Summary of modERN peak calling and batch effects

Dec. 5, 2016

wt fly

Untagged TF

eGFP   IRES   neomycin

Tagged TF

Collect x-linked sheared chromatin

IP using GFP-Ab   Total Chromatin Input   IP using TF-Ab   Total Chromatin Input   IP using GFP-Ab

b"/b'   a/a'   a/a'

wt Control Experiment
signal over chromatin input

Typical ChIP-seq Experiment
signal over chromatin input

Typical modERN Experiment
signal over chromatin input

a/b"

wt subtracted Experiment
signal over *wt* IP

Alec

Attempts on peak calling

- modERN standard (a/a' IDR) [IDR, by A.V.]

- a/b'' with SPP [A.V.]

- Refine the peaks called from a/a' IDR by comparing a/a' and b''/b' with fisher exact test [J.X.]

- Start from the a/b'' peak regions, with corrected b'' signal as background, then call p-value from regression [K.Y.]

Number of peaks from different methods

Embryonic	STD IDR		A.V. (a/b'')	J.X. (relative frequency of a/a' and b''/b')		K.Y (a/b'')			
	a/a'	b''/b'	a/b"	fdr<0.05	fdr<10^{-2}	p<10^{-2}	p<10^{-3}	p<10^{-4}	p<10^{-5}
ab	1547	batch1: 5927 batch2: 7022 (each batch included three biological reps and one input)	0	388	180	210	120	75	43
ems	3963		33	42	21	233	130	84	54
ERR	2930		284	1158	884	558	383	284	229
luna	1639		24	51	15	77	33	14	9
SuHw	5475		1949	2150	2010	2618	2432	2316	2229

ab	IDR	Jinrui	KK
IDR	-	408/1547	218/1547
Jinrui	387/388	-	111/388
KK	209/210	112/210	-

ERR	IDR	Jinrui	KK
IDR	-	1166/2930	1150/2930
Jinrui	558/1158	-	404/1158
KK	550/558	403/558	-

SuHw	IDR	Jinrui	KK
IDR	-	2021/5475	2484/5475
Jinrui	1995/2150	-	1936/2150
KK	2460/2618	1952/2618	-

ems	IDR	Jinrui	KK
IDR	-	47/3963	227/3963
Jinrui	42/42	-	16/42
KK	223/233	16/233	-

luna	IDR	Jinrui	KK
IDR	-	51/1639	77/1639
Jinrui	51/51	-	5/51
KK	77/77	5/77	-

Pair-wise overlap

Sample peaks of su(Hw)

a/a'

STD IDR

J.X.

K.Y.

b''/b'

Some background peaks were successfully removed after further filters were applied.

SuHw-GFP_WA_E0-24_IP.bppbp c.hit0.fdr0.05.bed

spp.optimal.Su...d.filtered.bed

Sample peaks of four TFs

Motifs via MEME

Su(Hw)

