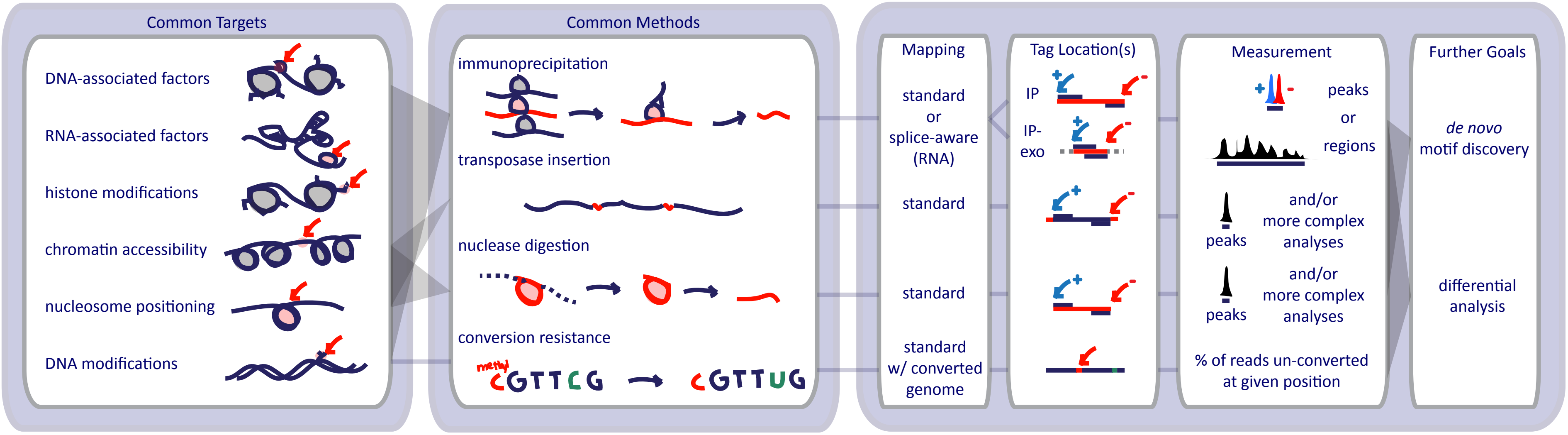
Epigenomics overview

BMI206 November 3, 2014



Common Targets

- DNA-associated factors
- RNA-associated factors
- histone modifications
- chromatin accessibility
- nucleosome positioning
- DNA modifications

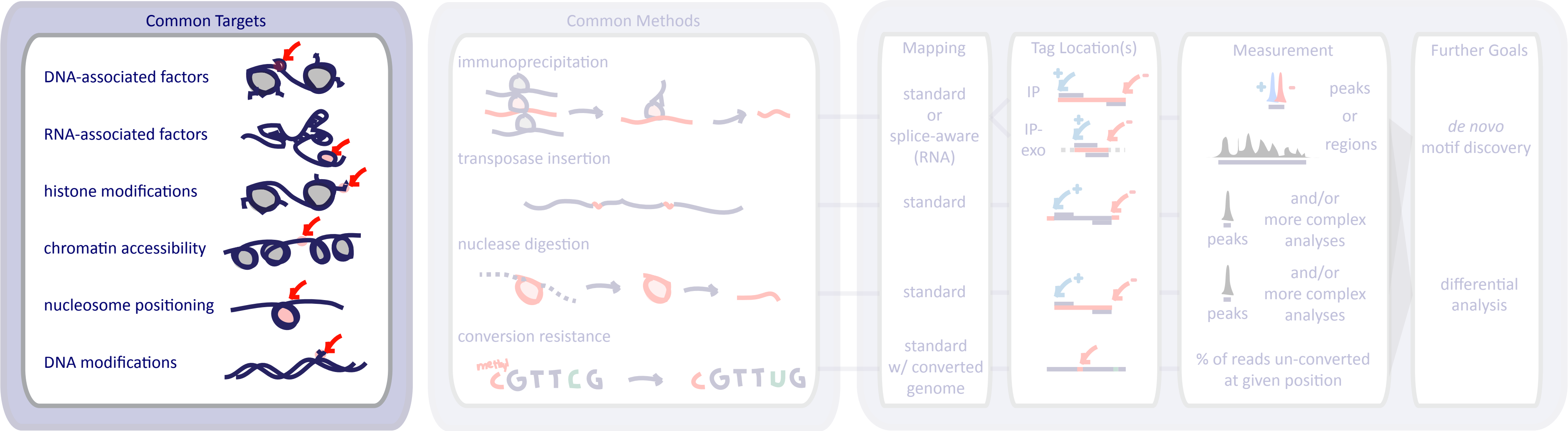
Common Methods

- immunoprecipitation
- transposase insertion
- nuclease digestion
- conversion resistance

Mapping	Tag Location(s)	Measurement	Further Goals
standard or splice-aware (RNA)	IP IP-exo	peaks or regions	*de novo* motif discovery
standard		peaks and/or more complex analyses	
standard		peaks and/or more complex analyses	differential analysis
standard w/ converted genome		% of reads un-converted at given position	

Epigenomics overview

BMI206 November 3, 2014



Common Targets

- DNA-associated factors
- RNA-associated factors
- histone modifications
- chromatin accessibility
- nucleosome positioning
- DNA modifications

Common Methods

- immunoprecipitation
- transposase insertion
- nuclease digestion
- conversion resistance

Mapping

- standard or splice-aware (RNA)
- standard
- standard
- standard w/ converted genome

Tag Location(s)

Measurement

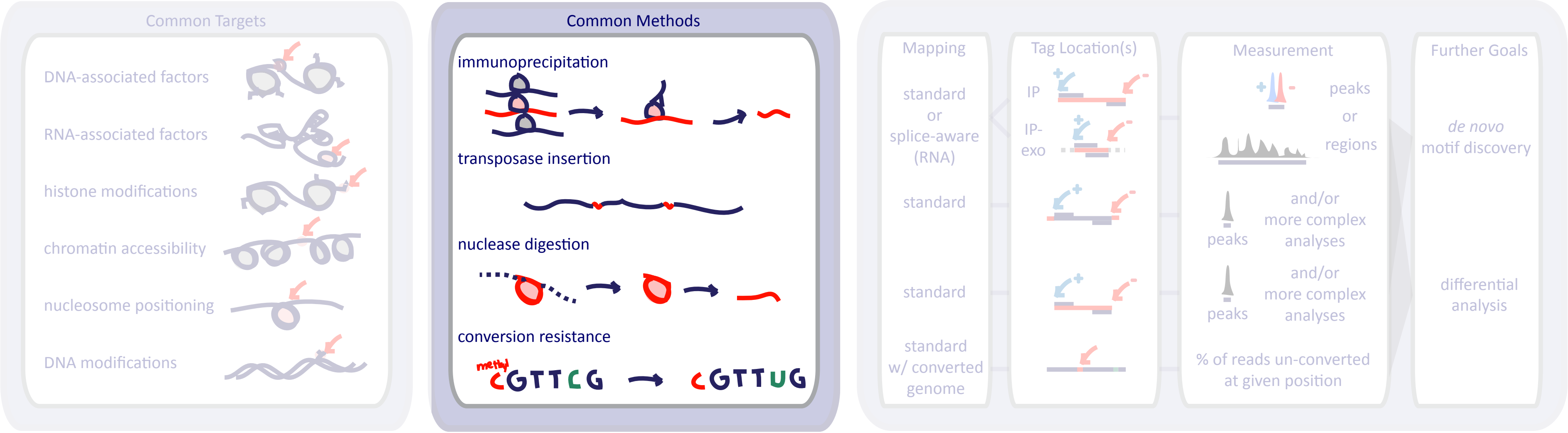
- peaks or regions
- peaks and/or more complex analyses
- peaks and/or more complex analyses
- % of reads un-converted at given position

