NYU School of Medicine
NYU LANGONE MEDICAL CENTER

ChIPSeqSpike: ChIP-Seq data scaling according to spike-in control

04/19/2018

Nicolas Descostes, Ph.D
Danny Reinberg laboratory

hhmi | Howard Hughes Medical Institute

ChIP-Seq protocol

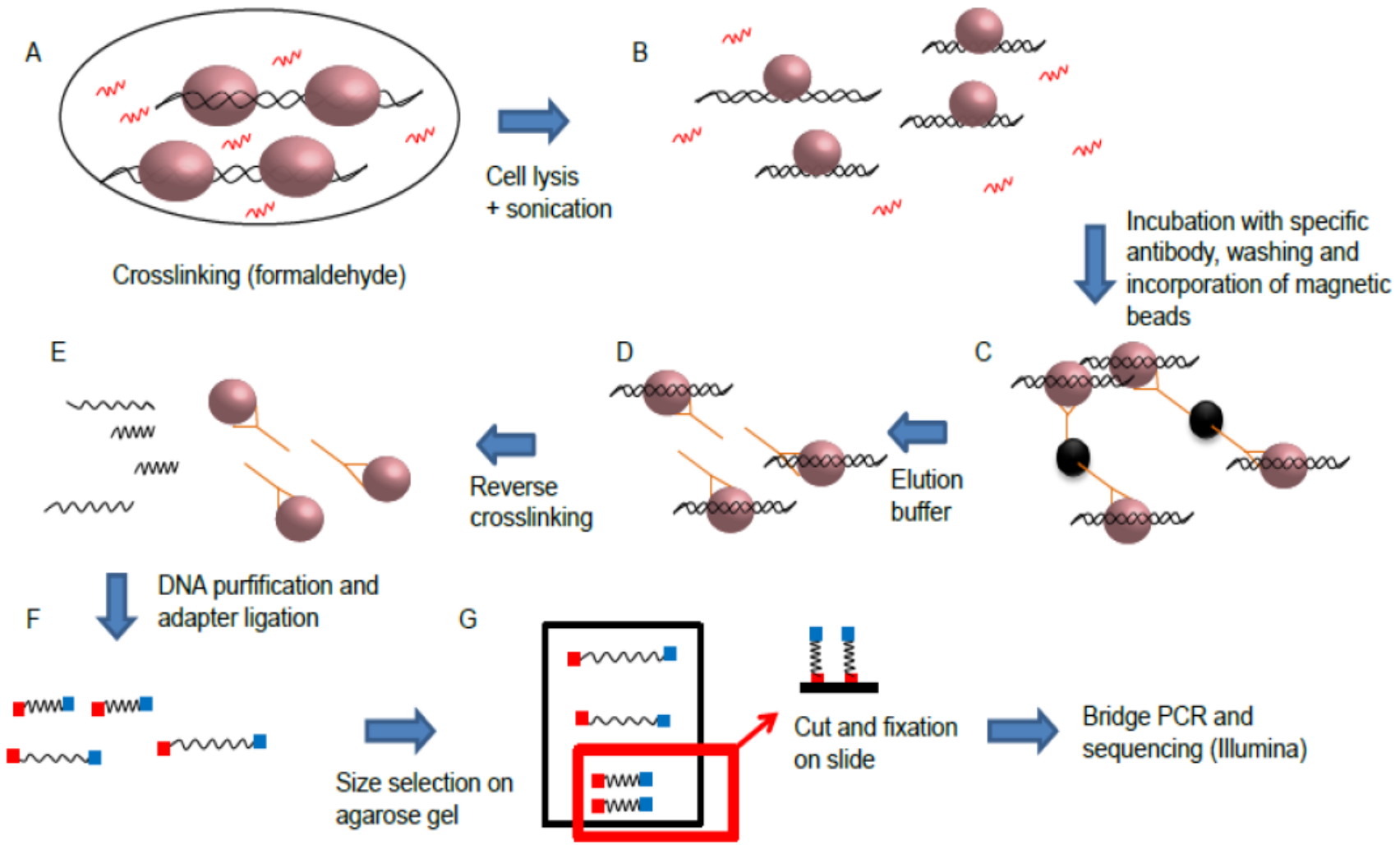
Why using spike-in controls?

ChIP-Seq data processing: From Fastq to BigWig files

The spike-in control scaling procedure

Visual exploration of the data

New York City - R/Bioconductor for Genomics:
Simplified General Principles of the ChIP-Seq Protocol



ChIP-Seq protocol

Why using spike-in controls?

