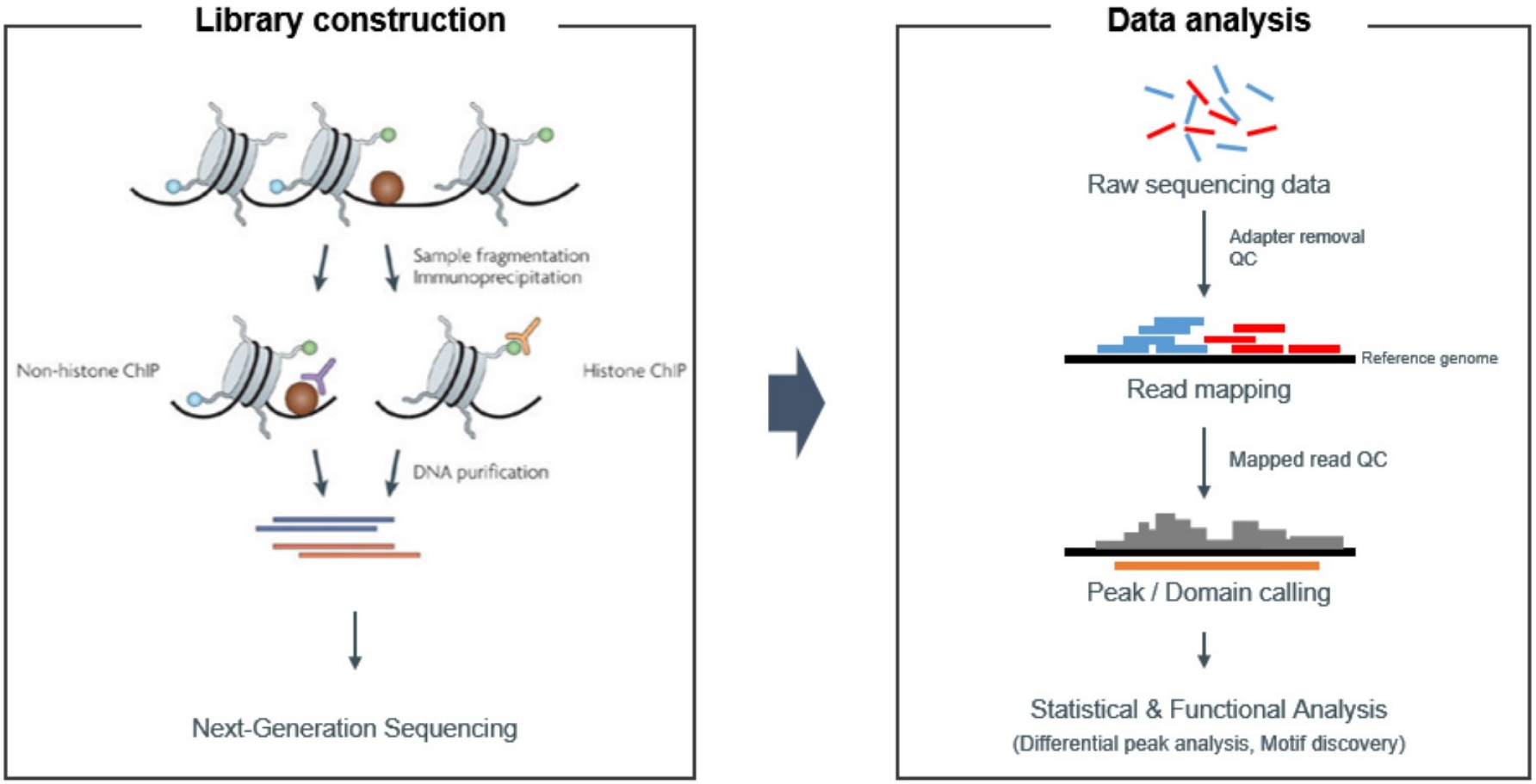
Peak enrichment evaluation of histone modification ChIP-Seq data

Tae-Young Roh
Laboratory of System Genomics
Div. of Integrative Biosciences and Biotechnology
Dept. of Life sciences
Pohang University of Science and Technology (POSTECH)

Overview of ChIP-Seq

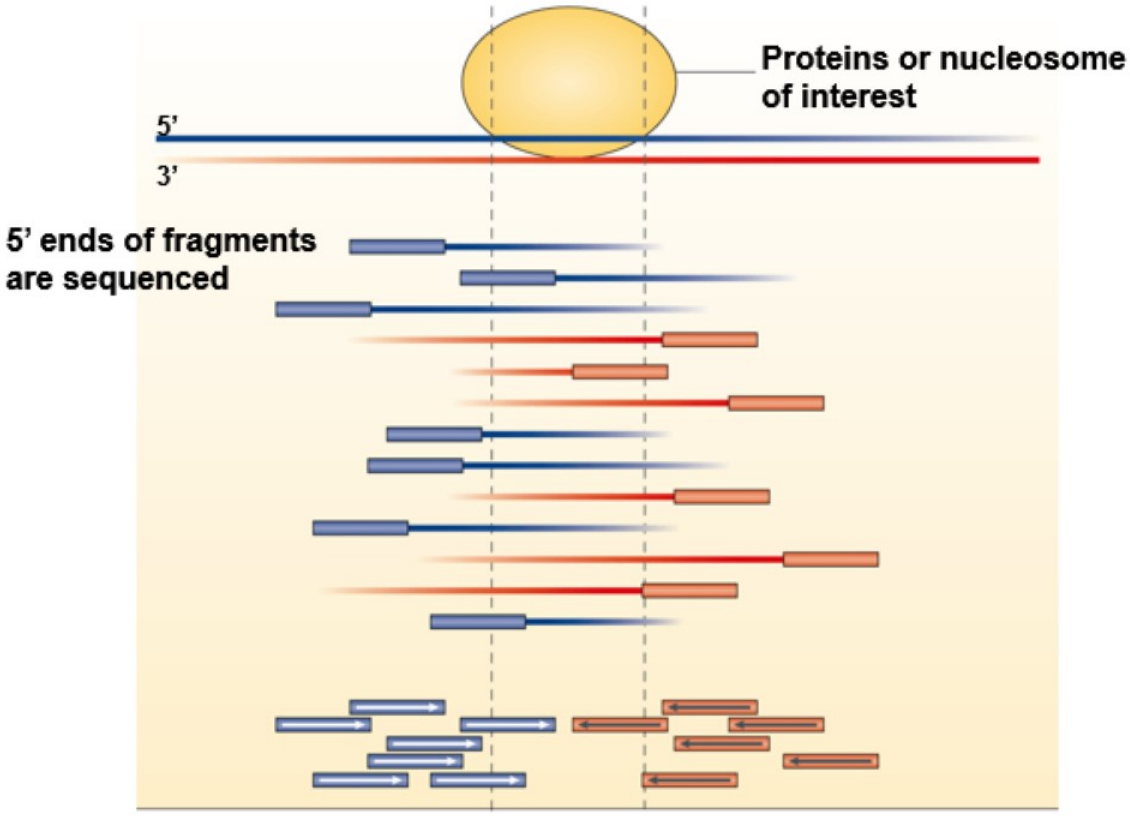
Chromatin ImmunoPrecipitation coupled with Sequencing



Park PJ. ChIP–seq: advantages and challenges of a maturing technology
Nature Reviews Genetics **10**, 669–680 (2009)

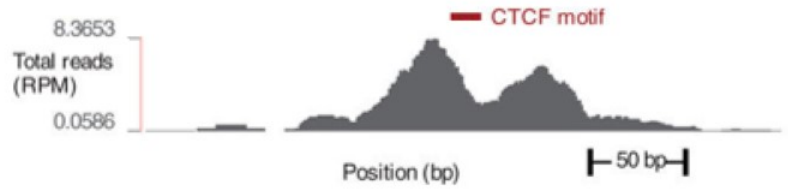
Profiles of Sequenced read at enriched sites

Sequenced reads from ChIP-Seq are distributed around binding positions



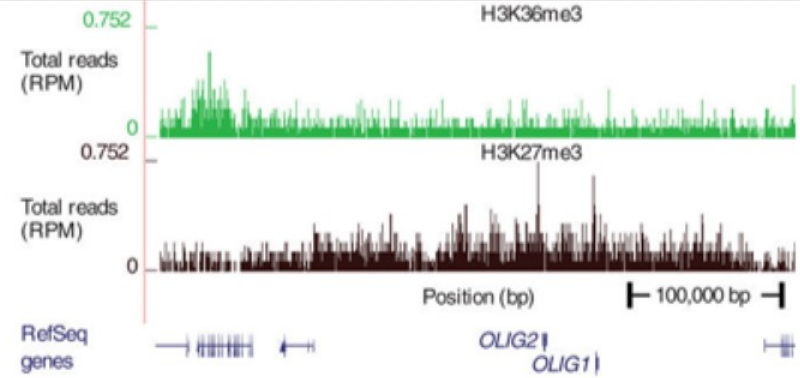
Park, Peter J, Nat. Rev. Gene., 2009

Classes of DNA-bound proteins



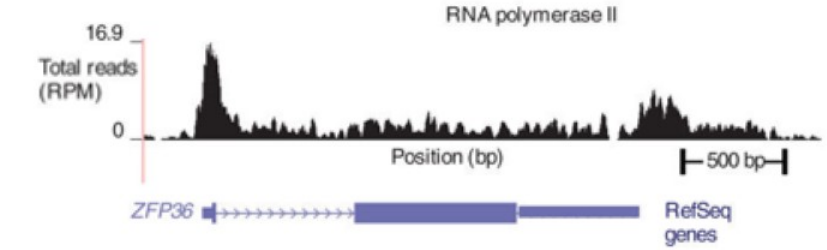
Point-source factors

Peak
(100bp ~ 1kbp)



