGAAGGAAGGAAGGAAGGAA
GGAAGGAAGGAAGGAAGGAAGGAA
```

institutCurie

MINES PARIS

Main steps of the MICSA pipeline



Mapped DNA tags

Identify candidate peaks in ChIP and control data

Remove peaks occurring in satellite and/or centromeric regions

Remove peaks identified both in ChIP and control data

Get DNA sequences for peaks in ChIP data

Extract overrepresented motifs from top area of several hundred significant peaks

Check motif presence in remaining peaks and calculate the motif p-values

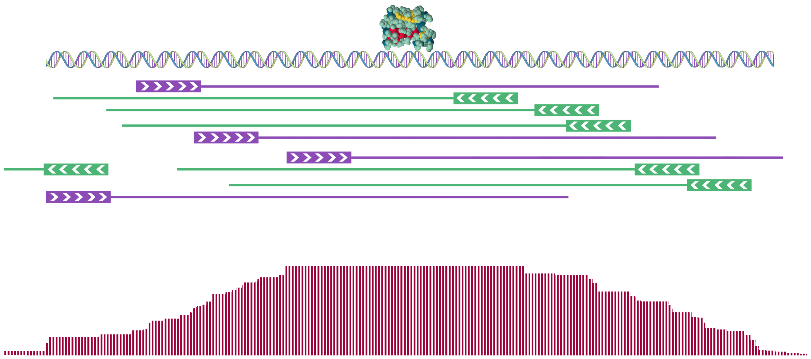
Run optimization to report maximal number of peaks within a given number of false positives

Identification of candidate peaks

Maped reads → FindPeaks

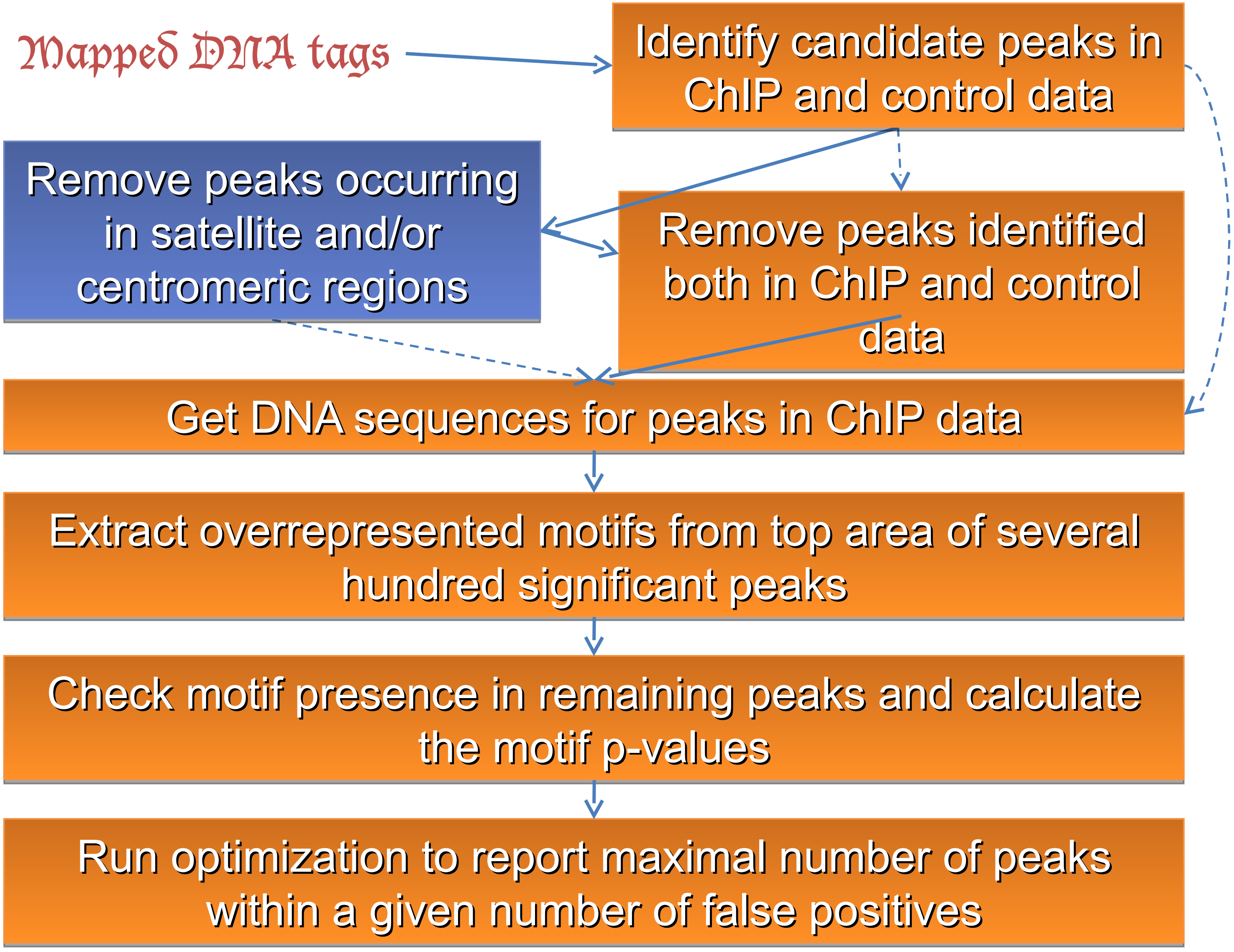
to detect putative binding sites

Main principle of FindPeaks:



FindPeaks automatically excludes duplicate reads from the analysis.

Main steps of the MICSA pipeline



Mapped DNA tags

Identify candidate peaks in ChIP and control data

Remove peaks occurring in satellite and/or centromeric regions

Remove peaks identified both in ChIP and control data

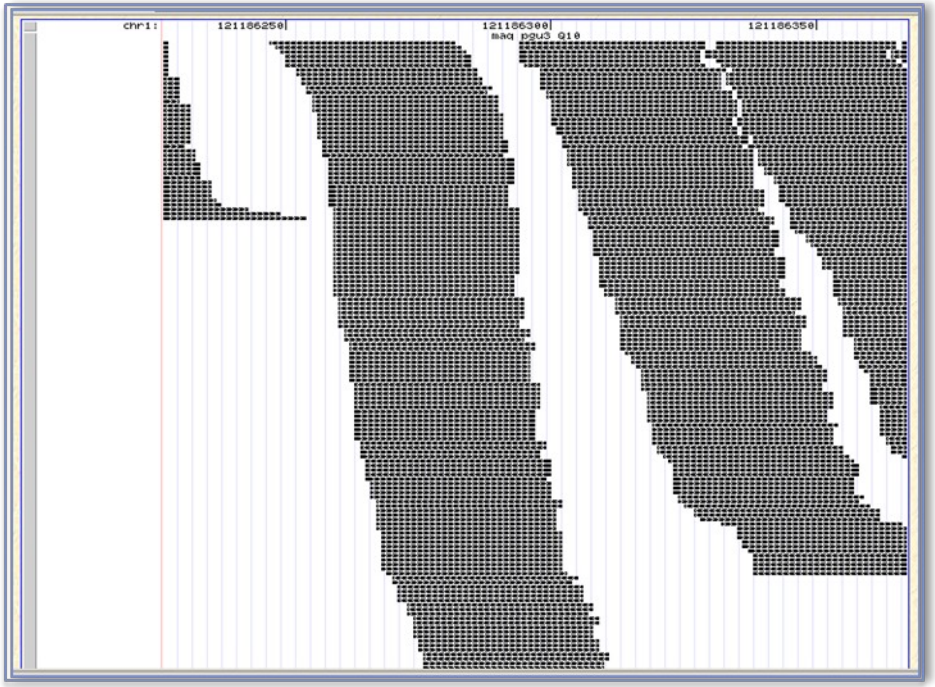