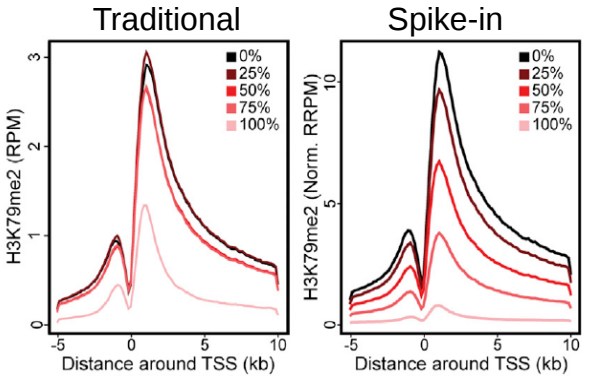
ChIP-Seq data processing: From Fastq to BigWig files

The spike-in control scaling procedure

Visual exploration of the data

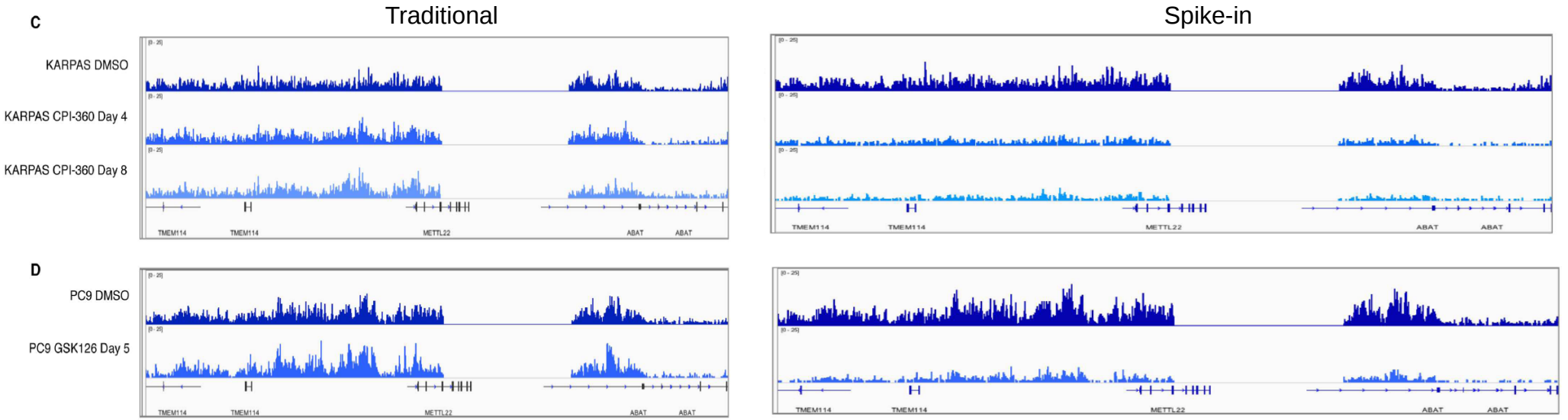
Why using spike-in controls?

Traditional data scaling methods do not catch global histone modification alterations