Further Goals

- *de novo* motif discovery
- differential analysis

Epigenomics overview

BMI206 November 3, 2014



Common Targets

DNA-associated factors

RNA-associated factors

histone modifications

chromatin accessibility

nucleosome positioning

DNA modifications

Common Methods

immunoprecipitation

transposase insertion

nuclease digestion

conversion resistance

Mapping

standard or splice-aware (RNA)

standard

standard

standard w/ converted genome

Tag Location(s)

IP

IP-exo

Measurement

peaks or regions

peaks

and/or more complex analyses

peaks

and/or more complex analyses

% of reads un-converted at given position

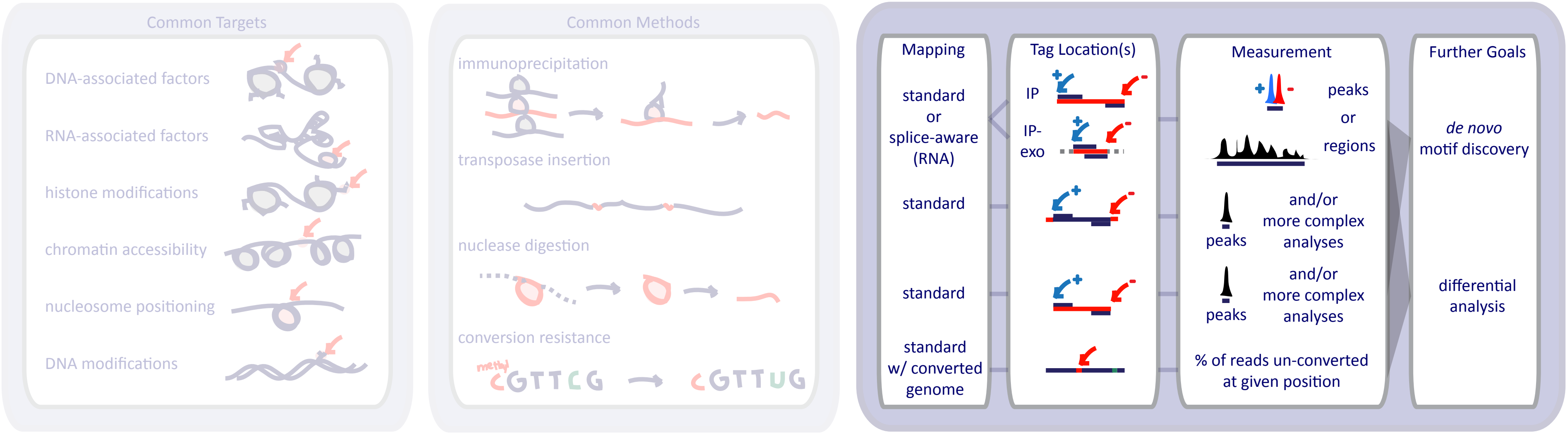
Further Goals

de novo motif discovery

differential analysis

Epigenomics overview

BMI206 November 3, 2014



Common Targets

DNA-associated factors

RNA-associated factors

histone modifications

chromatin accessibility

nucleosome positioning

DNA modifications

Common Methods

immunoprecipitation

transposase insertion

nuclease digestion

conversion resistance

Mapping

standard or splice-aware (RNA)

standard

standard

standard w/ converted genome

Tag Location(s)

IP

IP-exo

Measurement

peaks or regions

peaks

and/or more complex analyses

peaks

and/or more complex analyses

% of reads un-converted at given position

Further Goals

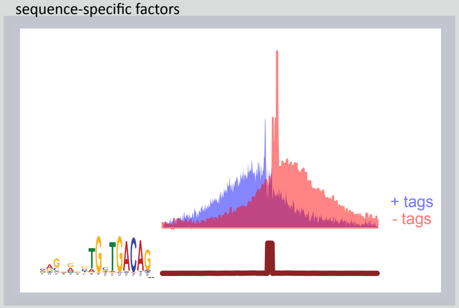
de novo motif discovery

differential analysis

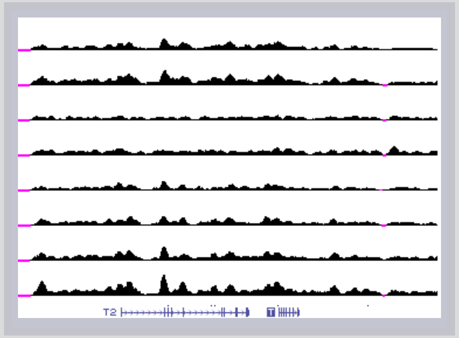
BMI206 November 3, 2014

DNA-associated factors

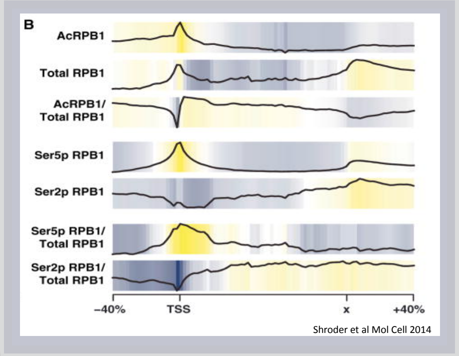
sequence-specific factors



chromatin associated



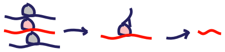
transcription machinery



Shroder et al Mol Cell 2014

methodology

immunoprecipitation

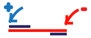


library construction

sequencing

mapping to genome reference (BWA or bowtie2)

identify tag locations



calculate median insert size (single end)

determine fragment midpoint (paired end)

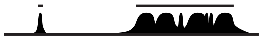
estimate actual binding location per tag

calculate genomic density of shifted tags

measure background density

calculate background-normalized binding density

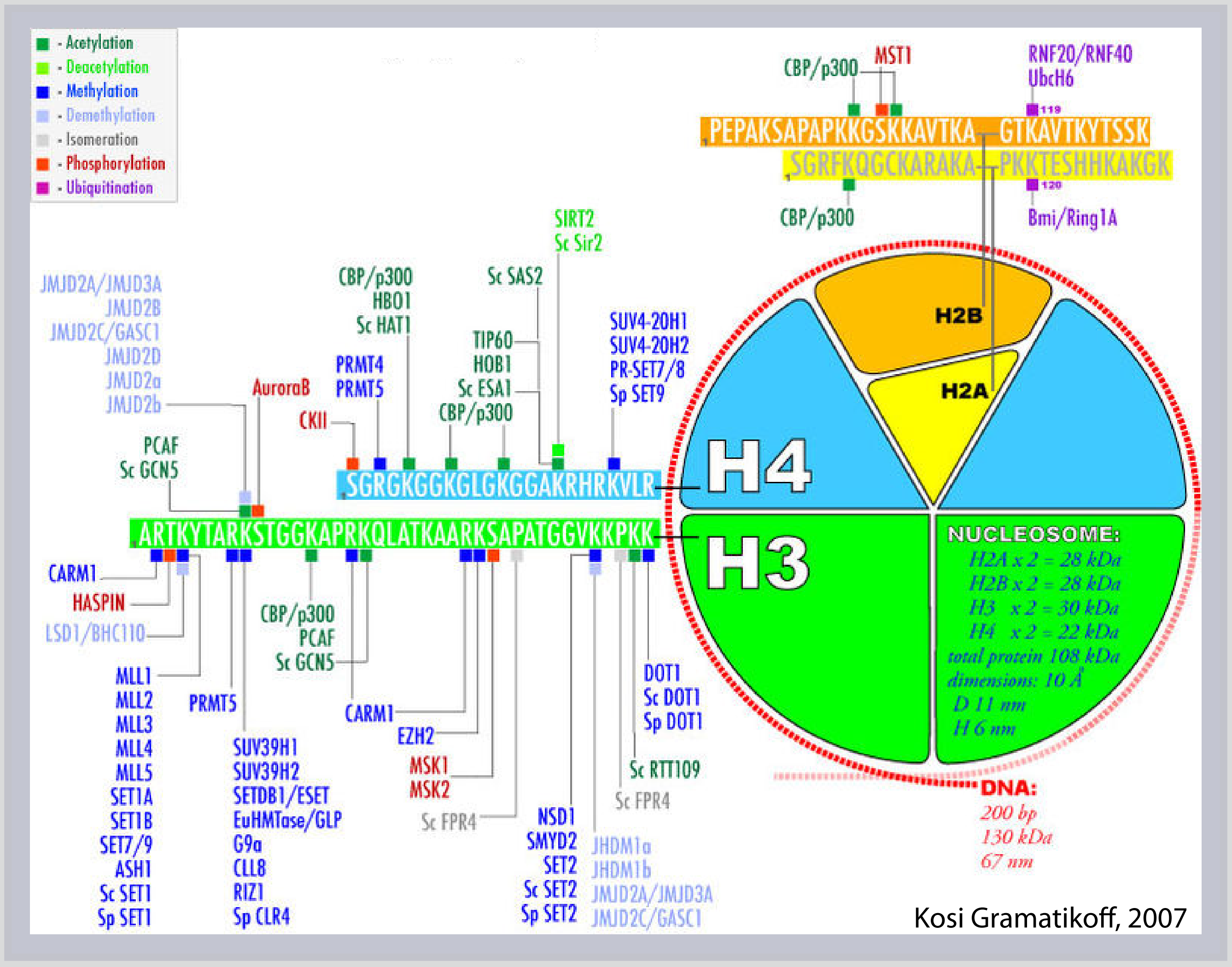
call peaks (punctate marks) or regions (broad marks)