Broad-source factors

Domain
(kbp ~ Mbp)



Mixed-source factors

Pepke S., Wold B., Mortazavi A. Computation for ChIP-seq and RNA-seq studies
Nature Methods **6**, S22–S32 (2009)

Peak calling programs

50 peak calling programs are available to identify enriched regions of interest protein binding sites

Design and analysis of ChIP-seq experiments for DNA-binding proteins

Peter V Kharchenko[1–3], Michael Y Tolstorukov[1,2] & Peter J Park[1–3]

An integrated software system for analyzing ChIP-chip and ChIP-seq data

Hongkai Ji[1], Hui Jiang[2], Wenxiu Ma[3], David S Johnson[4,8], Richard M Myers[5] & Wing H Wong[6,7]

FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology

Anthony P. Fejes[1,*], Gordon Robertson[1], Mikhail Bilenky[1], Richard Varhol[1], Matthew Bainbridge[2] and Steven J. M. Jones[1,*]

HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data

Zhaohui S Qin*[1,2,3], Jianjun Yu[3,4], Jincheng Shen[1], Christopher A Maher[2,3,4], Ming Hu[1], Shanker Kalyana-Sundaram[3,4], Jindan Yu[5] and Arul M Chinnaiyan[2,3,4,6,7,8]

Model-based Analysis of ChIP-Seq (MACS)

Yong Zhang[¤*], Tao Liu[¤*], Clifford A Meyer[*], Jérôme Eeckhoute[†], David S Johnson[‡], Bradley E Bernstein[§¶], Chad Nusbaum[¶], Richard M Myers[¥], Myles Brown[†], Wei Li[#] and X Shirley Liu[*]

PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls

Joel Rozowsky[1], Ghia Euskirchen[2], Raymond K Auerbach[3], Zhengdong D Zhang[1], Theodore Gibson[1], Robert Bjornson[4], Nicholas Carriero[4], Michael Snyder[1,2] & Mark B Gerstein[1,3,4]

BayesPeak: Bayesian analysis of ChIP-seq data

Christiana Spyrou*[1,3], Rory Stark[3], Andy G Lynch[4] and Simon Tavaré[2,4]

Genome-wide identification of *in vivo* protein–DNA binding sites from ChIP-Seq data

Raja Jothi, Suresh Cuddapah, Artem Barski, Kairong Cui and Keji Zhao*

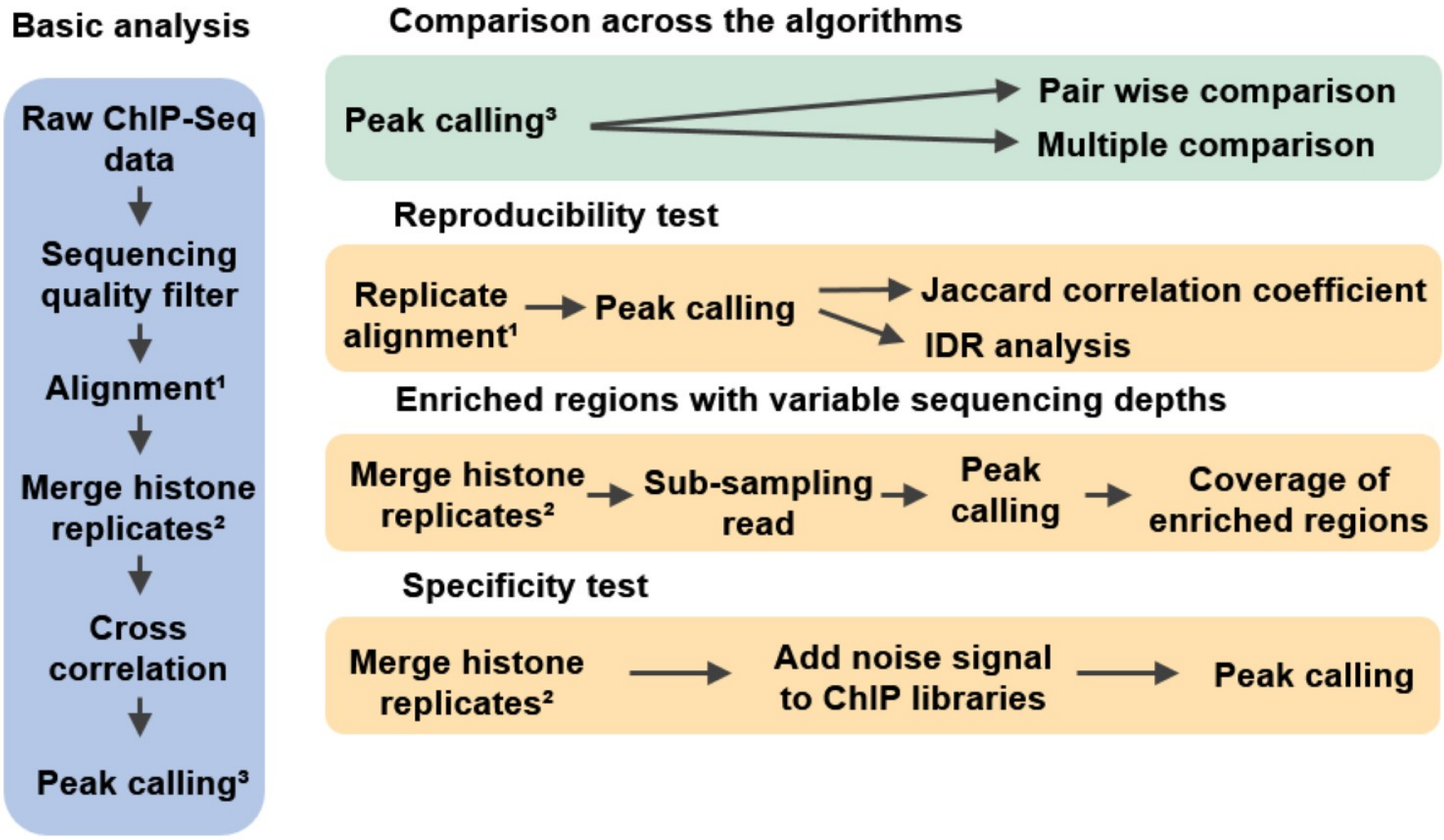
Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data

Kimberly R. Blahnik[1], Lei Dou[1], Henriette O'Geen[1], Timothy McPhillips[1], Xiaoqin Xu[1], Alina R. Cao[1], Sushma Iyengar[1], Charles M. Nicolet[1], Bertram Ludäscher[1,2], Ian Korf[1,3] and Peggy J. Farnham[1,4,*]

Peak calling programs

Program	Tag density profile	Peak definition	Significance
CisGenome (ver. 2.0)	Window sliding (bin size = 100bp)	Peak height and Fold enrichment	FDR (<0.01)
MACS1 (ver. 1.4.2)	Window sliding (predicted)	Peak height and Fold enrichment (10 < fold change < 30)	P-value (< 1e-5)
MACS2 (ver. 2.1.0)	Window sliding (predicted)	Peak height and Fold enrichment (5 < fold change < 50)	FDR (< 0.01)
MACS2 with the broad option (ver. 2.1.0)	Window sliding (predicted)	Peak height and Fold enrichment (5 < fold change < 50)	FDR (< 0.05)
PeakSeq (ver. 1.31)	Tag clustering	Peak height and Fold enrichment (Modeled threshold)	Compare to statistical model fitted with control data (FDR < 0.05)
SISSRs (ver. 1.4)	Window sliding (window size = 20)	Strand specific scoring (2 > number of directional reads)	Compare to normalized control data (FDR < 0.001)

Test Pipeline



Histone modification ChIP-Seq data

Basic analysis

- Raw ChIP-Seq data
- Sequencing quality filter
- Alignment[1]
- Merge histone replicates[2]
- Cross correlation
- Peak calling[3]

Cell line – H1 (human embryonic stem cell)

ChIP-Seq data	
H3K4ac	H3K27ac
H3K4me1	H3K27me3
H3K4me2	H3K36me3
H3K4me3	H3K56ac
H3K9ac	H3K79me1
H3K9me3	H3K79me2
	Input

ChIP quality assessment : Cross-correlation

Basic analysis

Raw ChIP-Seq data
↓
Sequencing quality filter
↓
Alignment[1]
↓
Merge histone replicates[2]
↓
Cross correlation
↓
Peak calling[3]