EPZ5676, DOT1L inhibitor

Orlando *et al.*, Cell Reports, 2014

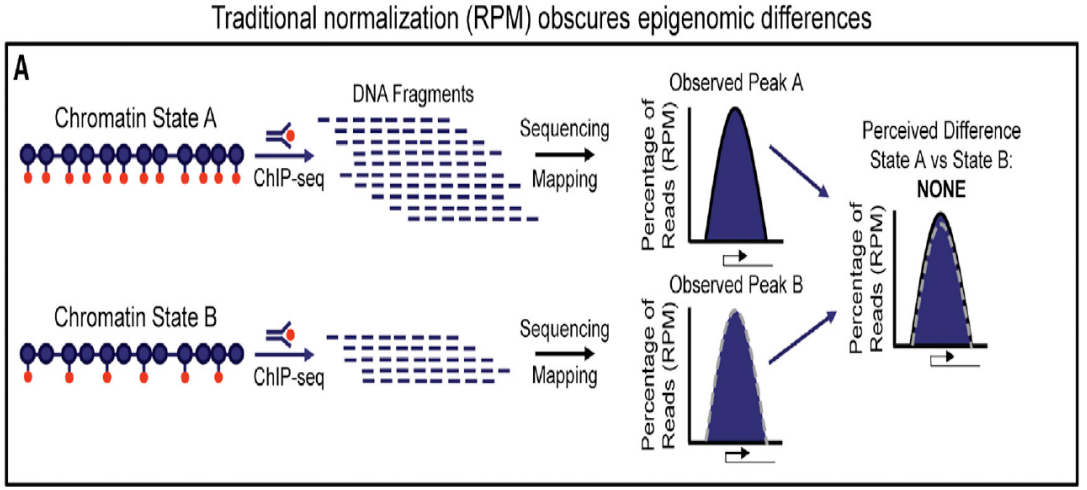


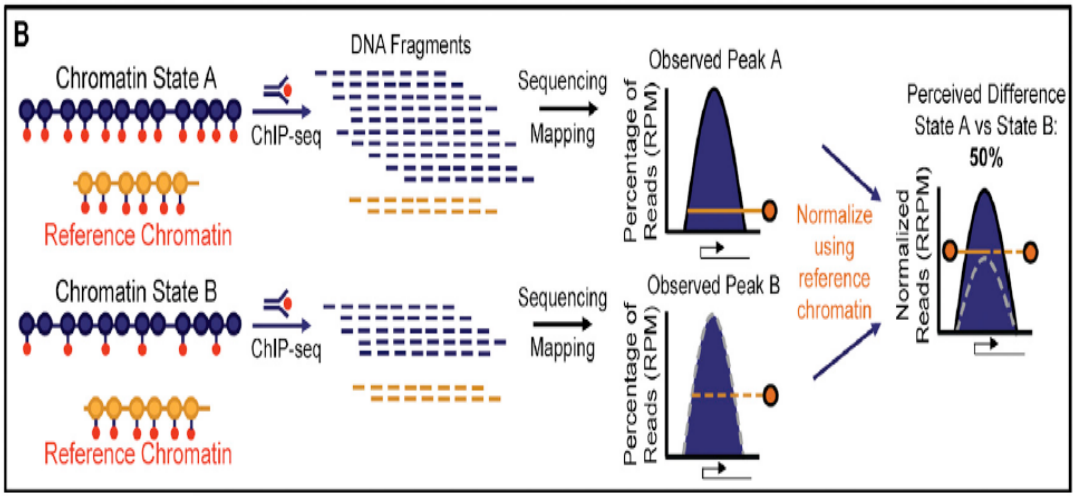
H3K27me3, EZH2 inhibitors

Egan *et al.*, Plos One, 2016

Why using spike-in controls?

Traditional data scaling methods do not catch global histone modification alterations





