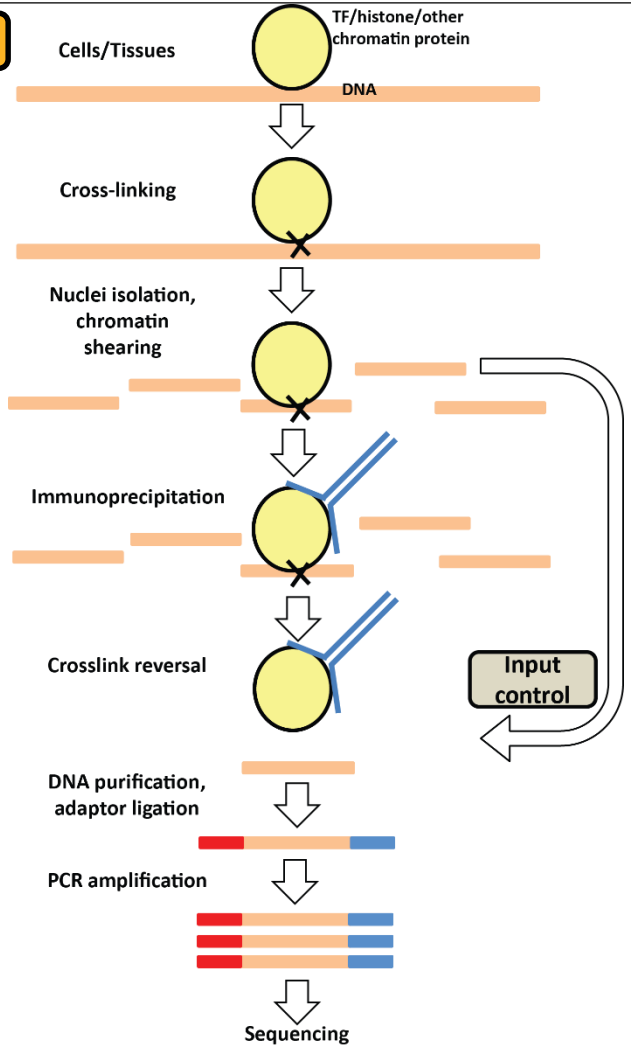
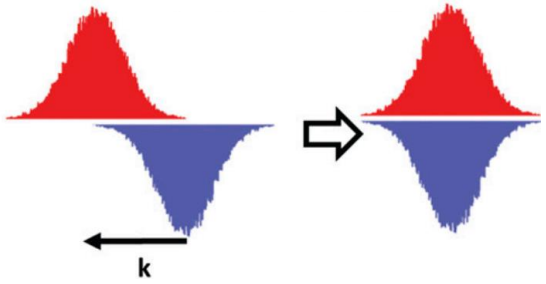
CHIP-SEQ VS. NEW TECHNOLOGIES

GEORGI K. MARINOV
AVANTI SHRIKUMAR

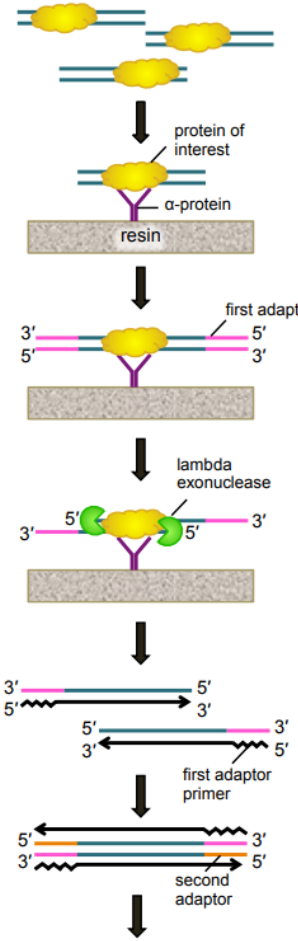
July 10th 2019

CHIP-SEQ


Cells/Tissues
TF/histone/other chromatin protein
DNA
Cross-linking
Nuclei isolation, chromatin shearing
Immunoprecipitation
Crosslink reversal
Input control
DNA purification, adaptor ligation
PCR amplification
Sequencing