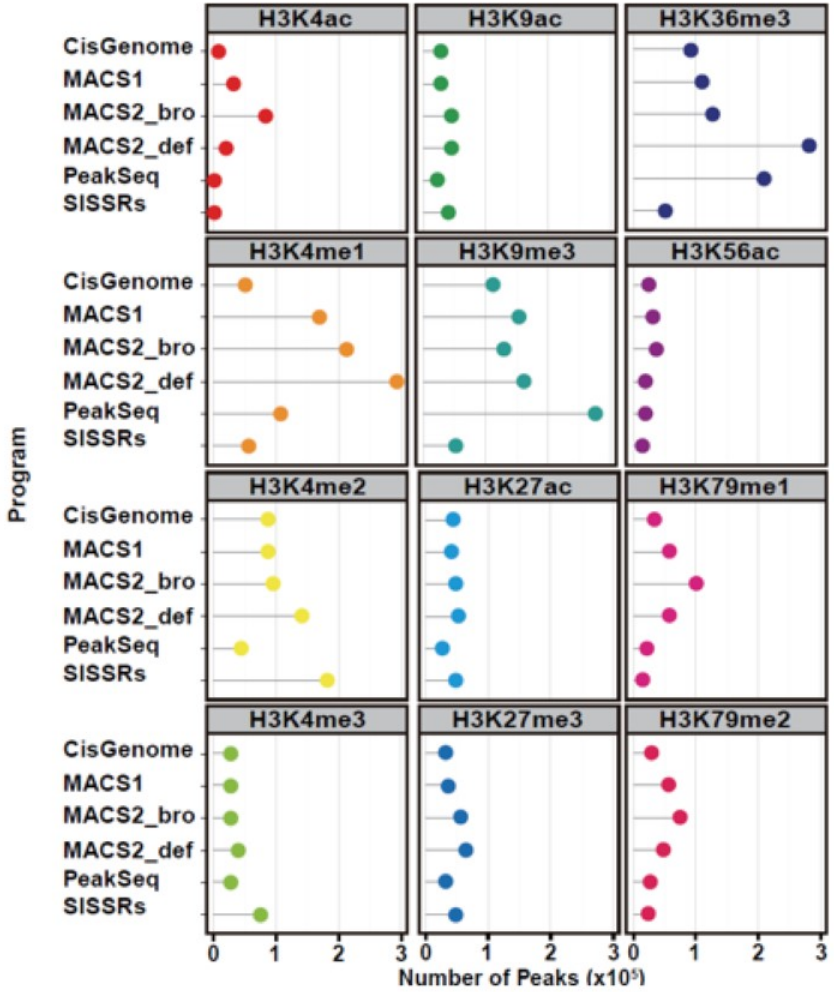
→

ID	NSC	RSC	SPP QC scores
H3K4ac	1.047	1.414	1.000
H3K4me1	1.105	1.581	2.000
H3K4me2	1.842	1.618	2.000
H3K4me3	2.418	1.266	1.000
H3K9ac	1.046	1.048	1.000
H3K9me3	1.057	1.307	1.000
H3K27ac	1.231	1.210	1.000
H3K27me3	1.234	1.625	2.000
H3K36me3	1.119	2.330	2.000
H3K56ac	1.181	1.331	1.000
H3K79me1	1.138	2.125	2.000
H3K79me2	1.050	1.354	1.000

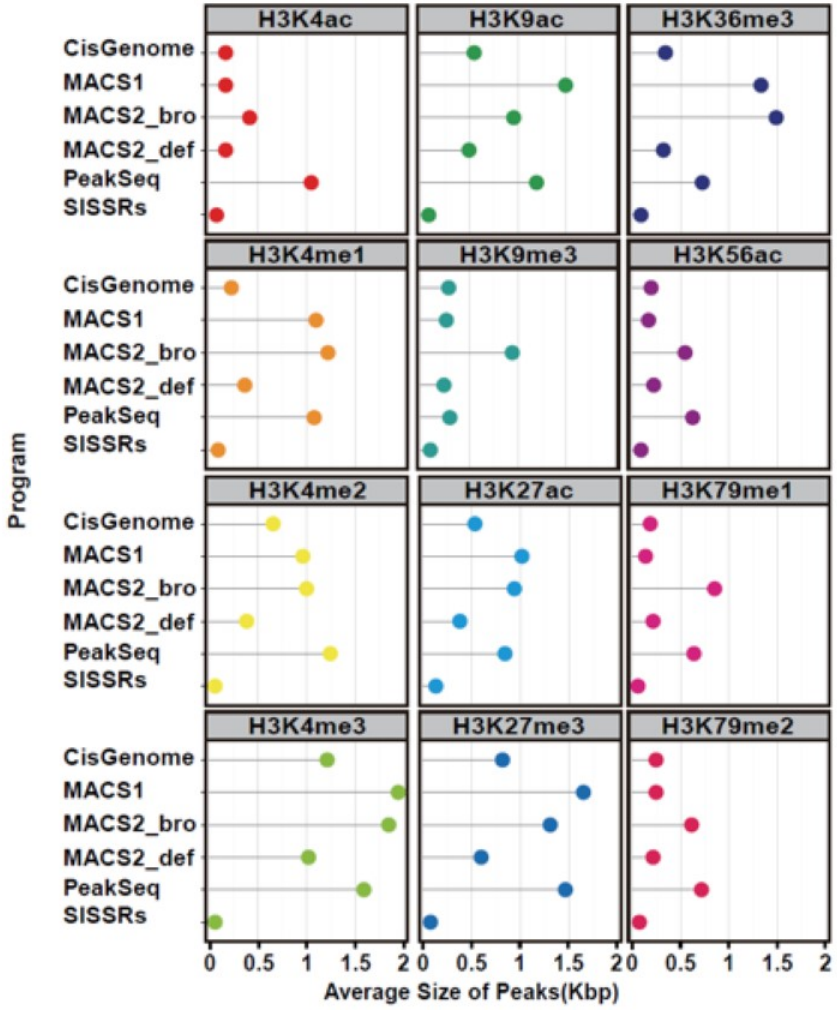
NSC : normalized strand coefficient
(=cc(fragment length) /cc(background)
RSC : relative strand correlation
(=cc(fragment length) - cc(background) / cc(read length) - cc(background)
ENCODE standard : NSC > 1.05
RSC > 0.8

Peak number and size

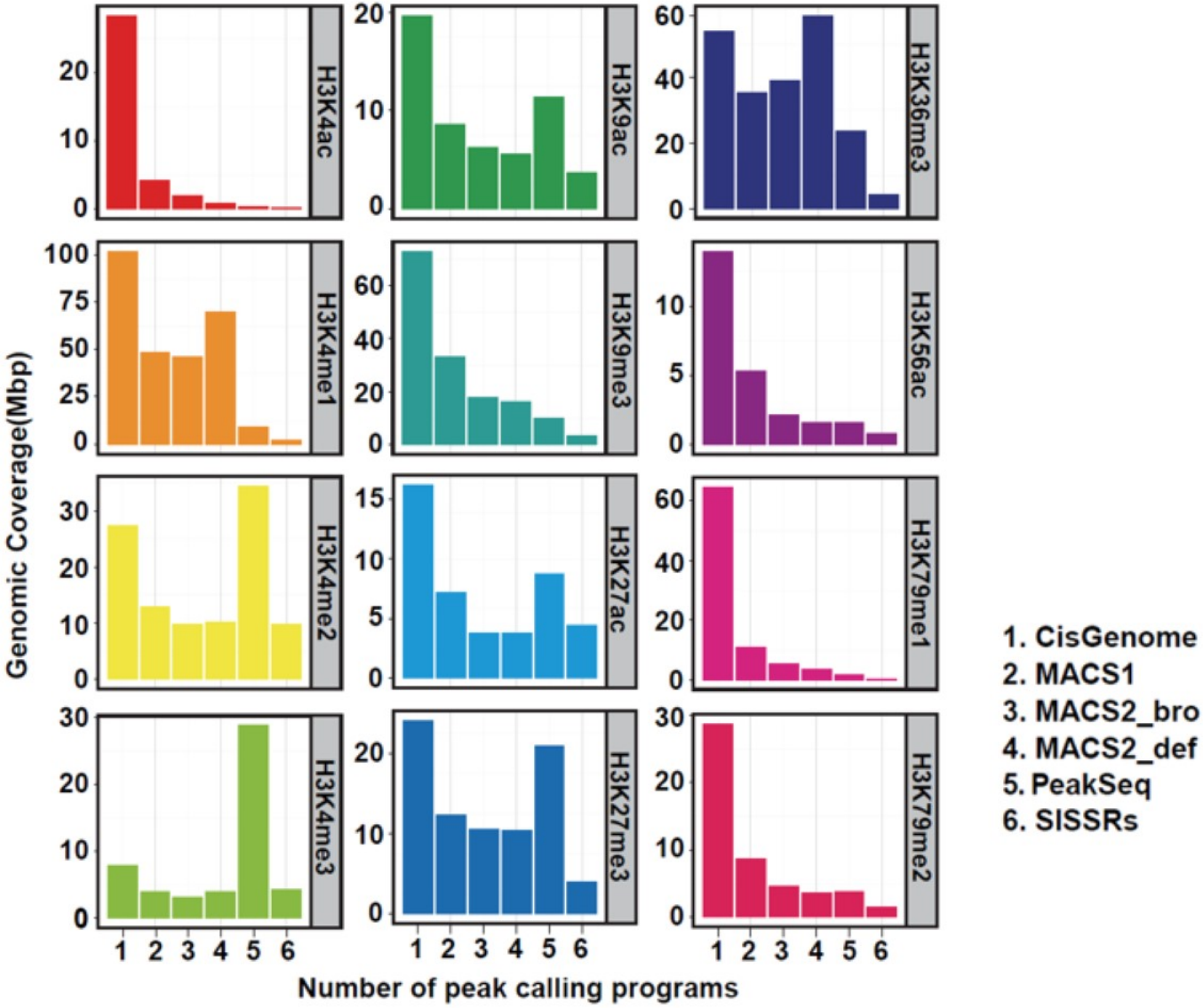
- **Number of peaks**



- **Size of peaks**



Genomic coverage

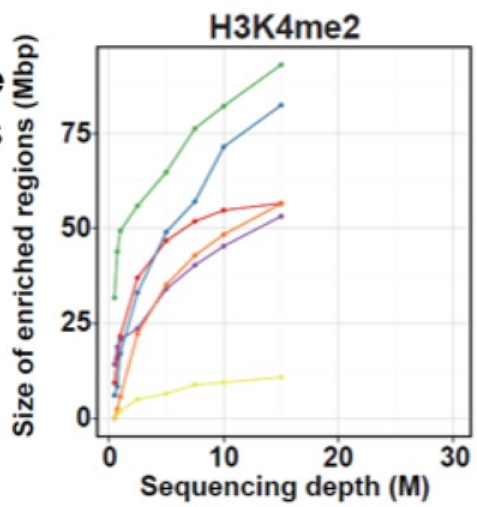


Coverage with sequencing depth

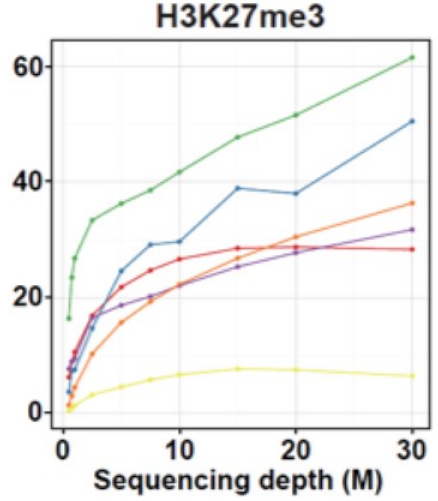
Enriched regions of point and broad source factors were dramatically increase at low sequencing depths