Orlando *et al.*, Cell Reports, 2014

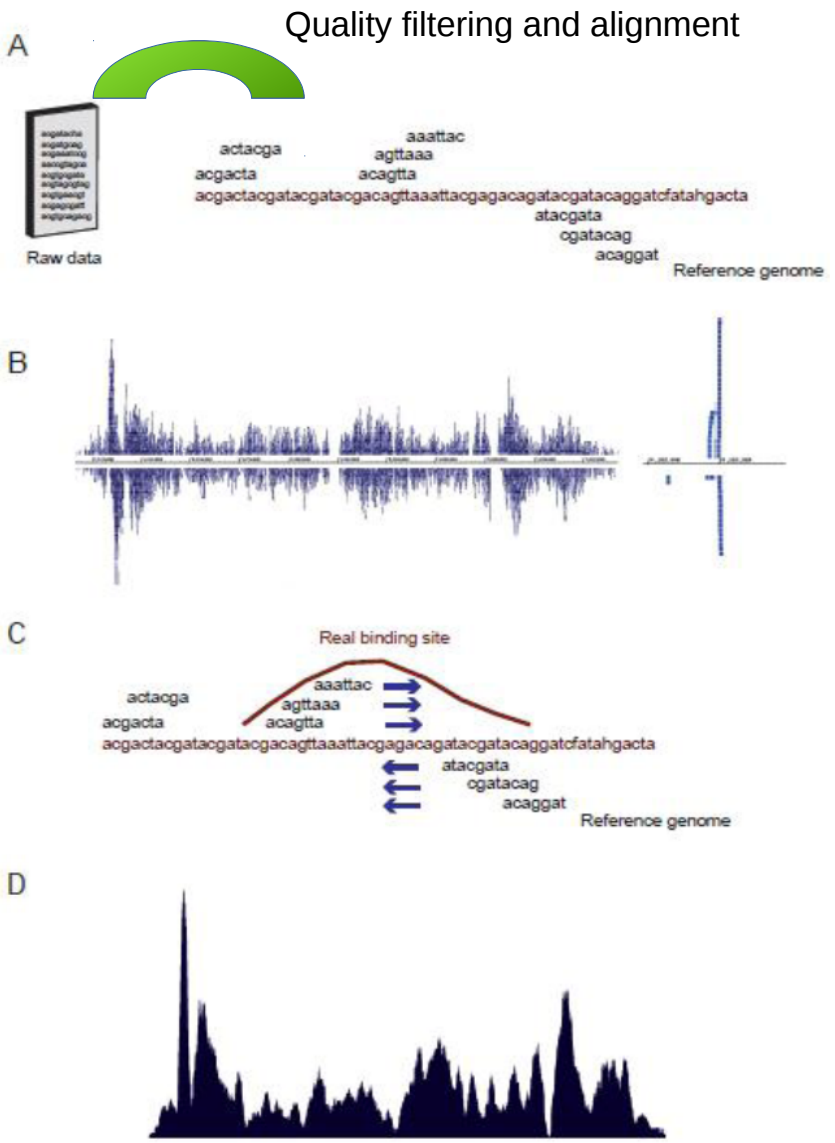
ChIP-Seq protocol

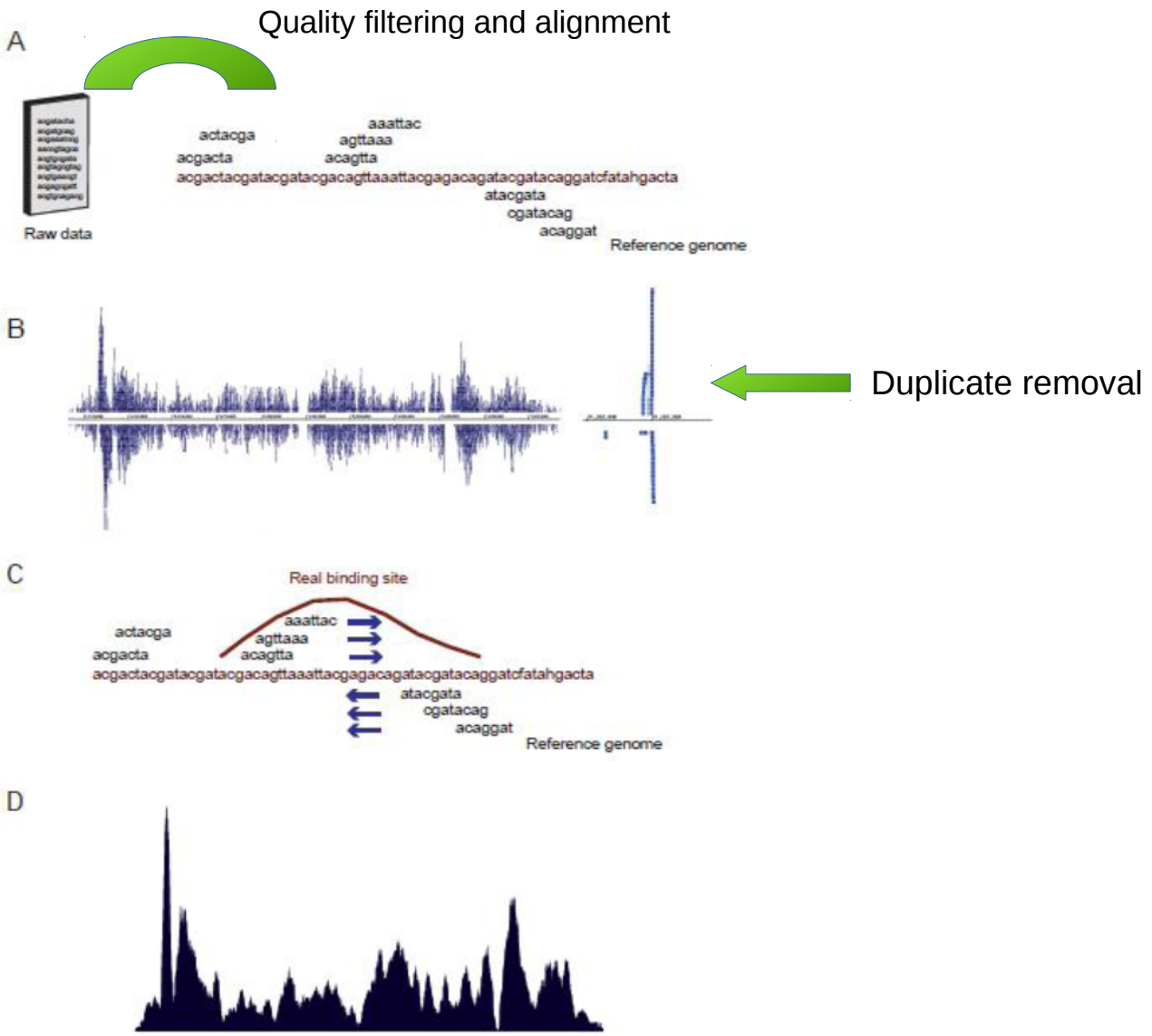
Why using spike-in controls?