histone modifications

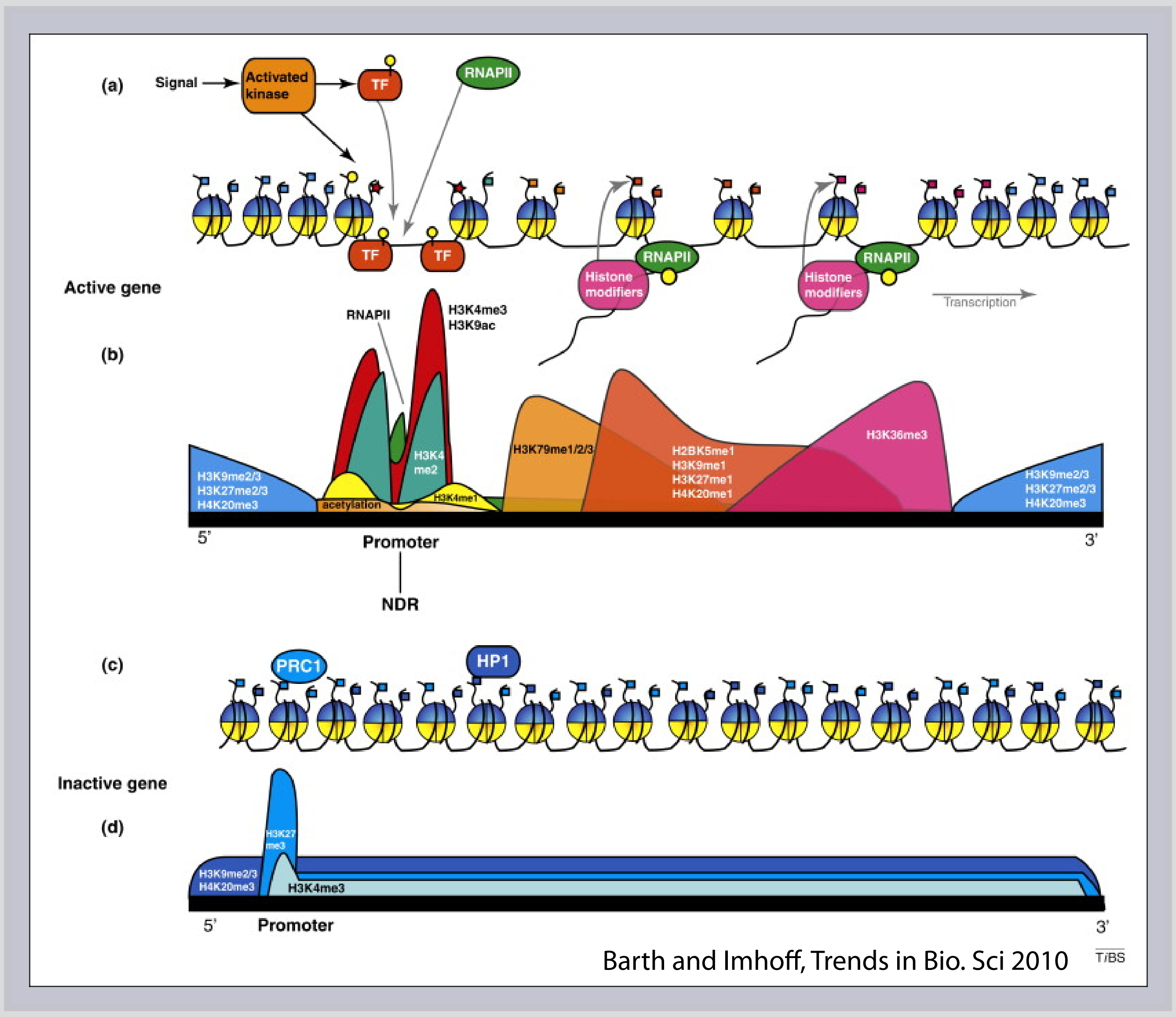
BMI206 November 3, 2014

catalogue



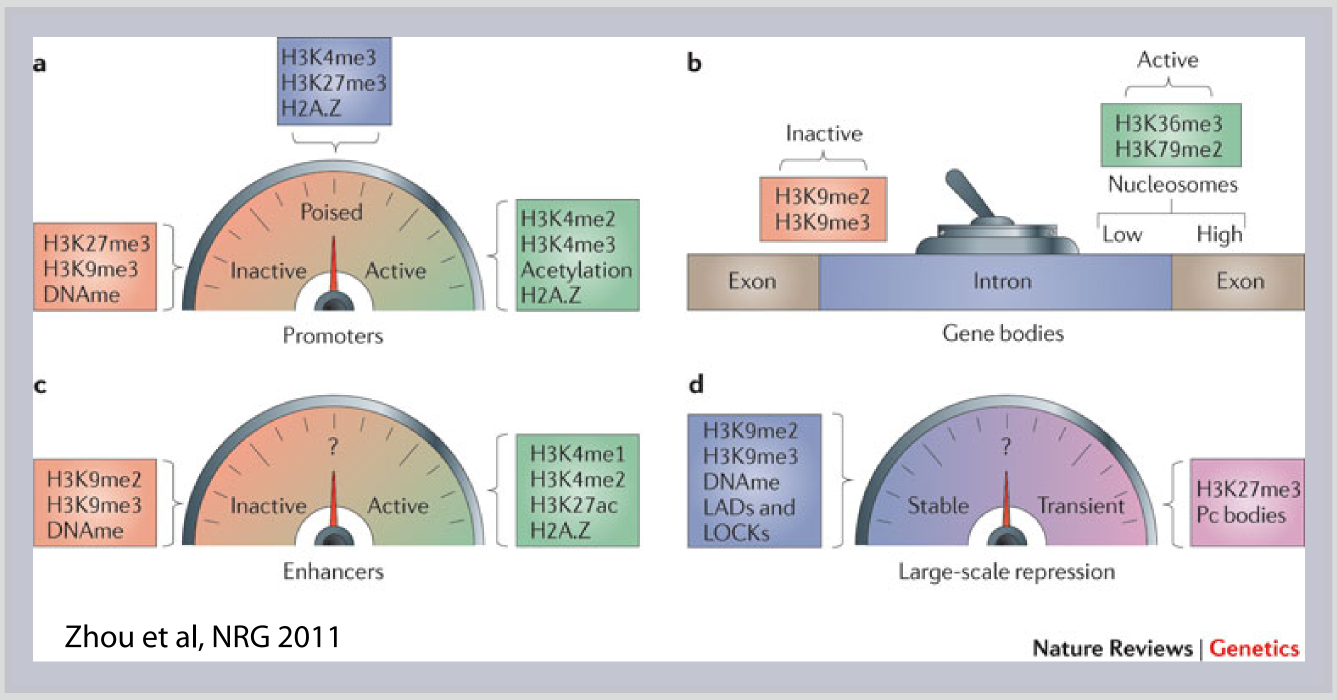
Kosi Gramatikoff, 2007

genomic distribution



Barth and Imhoff, Trends in Bio. Sci 2010

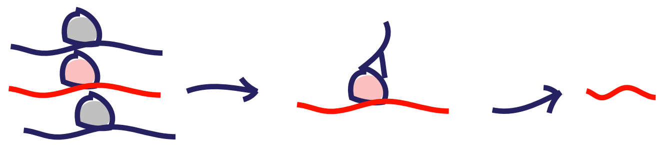
functional association



Zhou et al, NRG 2011

methodology

immunoprecipitation

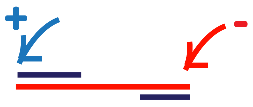


library construction

sequencing

mapping to genome reference (BWA or bowtie2)