Get DNA sequences for peaks in ChIP data

Extract overrepresented motifs from top area of several hundred significant peaks

Check motif presence in remaining peaks and calculate the motif p-values

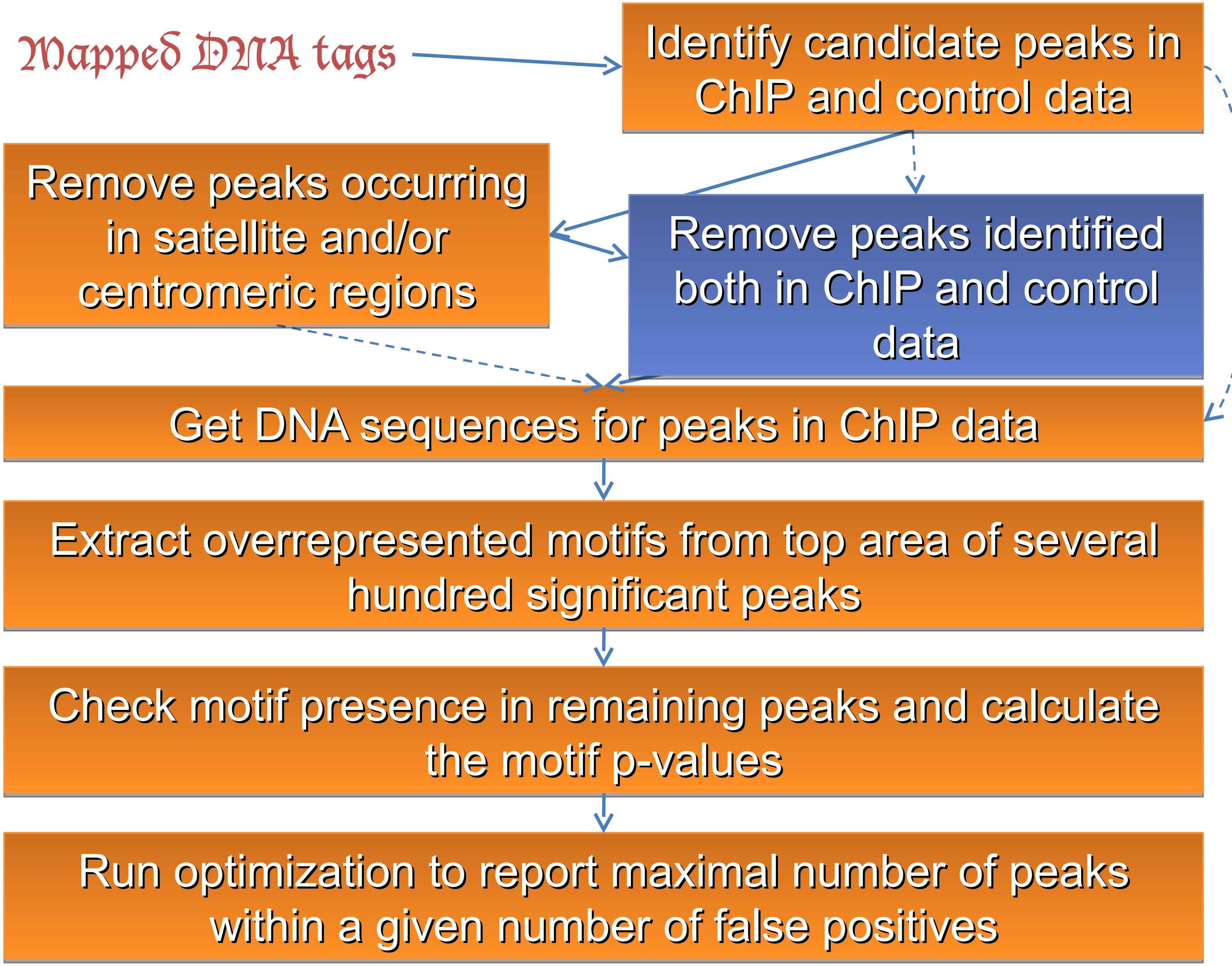
Run optimization to report maximal number of peaks within a given number of false positives

Alignment bias in satellites regions affects peak calling



- Tag enrichment in alpha-satellite peri-centromeric region in the control dataset
- Same enrichment observed in the ChIP dataset

Main steps of the MICSA pipeline



Mapped DNA tags

Identify candidate peaks in ChIP and control data

Remove peaks occurring in satellite and/or centromeric regions

Remove peaks identified both in ChIP and control data

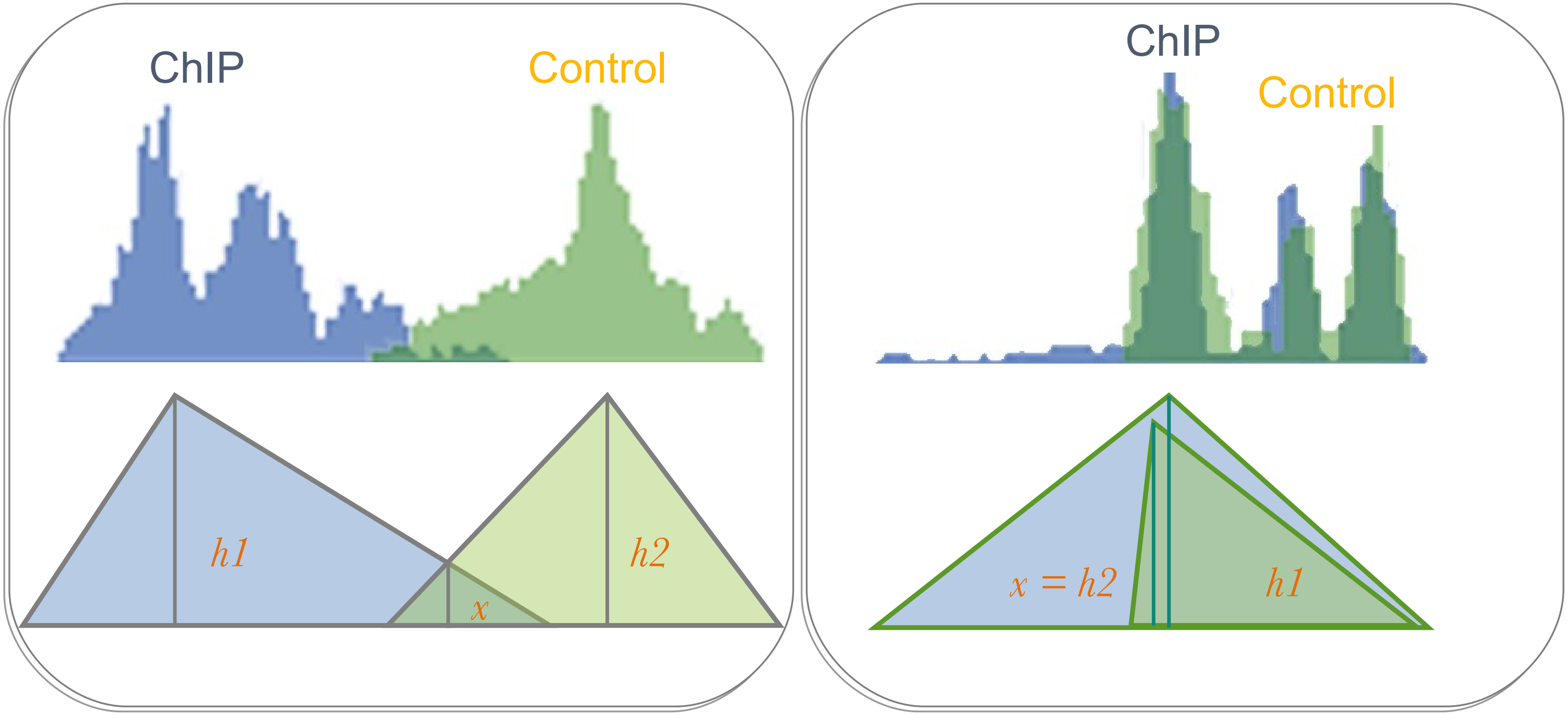
Get DNA sequences for peaks in ChIP data

Extract overrepresented motifs from top area of several hundred significant peaks

Check motif presence in remaining peaks and calculate the motif p-values

Run optimization to report maximal number of peaks within a given number of false positives

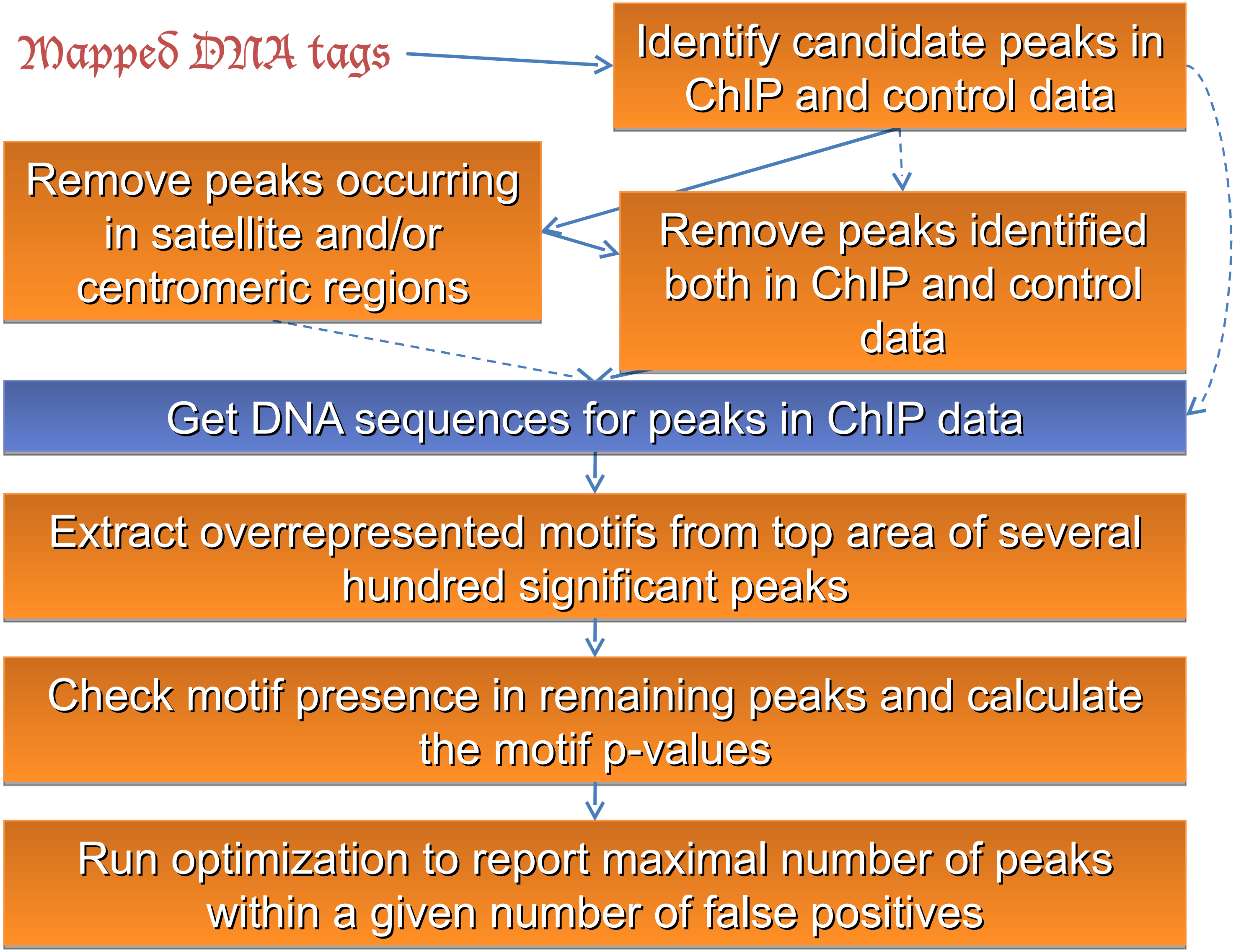
Filtering using control peaks.



