

Peak enrichment evaluation of histone modification ChIP-Seq data

Tae-Young Roh

Laboratory of System Genomics

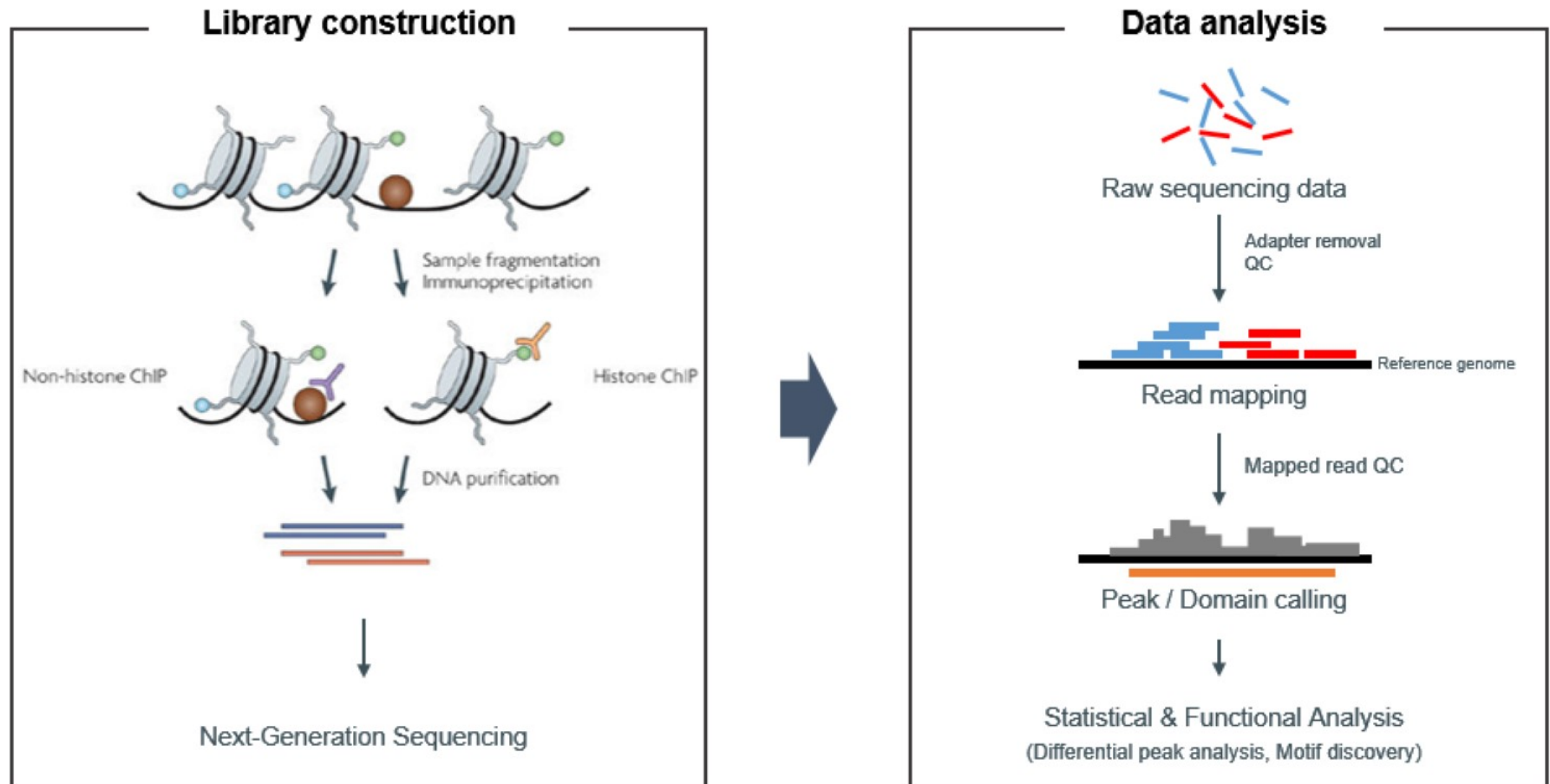
Div. of Integrative Biosciences and Biotechnology

Dept. of Life sciences

**Pohang University of Science and Technology
(POSTECH)**

Overview of ChIP-Seq

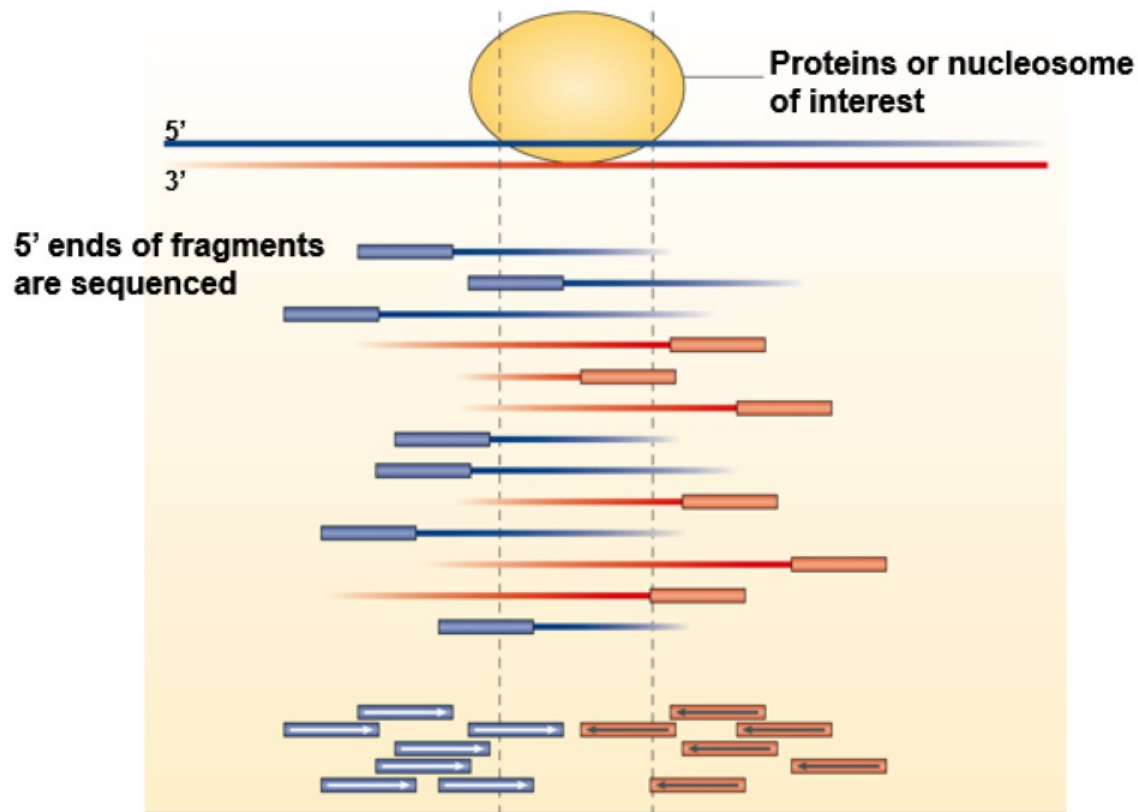
Chromatin ImmunoPrecipitation coupled with Sequencing