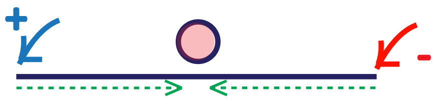
identify tag locations



calculate median insert size (single end)

determine fragment midpoint (paired end)

estimate actual binding location per tag

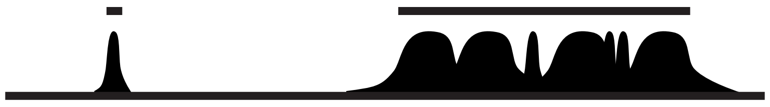


calculate genomic density of shifted tags

measure background density

calculate background-normalized binding density

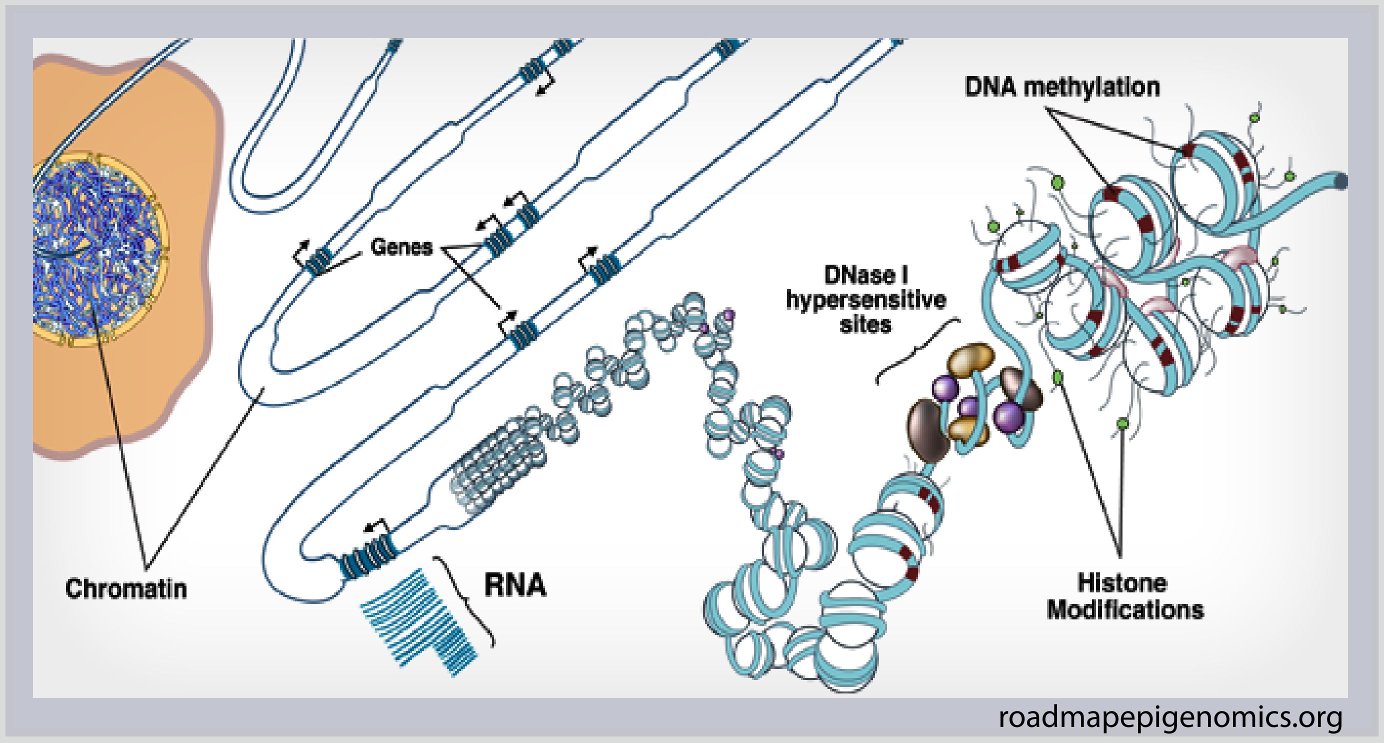
call peaks (punctate marks) or regions (broad marks)