However, the pattern of point and broad source factors are not detected in low fidelity factors

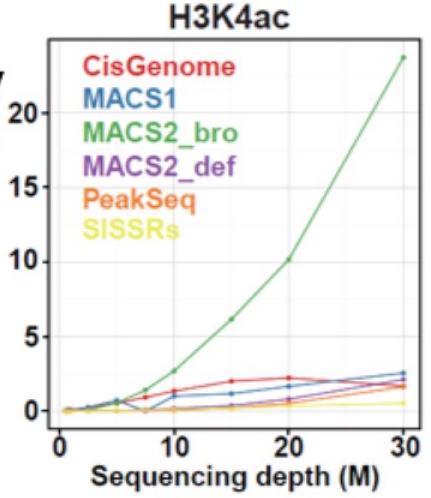
Point Source factors

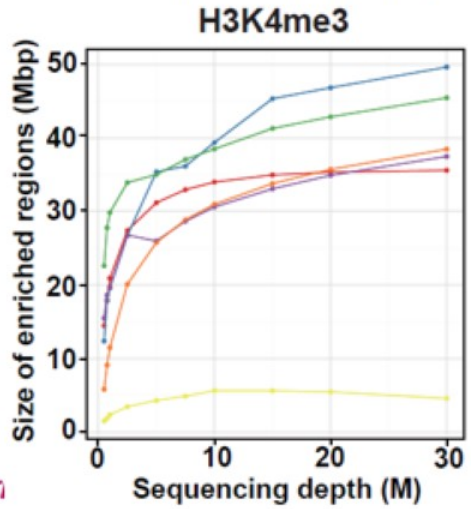


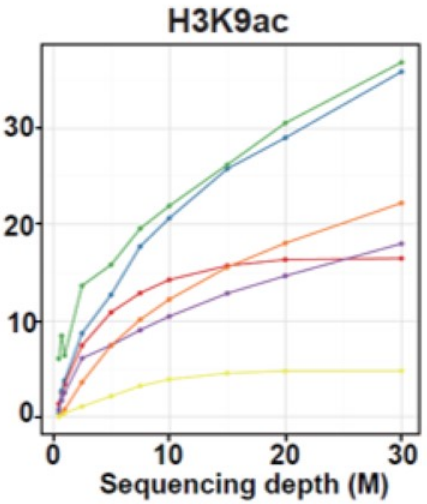
Broad Source factors

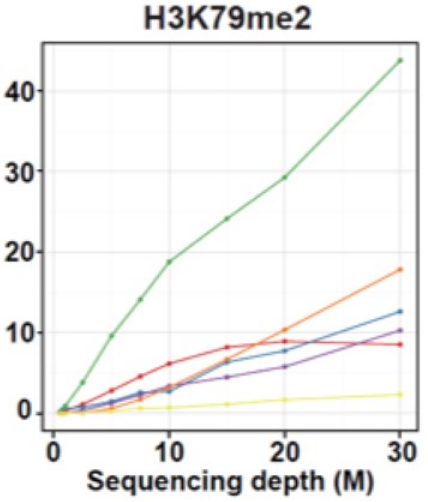


Low Fidelity factors

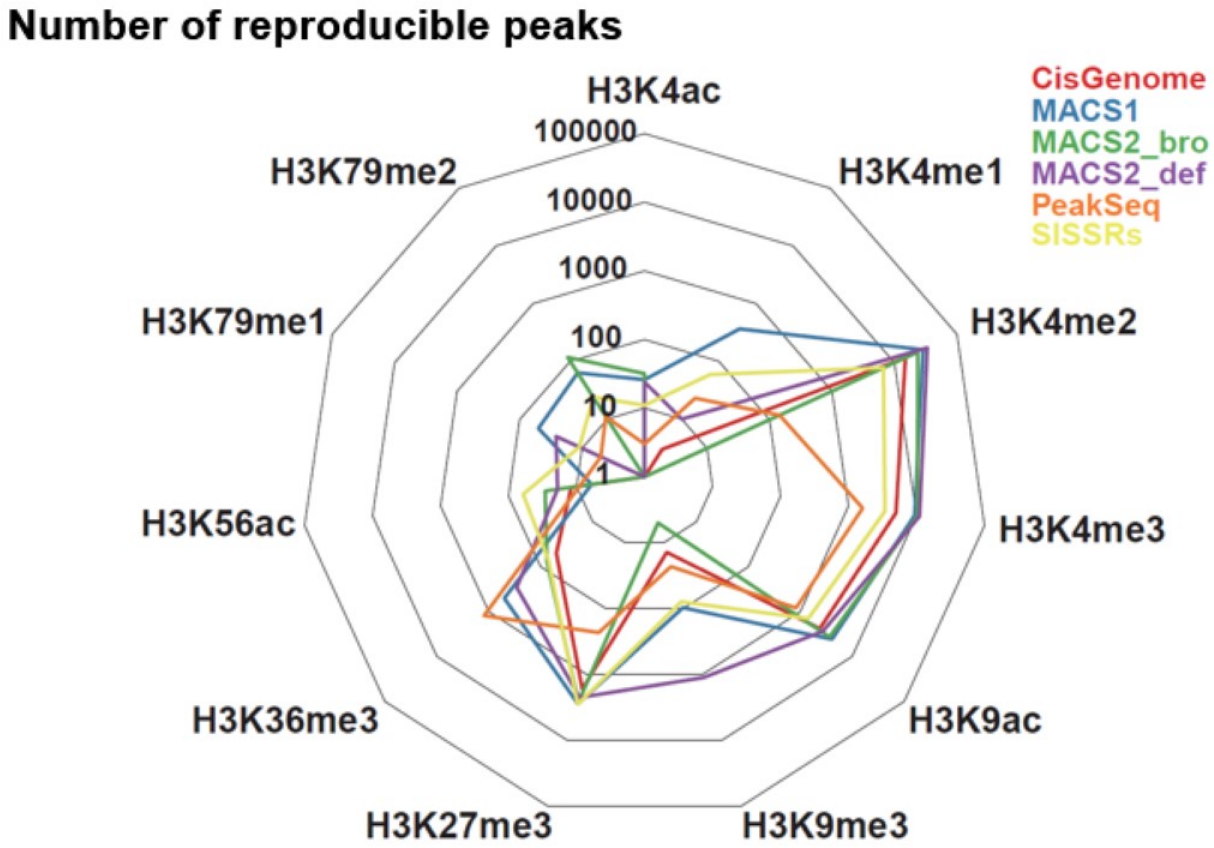








Reproducibility test



Summary I

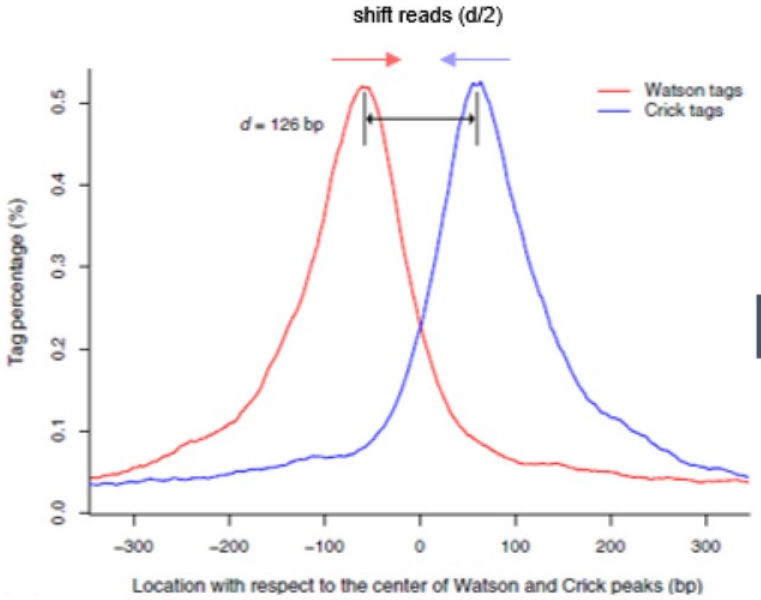
- We evaluate program's performance for ChIP-Seq data of 12 histone modifications
- The variability across peak callers is strongly influenced by histone types
- In point source factors, most of peak callers without SISSRs show similar performance.