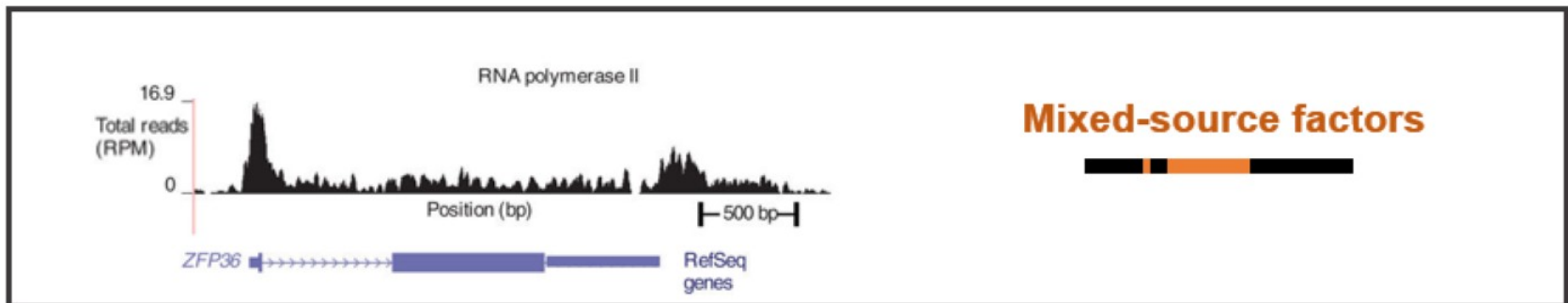
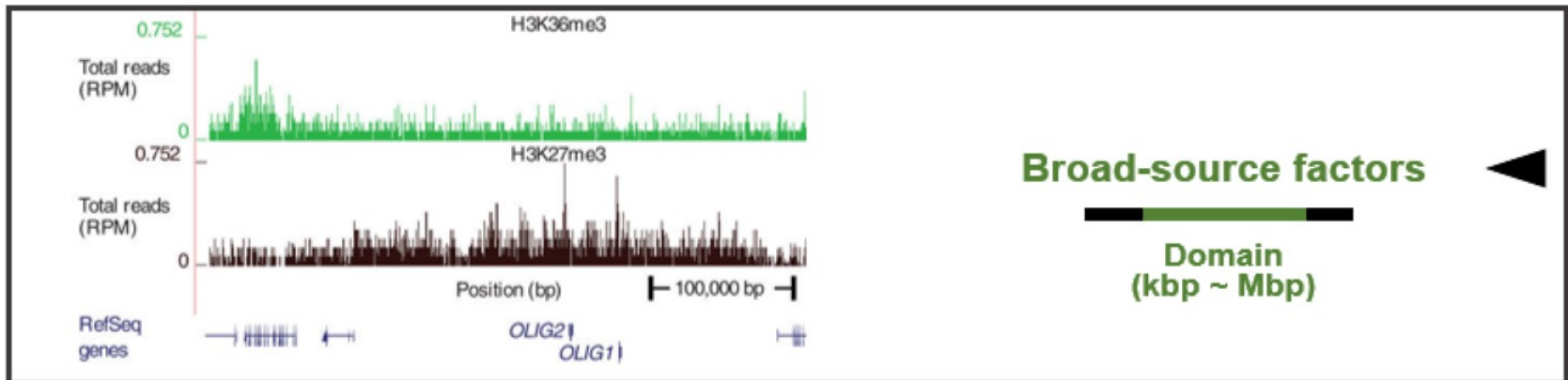
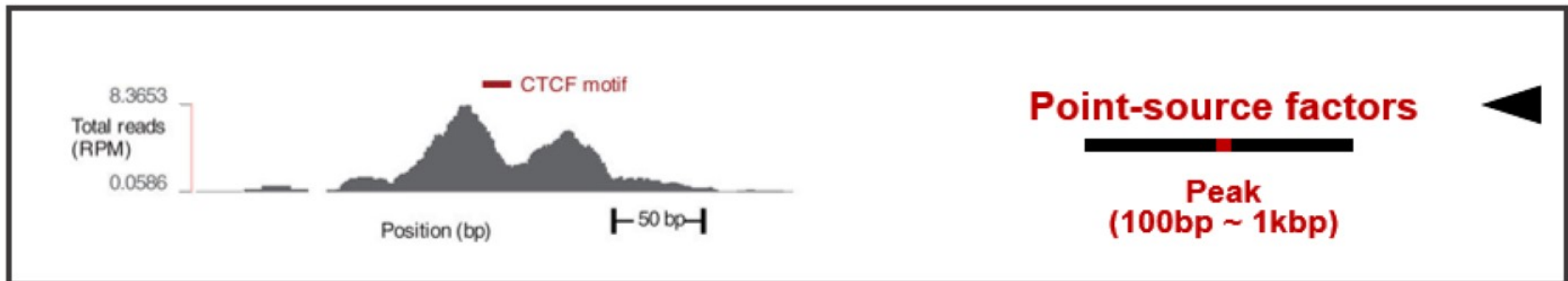
Park PJ. ChIP-seq: advantages and challenges of a maturing technology
Nature Reviews Genetics **10**, 669–680 (2009)

Profiles of Sequenced read at enriched sites

Sequenced reads from ChIP-Seq are distributed around binding positions



Classes of DNA-bound proteins



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 Kharchenko¹⁻³, Michael Y Tolstorukov^{1,2} & Peter J Park¹⁻³

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

Hongkai Ji¹, Hui Jiang², Wenxiu Ma³, David S Johnson^{4,8}, Richard M Myers⁵ & 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 Robertson¹, Mikhail Bilenky¹, Richard Varhol¹, Matthew Bainbridge² 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 Shen¹, Christopher A Maher^{2,3,4}, Ming Hu¹, Shanker Kalyana-Sundaram^{3,4}, Jindan Yu³ 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 Eeckhoutte[†], 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 Rozowsky¹, Ghia Euskirchen², Raymond K Auerbach³, Zhengdong D Zhang¹, Theodore Gibson¹, Robert Bjornson⁴, Nicholas Carriero⁴, Michael Snyder^{1,2} & Mark B Gerstein^{1,3,4}

BayesPeak: Bayesian analysis of ChIP-seq data

Christiana Spyrou^{*1,3}, Rory Stark³, Andy G Lynch⁴ 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

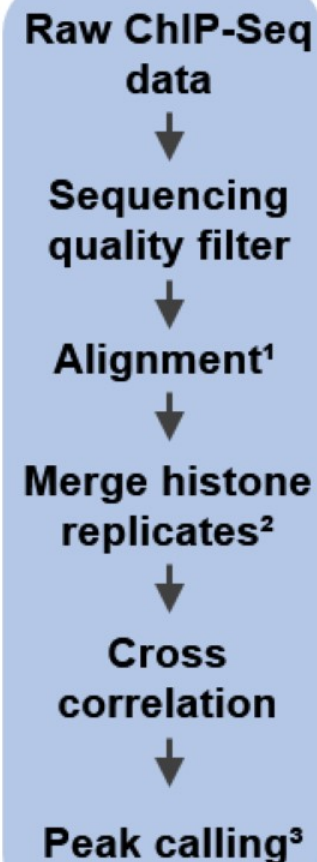
Kimberly R. Blahnik¹, Lei Dou¹, Henriette O'Geen¹, Timothy McPhillips¹, Xiaoqin Xu¹, Alina R. Cao¹, Sushma Iyengar¹, Charles M. Nicolet¹, Bertram Ludäscher^{1,2}, Ian Korf^{1,3} and Peggy J. Farnham^{1,4,*}

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 ($10 < \text{fold change} < 30$)	P-value ($< 1e-5$)
MACS2 (ver. 2.1.0)	Window sliding (predicted)	Peak height and Fold enrichment ($5 < \text{fold change} < 50$)	FDR (< 0.01)
MACS2 with the broad option (ver. 2.1.0)	Window sliding (predicted)	Peak height and Fold enrichment ($5 < \text{fold change} < 50$)	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 ($2 > \text{number of directional reads}$)	Compare to normalized control data (FDR < 0.001)

Test Pipeline

Basic analysis



Comparison across the algorithms



Reproducibility test



Enriched regions with variable sequencing depths



Specificity test



Histone modification ChIP-Seq data

Basic analysis

Raw ChIP-Seq
data



Sequencing
quality filter



Alignment¹



Merge histone
replicates²



Cross
correlation



Peak calling³



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



Alignment¹



Merge histone
replicates²



Cross
correlation



Peak calling³

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

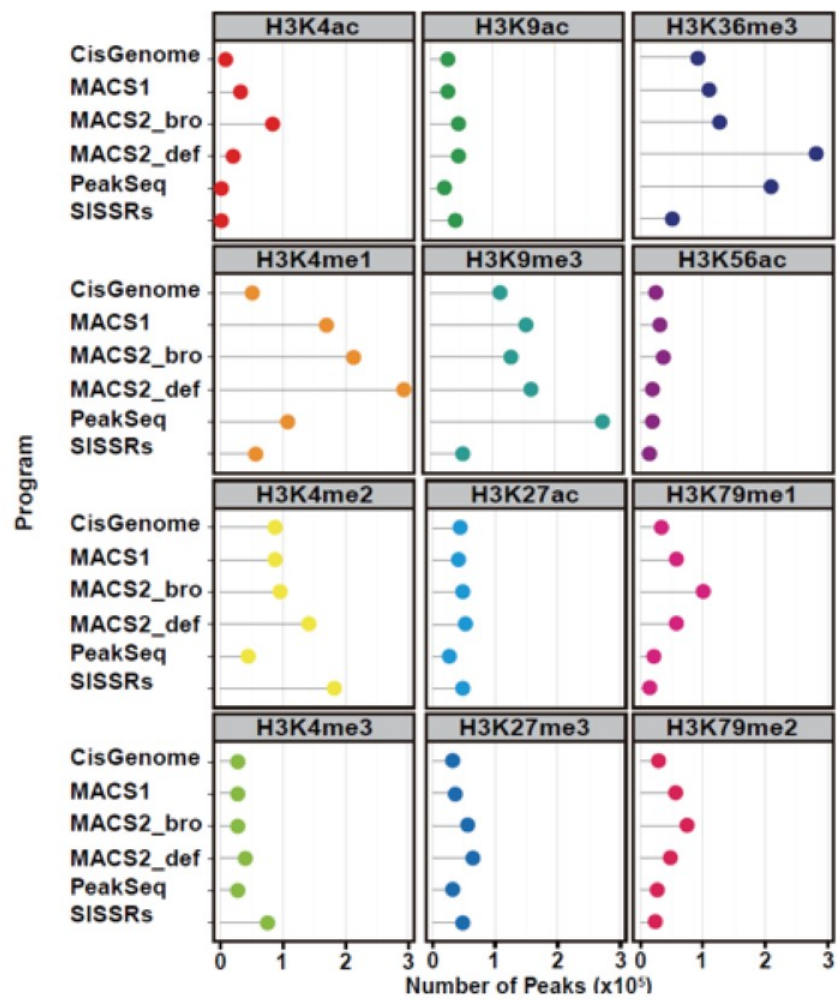
(=cc(fragment length) - cc(background) / cc(read length) - cc(background))