$h1/x > 2?$ ➡ Keep the peak

- The actual peak shapes → triangles (start, end, maximum and height).
- Then, the height (x) of maximal overlap is calculated.
- The ChIP peak is rejected if its height ($h1$) divided by x is less than or equal to 2.

Main steps of the MICSA pipeline



Mapped DNA tags

Identify candidate peaks in ChIP and control data

Remove peaks occurring in satellite and/or centromeric regions

Remove peaks identified both in ChIP and control data

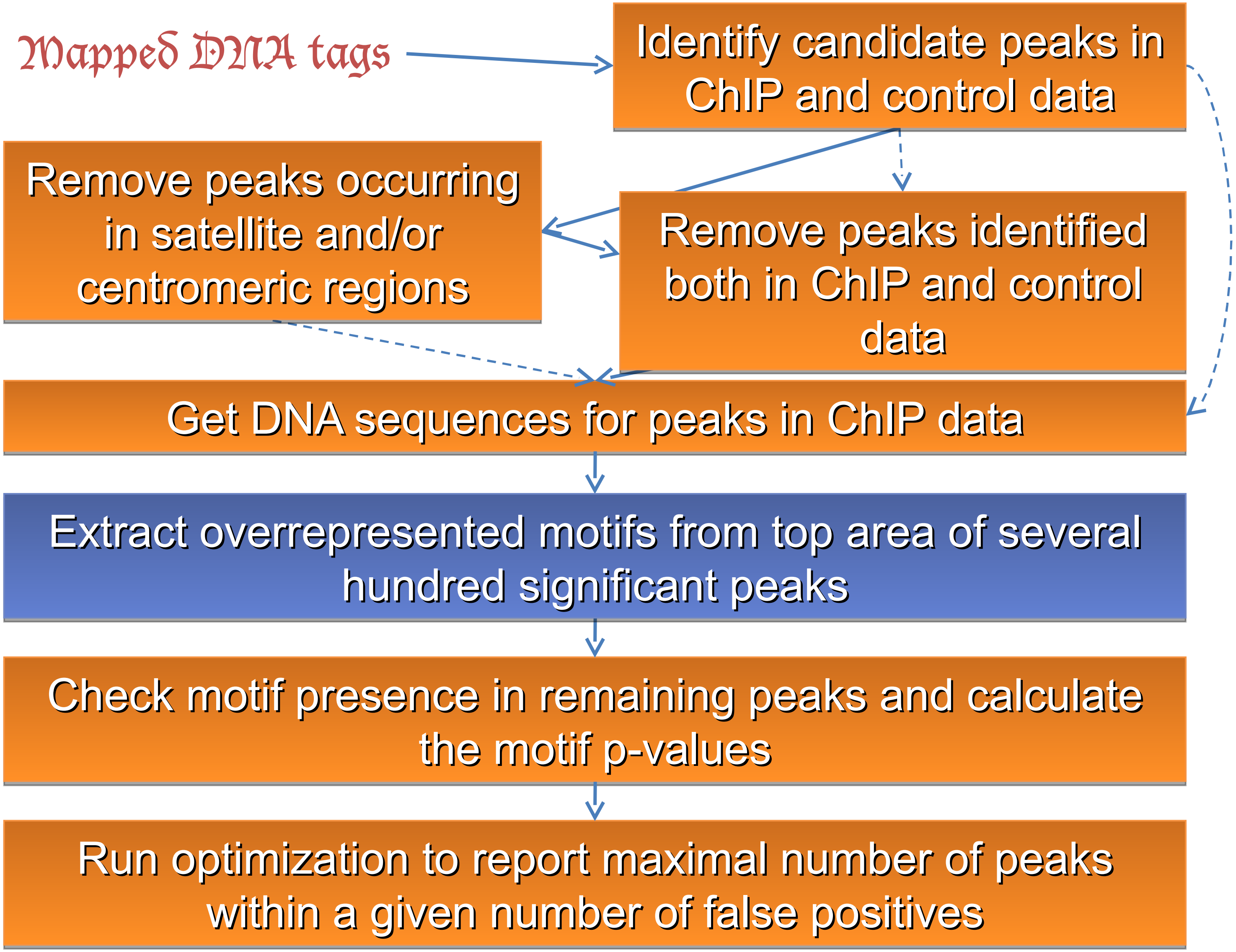
Get DNA sequences for peaks in ChIP data

Extract overrepresented motifs from top area of several hundred significant peaks

Check motif presence in remaining peaks and calculate the motif p-values

Run optimization to report maximal number of peaks within a given number of false positives

Main steps of the MICSA pipeline



Mapped DNA tags

Identify candidate peaks in ChIP and control data

Remove peaks occurring in satellite and/or centromeric regions

Remove peaks identified both in ChIP and control data