k

CHIP-EXO



Crosslink proteins to DNA, shear DNA

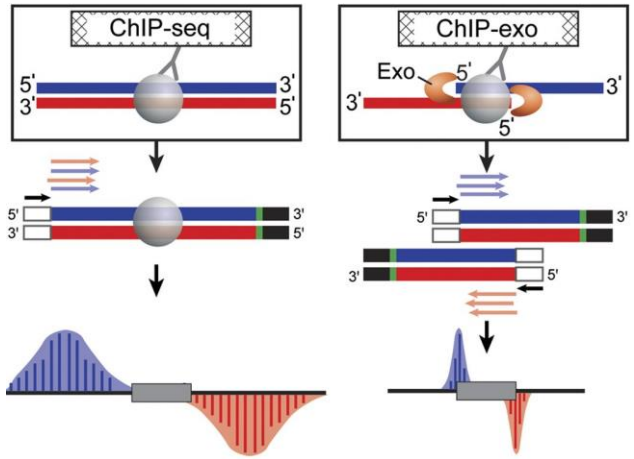
Chromatin-immunoprecipitation

Ligated first adaptor, fill in ends

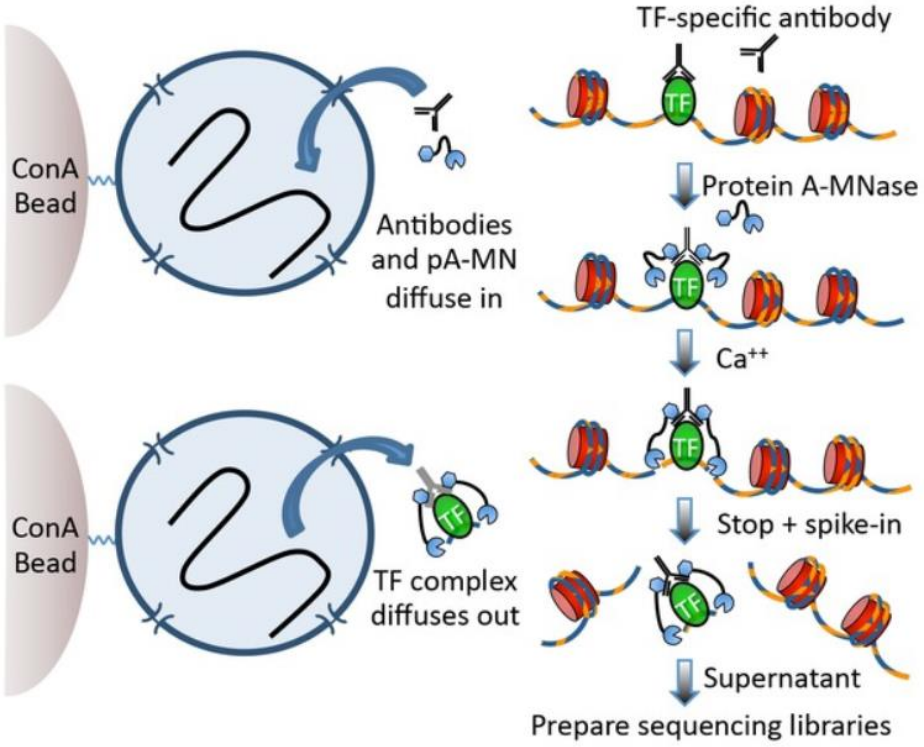
Exonuclease digestion

LM-PCR with primers to first adaptor

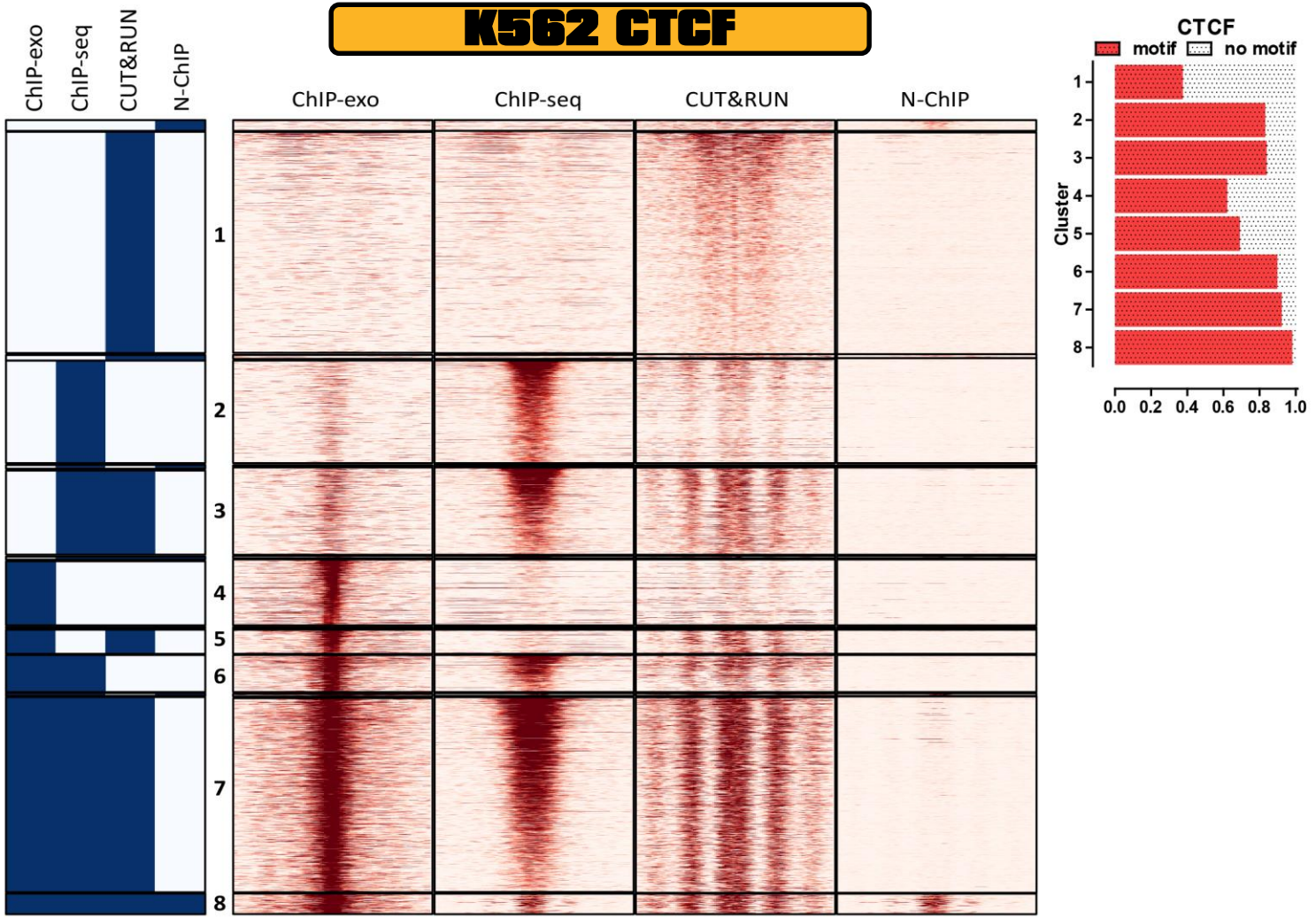
Ligate second adaptor, amplify DNA with LM-PCR



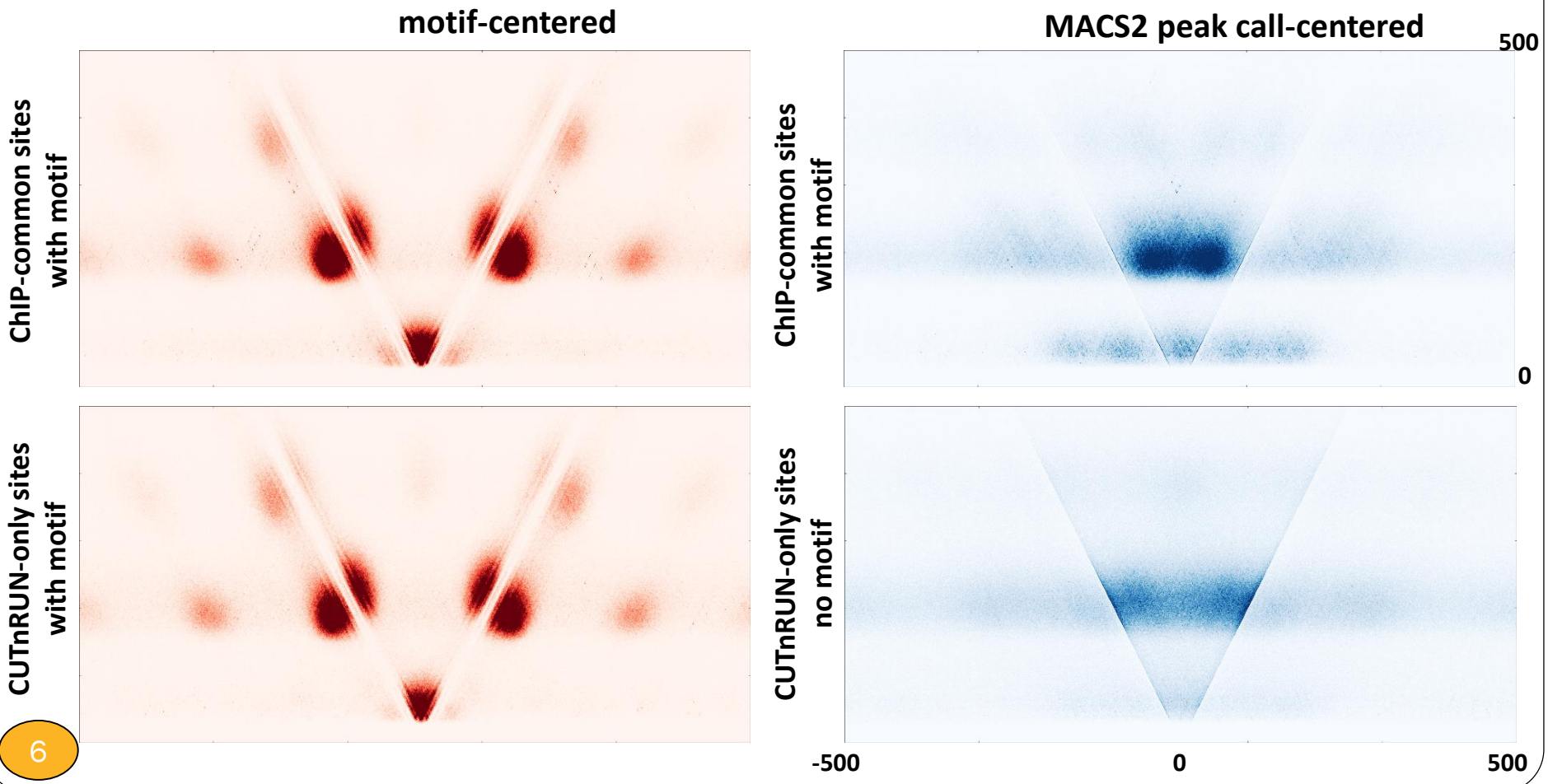
CHIP-EXO


TF-specific antibody
ConA Bead
Antibodies and pA-MN diffuse in
Protein A-MNase
Ca++
ConA Bead
TF complex diffuses out
Stop + spike-in
Supernatant
Prepare sequencing libraries

K562 CTCF


ChIP-exo
ChIP-seq
CUT&RUN
N-ChIP
1
2
3
4
5
6
7
8
CTCF
motif
no motif
Cluster
0.0
0.2
0.4
0.6
0.8
1.0