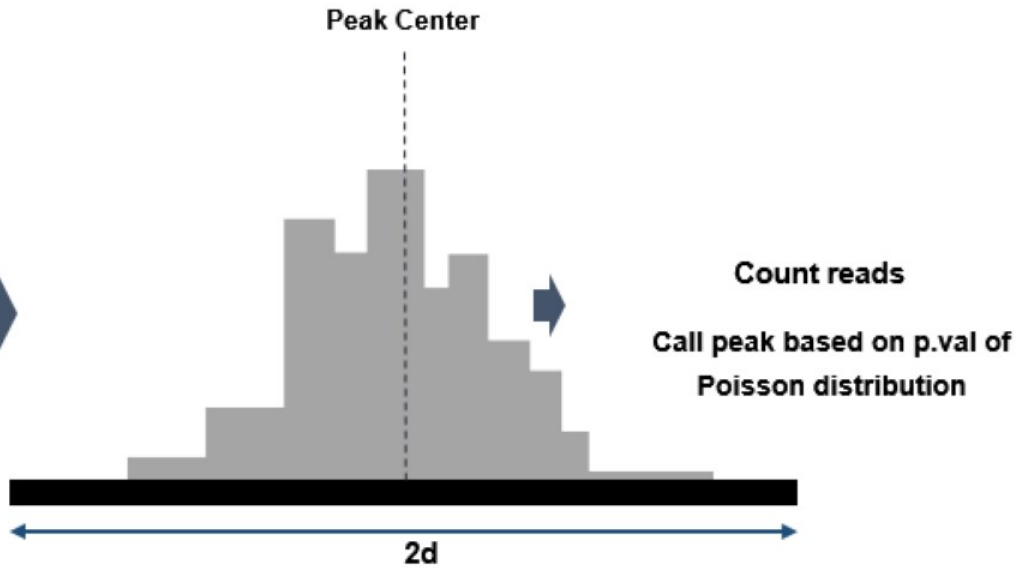
CisGenome
MACS1
MACS2_bro
MACS2_def
PeakSeq
SISSRs

Domain Assessment

Domain calling programs

Poisson distribution based programs - MACS2



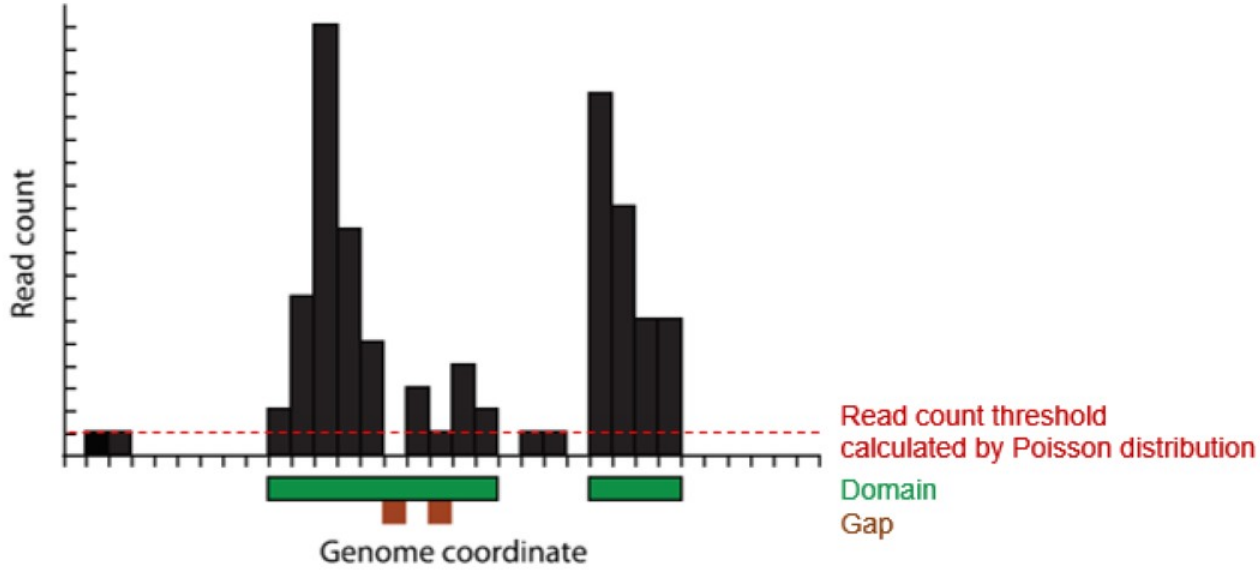


MACS2 merges called peaks within 4d when broad option is on

Zhang Y., Liu T., *et al*. Model-based Analysis of ChIP-Seq (MACS)
Genome Biology **9**, *R137* (2008)

Domain calling programs

Poisson distribution based programs - SICER

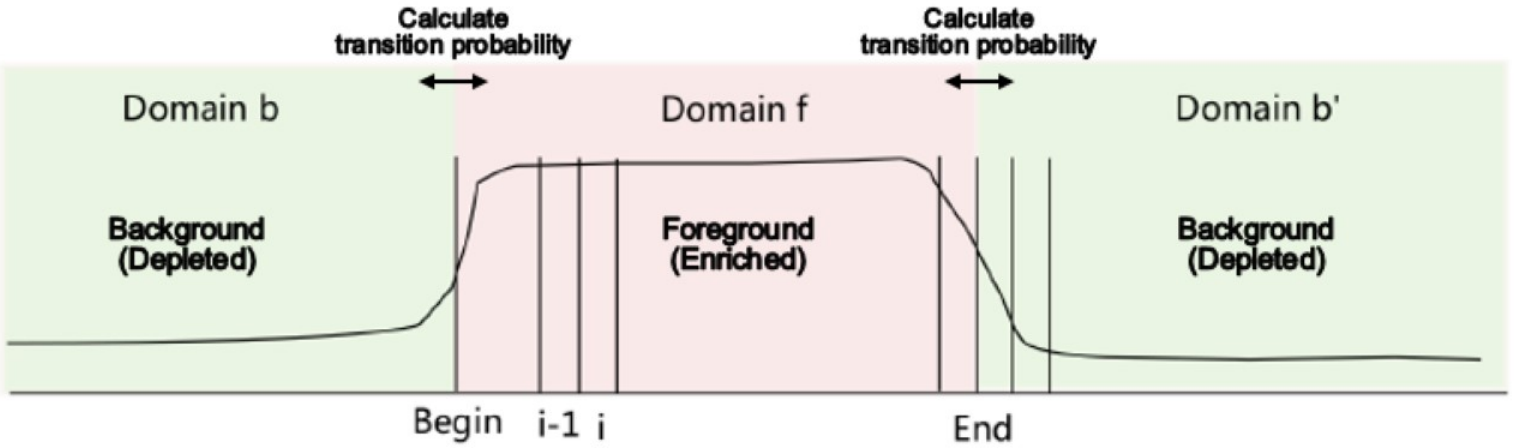


SICER estimates domain with island-based manner, allowing 3 gaps in domain

Xu S., Grullon S., Ge K., Peng W.
Spatial Clustering for Identification of ChIP-Enriched Regions
(SICER) to Map Regions of Histone Methylation Patterns in Embryonic Stem Cells
Methods Mol Biol **1150**, 97–111 (2014)

Domain calling programs

Hidden Markov Model (HMM) based programs - RSEG

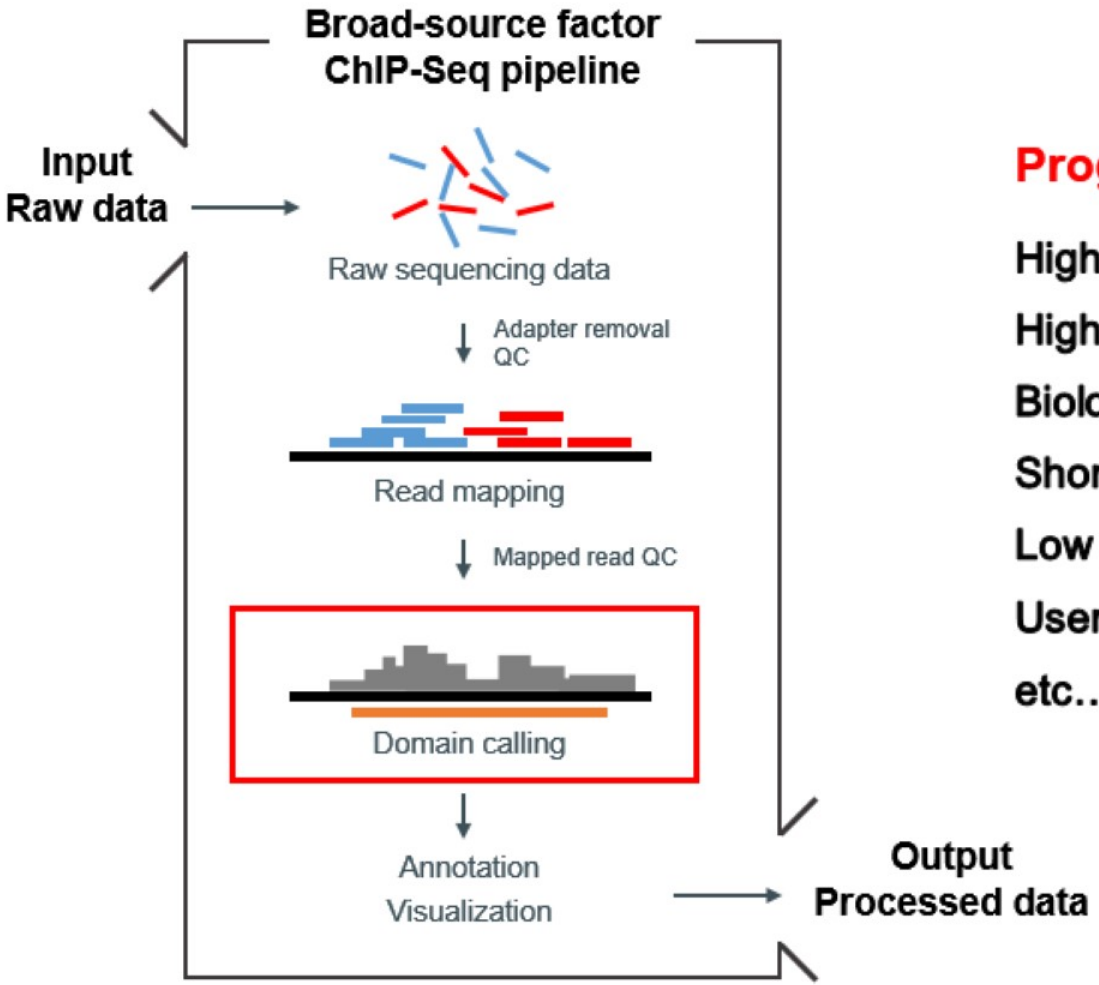


RSEG determines domain boundaries using finite state HMM

Song Q., Smith AD.
Identifying dispersed epigenomic domains from ChIP-Seq data
Bioinformatics **27**, 870-871 (2011)

Which program should we use?