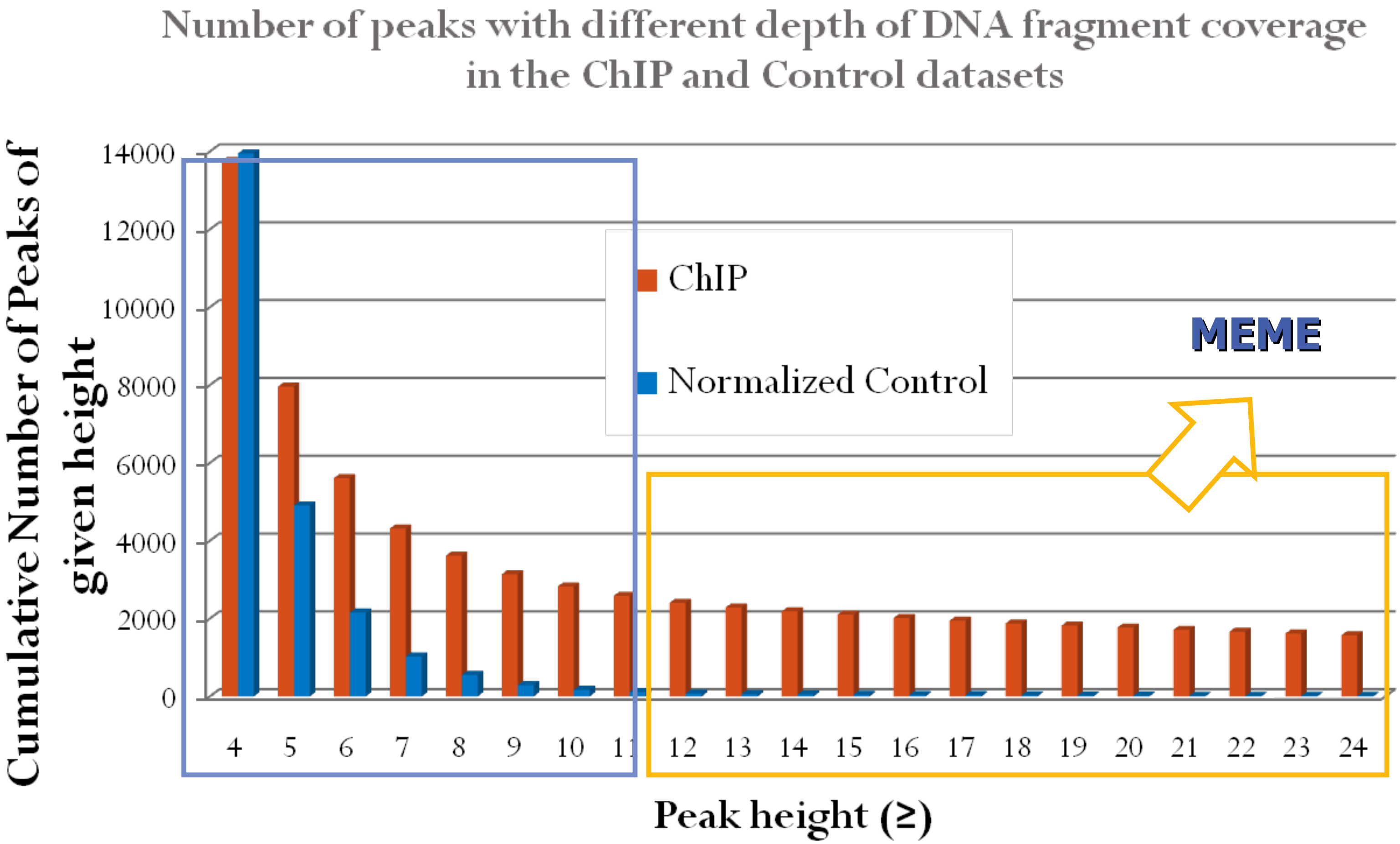
Get DNA sequences for peaks in ChIP data

Extract overrepresented motifs from top area of several hundred significant peaks

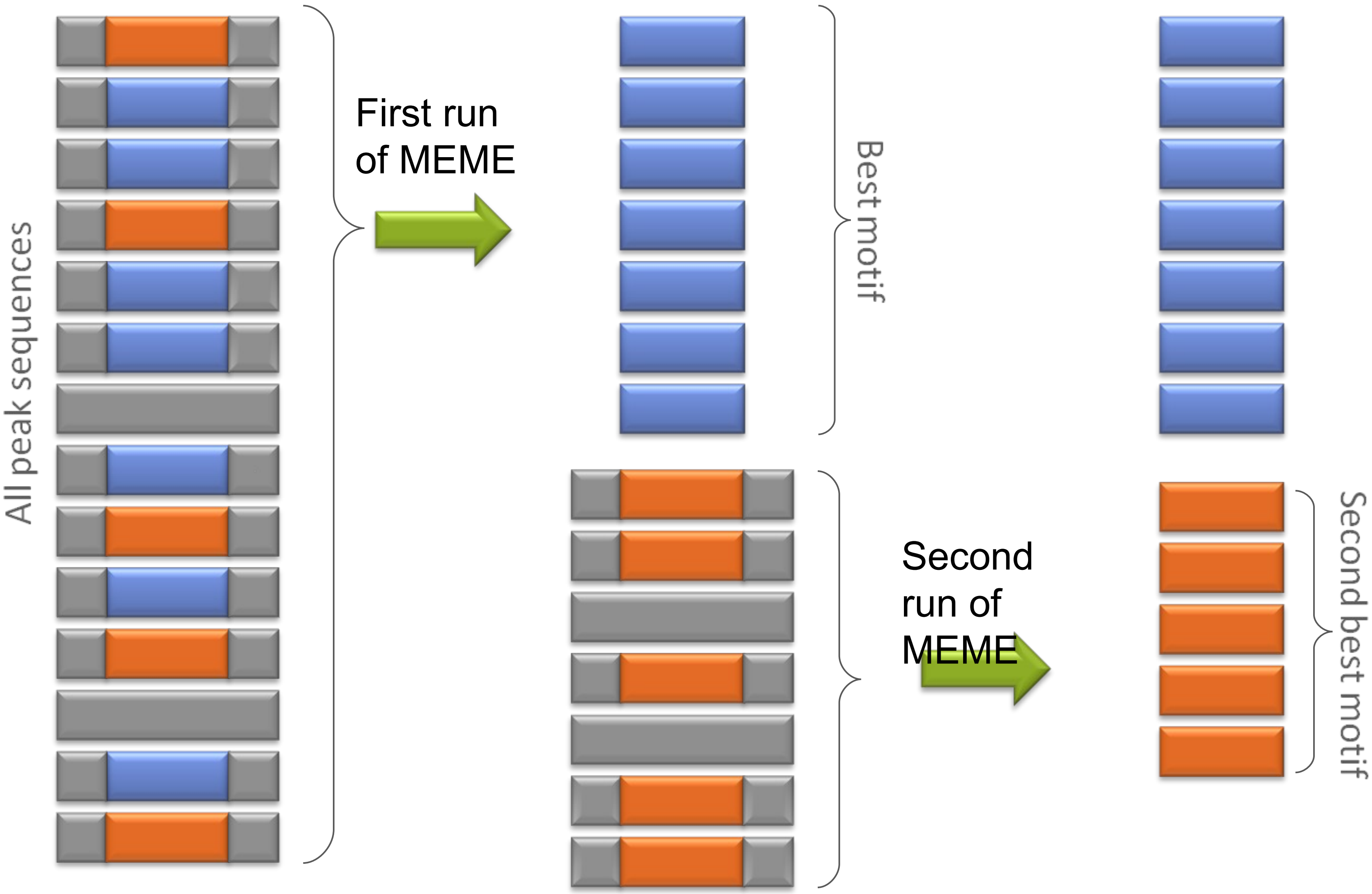
Check motif presence in remaining peaks and calculate the motif p-values

Run optimization to report maximal number of peaks within a given number of false positives

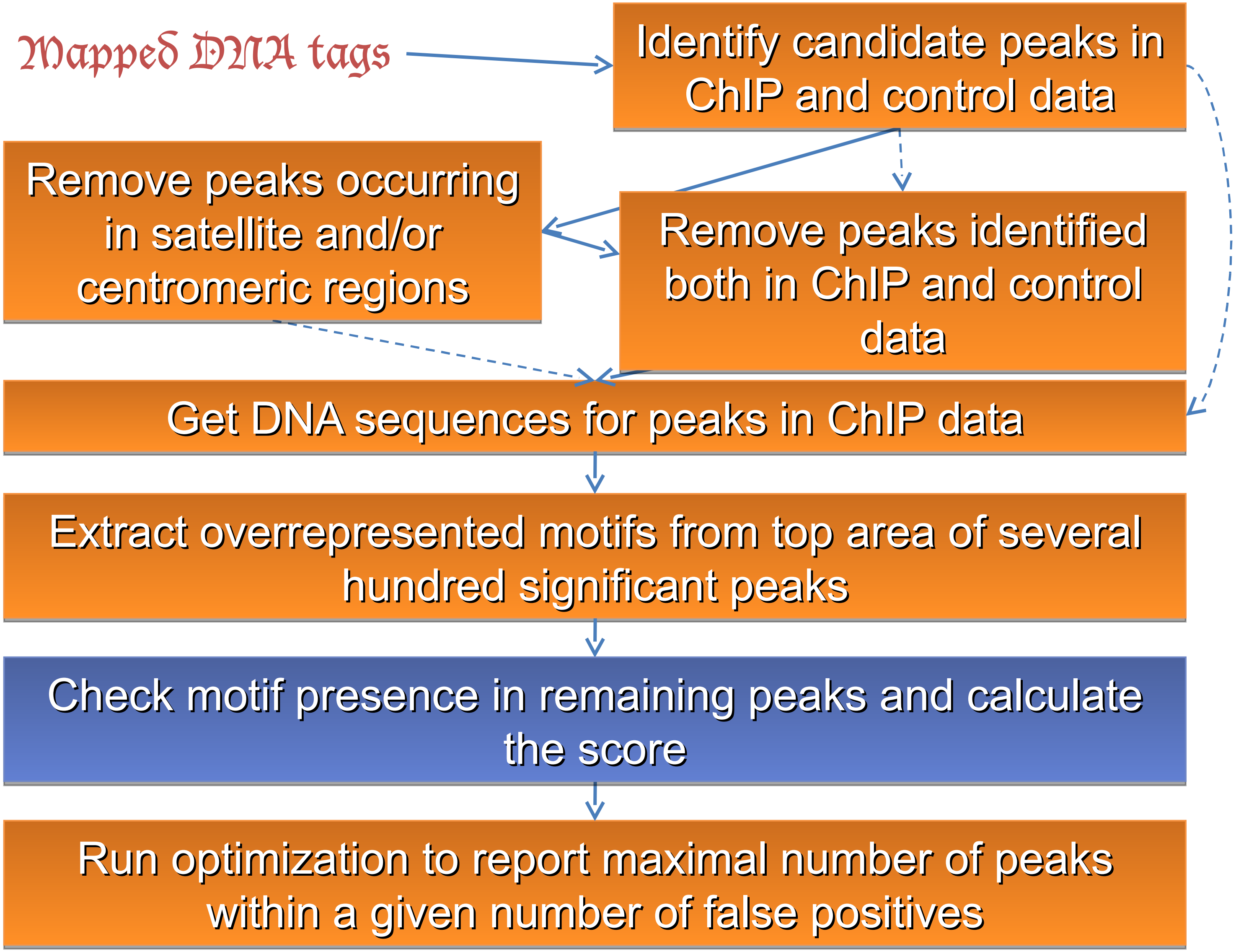
High peaks are confident while low peaks are more likely to be false



Running MEME to identify multiple motifs



Main steps of the MICSA pipeline



Mapped DNA tags

Identify candidate peaks in ChIP and control data

Remove peaks occurring in satellite and/or centromeric regions

Remove peaks identified both in ChIP and control data

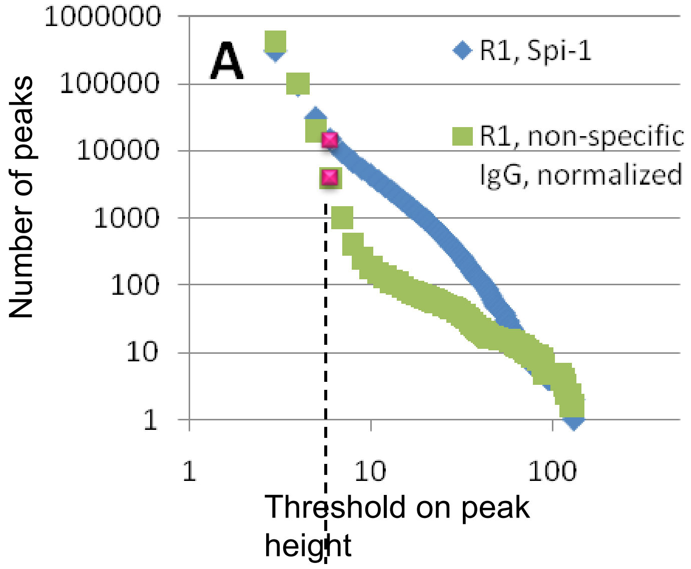
Get DNA sequences for peaks in ChIP data

Extract overrepresented motifs from top area of several hundred significant peaks

Check motif presence in remaining peaks and calculate the score