CTCF


motif-centered
MACS2 peak call-centered
ChIP-common sites with motif
ChIP-common sites with motif
CUTnRUN-only sites with motif
CUTnRUN-only sites no motif
500
0
-500
0
500

DATASETS:

1. ChIP-seq/ChIP-exo/CUT&RUN for CTCF in K562 cells
2. ChIP-seq/CUT&RUN for Myc/Max/GATA1 in K562 cells
3. ChIP-seq/CUT&RUN for FOSL2 in A549 cells
4. ChIP-seq/CUT&RUN for CTCF/Nanog/Oct4 in H1-hESC cells

(SOME) GOALS FOR WORKSHOP:

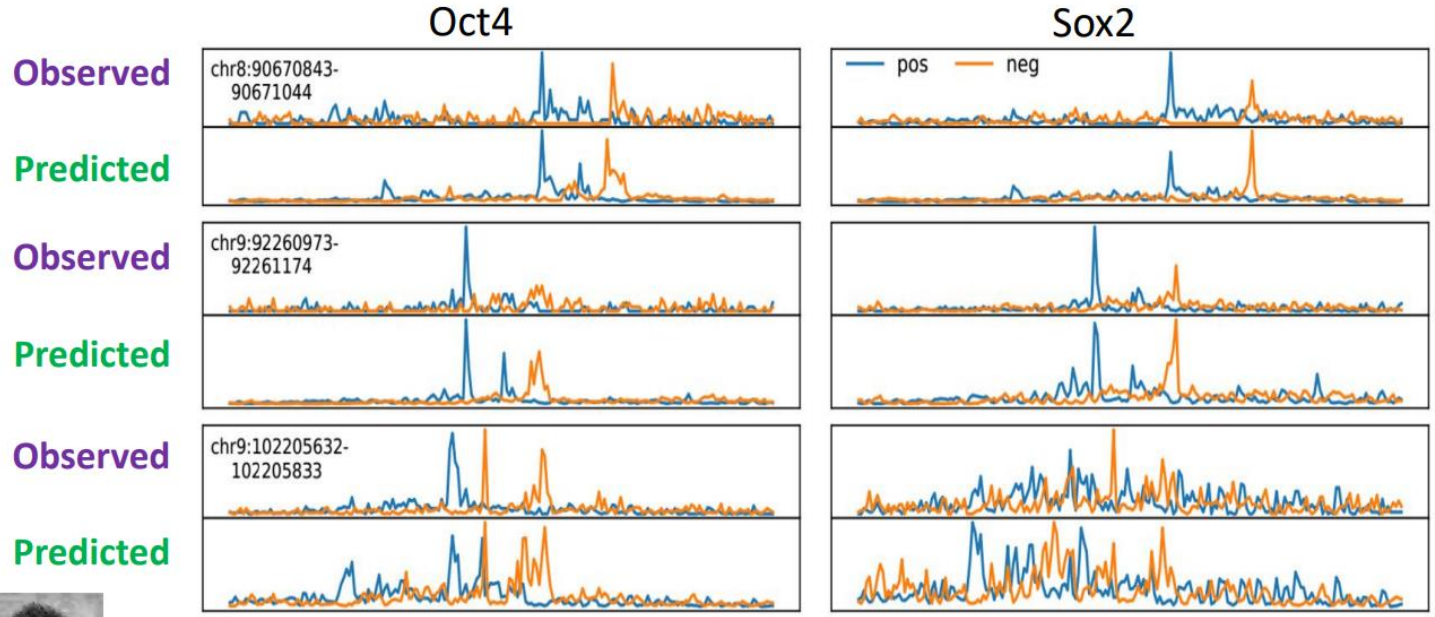
1. Can we develop predictive models that relate sequence to the signal profile of each assay, especially CUT&RUN?

2. Can we apply these models to understand the differences between the occupancy sets derived from the different assays at the sequence level?

3. Can we identify epigenetic properties that distinguish assay-specific sites based on other ENCODE assays?

PROFILE DNN MODELS

DNNs can accurately model bp. resolution TF footprints ChIP-nexus/exo



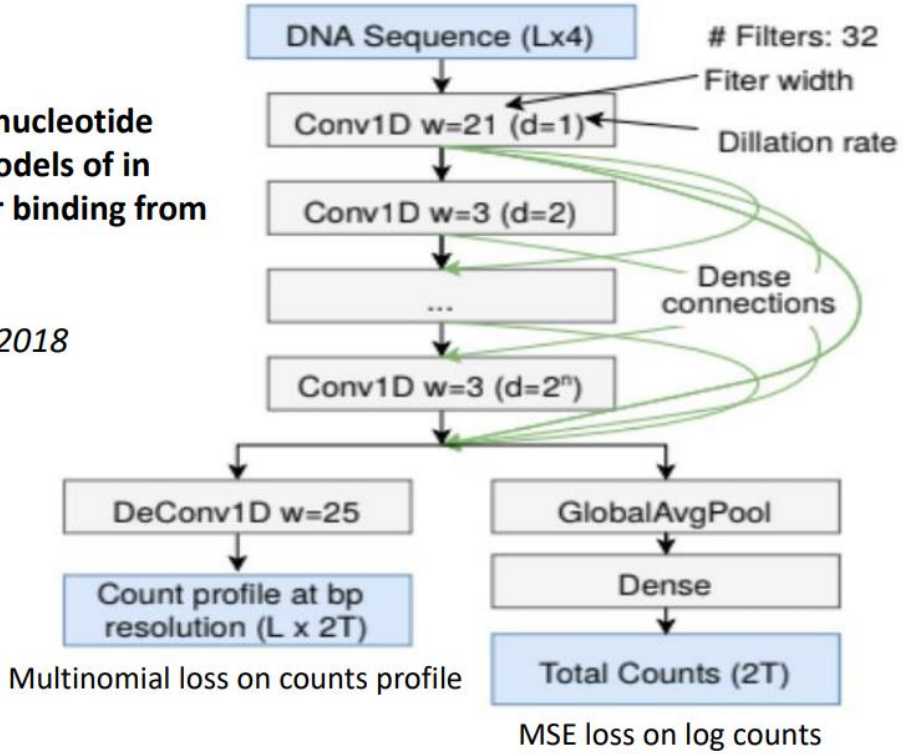
Žiga Avsec

“BPNet” architecture from Avsec et al., 2018

BPNet: Learning single-nucleotide resolution predictive models of in vivo transcription factor binding from ChIP-nexus data