Program selection for automated broad-source ChIP-Seq analysis pipeline



Programs with

- High sensitivity
- High specificity
- Biologically meaningful results
- Short running time
- Low resource requirements
- User-friendly options
- etc…

Data set

Simulated H3K36me3 ChIP-Seq data

(Song Q., Smith AD., 2011)

Sample	Read length	Read number
Simulated H3K36me3 ChIP	27~36	7,399,225
Simulated Input	27~36	7,499,999

H1 ES cell line Histone ChIP-Seq data

Data from NIH roadmap epigenomics project

(Bernstein BE., Stamatoyannopoulos JA., Costello JF., Ren B., et al., 2011)

Sample	Read length	Read number
H1 H3K27me3 ChIP	36~76	115,655,373
H1 H3K36me3 ChIP	36~76	192,914,242
H1 Histone Input	36~76	302,343,443

Summary II

	RSEG	SICER	hiddenDomains	PeakRanger-BCP
Running time	Very long	Long	Short	Very short
Output file	Good for further analysis	Good for further analysis	Bad for further analysis	Good for further analysis
Called domain number in simulated data	Reasonable (difference <10%)	More called domains (difference >50%)	Reasonable (difference <10%)	Reasonable (difference <10%)
Called domain size in simulated data	Slightly big (difference>20%)	Slightly small (difference>20%)	Reasonable (difference <10%)	Slightly big (difference>20%)
Called domain size in H1 H3K27me3 data	Reasonable (Avr. size > 10 kbp)	Slightly small (Avr. size < 10 kbp)	Slightly small (Avr. size < 10 kbp)	Reasonable (Avr. size > 10 kbp)
Called domain size in H1 H3K36me3 data	Large (Avr. size > 30 kbp)	Slightly small (Avr. size < 10 kbp)	Reasonable (Avr. size > 10 kbp)	Reasonable (Avr. size > 10 kbp)
Jaccard index in simulated data	High (> 60%)	High (> 60%)	Very high (> 80%)	High (> 60%)
Recall in simulated data	Very high (> 80%)	Very high (> 80%)	Very high (> 80%)	Very high (> 80%)
False positive rate in simulated data	Very low (7% called domains)	Slightly high (40% called domains)	Low (17% called domains)	Low (14% called domains)
Domains with CTCF in experimental data	Reasonable (7% < domain < 9%)	Very low (< 5%)	Slightly low (< 7%)	Reasonable (7% < domain < 9%)