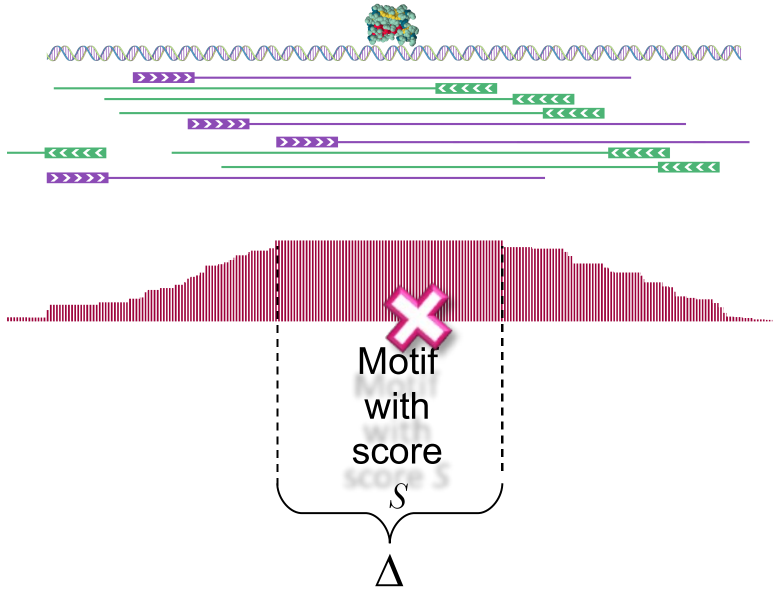
Run optimization to report maximal number of peaks within a given number of false positives

Score calculation in peaks



Threshold on peak height

$$FDR = \frac{\text{\# peaks in the contol}}{\text{\# peaks in the sample}}$$



Motif with score S

Δ

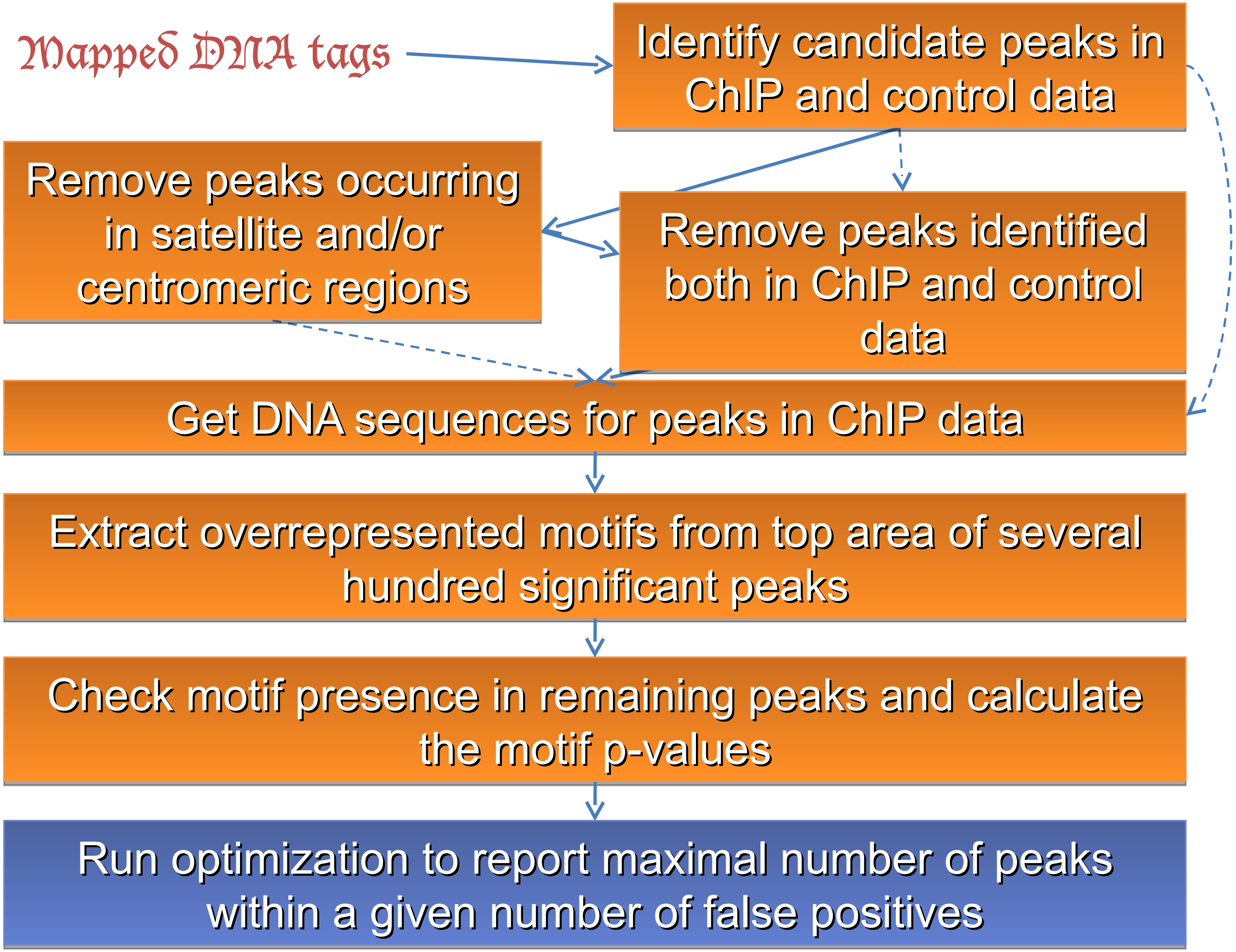
Under random model of nucleotide distribution:

$$p\text{-}value \approx 1 - \left(1 - \text{MotifProbability}(S)\right)^{\Delta - MotifLength + 1}$$

×

Final score = -log(FDR x p-value)

Main steps of the MICSA pipeline



Mapped DNA tags

Identify candidate peaks in ChIP and control data

Remove peaks occurring in satellite and/or centromeric regions

Remove peaks identified both in ChIP and control data

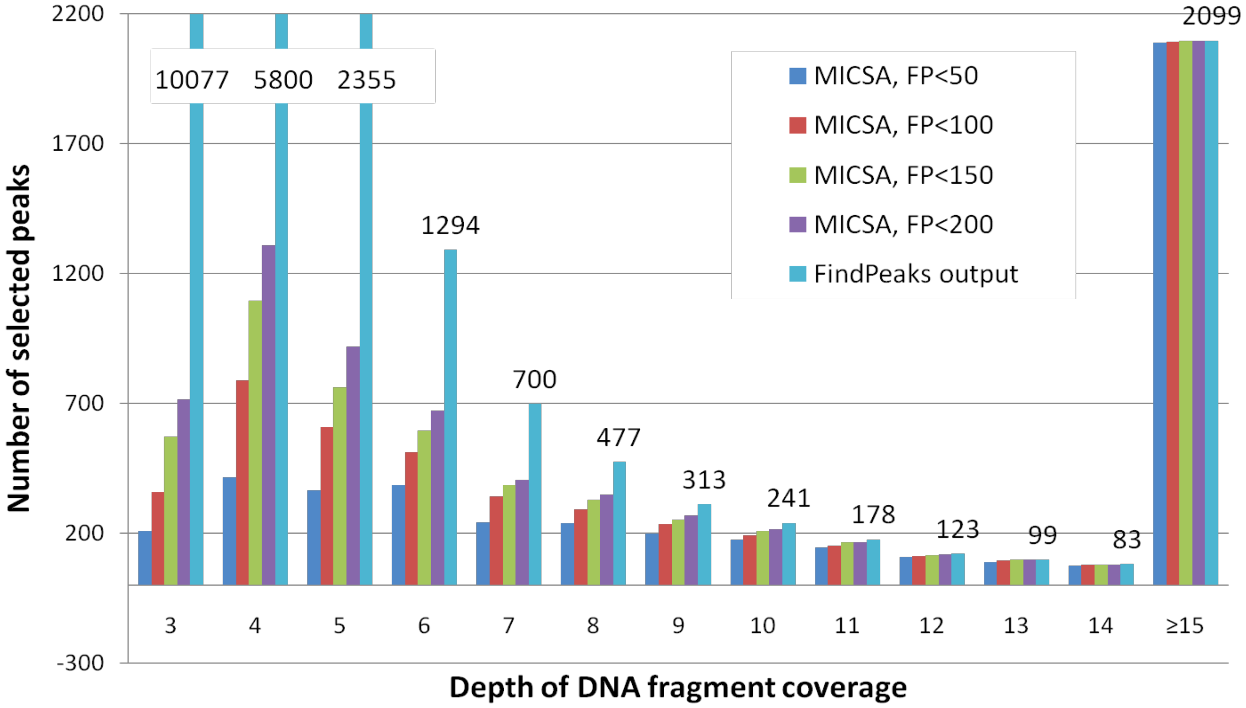
Get DNA sequences for peaks in ChIP data

Extract overrepresented motifs from top area of several hundred significant peaks

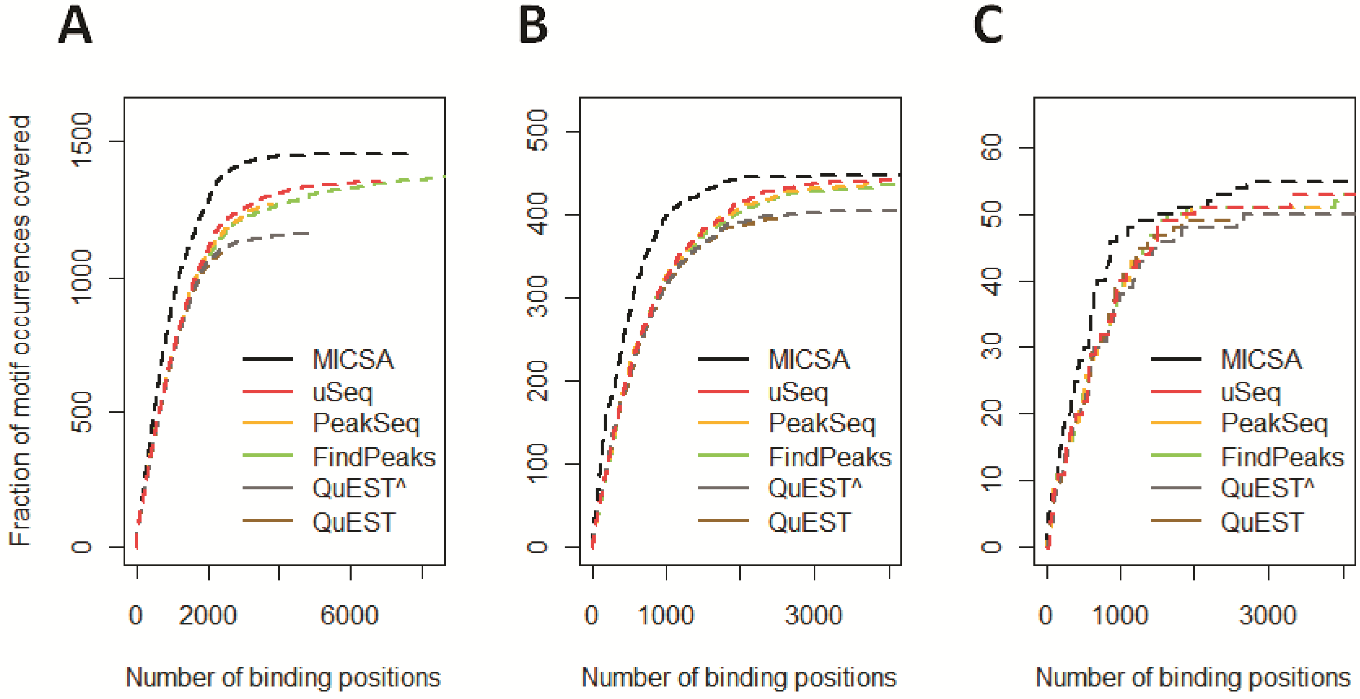
Check motif presence in remaining peaks and calculate the motif p-values