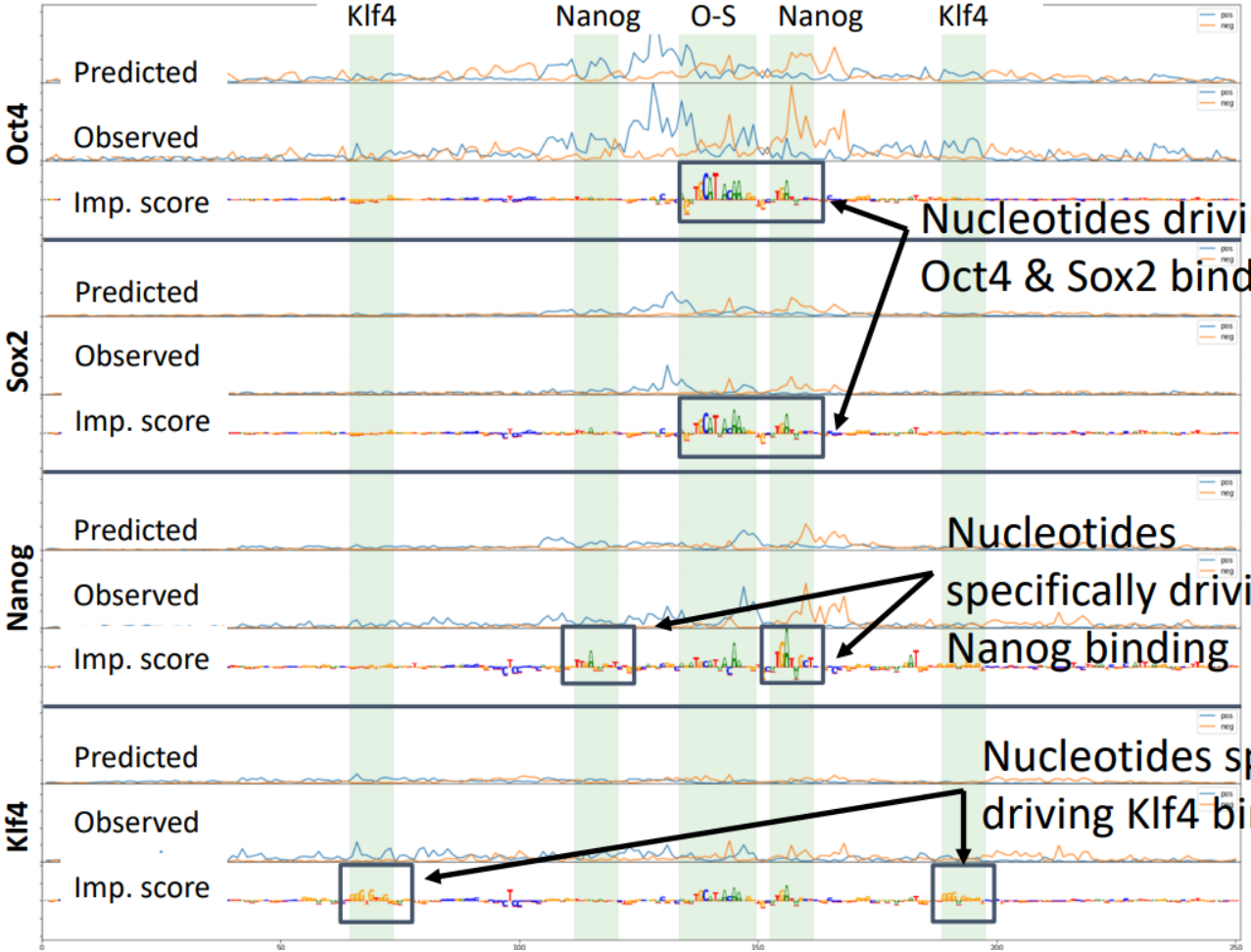
Avsec et al., ICML WCB, 2018

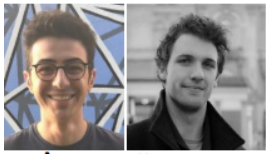
Žiga Avsec



Oct4 Distal Enhancer

Model interpretation reveals TF-specific regulatory sequence code, supported by footprints





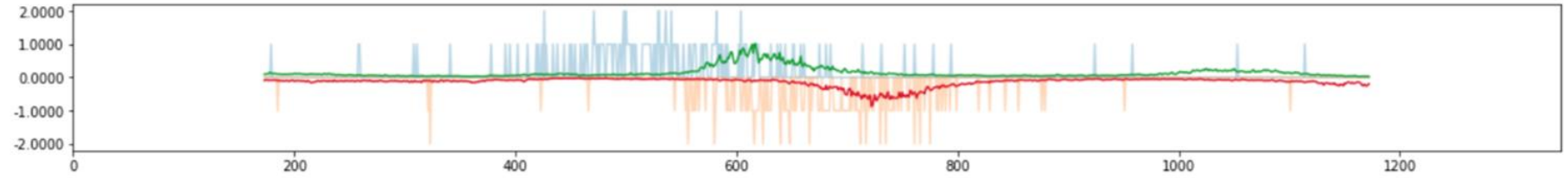
Amr Alexandari

Žiga Avsec

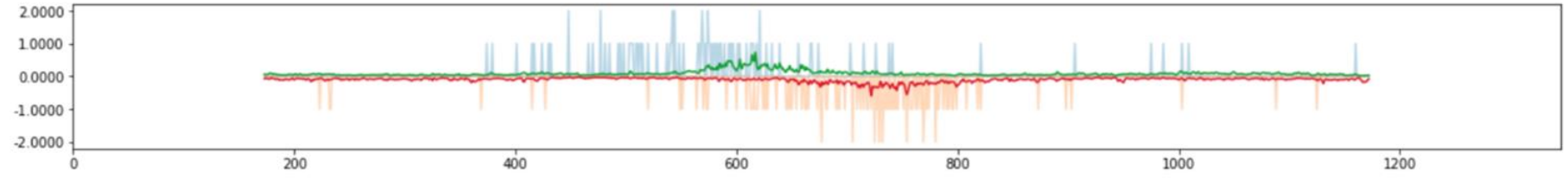
Initial model: ChIP-seq and CUT-N-RUN in H1ESCs

- CUT-N-RUN from “Pioneer Factor-Nucleosome Binding Events during Differentiation Are Motif Encoded” (Meers et al., Henikoff lab)
- ChIP-seq from ENCODE
- Started with model predicting POU5F1/OCT4 and NANOG ChIP-seq, and POU5F1/OCT4 CUT-N-RUN, given **only dna sequence**