PeakRanger-BCP is the best program for domain calling process

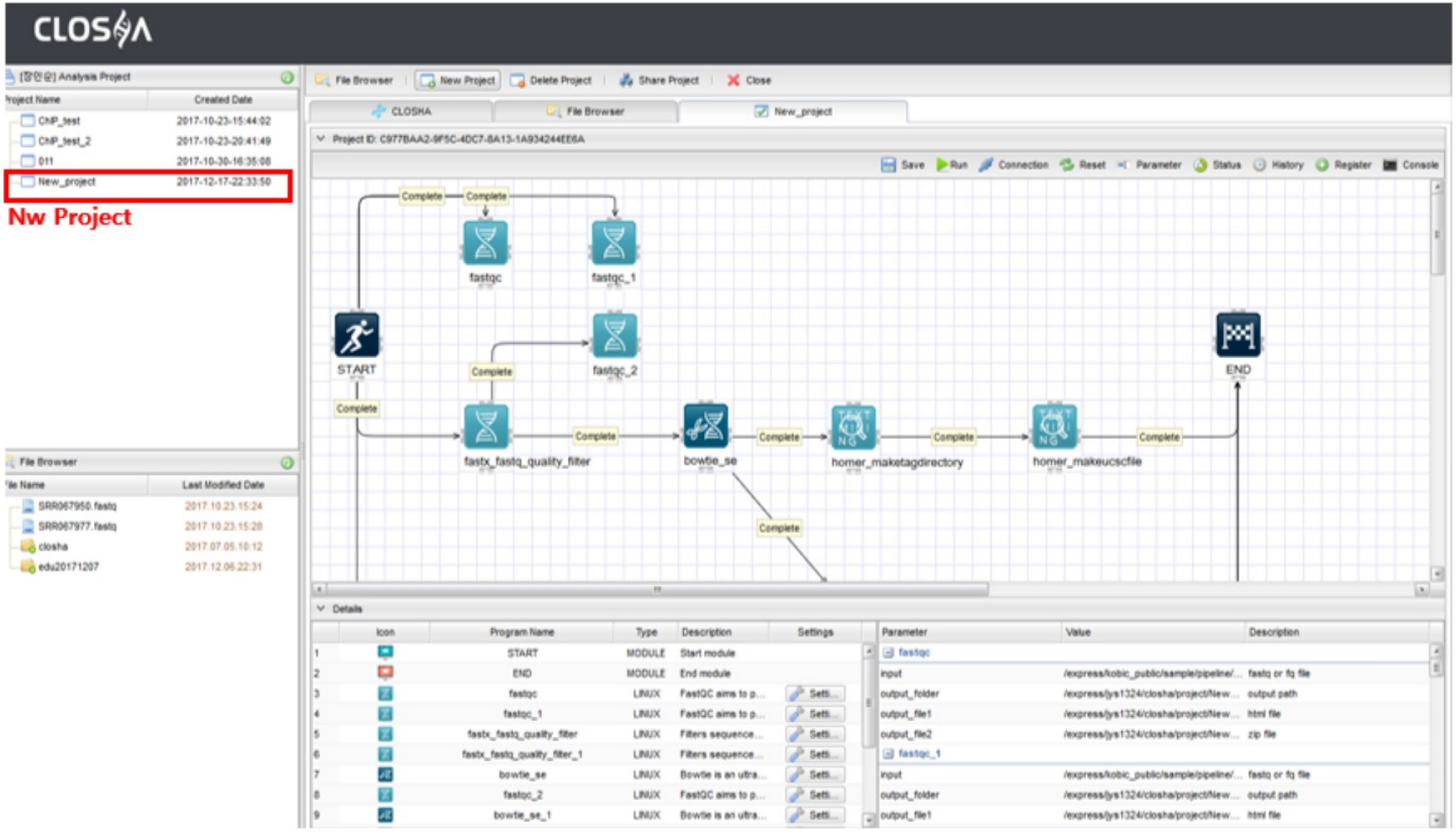
BioExpress @ KOBIC

https://www.bioexpress.re.kr/

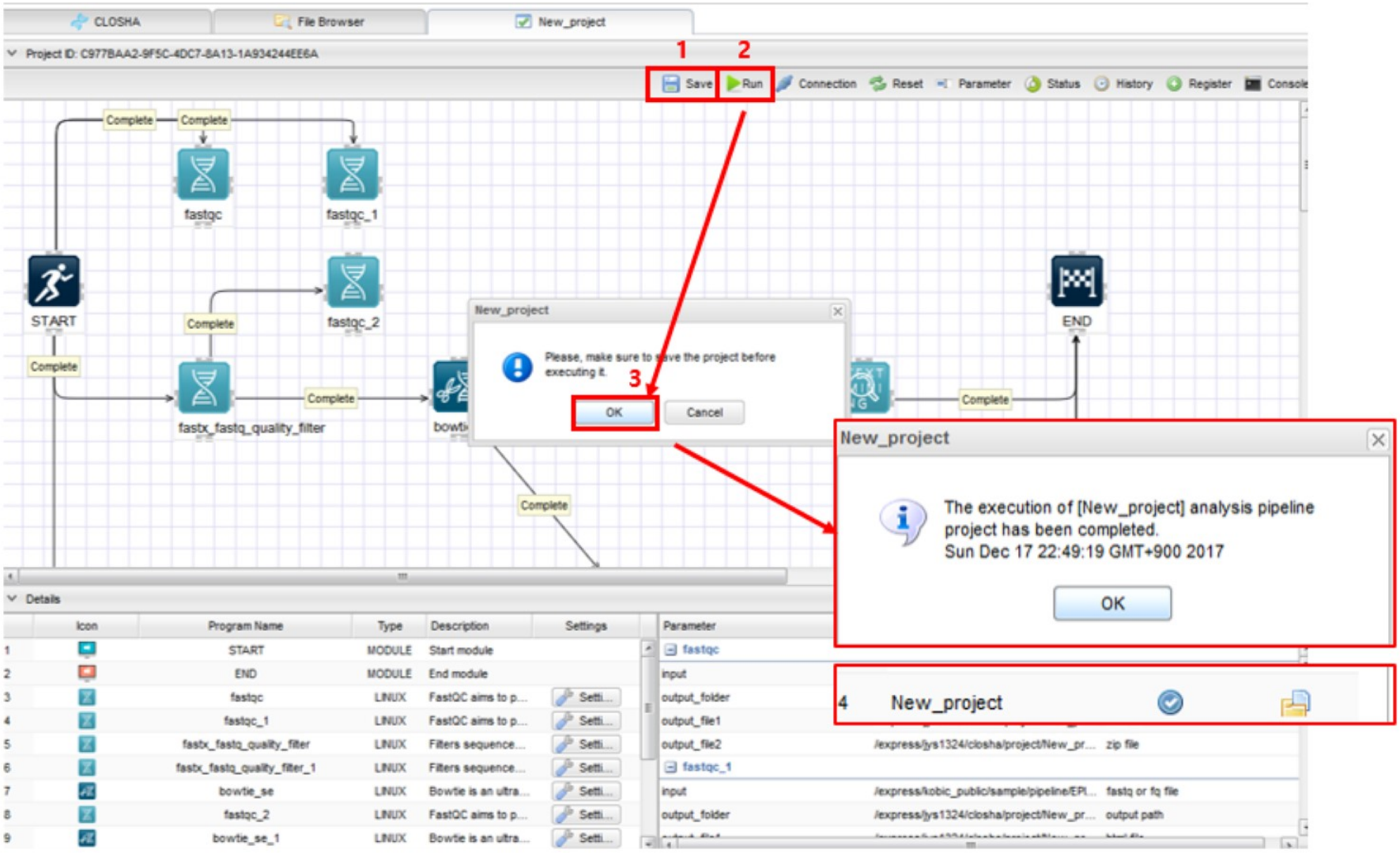
CLOSHA under BioExpress @ KOBIC

https://closha.kobic.re.kr/

ChIP-Seq module in CLOSHA

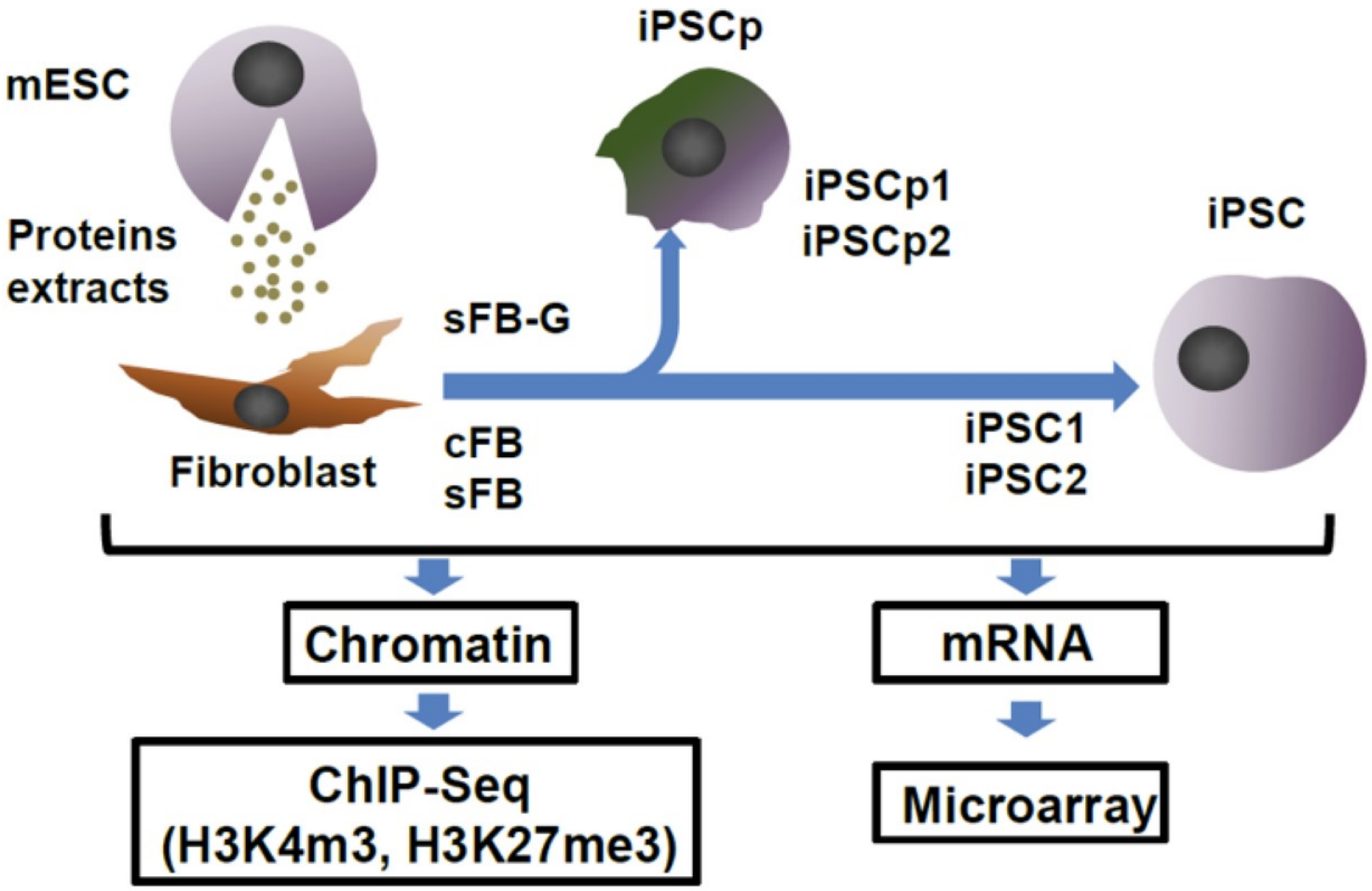


ChIP-Seq module in CLOSHA

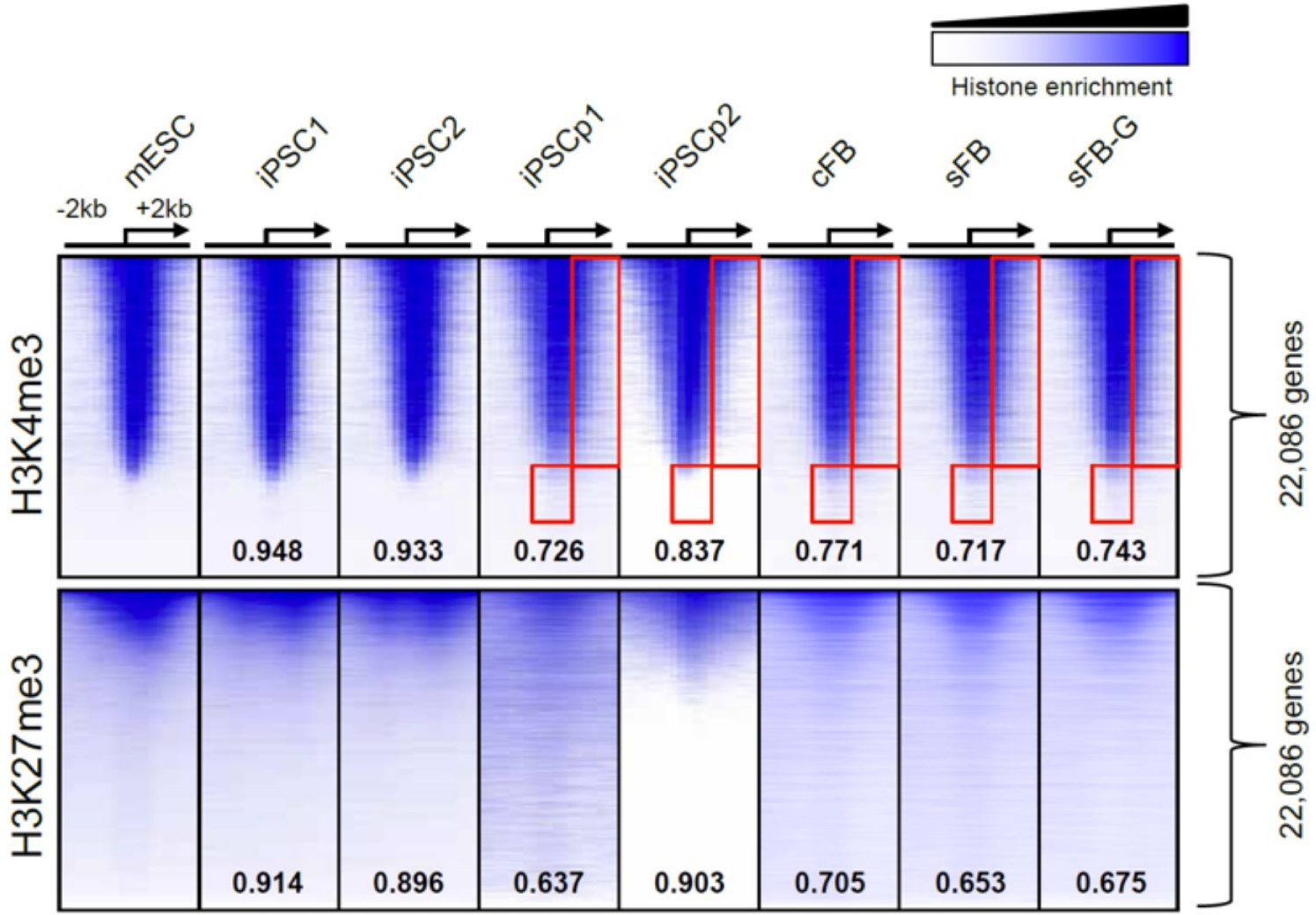


iPSC vs. mES epigenome validation

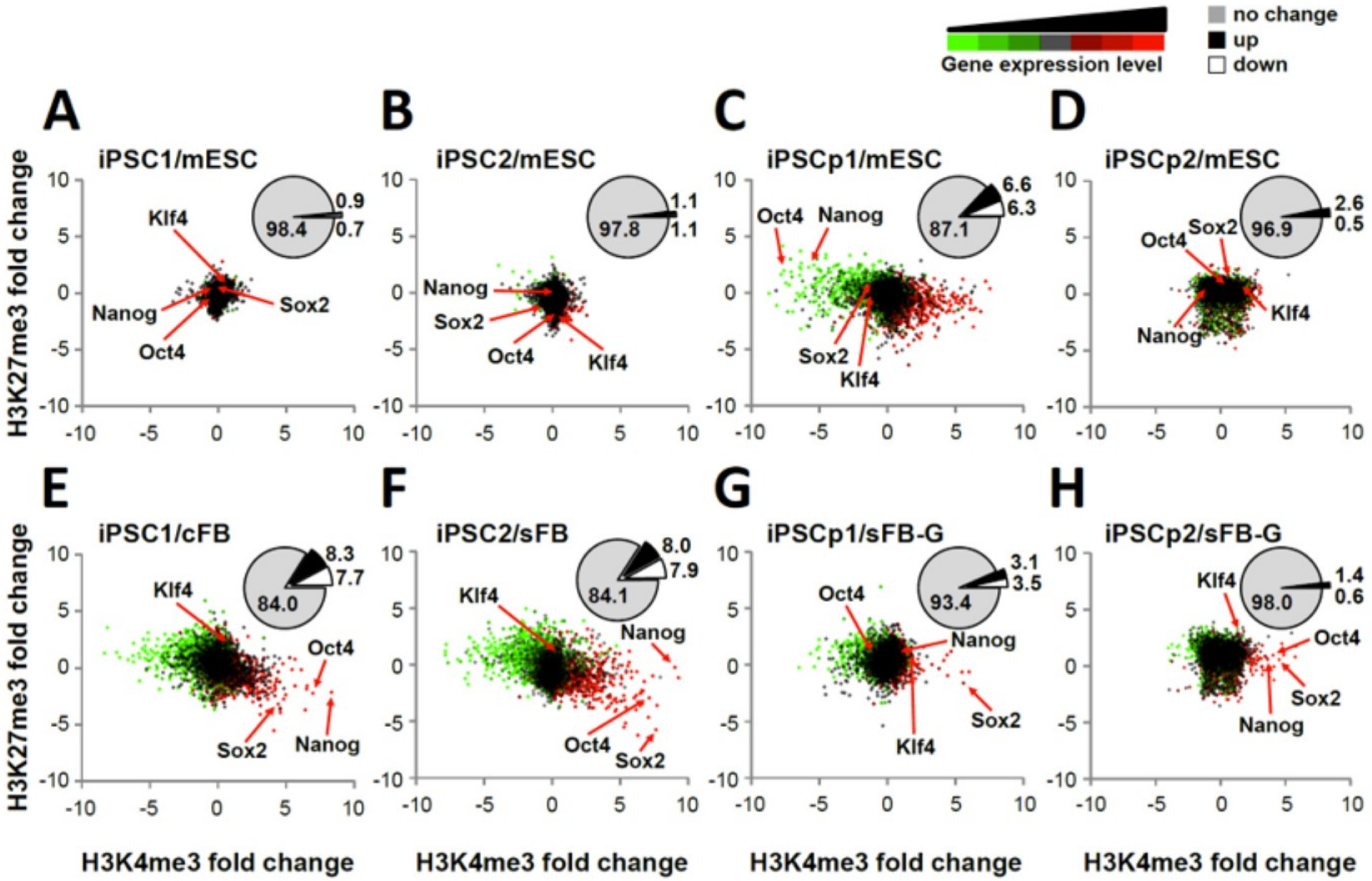
Experimental design



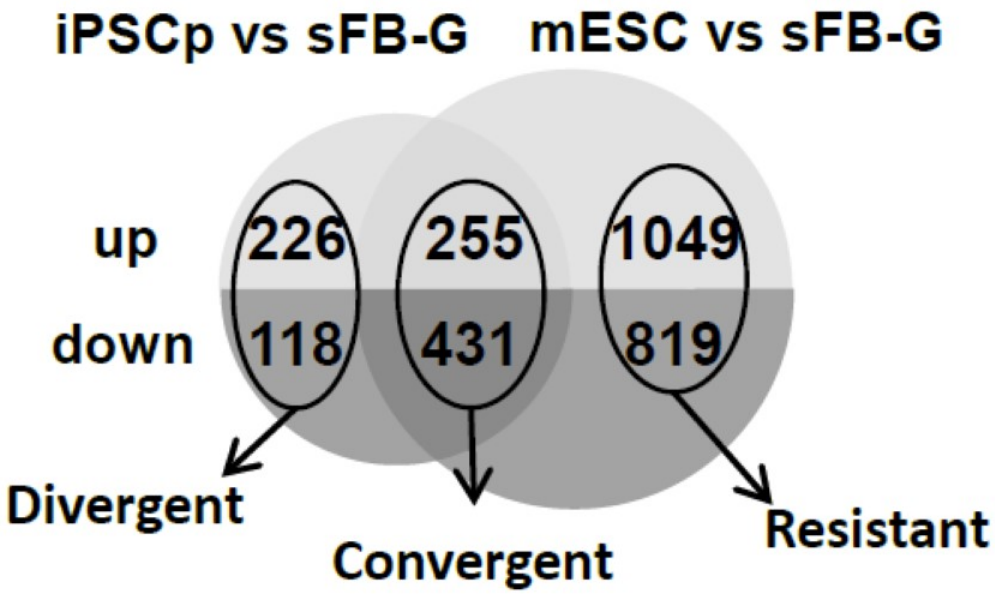