chromatin accessibility

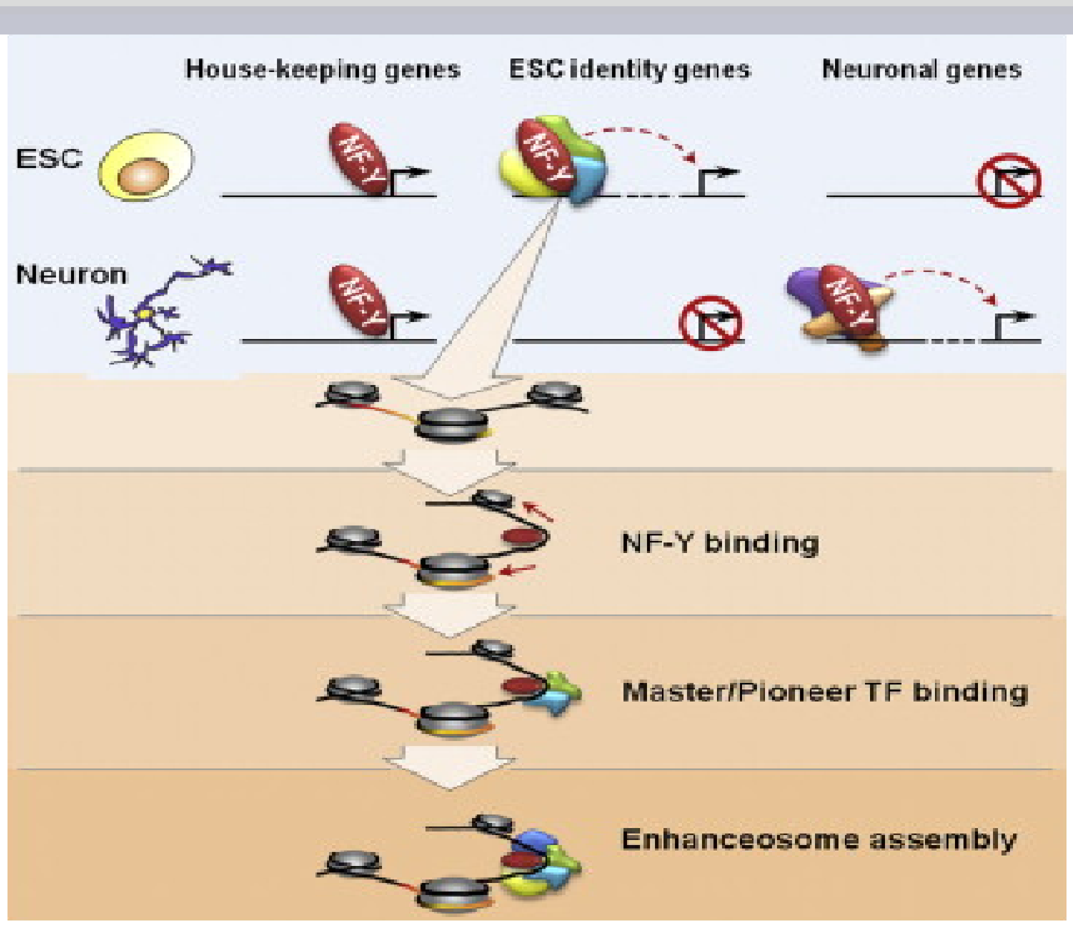
BMI206 November 3, 2014

functional regulatory elements are accessible



roadmapepigenomics.org

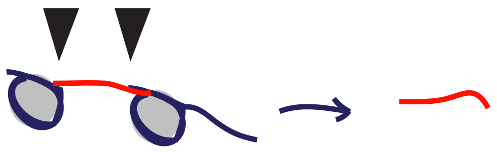
formation of accessible chromatin



Oldfield et al, Mol Cell 2014

methodology (digestion or transposase)

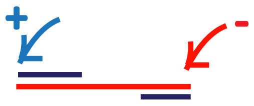
DNase I or Tn5 release fragments of open chromatin



library construction and sequencing

mapping to genome reference (BWA or bowtie2)

identify cleavage/insertion location

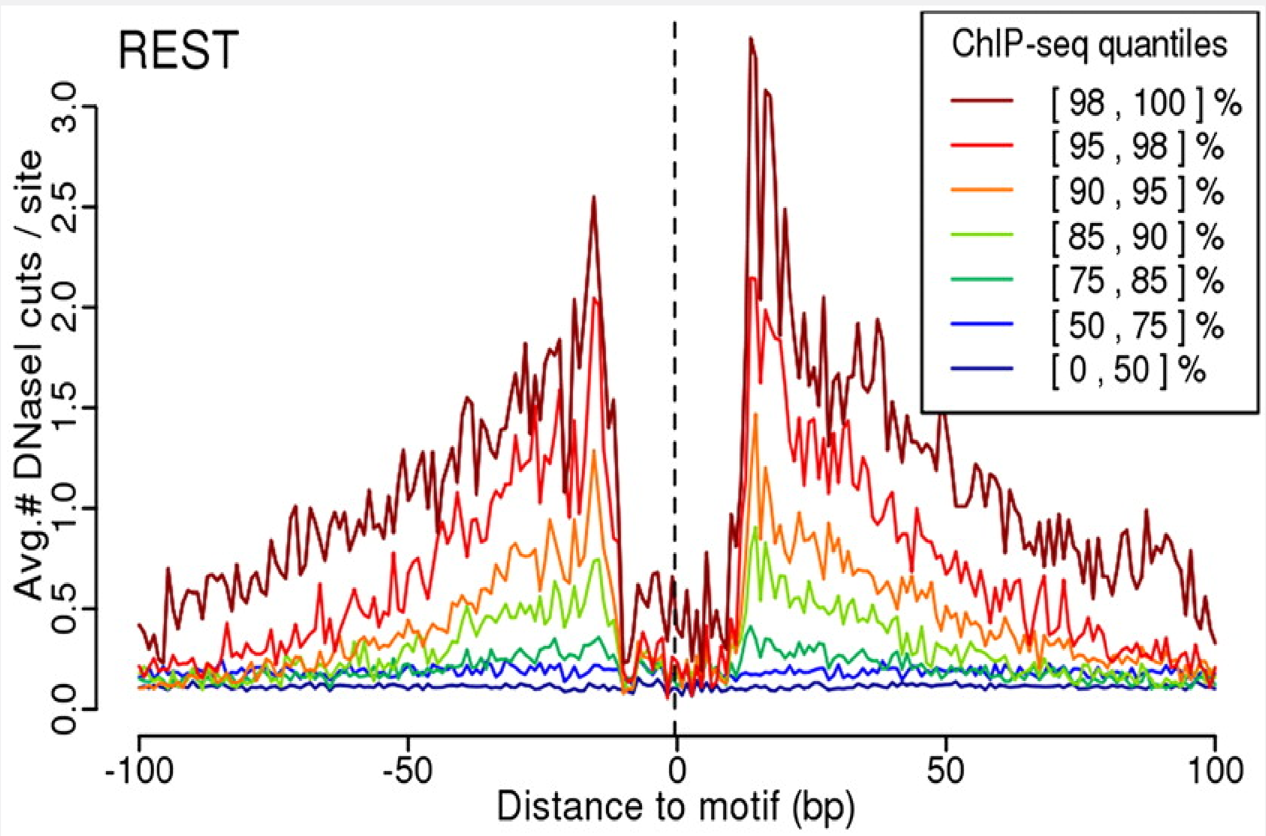


calculate density of cleavage/insertion events

call regions of accessibility

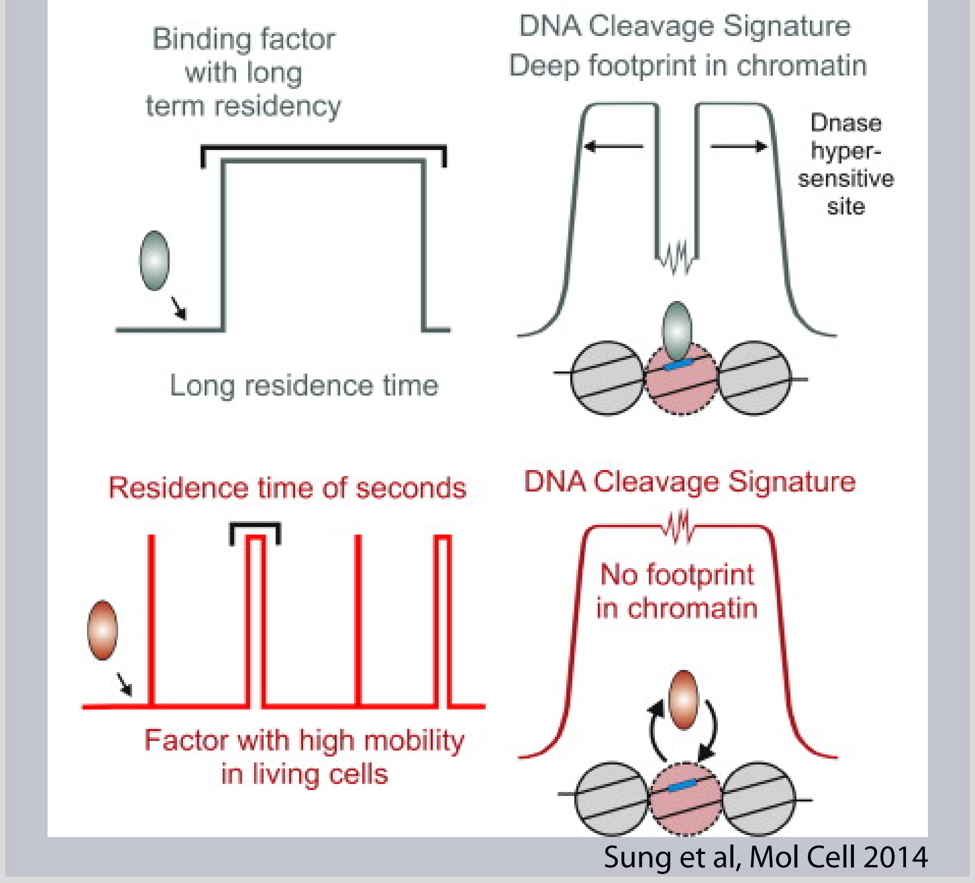
call footprints within accessible chromatin