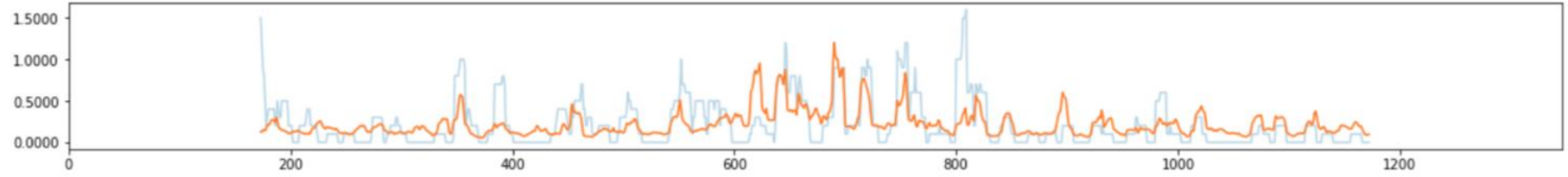
NANOG ChIPseq preds



POU5F1 ChIPseq preds

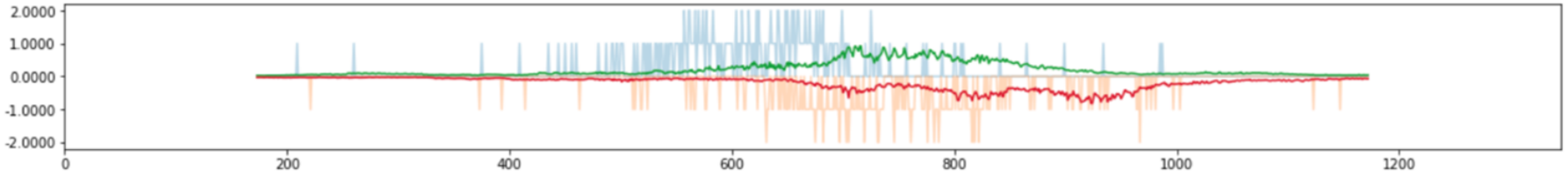


POU5F1 CUTNRUN preds

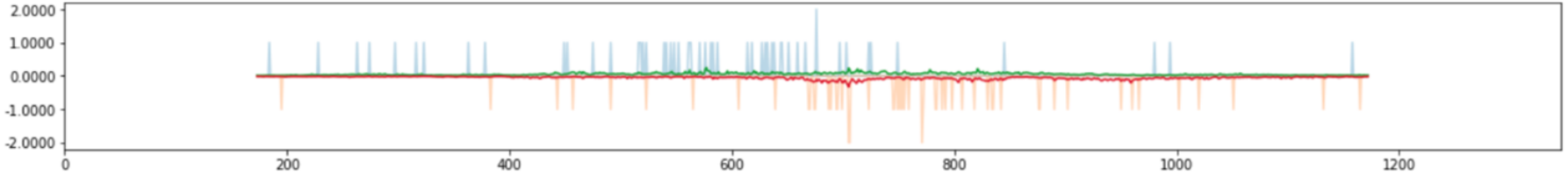


1923

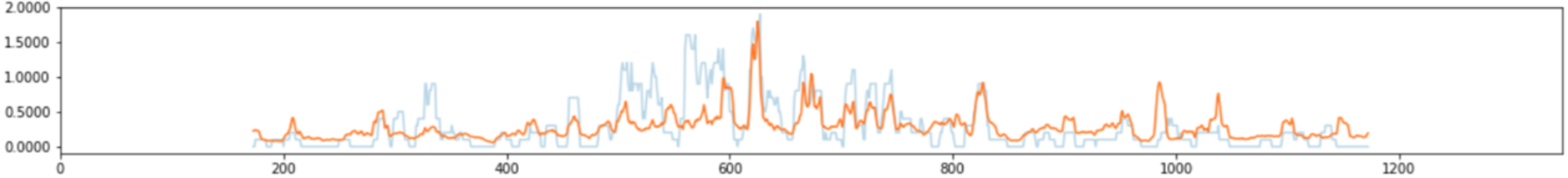
NANOG ChIPseq preds



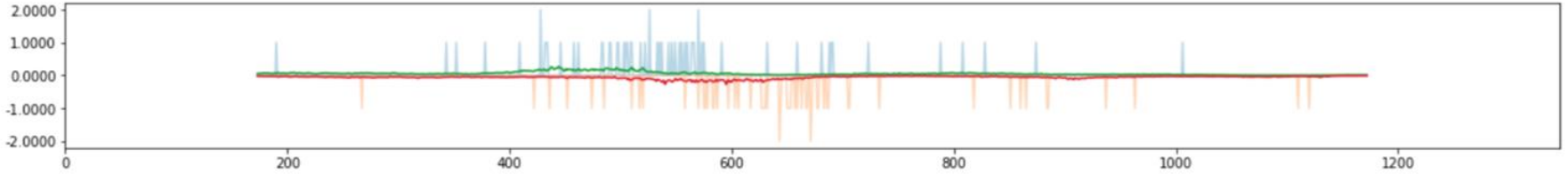
POU5F1 ChIPseq preds



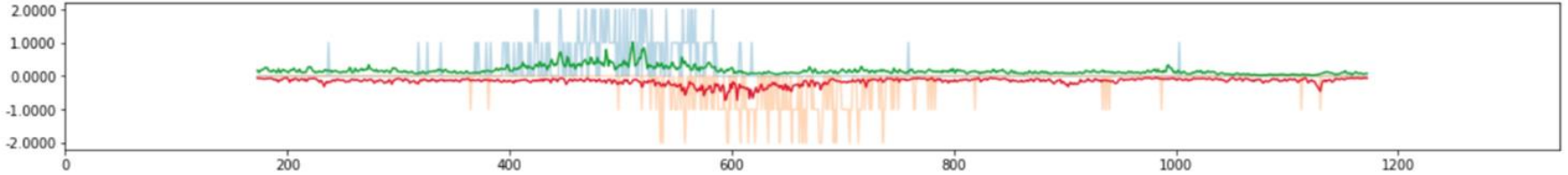
POU5F1 CUTNRUN preds