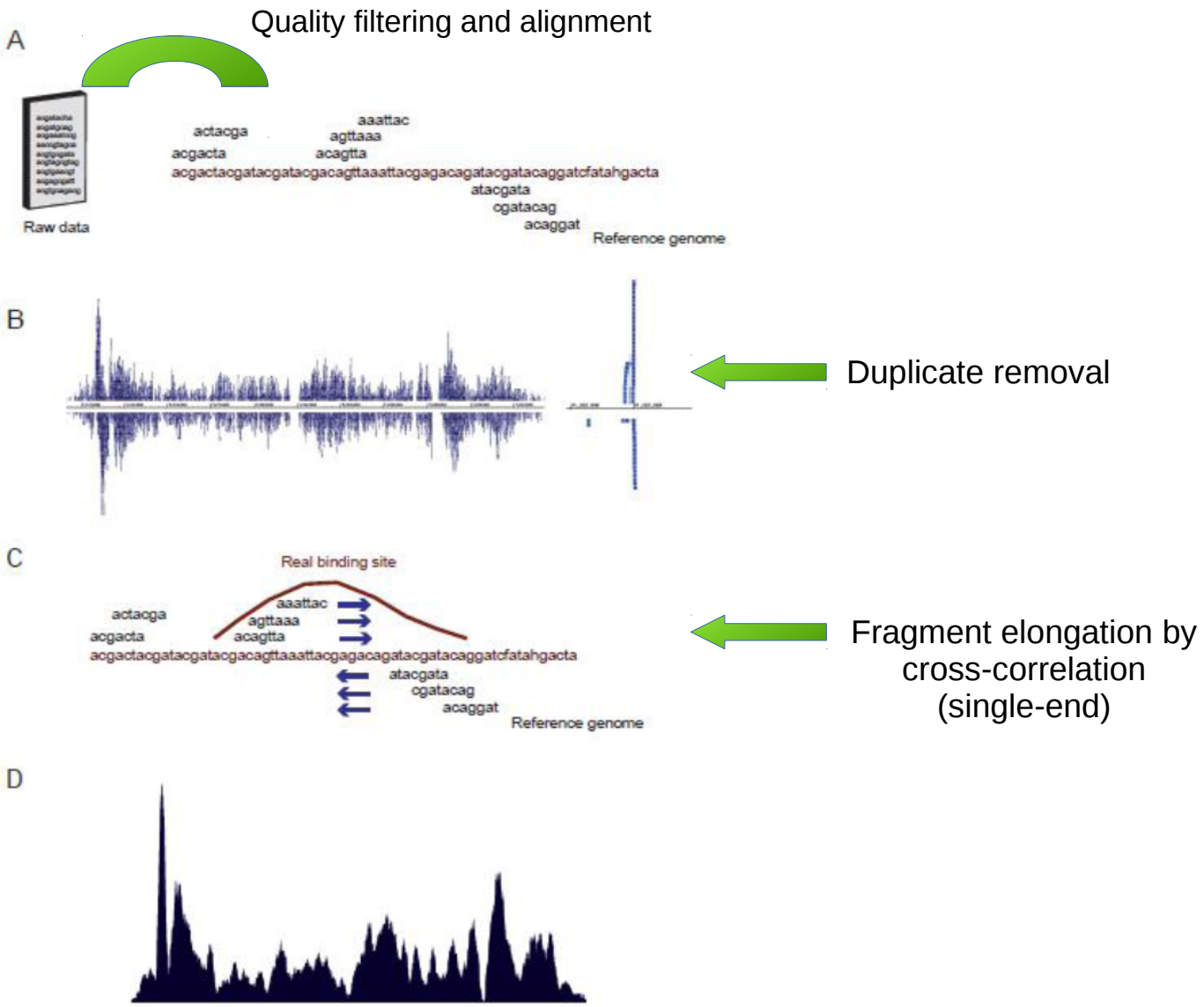
ChIP-Seq data processing: From Fastq to BigWig files