ENCODE standard : NSC > 1.05

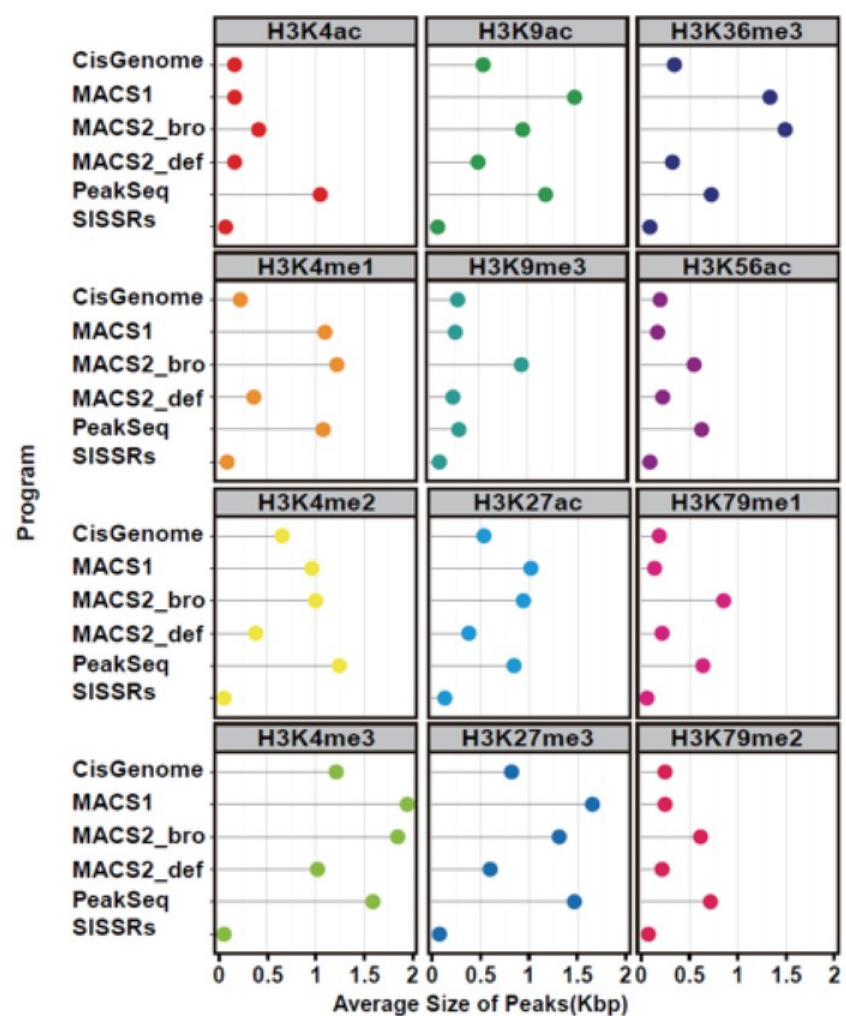
RSC > 0.8

Peak number and size

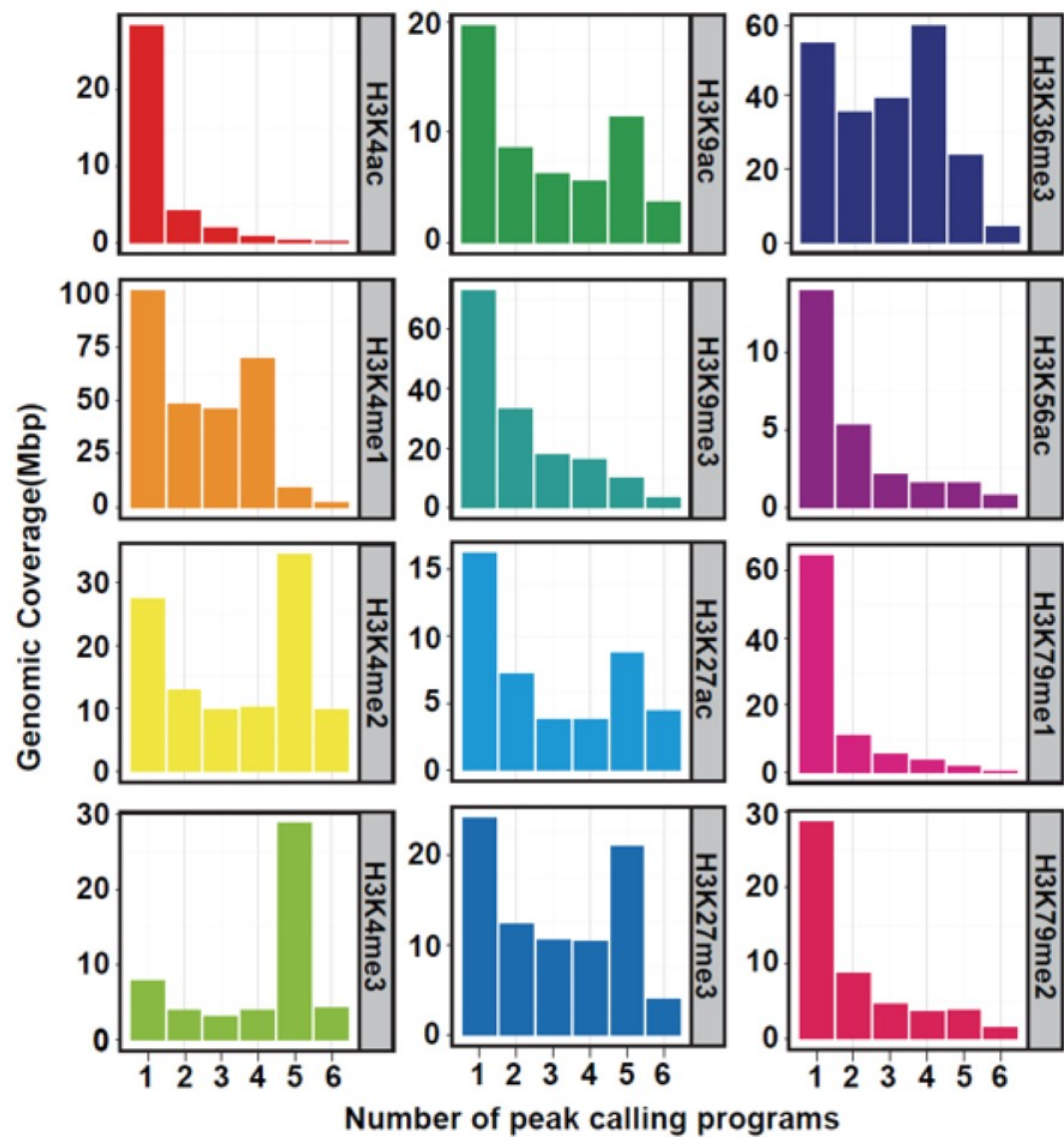
- Number of peaks



- Size of peaks



Genomic coverage



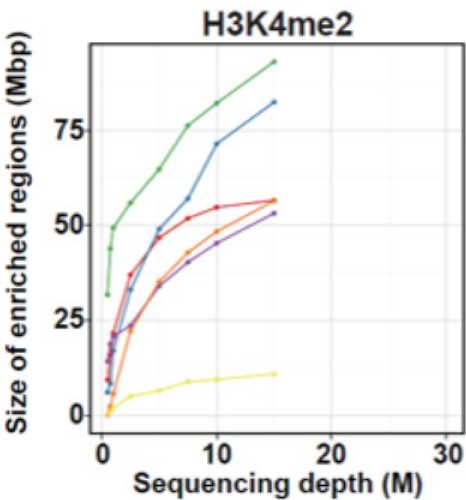
- 1. CisGenome
- 2. MACS1
- 3. MACS2_bro
- 4. MACS2_def
- 5. PeakSeq
- 6. SISSRs