Run optimization to report maximal number of peaks within a given number of false positives

MICSA keeps all high peaks but eliminate low peaks without motifs



MICSA's performance in identification of binding sites with motifs (ChIP-Seq data for NRSF)

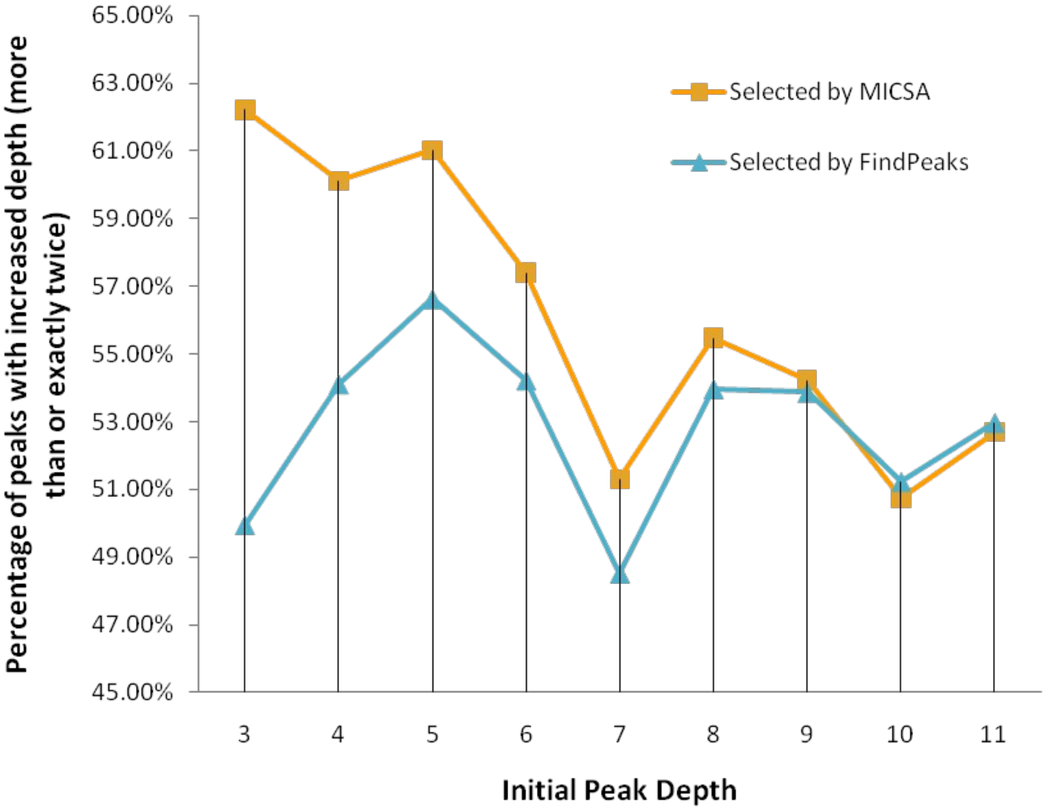


Positive set of binding sites of NRSF:

A. 3,000 best matches of the canonical NRSF matrix in the human genome
B. 500 best matches of the canonical NRSF matrix in the human genome
C. 83 q-PCR verified NRSF binding sites in the human genome

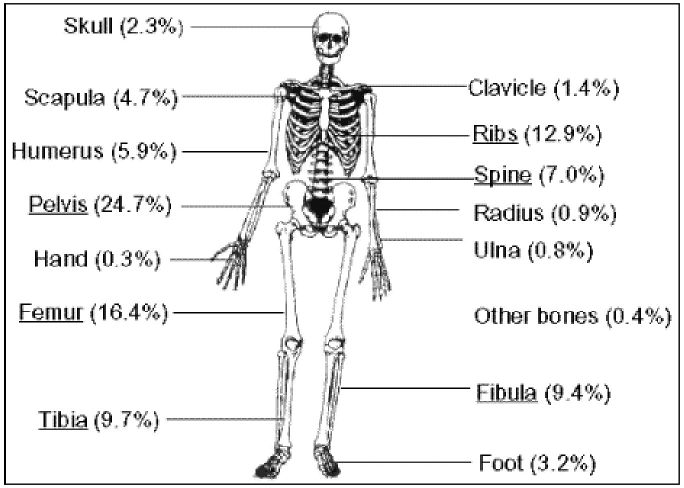
Peaks selected by MICSA are more reproducible than those selected by FindPeaks

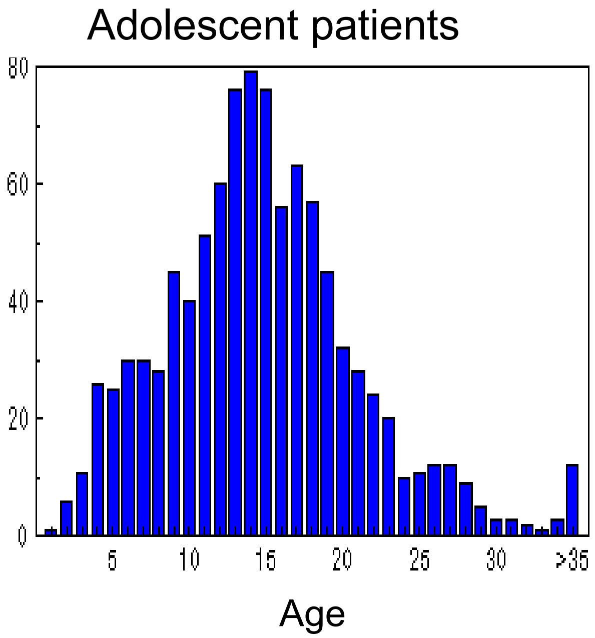
- Low depth of sequencing NRSF dataset *vs* high depth of sequencing NRSF dataset
- How many peaks



An example of ChIP-Seq analysis: EWS-FLI1 (O.Delattre team)

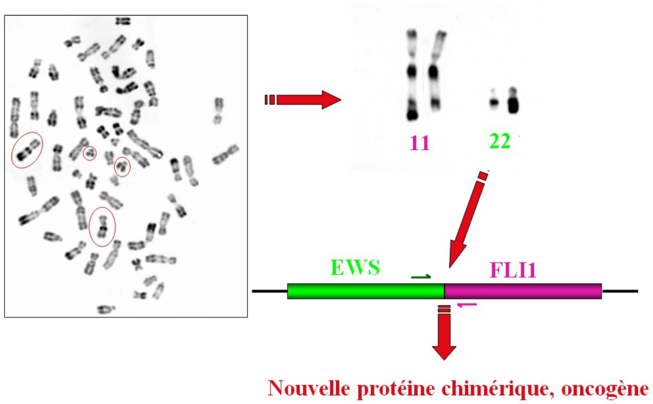
- EWS-FLI1 oncogenic transcription factor – cause of Ewing sarcoma.