footprinting



Pique-Regi et al Genome Research 2011

Footprints reflect residence time in chromatin

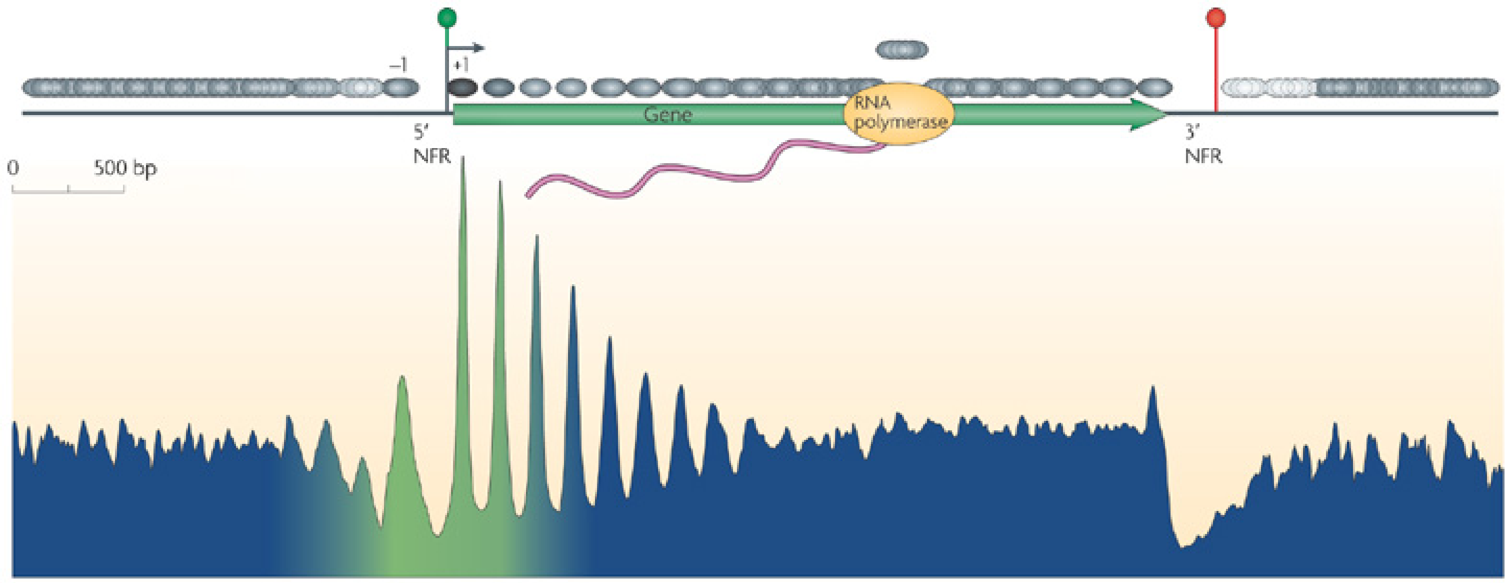

Binding factor with long term residency
DNA Cleavage Signature Deep footprint in chromatin
Dnase hyper-sensitive site
Long residence time
Residence time of seconds
DNA Cleavage Signature
No footprint in chromatin
Factor with high mobility in living cells

Sung et al, Mol Cell 2014

nucleosome position

BMI206 November 3, 2014

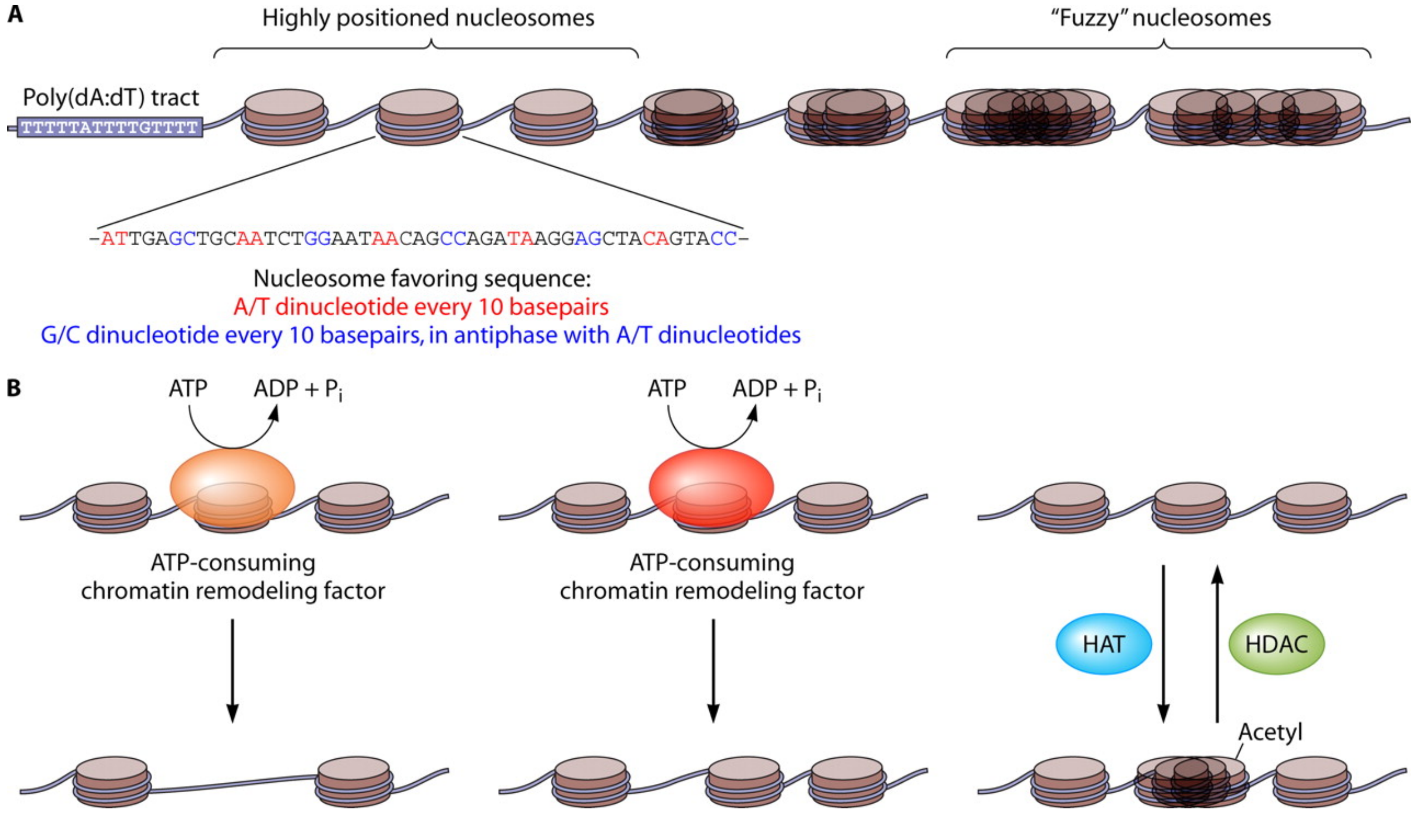
nucleosome positioning during transcription