Coverage with sequencing depth

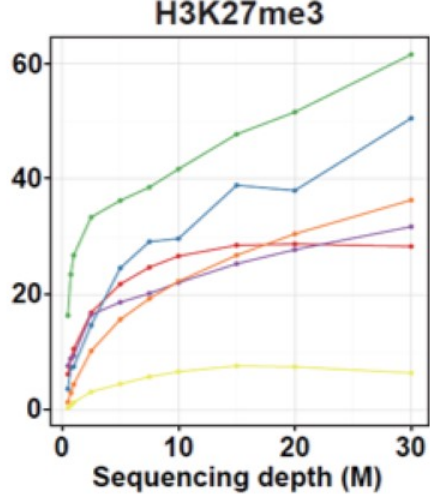
Enriched regions of point and broad source factors were dramatically increase at low sequencing depths

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

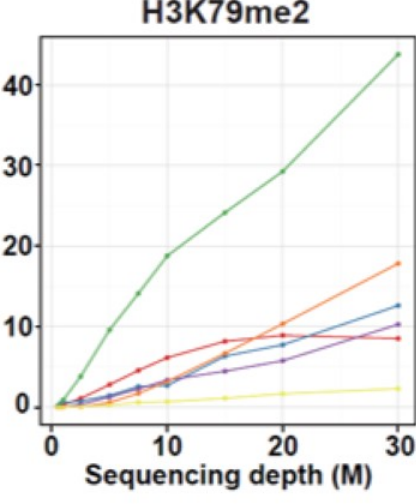
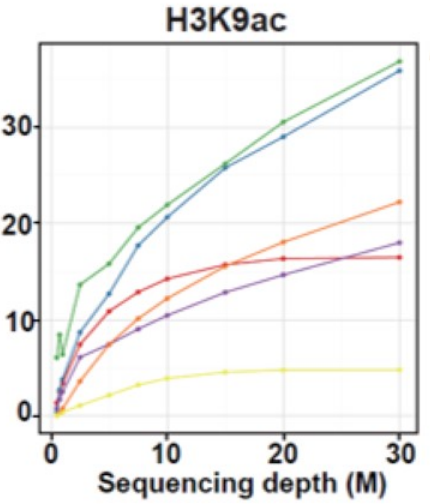
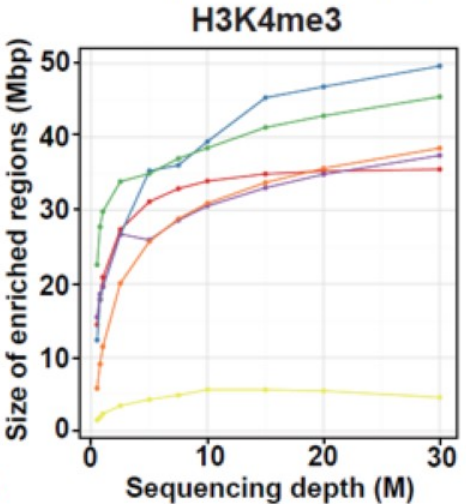
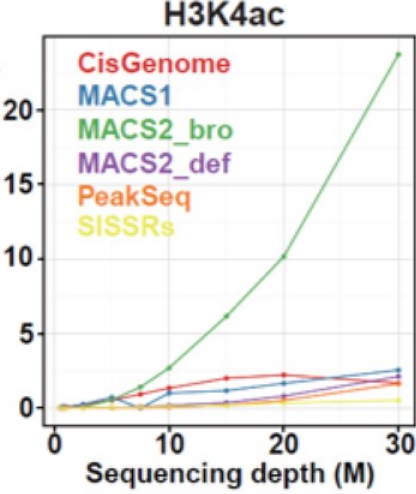
**Point
Source
factors**



**Broad
Source
factors**

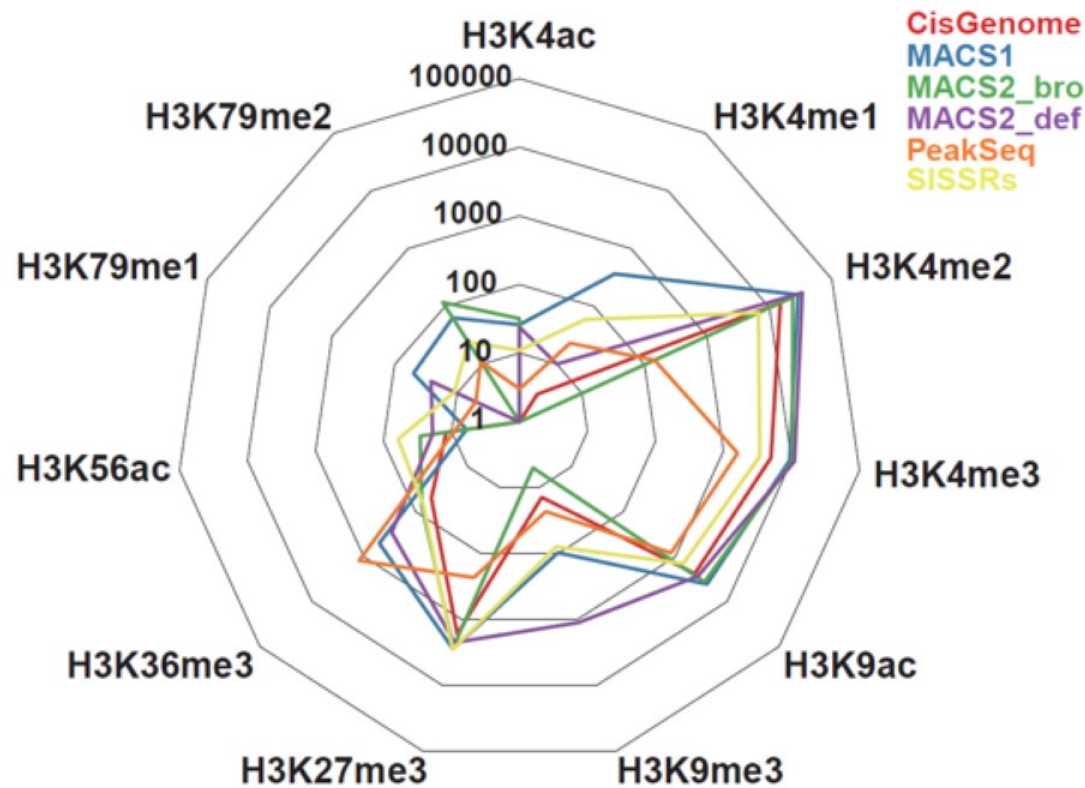


**Low
Fidelity
factors**



Reproducibility test

Number of reproducible peaks



Summary I

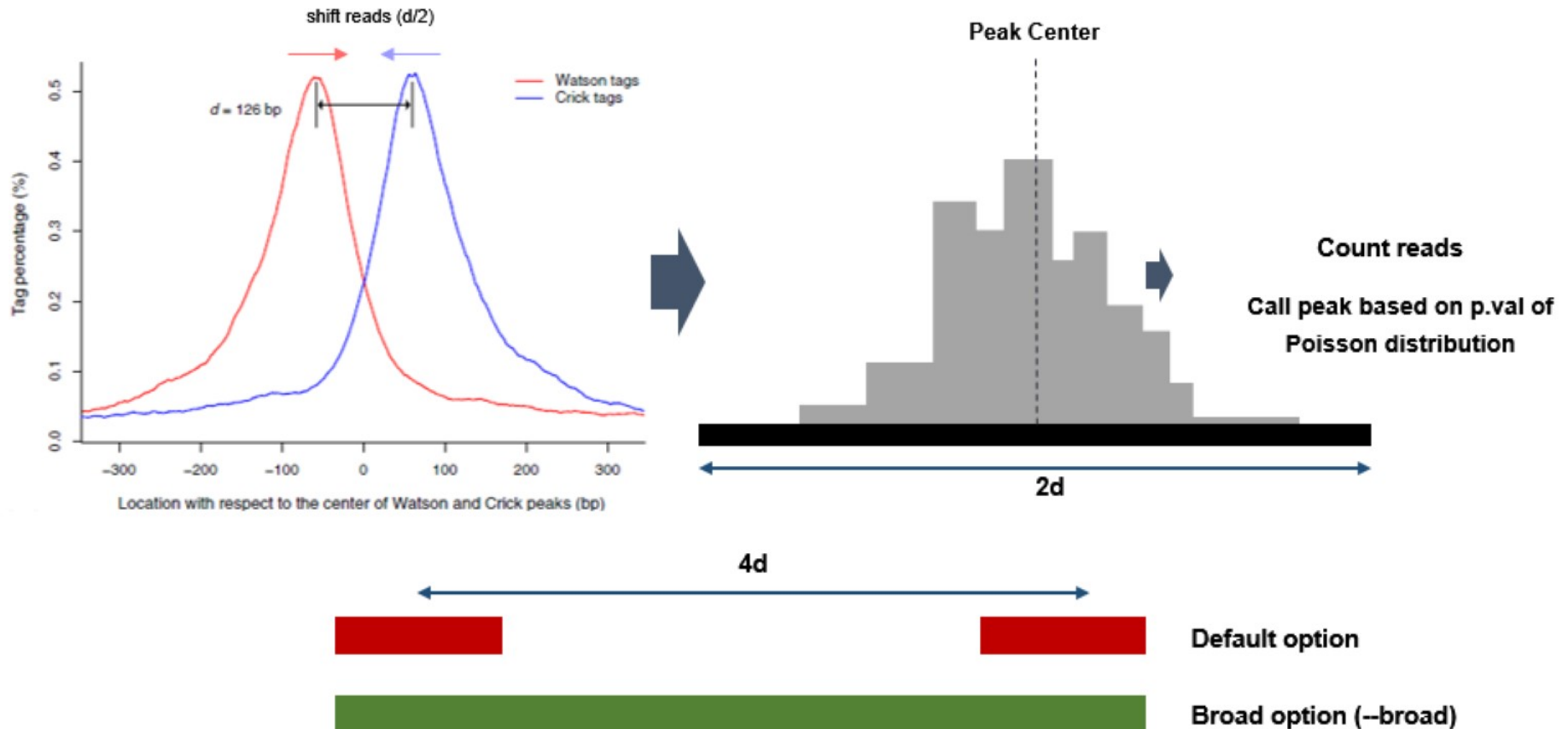
- We evaluate program's performance for ChIP-Seq data of 12 histone modifications
- The variability across peak callers is strongly influenced by histone types
- In point source factors, most of peak callers without SISSRs show similar performance.