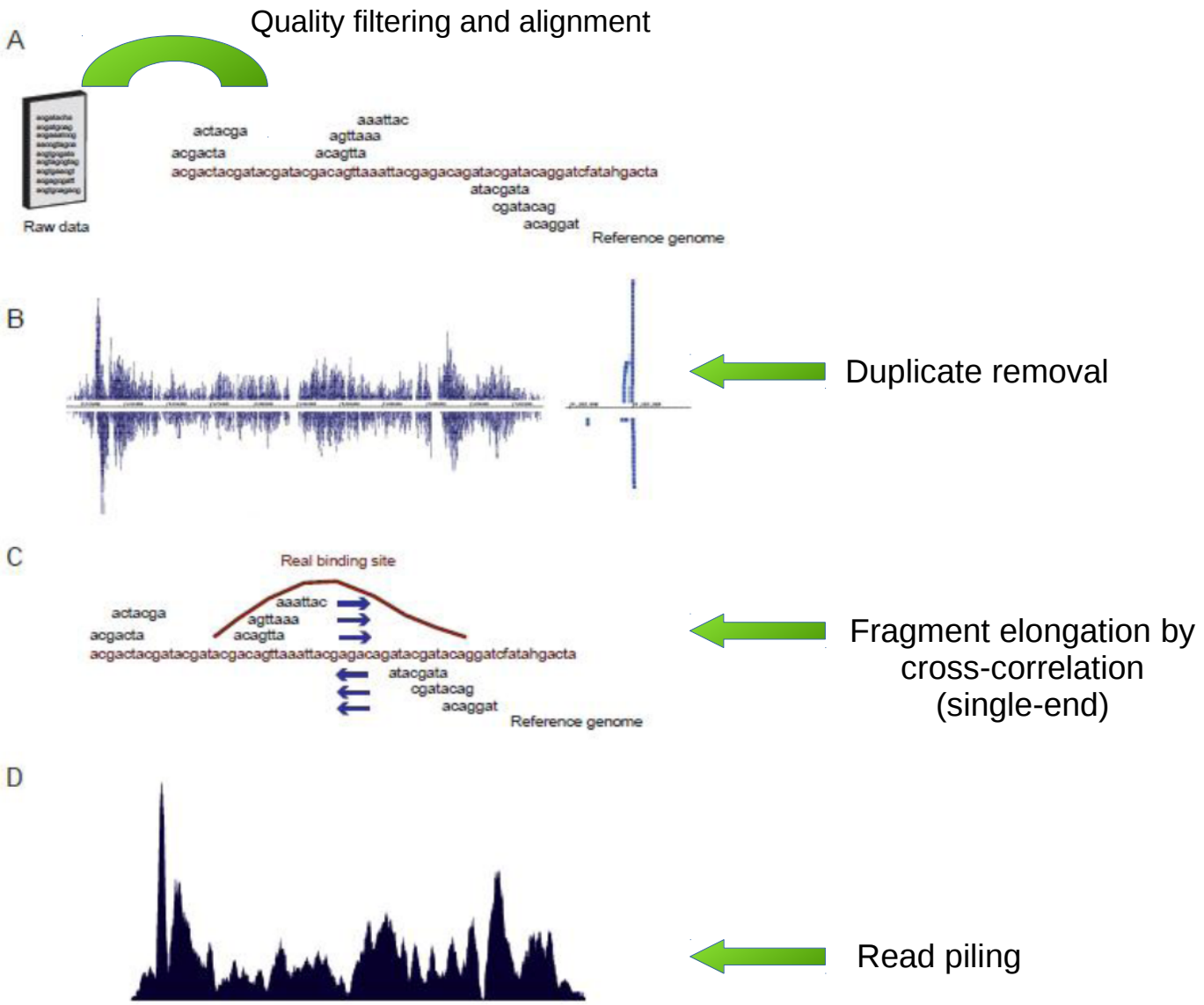
The spike-in control scaling procedure

Visual exploration of the data

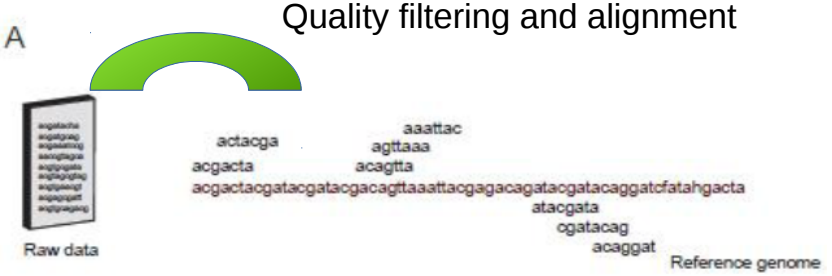

Quality filtering and alignment
A
Raw data
actacga
acgacta
aaattac
agttaaa
acagtta
acgactacgatacgatacgacagttaaattacgagacagatacgatacaggatcfatahgacta
atacgata
ogatacag
acaggat
Reference genome
B
C
Real binding site
actacga
acgacta
aaattac
agttaaa
acagtta
acgactacgatacgatacgacagttaaattacgagacagatacgatacaggatcfatahgacta
atacgata
ogatacag
acaggat
Reference genome
D