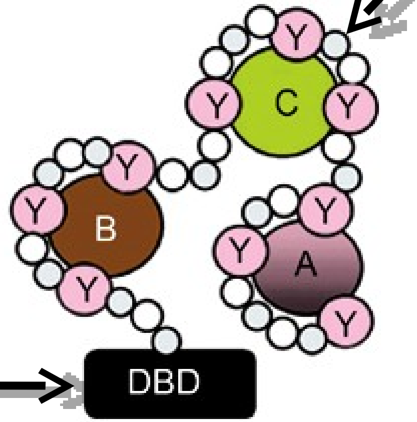


An example of ChIP-Seq analysis: EWS-FLI1 (O.Delattre team)

- EWS-FLI1 oncogenic transcription factor – cause of Ewing sarcoma.



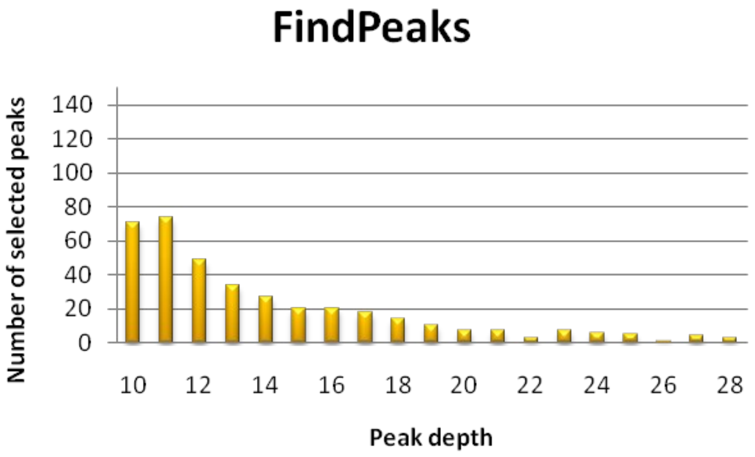
EWS-activation-domain

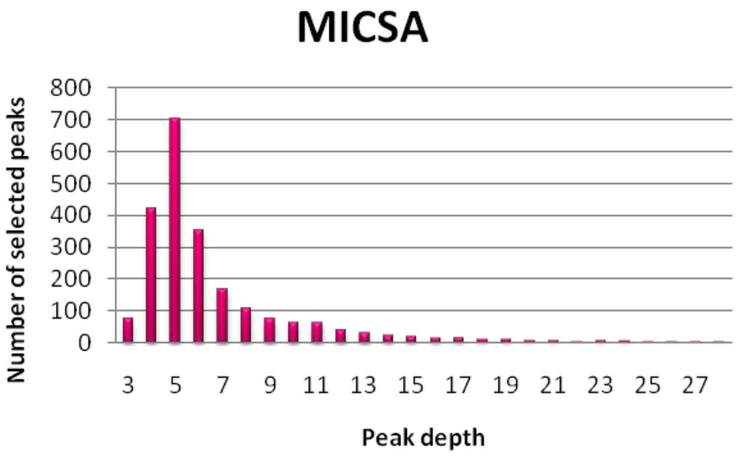


FLI1 DNA binding domain

Adopted from: Ewings family oncoproteins: drunk, disorderly and in search of partners
Kevin A. W. Lee

With the same false positive rate MICSA selected more peaks than FindPeaks



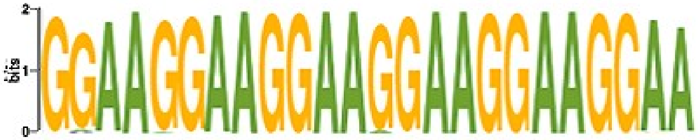


- 412 peaks
- with 20% false discovery rate

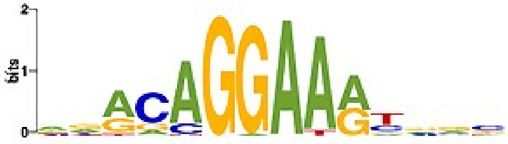
- 2264 peaks
- with 5% false discovery rate

MICSA identified two binding motifs for EWS-FLI1

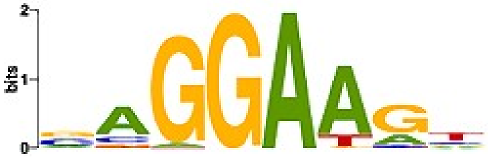
2 motifs
by MICSA:



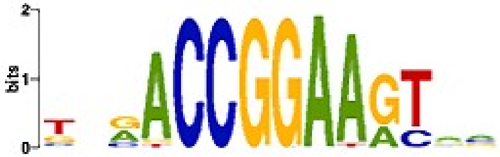
EWS-FLI1



EWS-FLI1



Canonical ETSF

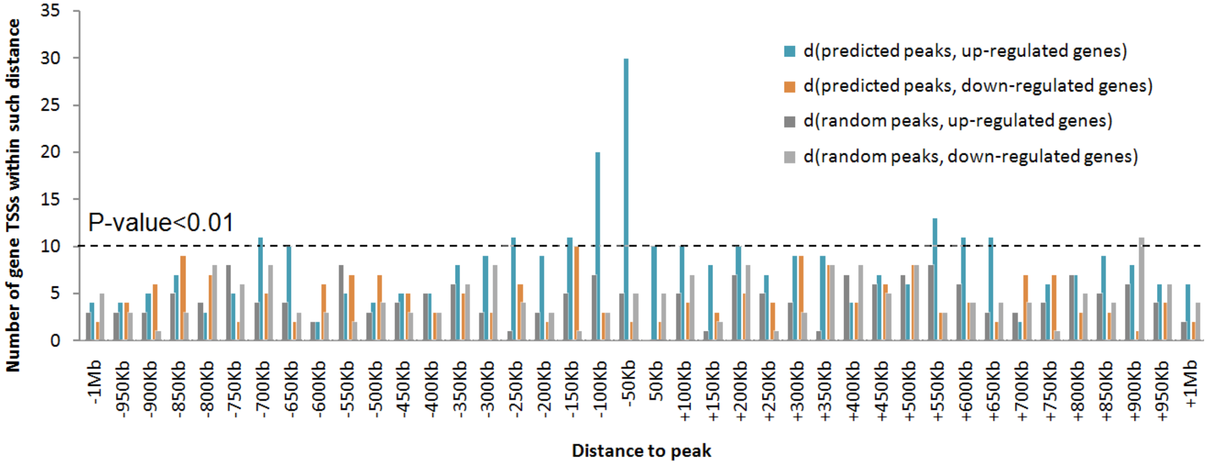


Canonical FLI1

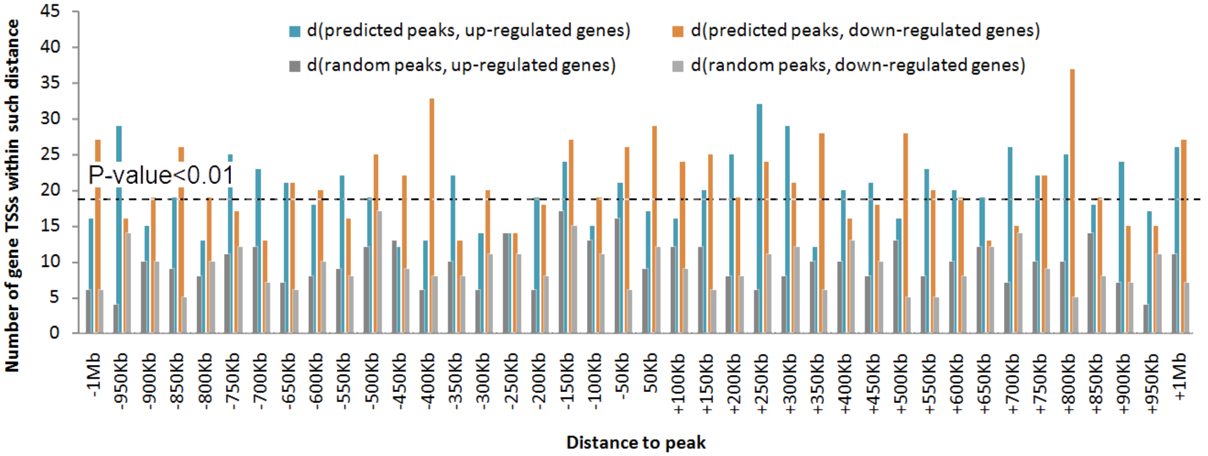
Crossing peaks with gene expression data

Distances between predicted/random peaks and genes up-/down-regulated by EWS-FLI1.