CisGenome
MACS1
MACS2_bro
MACS2_def
PeakSeq
SISSRs

Domain Assessment

Domain calling programs

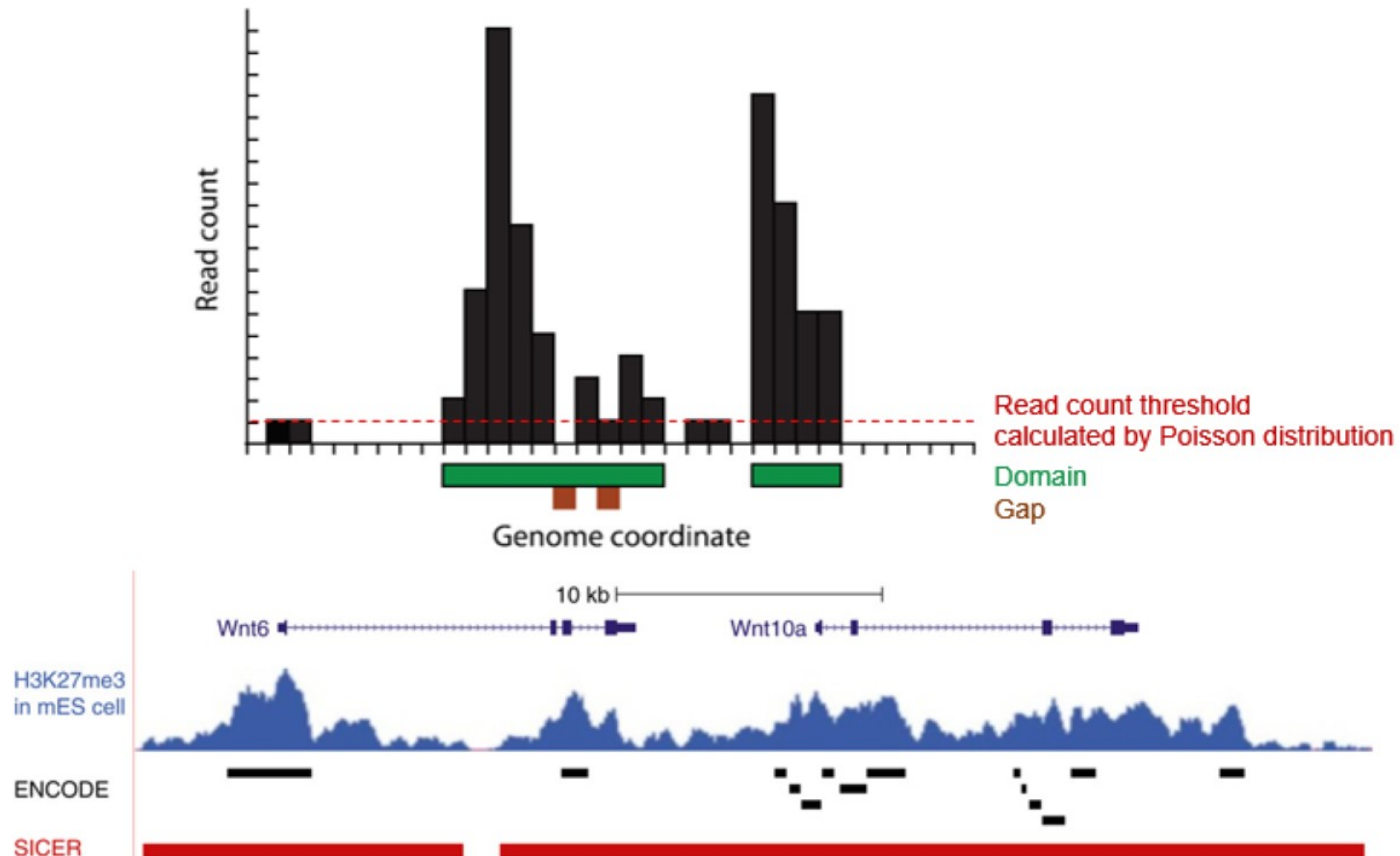
Poisson distribution based programs - MACS2



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

Domain calling programs

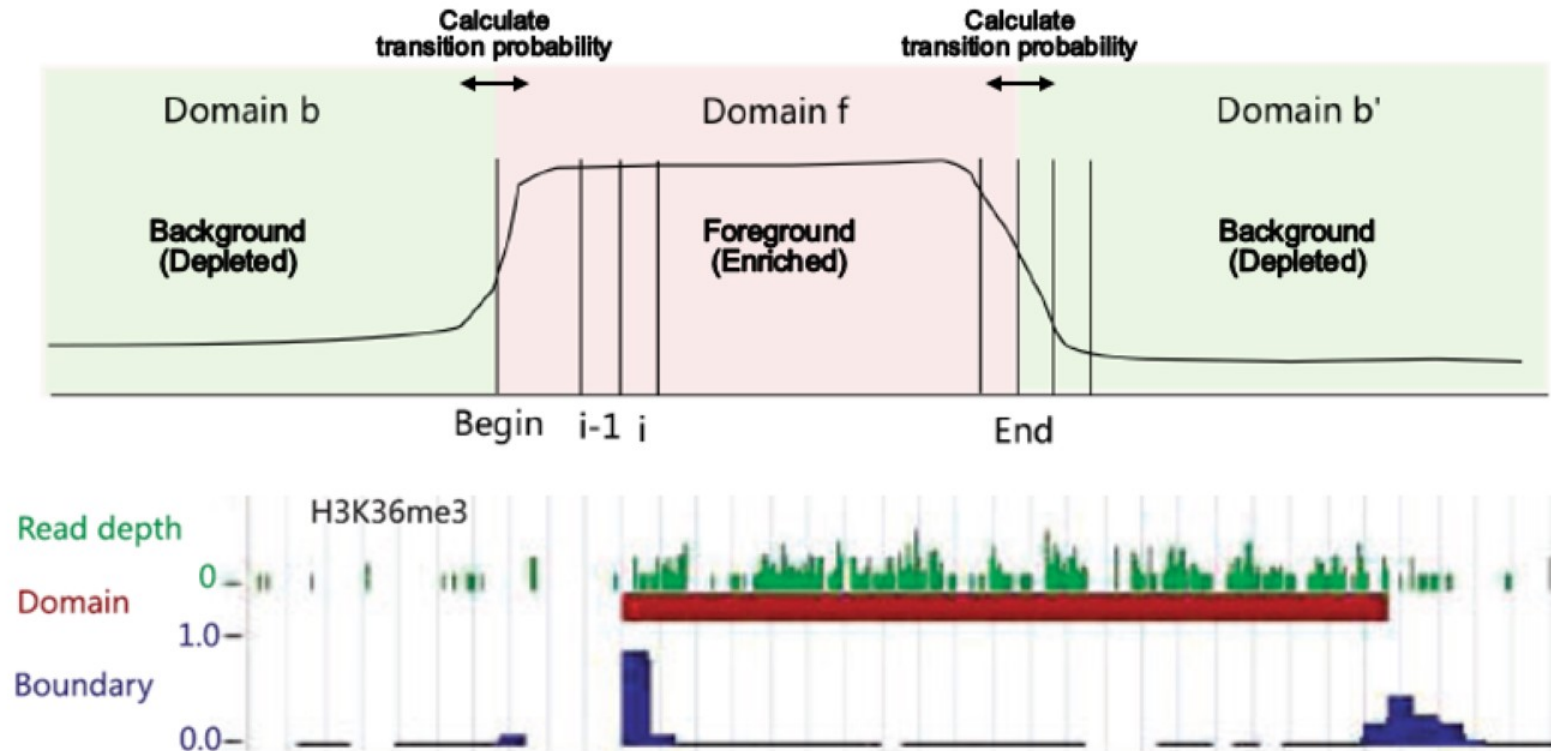
Poisson distribution based programs - SICER