MA0533.1
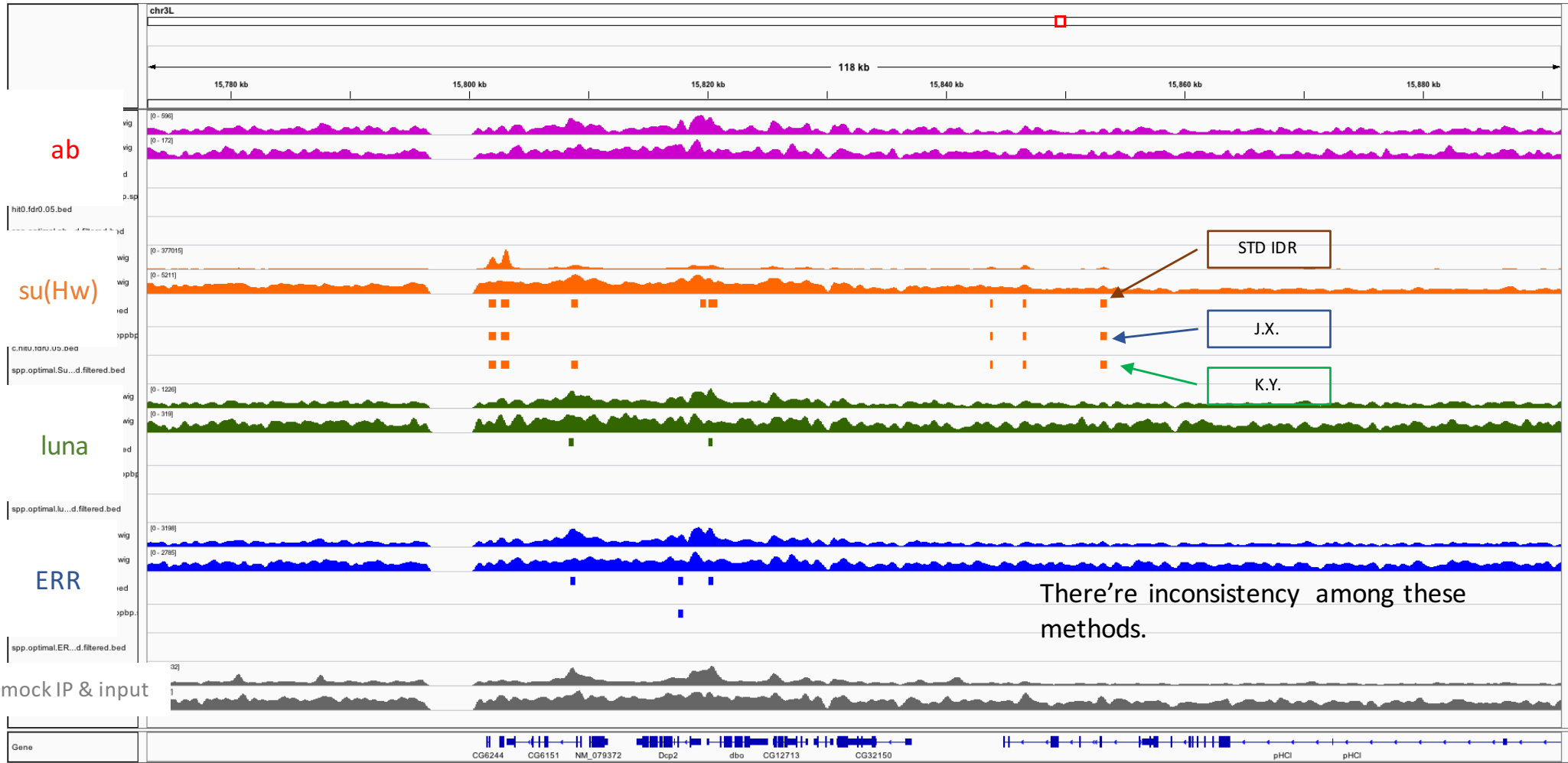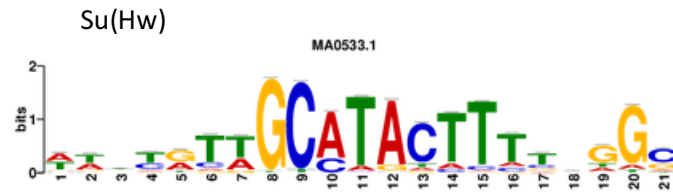
	AV		KK		JX			
					FDR<0.05		FDR<0.01	
	E-value	Seq_with_motif/seq_as_query	E-value	Seq_with_motif/seq_as_query	Evalue	Seq_with_motif/seq_as_query	Evalue	Seq_with_motif/seq_as_query
ab	-	-	2e-43	49/200	1e-69	101/372	4.9e-32	66/171
ems	-	-	7e-87	63/219	3e-8	35/42	4.9e-3	17/21
ERR	1e-55	102/284	5.8e-92	153/500	1e-66	166/1116	1e-82	159/846
luna	-	-	1e-6	77/77	1e-10	26/51	-	-
Su(Hw)	6.8e-1320	494/500	4e-582	359/500	6e-916	424/500	7e-606	436/500

meme.html files:
 AV: http://igsbmod.uchicago.edu/~avictorsen/zoops/
 KK: http://igsbmod.uchicago.edu/~avictorsen/YALE_2/
 JX 0.05: http://igsbmod.uchicago.edu/~avictorsen/YALE_3/0.05/
 JX 0.01: http://igsbmod.uchicago.edu/~avictorsen/YALE_3/0.01/

Alec

Thoughts

- Good quality data have better consistency across different peak calling approaches (e.g. su(Hw)).

- The overall overlap between Jinrui and Koon-Kiu's post-filtering peaks are reasonablely well.

- But in some dataset, the number of post-filtering peaks are very small.

- Suggestion: Define two sets of peaks
    1. Regular SPP-IDR peaks
    2. High-confidence/high-specificity peaks

Color Key

0.4   0.8
correlation coefficient

■ EM
■ W3L/WPP
■ ADULT
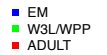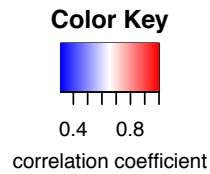
Purpose

1. To look up the variations between input data. *Is it driven by developmental stage or the date that the sample was collected/processed?*

Clustering of input

1. Correlation coefficient of input data was calculated pair-wise.
2. **The developmental stages** were color coded on the top.
3. **The dates of sample** processing were coded on the left.