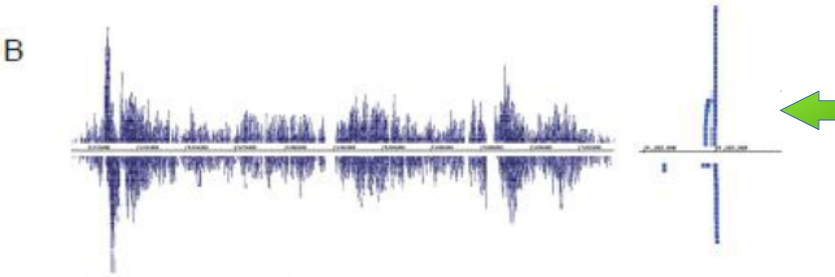

Quality filtering and alignment
A
Raw data
actacga
acgacta
aaattac
agttaaa
acagtta
acgactacgatacgatacgacagttaaattacgagacagatacgatacaggatcfatahgacta
atacgata
ogatacag
acaggat
Reference genome
B
Duplicate removal
C
Real binding site
actacga
acgacta
aaattac
agttaaa
acagtta
acgactacgatacgatacgacagttaaattacgagacagatacgatacaggatcfatahgacta
atacgata
ogatacag
acaggat
Reference genome
D


Quality filtering and alignment
A
Raw data
actacga
acgacta
aaattac
agttaaa
acagtta
acgactacgatacgatacgacagttaaattacgagacagatacgatacaggatcfatahgacta
atacgata
ogatacag
acaggat
Reference genome
B
Duplicate removal
C
Real binding site
actacga
acgacta
aaattac
agttaaa
acagtta
acgactacgatacgatacgacagttaaattacgagacagatacgatacaggatcfatahgacta
atacgata
ogatacag
acaggat
Reference genome
Fragment elongation by cross-correlation (single-end)
D

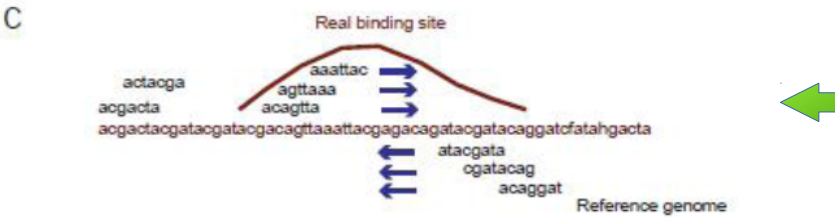

Quality filtering and alignment
A
Raw data
actacga
acgacta
aaattac
agttaaa
acagtta
acgactacgatacgatacgacagttaaattacgagacagatacgatacaggatcfatahgacta
atacgata
ogatacag
acaggat
Reference genome
B
Duplicate removal
C
Real binding site
actacga
acgacta
aaattac
agttaaa
acagtta
acgactacgatacgatacgacagttaaattacgagacagatacgatacaggatcfatahgacta
atacgata
ogatacag
acaggat
Reference genome
Fragment elongation by cross-correlation (single-end)
D
Read piling

Quality filtering and alignment

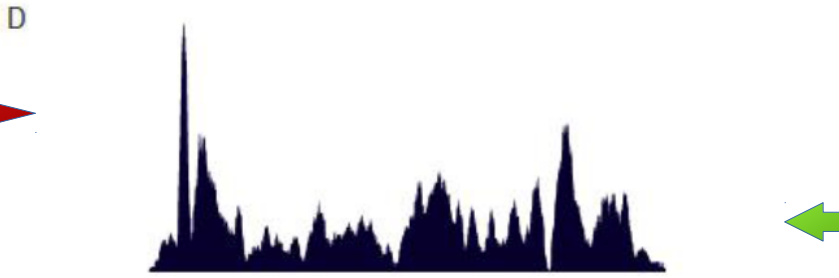




Duplicate removal



Fragment elongation by cross-correlation (single-end)