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

Domain calling programs

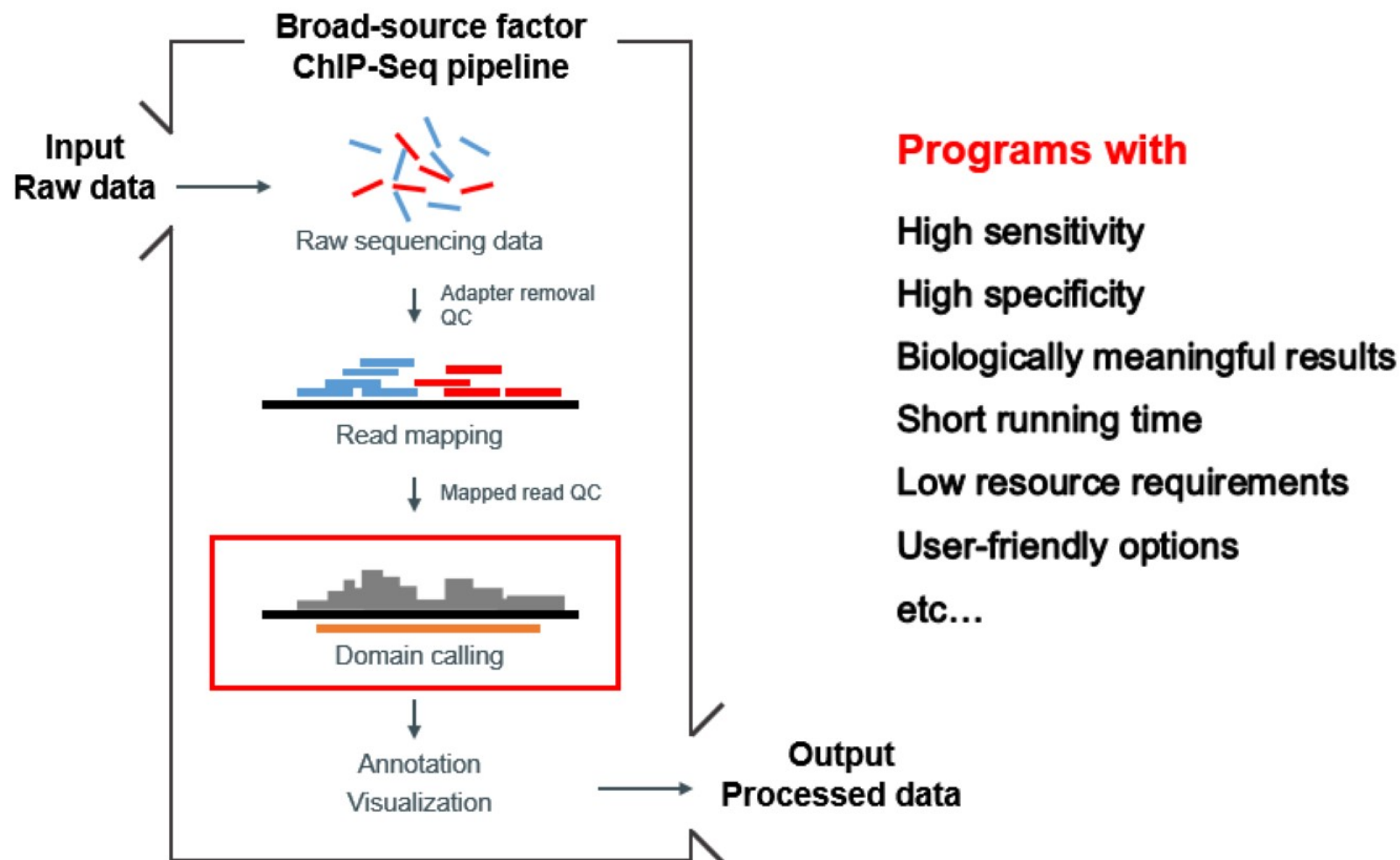
Hidden Markov Model (HMM) based programs - RSEG



RSEG determines domain boundaries using finite state HMM

Which program should we use?

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



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

국가생명연구자원정보센터

국가생명연구자원통합정보시스템

클라우드서비스

생명정보연구성과물등록시스템

유전체정보연계시스템

BIOEXPRESS

KORENG

검색

로그인

회원가입

서비스 소개

사용 방법

시연영상

교육 정보

교육 신청

분석 파이프라인

프로그램 요청

사용 통계

지원문의

Bio-Express

국가생명연구자원정보센터(KOBIC)에서는
대용량 유전체 데이터 분석이 필요한 국내 연구자를 위해
클라우드 기반의 Bio-Express 서비스를 제공하고 있습니다.

회원 가입

사용방법

매뉴얼

매뉴얼 PDF 다운로드

CLOSA

통합 자동 분석 시스템 접속

고속 전송 시스템 KODS 3.5 다운로드

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

CLOSHA under BioExpress @ KOBIC

Single sign-on for CLOSHA

User ID :

Password :

LOGIN

REGISTER ID/PW

HYBRID INFRA SYSTEM KoDS NGS DATA ANALYSIS CLOUD SERVICE PIPELINE DESIGN

<https://closa.kobic.re.kr/>

ChIP-Seq module in CLOSHA

CLOSHA

Analysis Project

Project Name | Created Date

ChIP_test	2017-10-23-15:44:02
ChIP_test_2	2017-10-23-20:41:49
011	2017-10-30-16:35:00
New_project	2017-12-17-22:33:50

Nw Project

File Browser

File Name	Last Modified Date
SRR067950.fastq	2017.10.23.15.24
SRR067977.fastq	2017.10.23.15.28
closa	2017.07.05.10.12
edu20171207	2017.12.06.22.31

File Browser | New Project | Delete Project | Share Project | Close

CLOSHA | File Browser | New_project

Project ID: C977BAA2-9F5C-4DC7-8A13-1A934244EE6A

Save | Run | Connection | Reset | Parameter | Status | History | Register | Console

START | Complete | fastqc | Complete | fastqc_1 | Complete | fastqc_2 | Complete | fastx_fastq_quality_filter | Complete | bowtie_se | Complete | homer_maketagdirectory | Complete | homer_makeucscfile | Complete | END

Details

Icon	Program Name	Type	Description	Settings	Parameter	Value	Description
1	START	MODULE	Start module				
2	END	MODULE	End module				
3	fastqc	Linux	FastQC aims to p...	Set...	input	/express/kobic_public/sample/pipeline/...	fastq or fq file
4	fastqc_1	Linux	FastQC aims to p...	Set...	output_folder	/express/lys1324/closa/project/New...	output path
5	fastx_fastq_quality_filter	Linux	Filters sequence...	Set...	output_file1	/express/lys1324/closa/project/New...	html file
6	fastx_fastq_quality_filter_1	Linux	Filters sequence...	Set...	output_file2	/express/lys1324/closa/project/New...	zip file
7	bowtie_se	Linux	Bowtie is an ultra...	Set...	fastqc_1		
8	fastqc_2	Linux	FastQC aims to p...	Set...	input	/express/kobic_public/sample/pipeline/...	fastq or fq file
9	bowtie_se_1	Linux	Bowtie is an ultra...	Set...	output_folder	/express/lys1324/closa/project/New...	output path
					output_file1	/express/lys1324/closa/project/New...	html file