Histone modification comparison



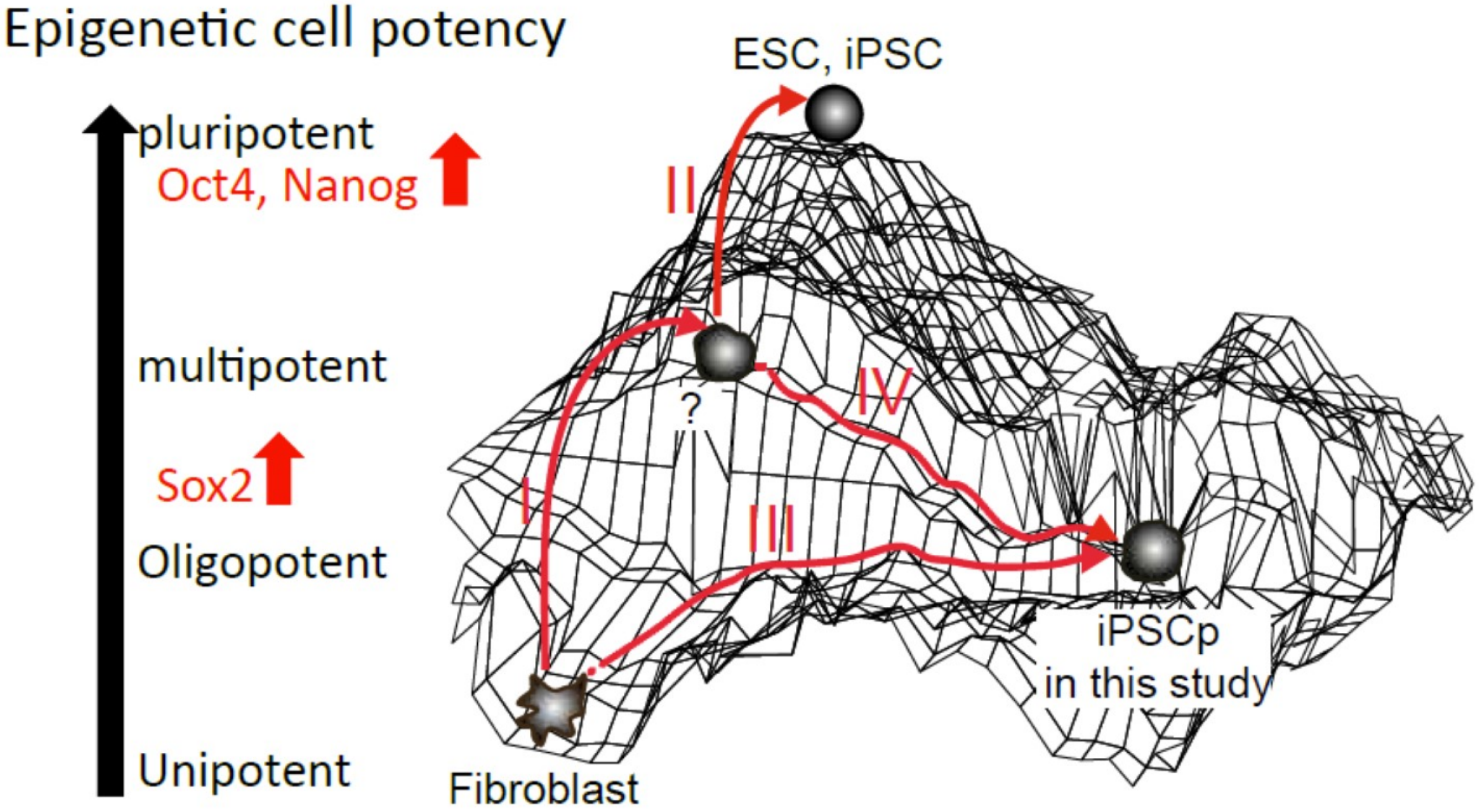
Integration of expression data with histone modification



Gene-level comparsion



Proposed epigenetic state during de-differentiation



Summary III

- We set up the optimal condition for identification of ChIP-Seq peaks
- The comparative analysis of iPSC and mES revealed that the epigenomics state was closely related as expected.
- The incorporation of gene expression data with histone modification profiles, we proposed Sox2 is the early responder to the de-differentiation signal.