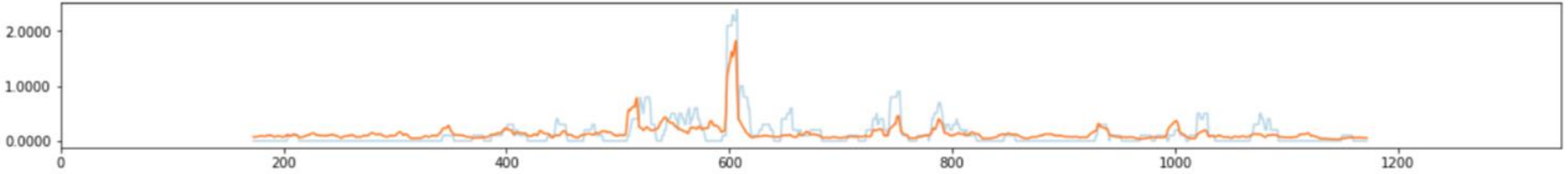
NANOG ChIPseq preds

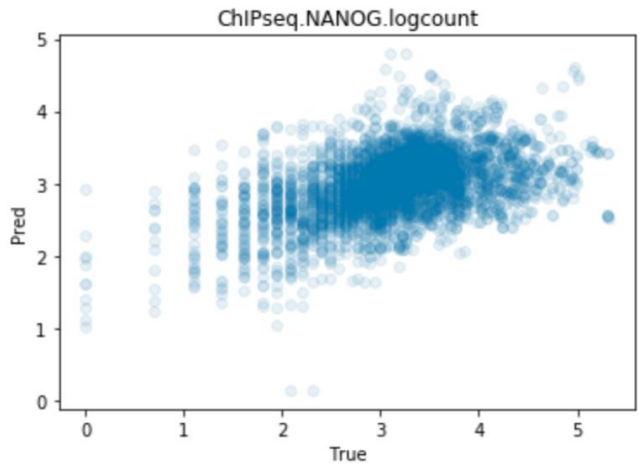


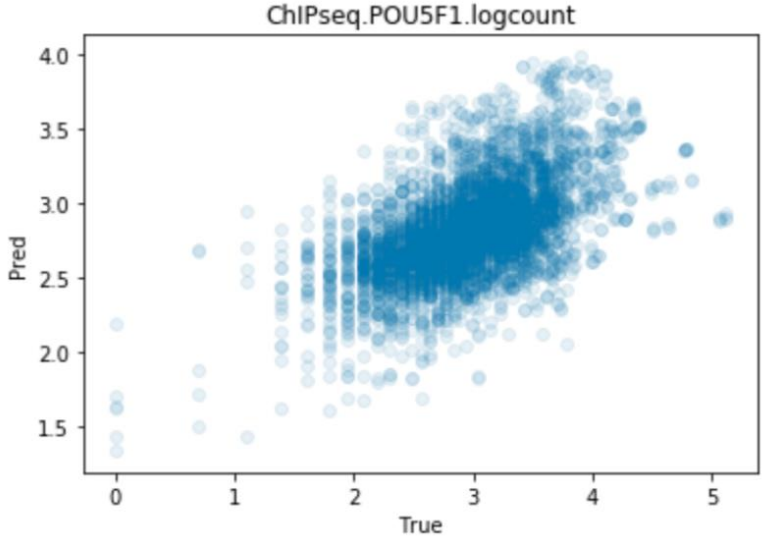
POU5F1 ChIPseq preds

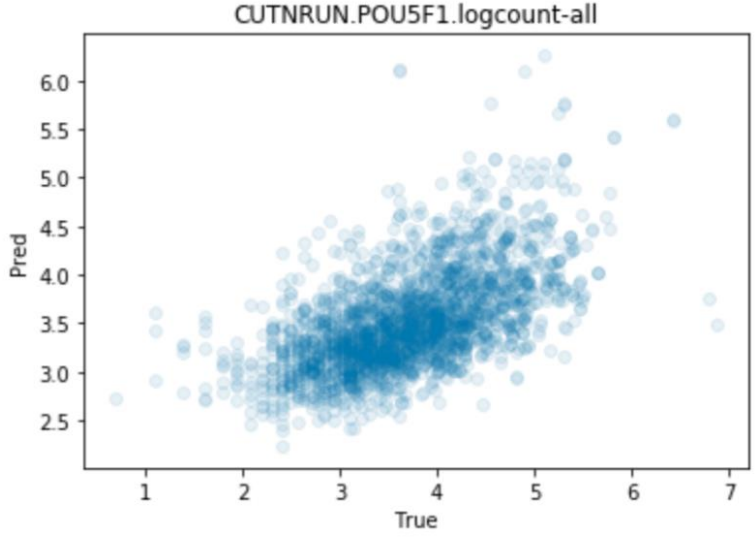


POU5F1 CUTNRUN preds









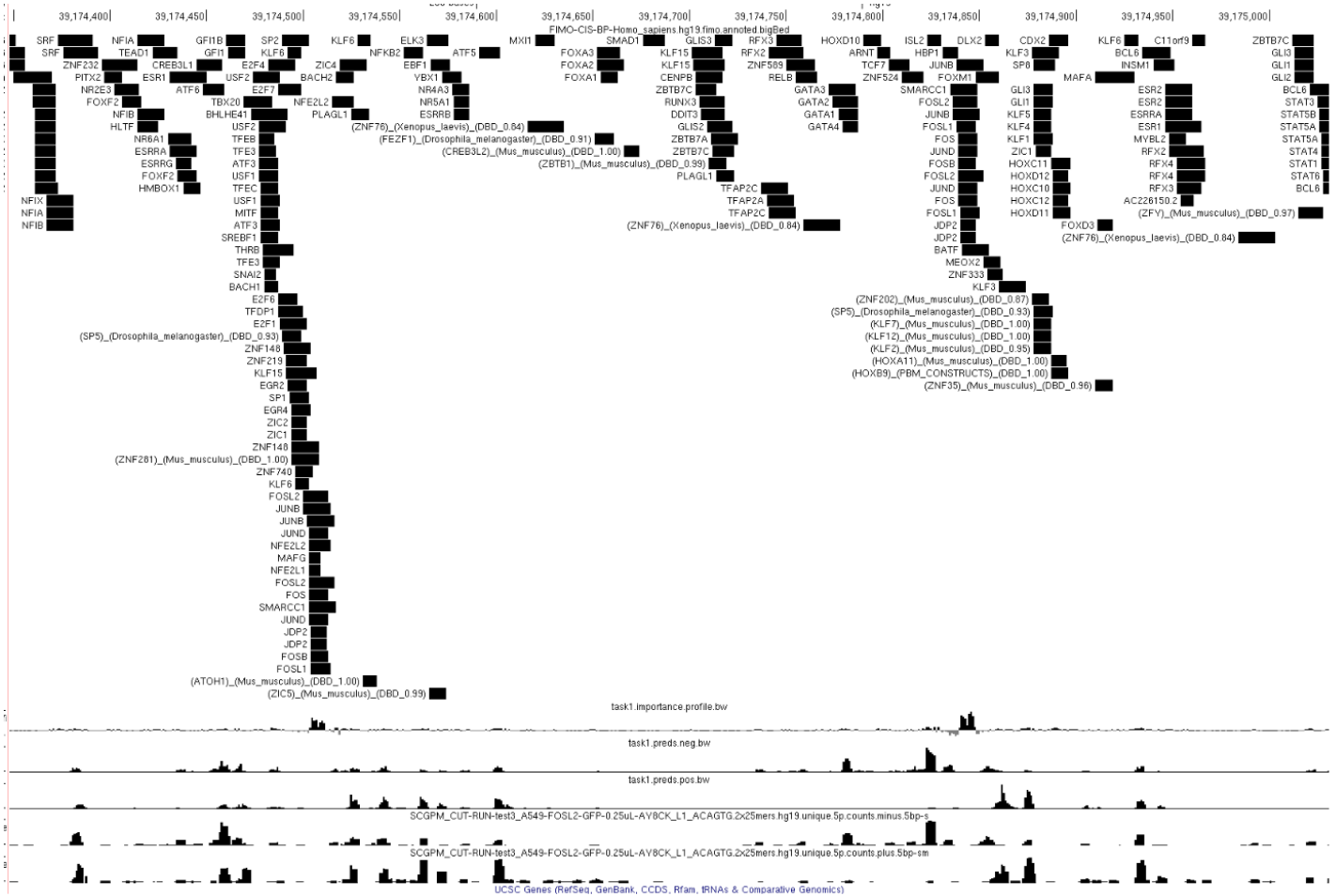
SpearmanrResult(correlation=0.5488101936918638,

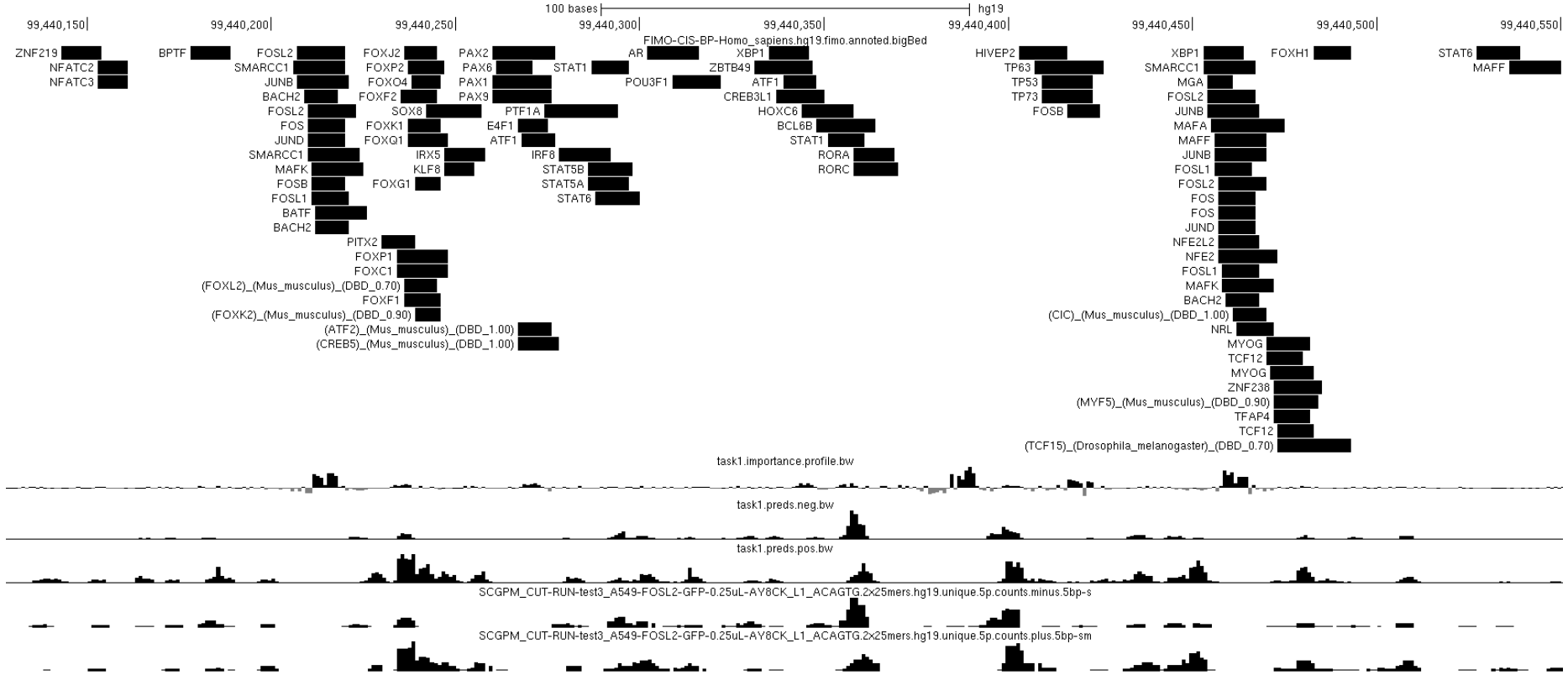
SpearmanrResult(correlation=0.5945533452203494,

More models:

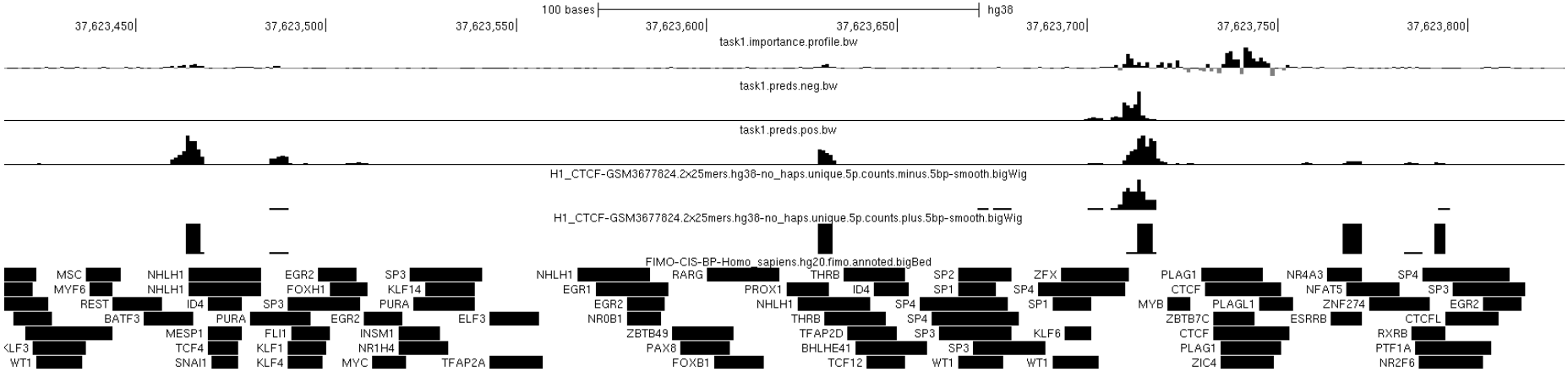
- - CTCF, FOSL1, Oct4, Nanog
- - Different smoothing strategies
- - Different fragment subsetting strategies (nucleosomal vs. subnucleosomal, etc.)