- Sites containing (GGAA)n microsatellites :



- ETS-like sites (site without microsatellites):



We identified putative direct targets of EWS-FLI1

Gene	Distance to TSS	Fold change	Site type	Gene	Distance to TSS	Fold change	Site type	Gene	Distance to TSS	Fold change	Site type
ABHD6	-6293	8.63	microsatellite	HIPK1	-8866.5	2.08	microsatellite	RAD23A	-28705	2.60	ETS-like
ACTN2	-31613.5	2.20	microsatellite	HSPA14	15762	2.80	ETS-like	RBM28	-21792	2.73	microsatellite
AKAP7	-20331	18.63	microsatellite	IGFBP5	83419	-34.78	ETS-like	RCC1	-126622	2.38	ETS-like
ANGPTL4	-146409	-2.47	ETS-like	IL1RAP	8451.5	2.93	microsatellite	RCN3	-16133.5	2.69	microsatellite
AQP1	-75750	-2.51	ETS-like	IL6ST	-118855	-2.78	ETS-like	RDX	-58104	2.27	microsatellite
BCHE	101740	-7.11	ETS-like	INSIG1	116923	-2.87	ETS-like	RFXAP	-124733	2.04	ETS-like
BCL6	-81669	5.48	ETS-like	ISG15	95685	-4.87	ETS-like	RGS3	-139611	-6.77	ETS-like
BCL7A	98390	2.47	ETS-like	JARID2	-137865.5	2.41	microsatellite	RNASEH2A	110521	2.75	ETS-like
C12orf11	134494	2.29	ETS-like	KDELR1	-101340	-2.67	ETS-like	RRAGD	-71287	-3.93	ETS-like
C13orf34	-28718	2.04	ETS-like	KLHL23	110393	2.17	ETS-like	RRM1	47954	2.50	ETS-like
C16orf68	-56414	2.16	microsatellite	LBH	-8382.5	12.92	microsatellite	RRN3	85402	2.19	ETS-like
C1orf112	53912	3.52	ETS-like	LBH	-104152	12.92	microsatellite	S100A13	-136384	-22.80	ETS-like
C6orf130	63924	2.10	ETS-like	LBR	40276	2.92	ETS-like	SALL2	-120208	8.24	microsatellite
CA12	71941	-2.25	ETS-like	LMO2	-144250	-4.07	ETS-like	SAT1	21894	-5.55	ETS-like
CADPS2	-84683.5	2.23	microsatellite	LMO3	-118377.5	5.68	microsatellite	SDC2	62422	-4.65	ETS-like
CAND1	99147	3.10	ETS-like	LTBP1	93962	-3.35	ETS-like	SERP1	-53971	3.00	ETS-like
CAV2	24086.5	6.56	microsatellite	MAN2A1	-78698	4.99	microsatellite	SFRS10	-77123.5	2.72	microsatellite
CAV2	-30668.5	6.56	microsatellite	METTL3	-146088	2.61	microsatellite	SGCB	-5466	-2.59	ETS-like
CCK	-336	6.84	ETS-like	MMP1	-89598	-2.76	ETS-like	SHFM1	-37867.5	2.04	microsatellite
CCND1	-18880.5	3.75	microsatellite	MMP2	-47451	-28.34	ETS-like	SHFM1	-93935	2.04	ETS-like
CCNF	-100162	2.25	ETS-like	MPHOSPH10	-112643	2.24	ETS-like	SKP2	-83175	3.64	microsatellite
CD58	-123809	-4.83	ETS-like	MPP5	39525.5	3.91	microsatellite	SLC24A3	10004.5	9.12	microsatellite
CDC25A	-634	2.58	ETS-like	MRPS15	39498	2.13	ETS-like	SLC24A3	-66227	9.12	ETS-like
CDC34	91307	-2.49	ETS-like	MVP	-63949	-6.33	ETS-like	SLC26A2	-24353	14.13	microsatellite
CDC34	-96707	-2.49	ETS-like	MYBL1	-49592	-8.96	ETS-like	SLC26A2	-39409	14.13	microsatellite
CENPA	-135353	3.44	ETS-like	MYST3	-144722	2.57	ETS-like	SLC2A4RG	6763	-2.37	ETS-like
CENPE	124534	2.08	ETS-like	NAGK	-50646	-3.96	ETS-like	SLCO5A1	-36534	2.50	microsatellite
CGGBP1	32447	2.31	ETS-like	NDUFB5	43965	2.19	ETS-like	SMARCA4	76205	2.21	ETS-like
CLEC11A	15675	2.36	microsatellite	NDUFS1	-17895	2.73	ETS-like	SMARCC1	-125008	4.26	ETS-like
CPB2	-11862	2.75	microsatellite	NETO2	-15118	-4.33	ETS-like	SNAPC1	42994	-2.25	ETS-like
CXADR	124621	3.99	ETS-like	NEU1	66972	-2.35	ETS-like	SNW1	-25845	2.33	ETS-like
CYP1B1	148930	3.81	ETS-like	NGDN	86534	2.26	ETS-like	SNW1	-81327	2.33	microsatellite
DAPK1	-15271.5	11.04	microsatellite	NKX2-2	-62388.5	15.57	microsatellite	SORD	14945.5	2.68	microsatellite
DAPK1	-94919.5	11.04	microsatellite	NMI	-40044.5	4.28	microsatellite	SORD	-111781.5	2.68	microsatellite
DAZAP1	-126495	2.01	ETS-like	NOLC1	-35363	2.67	ETS-like	SSBP2	-131016	-2.64	ETS-like
DCLRE1A	-131460	5.68	microsatellite	NR3C1	6879	-3.91	ETS-like	STOM	-34978	-3.27	ETS-like
DDAH2	-65671	-3.15	ETS-like	NRP1	21821	-17.44	ETS-like	TARDBP	-90110	2.16	ETS-like
DHCR24	37090	4.83	ETS-like	NUDT11	-20951	2.91	ETS-like	TBC1D15	37916	4.49	microsatellite
DHX29	29399	4.34	ETS-like	NUDT3	58541	2.24	ETS-like	TCERG1	-82817.5	3.16	microsatellite
DHX29	-64608	4.34	ETS-like	NUP205	108706	2.36	ETS-like	TCF12	-16519.5	2.22	microsatellite
DHX29	-64608	4.34	ETS-like	OLFML3	-129152.5	3.77	microsatellite	TFPI	-103495	-5.03	ETS-like
DKK1	-77194	-11.55	ETS-like	PAPD1	-34891	2.13	ETS-like	THY1	-102883	-2.47	ETS-like
DLGAP4	40674	-2.01	ETS-like	PAPPA	-30473	3.06	microsatellite	TJP2	-16898	-9.72	ETS-like
ECT2	-58486	5.43	ETS-like	PCCB	-33830	4.73	ETS-like	TMEM106C	39488	-3.68	ETS-like
EHD2	27405	-4.57	ETS-like	PCSK2	-57737	27.36	microsatellite	TMEM48	-71586	3.67	ETS-like
EMP1	29531	-4.59	ETS-like	PFKM	-104281	4.09	ETS-like	TMSL8	-87858.5	11.89	microsatellite
EPB41L2	-61806	3.98	microsatellite	PHF16	108126	5.24	ETS-like	TNC	105435	-7.90	ETS-like
EXOSC7	-13646	5.18	microsatellite	PIR	-133116	5.63	ETS-like	TNFAIP6	-27642.5	22.33	microsatellite

Conclusions

- MICSA: specially developed to analyze Chip-Seq data for transcription factors
- allows identification of binding sites with greater sensitivity
- Can identify several binding motifs
- Has user friendly graphical interface

Authors

Valentina Boeva U830, U900 Institut Curie/INSERM/Ecole des Mines

Institut Curie, INSERM, U830, Génétique et Biologie des Cancers:

Didier Surdez Noëlle Guillon Franck Tirode Olivier Delattre

Institut Curie, INSERM, U900, Cancer et Génome : bioinformatique, biostatistiques et épidémiologie d'un système complexe:

Emmanuel Barillot

Genome Sciences Centre, BC Cancer Agency, Canada: Anthony Fejes

Thanks

Andrei Zinovyev, Peter Kharchenko