Nature Reviews | Genetics

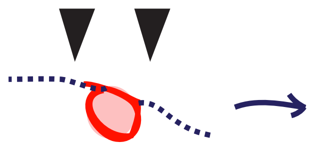
Jiang et al NRG 2009

determinants of positioning stability

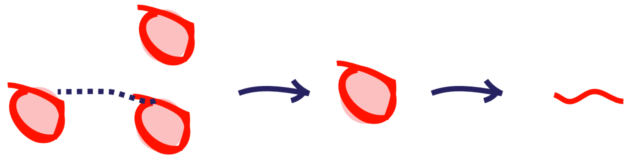

A
Highly positioned nucleosomes
"Fuzzy" nucleosomes
Poly(dA:dT) tract
TTTTTATTTTGTTTT
–ATTGAGCTGCAATCTGGAATAACAGCCAGATAAGGAGCTACAGTACC–
Nucleosome favoring sequence:
A/T dinucleotide every 10 basepairs
G/C dinucleotide every 10 basepairs, in antiphase with A/T dinucleotides
B
ATP
ADP + P_i
ATP-consuming chromatin remodeling factor
ATP
ADP + P_i
ATP-consuming chromatin remodeling factor
HAT
HDAC
Acetyl

methodology (MNase digestion)

MNase I digests linker DNA, releasing multisomes



purify mononucleosome-bound DNA



library construction and sequencing

mapping to genome reference (BWA or bowtie2)

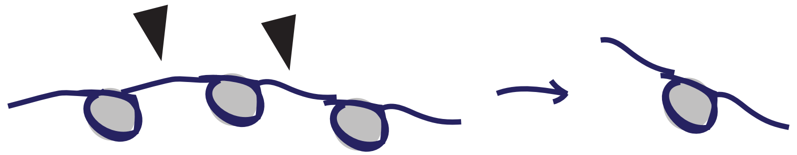
identify cleavage location

calculate nucleosomal density

call positioned nucleosomes

methodology (transposon insertion)

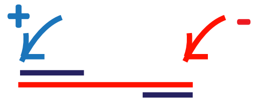
Tn5 integrates into accessible chromatin, and releases multisomes



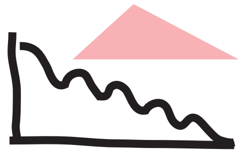
library construction and sequencing

mapping to genome reference (BWA or bowtie2)

identify cleavage location



identify reads that span at least one nucleosome



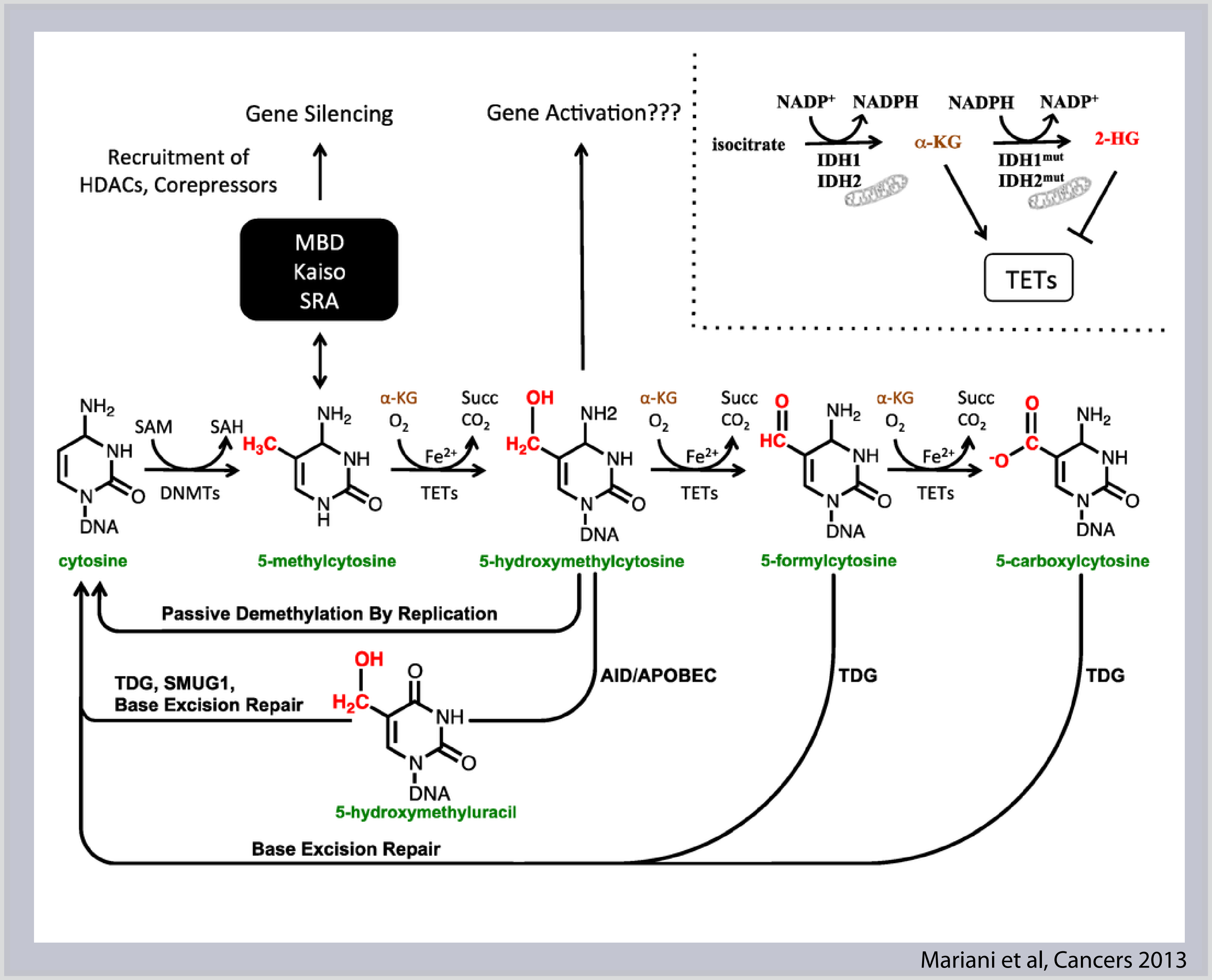
read length

calculate density of tags / number of insertion events

call positioned nucleosomes

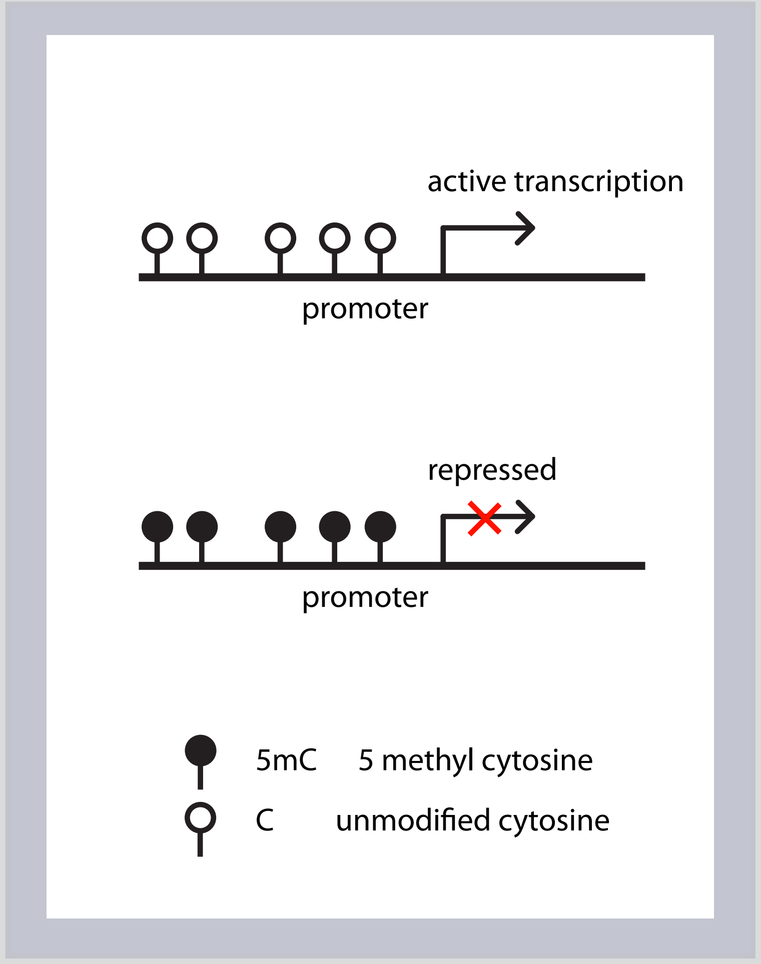
DNA modifications

catalogue of common base modifications (in mammals)