ChIP-Seq module in CLOSHA

Project ID: C977BAA2-9F5C-4DC7-8A13-1A934244EE6A

1 Save 2 Run

3 OK

New_project

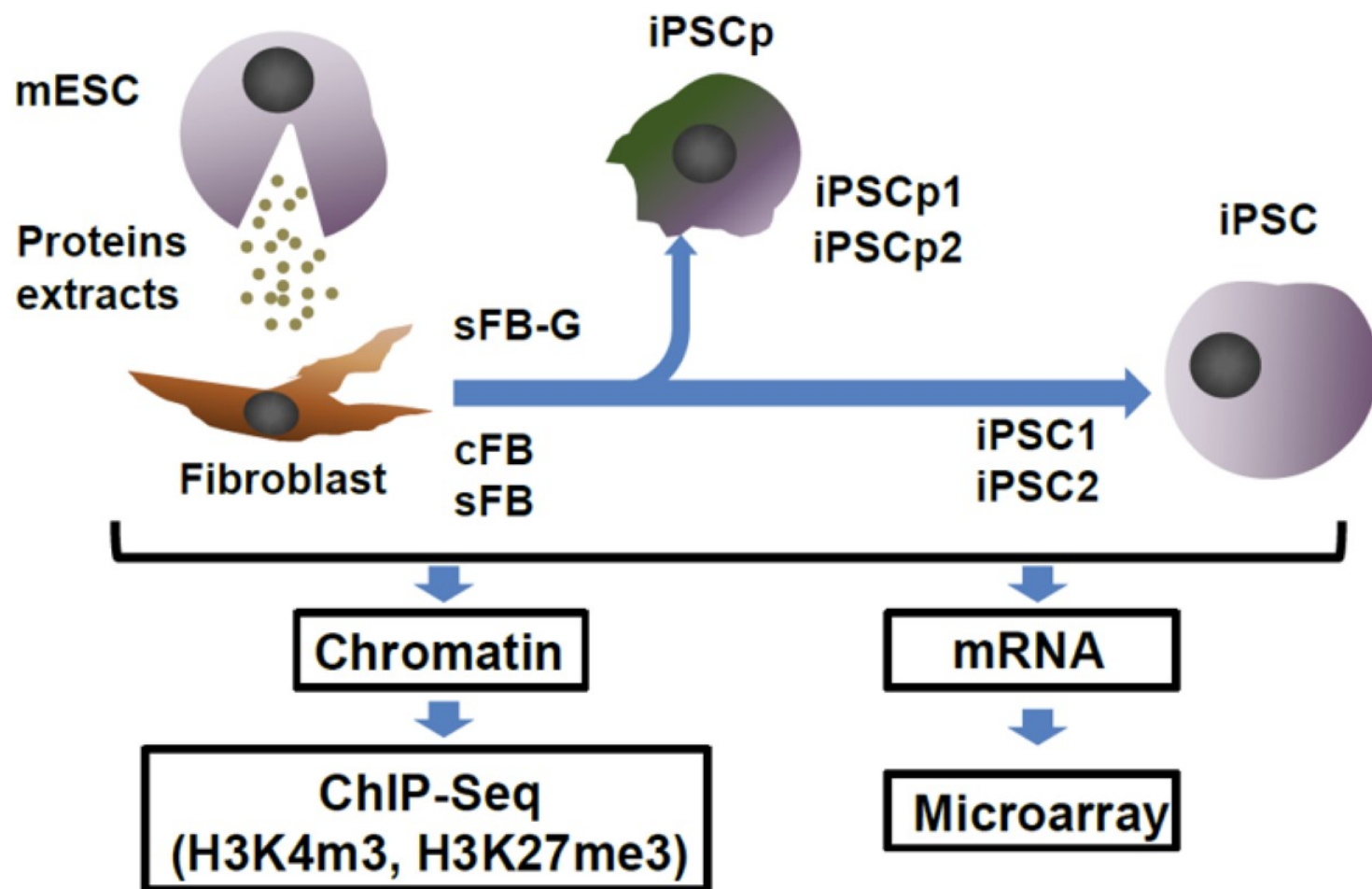
The execution of [New_project] analysis pipeline project has been completed.
Sun Dec 17 22:49:19 GMT+900 2017

4 New_project

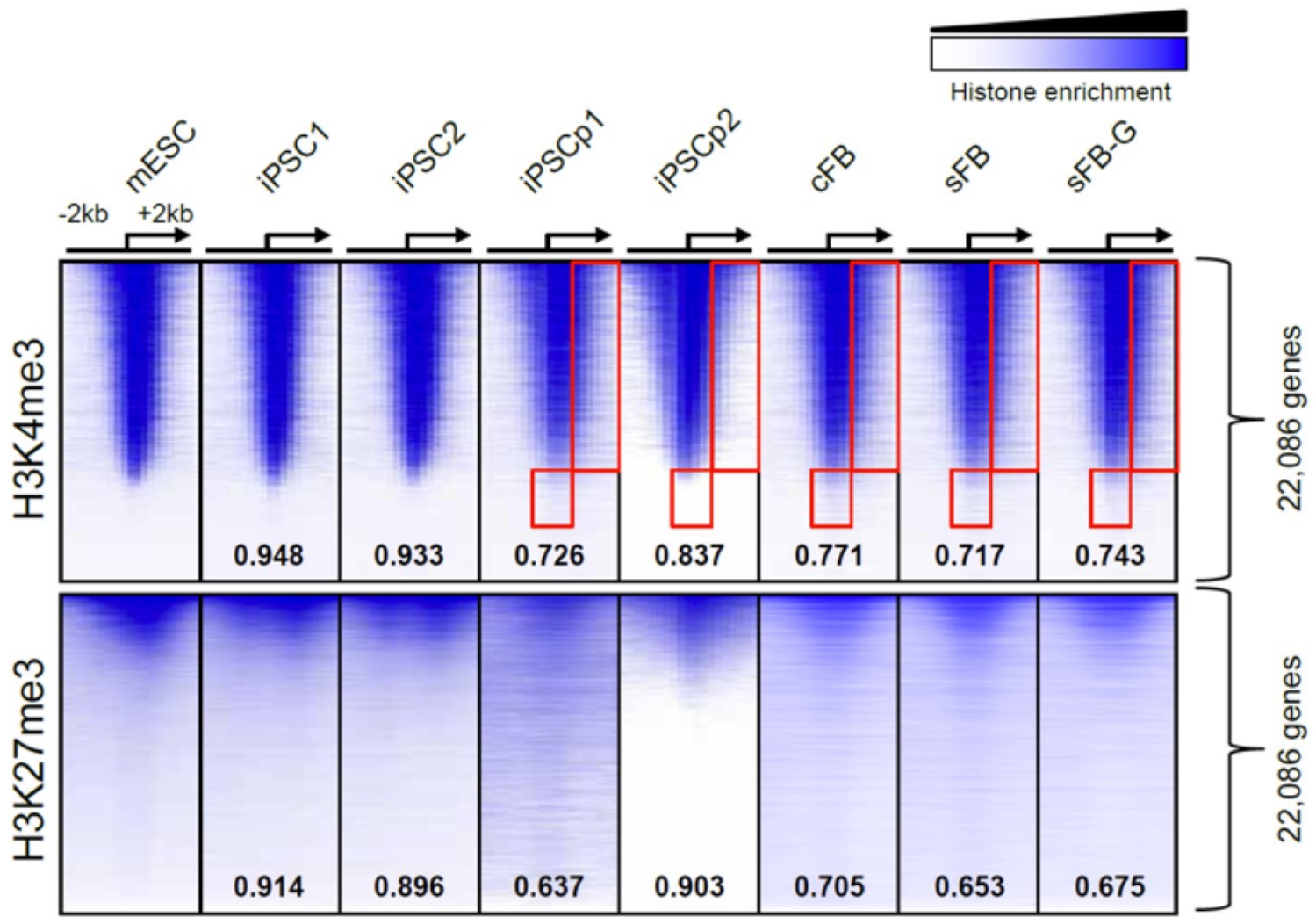
	Icon	Program Name	Type	Description	Settings	Parameter
1		START	MODULE	Start module		
2		END	MODULE	End module		
3		fastqc	LINUX	FastQC aims to p...		input
4		fastqc_1	LINUX	FastQC aims to p...		output_folder output_file1
5		fastx_fastq_quality_filter	LINUX	Filters sequence...		output_file2
6		fastx_fastq_quality_filter_1	LINUX	Filters sequence...		fastqc_1
7		bowtie_se	LINUX	Bowtie is an ultra...		input output_folder
8		fastqc_2	LINUX	FastQC aims to p...		output_folder
9		bowtie_se_1	LINUX	Bowtie is an ultra...		output_folder