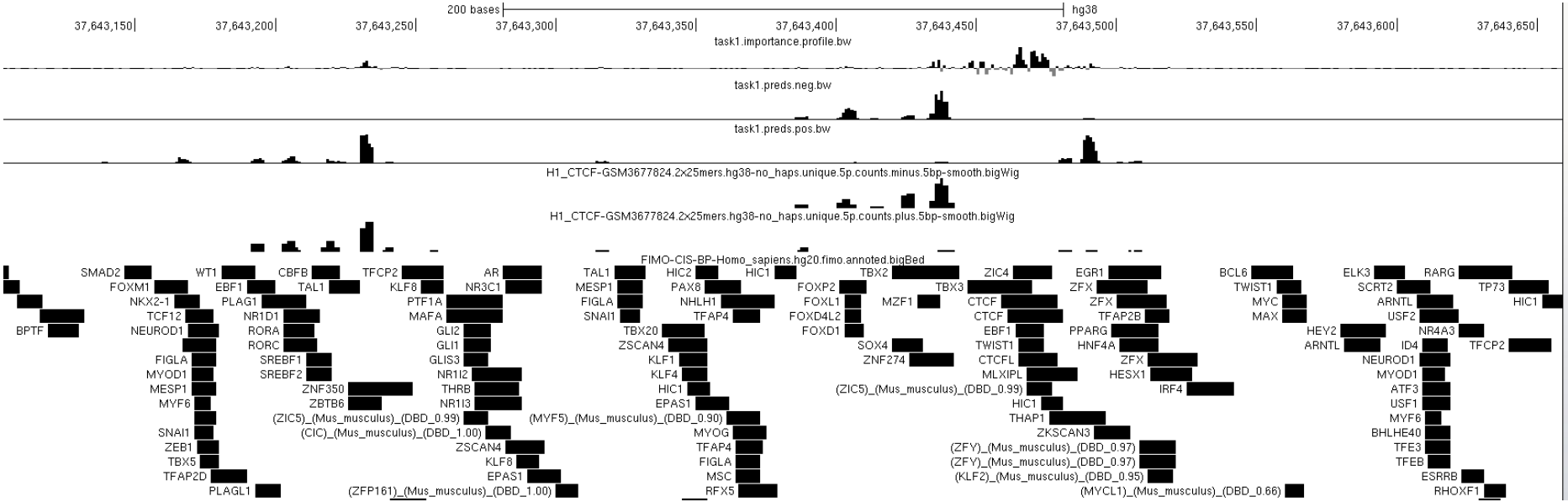
FOSL2 PROFILE MODELS

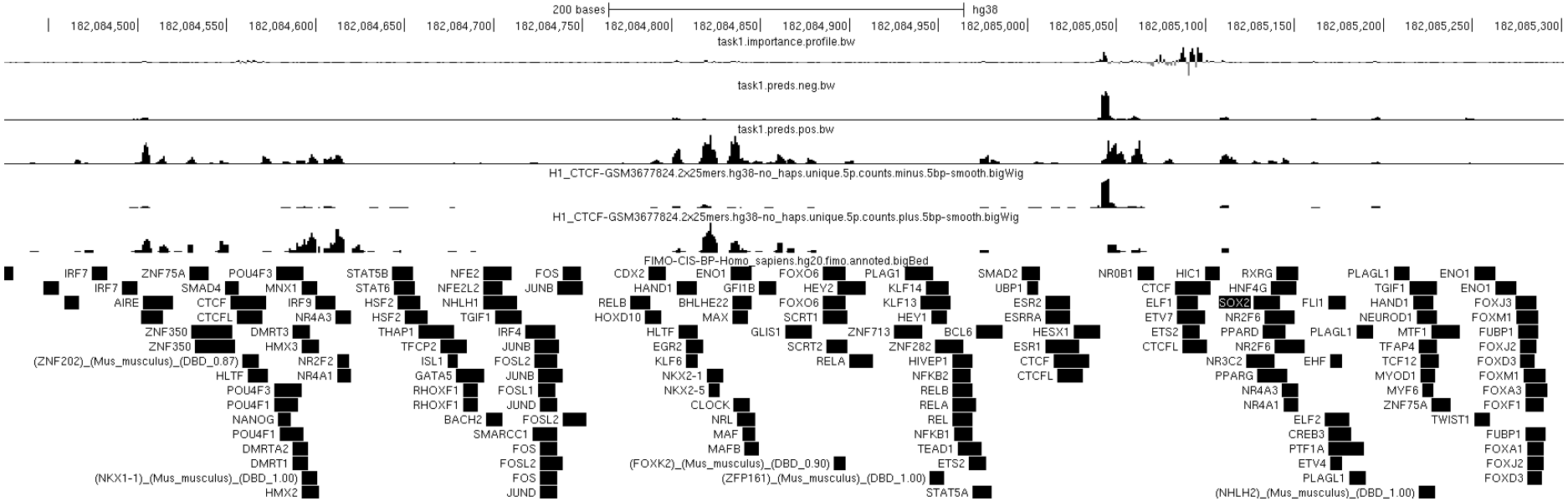



100 bases
hg19
99,440,150
99,440,200
99,440,250
99,440,300
99,440,350
99,440,400
99,440,450
99,440,500
99,440,550
FIMO-CIS-BP-Homo_sapiens.hg19.fimo.annotated.bigBed
ZNF219
NFATC2
NFATC3
BPTF
FOSL2
SMARCC1
JUNB
BACH2
FOSL2
FOS
JUND
SMARCC1
MAFK
FOSB
FOSL1
BATF
BACH2
PITX2
FOXP1
FOXC1
(FOXL2)_(Mus_musculus)_(DBD_0.70)
FOXF1
(FOXK2)_(Mus_musculus)_(DBD_0.90)
(ATF2)_(Mus_musculus)_(DBD_1.00)
(CREB5)_(Mus_musculus)_(DBD_1.00)
FOXJ2
FOXP2
FOXO4
FOXF2
SOX8
FOXK1
FOXQ1
IRX5
KLF6
FOXG1
PAX2
PAX6
PAX1
PAX9
PTF1A
E4F1
ATF1
IRF8
STAT5B
STAT5A
STAT6
AR
STAT1
POU3F1
XBP1
ZBTB49
ATF1
CREB3L1
HOXC6
BCL6B
STAT1
RORA
RORC
HIVEP2
TP63
TP53
TP73
FOSB
XBP1
SMARCC1
MGA
FOSL2
JUNB
MAFA
MAFF
JUNB
FOSL1
FOSL2
FOS
FOS
JUND
NFE2L2
NFE2
FOSL1
MAFK
BACH2
(CIC)_(Mus_musculus)_(DBD_1.00)
NRL
MYOG
TCF12
MYOG
ZNF238
(MYF5)_(Mus_musculus)_(DBD_0.90)
TFAP4
TCF12
(TCF15)_(Drosophila_melanogaster)_(DBD_0.70)
FOXH1
STAT6
MAFF
task1.importance.profile.bw
task1.preds.neg.bw
task1.preds.pos.bw
SCGPM_CUT-RUN-test3_A549-FOSL2-GFP-0.25uL-AY8CK_L1_ACAGTG.2x25mers.hg19.unique.5p.counts.minus.5bp-s
SCGPM_CUT-RUN-test3_A549-FOSL2-GFP-0.25uL-AY8CK_L1_ACAGTG.2x25mers.hg19.unique.5p.counts.plus.5bp-sm

CTCF PROFILE MODELS




200 bases
hg38
37,643,150
37,643,200
37,643,250
37,643,300
37,643,350
37,643,400
37,643,450
37,643,500
37,643,550
37,643,600
37,643,650
task1.importance.profile.bw
task1.preds.neg.bw
task1.preds.pos.bw
H1_CTCF-GSM3677824.2x25mers.hg38-no_haps.unique.5p.counts.minus.5bp-smooth.bigWig
H1_CTCF-GSM3677824.2x25mers.hg38-no_haps.unique.5p.counts.plus.5bp-smooth.bigWig
FIMO-CIS-BP-Homo_sapiens.hg20.fimo.annoted.bigBed
SMAD2
FOXM1
NKX2-1
TCF12
BPTF
NEUROD1
FIGLA
MYOD1
MESP1
MYF6
SNAI1
ZEB1
TBX5
TFAP2D
PLAGL1
WT1
EBF1
PLAG1
NR1D1
RORA
RORC
SREBF1
SREBF2
ZNF350
ZBTB6
(ZIC5)_(Mus_musculus)_(DBD_0.99)
(CIC)_(Mus_musculus)_(DBD_1.00)
ZSCAN4
KLF8
EPAS1
(ZFP161)_(Mus_musculus)_(DBD_1.00)
CBFB
TAL1
CBFB
TFCP2
KLF8
PTF1A
MAFA
GLI2
GLI1
GLIS3
NR1I2
THRB
NR1I3
AR
NR3C1
TAL1
MESP1
FIGLA
SNAI1
TBX20
ZSCAN4
KLF1
KLF4
HIC1
EPAS1
(MYF5)_(Mus_musculus)_(DBD_0.90)
MYOG
TFAP4
FIGLA
MSC
RFX5
HIC2
PAX8
NHLH1
TFAP4
HIC1
TBX2
FOXP2
FOXL1
FOXD4L2
FOXD1
SOX4
ZNF274
(ZIC5)_(Mus_musculus)_(DBD_0.99)
ZIC4
TBX3
MZF1
CTCF
CTCF
EBF1
TWIST1
CTCFL
MLXIPL
HIC1
THAP1
ZKSCAN3
(ZFY)_(Mus_musculus)_(DBD_0.97)
(ZFY)_(Mus_musculus)_(DBD_0.97)
(KLF2)_(Mus_musculus)_(DBD_0.95)
EGR1
ZFX
ZFX
TFAP2B
PPARG
HNF4A
ZFX
HESX1
IRF4
(MYCL1)_(Mus_musculus)_(DBD_0.66)
BCL6
TWIST1
MYC
MAX
HEY2
ARNTL
ELK3
SCRT2
ARNTL
USF2
NR4A3
ID4
NEUROD1
MYOD1
ATF3
USF1
MYF6
BHLHE40
TFE3
TFEB
ESRRB
RHOXF1
RARG
TP73
HIC1
TFCP2


200 bases
hg38
182,084,500
182,084,550
182,084,600
182,084,650
182,084,700
182,084,750
182,084,800
182,084,850
182,084,900
182,084,950
182,085,000
182,085,050
182,085,100
182,085,150
182,085,200
182,085,250
182,085,300
task1.importance.profile.bw
task1.preds.neg.bw
task1.preds.pos.bw
H1_CTCF-GSM3677824.2x25mers.hg38-no_haps.unique.5p.counts.minus.5bp-smooth.bigWig
H1_CTCF-GSM3677824.2x25mers.hg38-no_haps.unique.5p.counts.plus.5bp-smooth.bigWig
FIMO-CIS-BP-Homo_sapiens.hg20.fimo.annoted.bigBed
IRF7
ZNF75A
POU4F3
IRF7
SMAD4
MNX1
AIRE
CTCF
IRF9
CTCFL
NR4A3
ZNF350
DMRT3
ZNF350
HMX3
(ZNF202)_(Mus_musculus)_(DBD_0.67)
NR2F2
HLTF
NR4A1
POU4F3
POU4F1
NANOG
POU4F1
DMRTA2
DMRT1
(NKX1-1)_(Mus_musculus)_(DBD_1.00)
HMX2
STAT5B
NFE2
FOS
STAT6
NFE2L2
JUNB
HSF2
NHLH1
HSF2
TGIF1
THAP1
IRF4
TFCP2
JUNB
ISL1
FOSL2
GATA5
JUNB
RHOXF1
FOSL1
RHOXF1
JUND
BACH2
FOSL2
SMARCC1
FOS
FOSL2
FOS
JUND
CDX2
ENO1
FOXO6
PLAG1
HAND1
GFI1B
HEY2
KLF14
RELB
BHLHE22
FOXO6
KLF13
HOXD10
MAX
SCRT1
HEY1
HLTF
GLIS1
ZNF713
BCL6
EGR2
SCRT2
ZNF282
KLF6
RELA
HIVEP1
NKX2-1
NFKB2
NKX2-5
RELB
CLOCK
RELA
NRL
REL
MAF
NFKB1
MAFB
TEAD1
(FOXK2)_(Mus_musculus)_(DBD_0.90)
ETS2
(ZFP161)_(Mus_musculus)_(DBD_1.00)
STAT5A
SMAD2
UBP1
ESR2
ESRRA
HESX1
ESR1
CTCF
CTCFL
NR0B1
HIC1
RXRG
CTCF
HNF4G
ELF1
SOX2
FLI1
ETV7
NR2F6
ETS2
PPARD
CTCFL
NR2F6
NR3C2
EHF
PPARG
NR4A3
NR4A1
ELF2
CREB3
PTF1A
ETV4
PLAGL1
(NHLH2)_(Mus_musculus)_(DBD_1.00)
PLAGL1
ENO1
TGIF1
ENO1
HAND1
FOXJ3
NEUROD1
FOXM1
PLAGL1
MTF1
FUBP1
TFAP4
FOXJ2
TCF12
FOXD3
MYOD1
FOXM1
MYF6
FOXA3
ZNF75A
FOXF1
TWIST1
FUBP1
FOXA1
FOXJ2
FOXD3

MOVING FORWARD

1. Refine and improve profile models for CUT&RUN datasets

2. Work out V-plot based models

3. Identify determinants of assay-specific enriched regions