Read piling

Where we start from

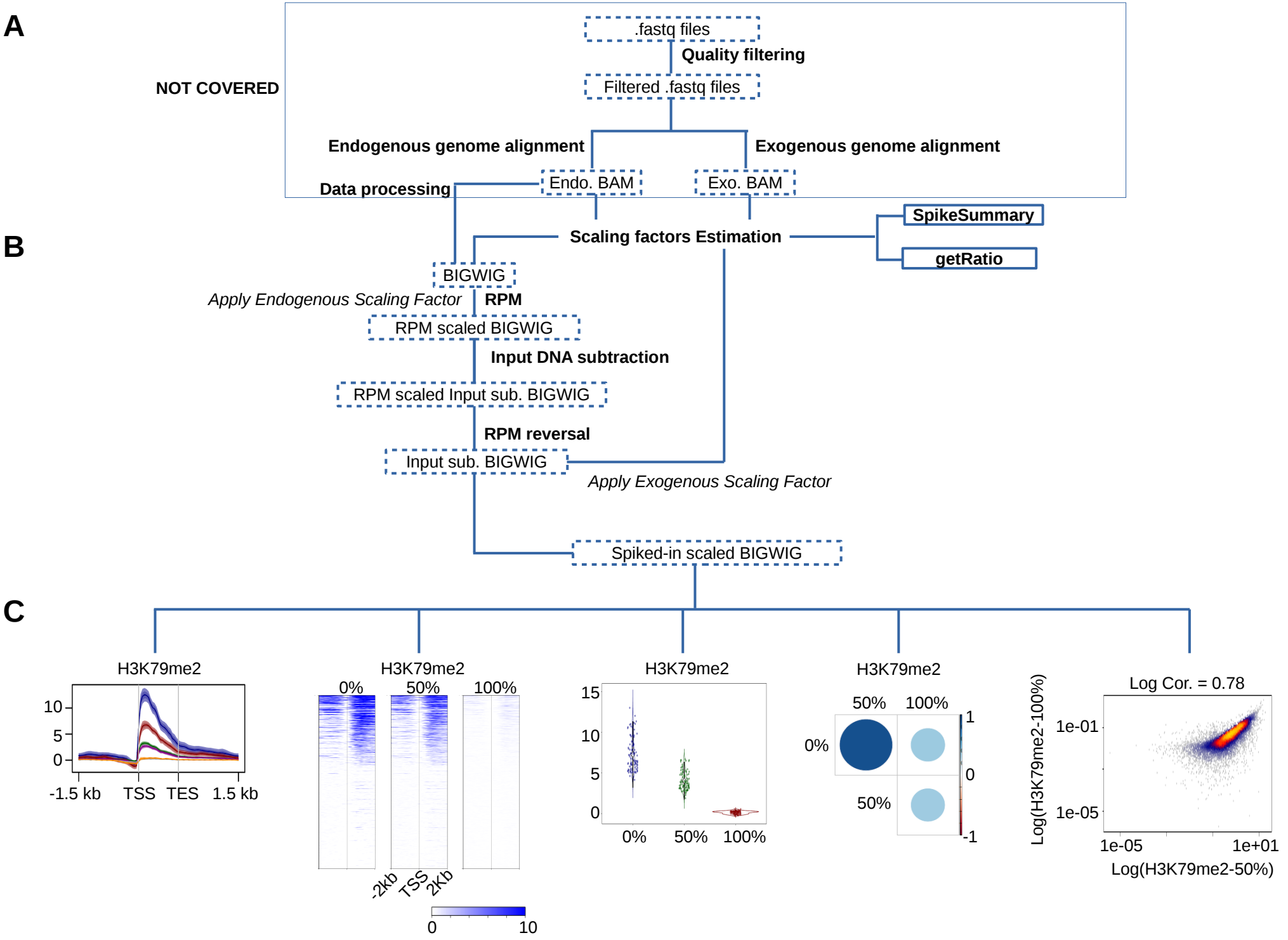
ChIP-Seq protocol

Why using spike-in controls?

ChIP-Seq data processing: From Fastq to BigWig files

Tutorial overview:

The spike-in control scaling procedure
&
Visual exploration of the data


A
NOT COVERED
.fastq files
Quality filtering
Filtered .fastq files
Endogenous genome alignment
Exogenous genome alignment
Data processing
Endo. BAM
Exo. BAM
B
Scaling factors Estimation
SpikeSummary
getRatio
BIGWIG
Apply Endogenous Scaling Factor
RPM
RPM scaled BIGWIG
Input DNA subtraction
RPM scaled Input sub. BIGWIG
RPM reversal
Input sub. BIGWIG
Apply Exogenous Scaling Factor
Spiked-in scaled BIGWIG
C
H3K79me2
10
5
0
-1.5 kb
TSS
TES
1.5 kb
H3K79me2
0%
50%
100%
-2kb
TSS
2Kb
0
10
H3K79me2
15
10
5
0
0%
50%
100%
H3K79me2
50%
100%
0%
50%
1
0
-1
Log Cor. = 0.78
1e-01
1e-05
1e-05
1e+01
Log(H3K79me2-100%)
Log(H3K79me2-50%)