iPSC vs. mES epigenome validation

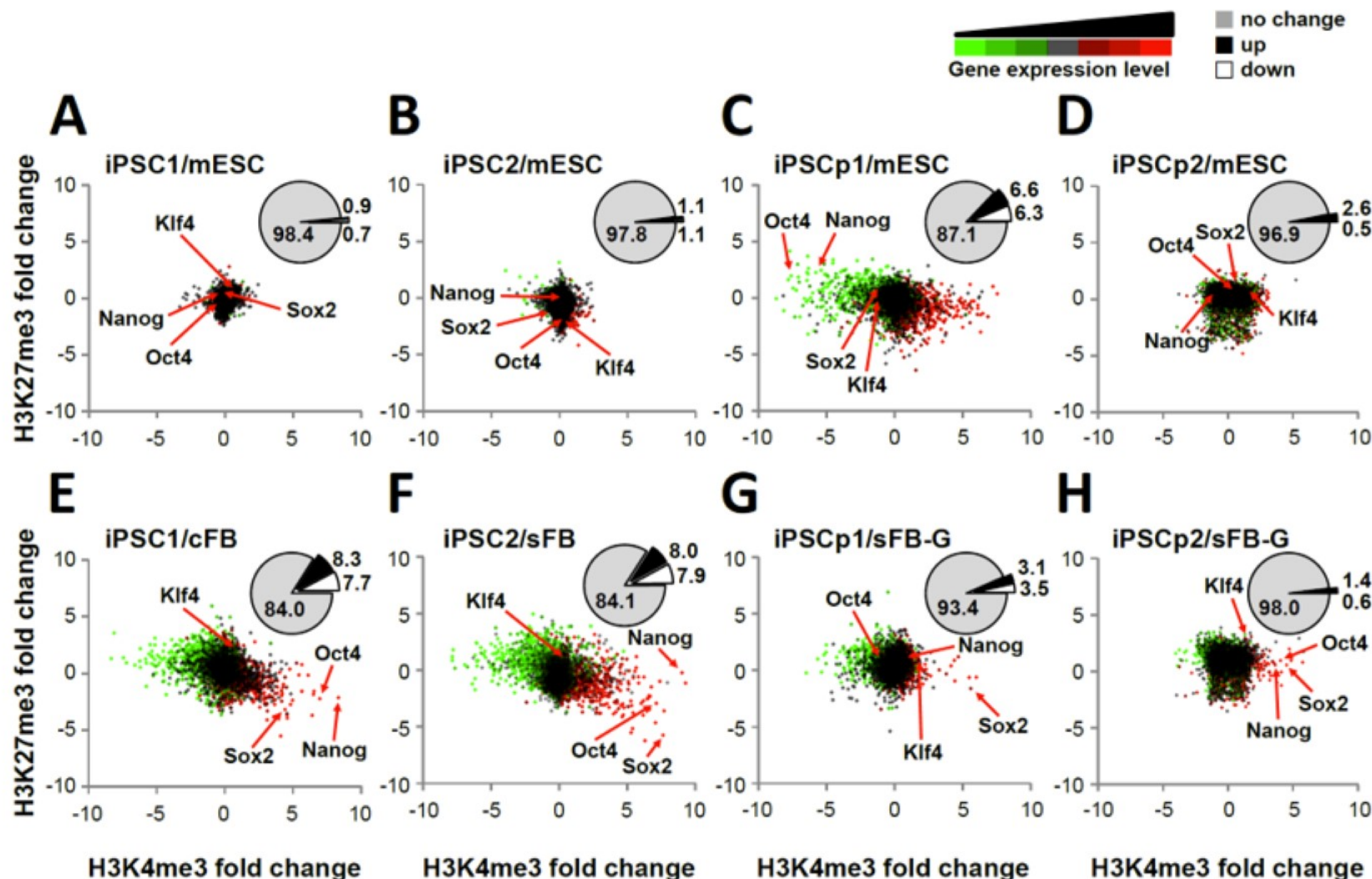
Experimental design



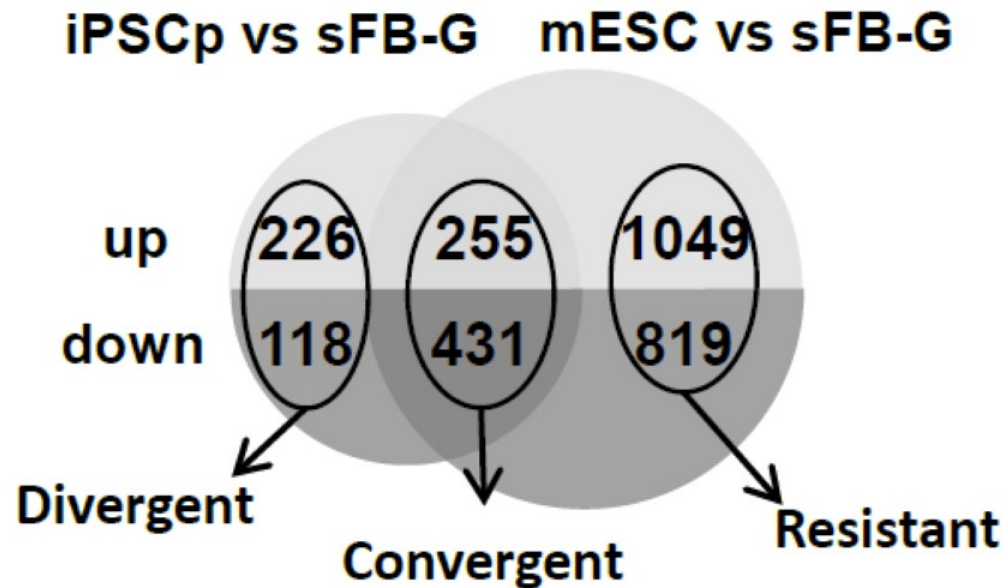
Histone modification comparison



Integration of expression data with histone modification

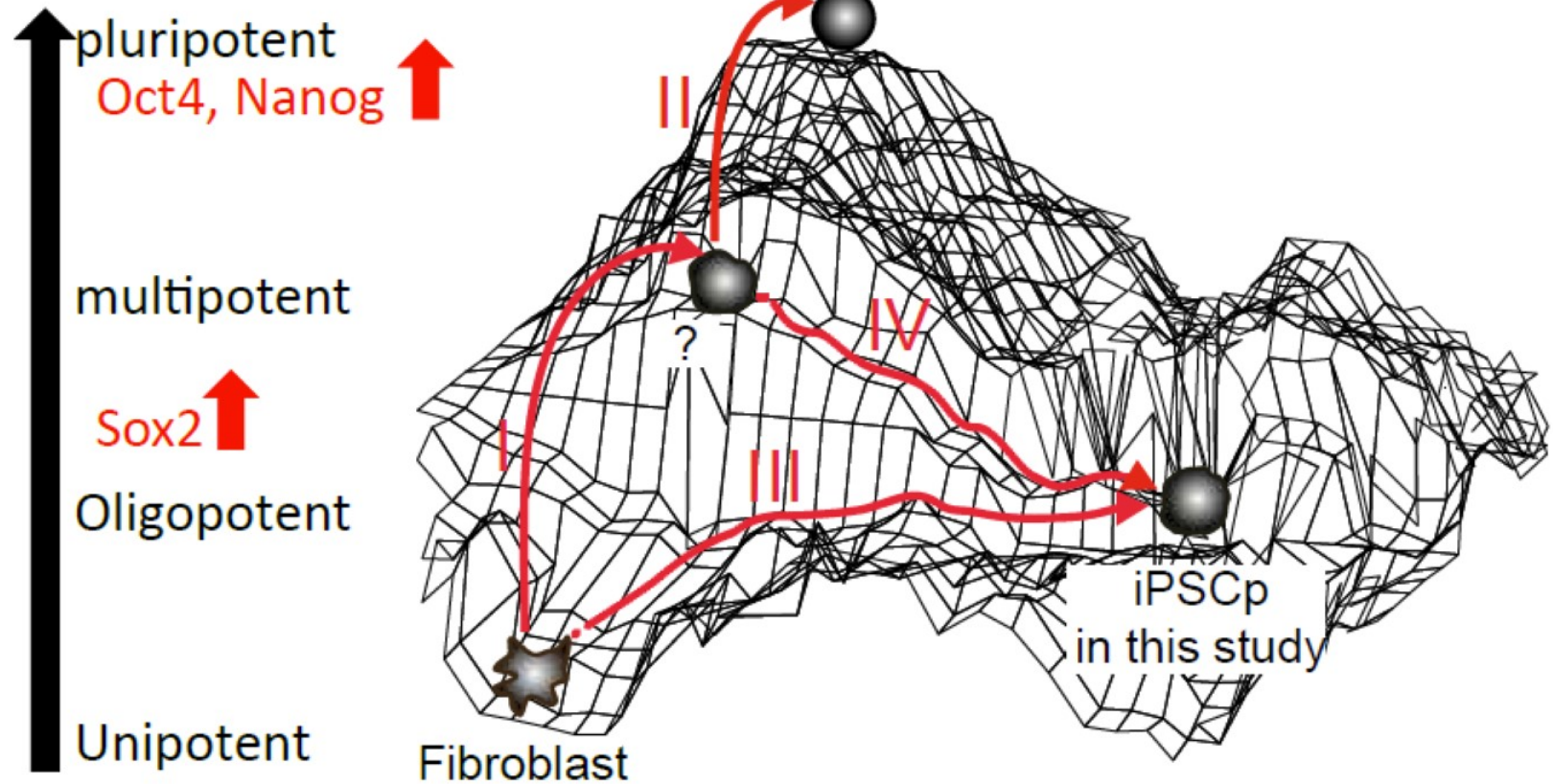


Gene-level comparson



Proposed epigenetic state during de-differentiation

Epigenetic cell potency



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.**