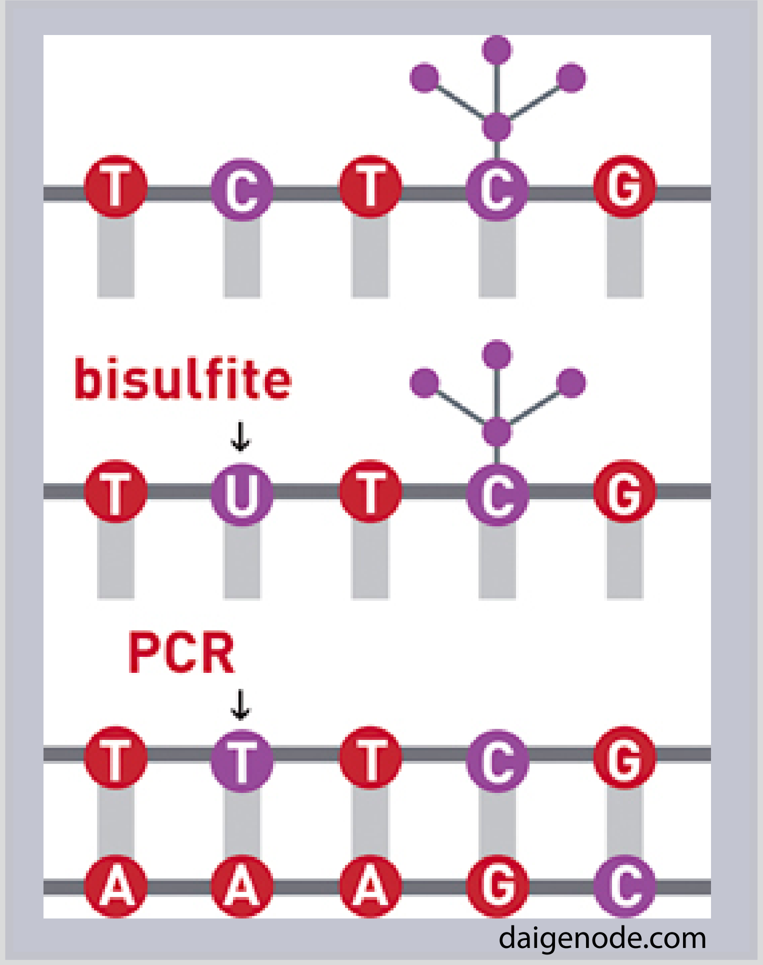


Mariani et al, Cancers 2013

5mC repression of gene expression



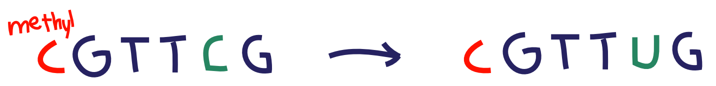
bisulfite conversion / protection



daigenode.com

methodology

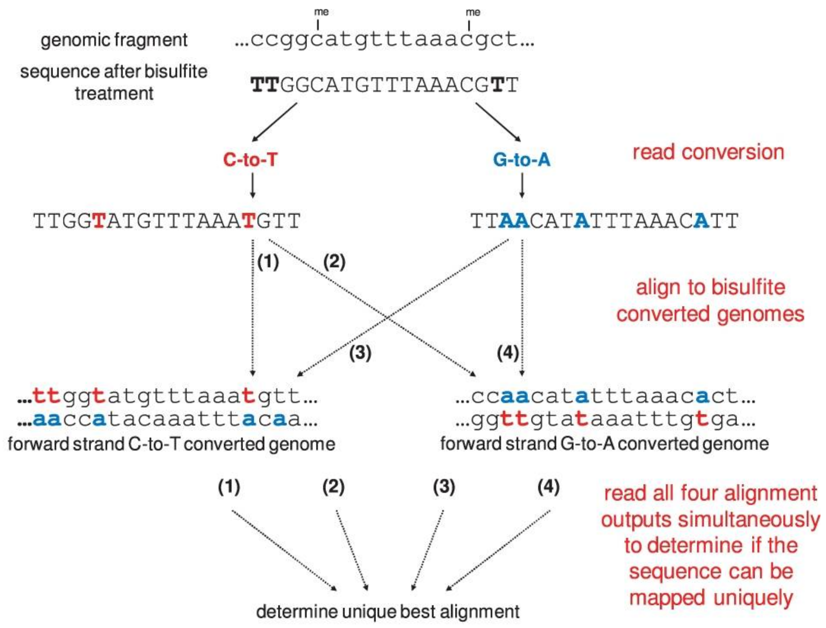
convert DNA with bisulfite



fragmentation

library construction and sequencing

mapping


genomic fragment
me
me
...ccggcatgtttaaacgct...
sequence after bisulfite treatment
TTGGCATGTTTAAACGTT
C-to-T
G-to-A
read conversion
TTGGTATGTTTAAATGTT
TTAACATATTTAAACATT
(1)
(2)
(3)
(4)
align to bisulfite converted genomes
...ttggtatgtttaaatgtt...
...aaccatacaaatttacaa...
forward strand C-to-T converted genome
...ccaacatatttaaacact...
...ggttgtataaatttgtga...
forward strand G-to-A converted genome
(1)
(2)
(3)
(4)
read all four alignment outputs simultaneously to determine if the sequence can be mapped uniquely
determine unique best alignment

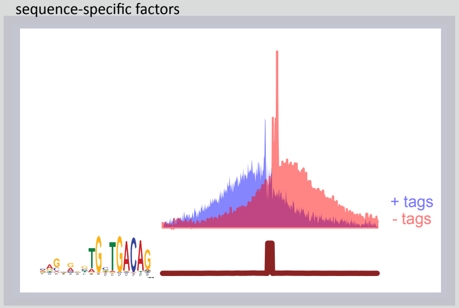
Krueger et al Bioinformatics 2011

calculate % of tags methylated per genomic position

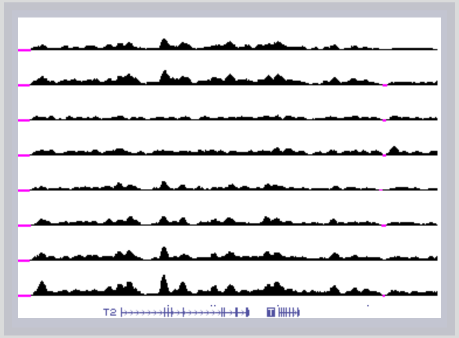
DNA-associated factors

BMI206 November 3, 2014

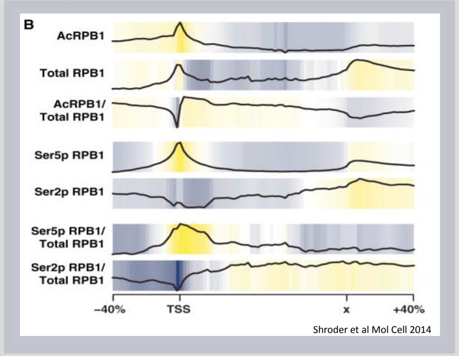
sequence-specific factors



chromatin associated



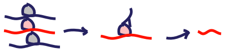
transcription machinery



Shroder et al Mol Cell 2014

methodology

immunoprecipitation

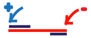


library construction

sequencing

mapping to genome reference (BWA or bowtie2)

identify tag locations



calculate median insert size (single end)

determine fragment midpoint (paired end)

estimate actual binding location per tag

calculate genomic density of shifted tags

measure background density

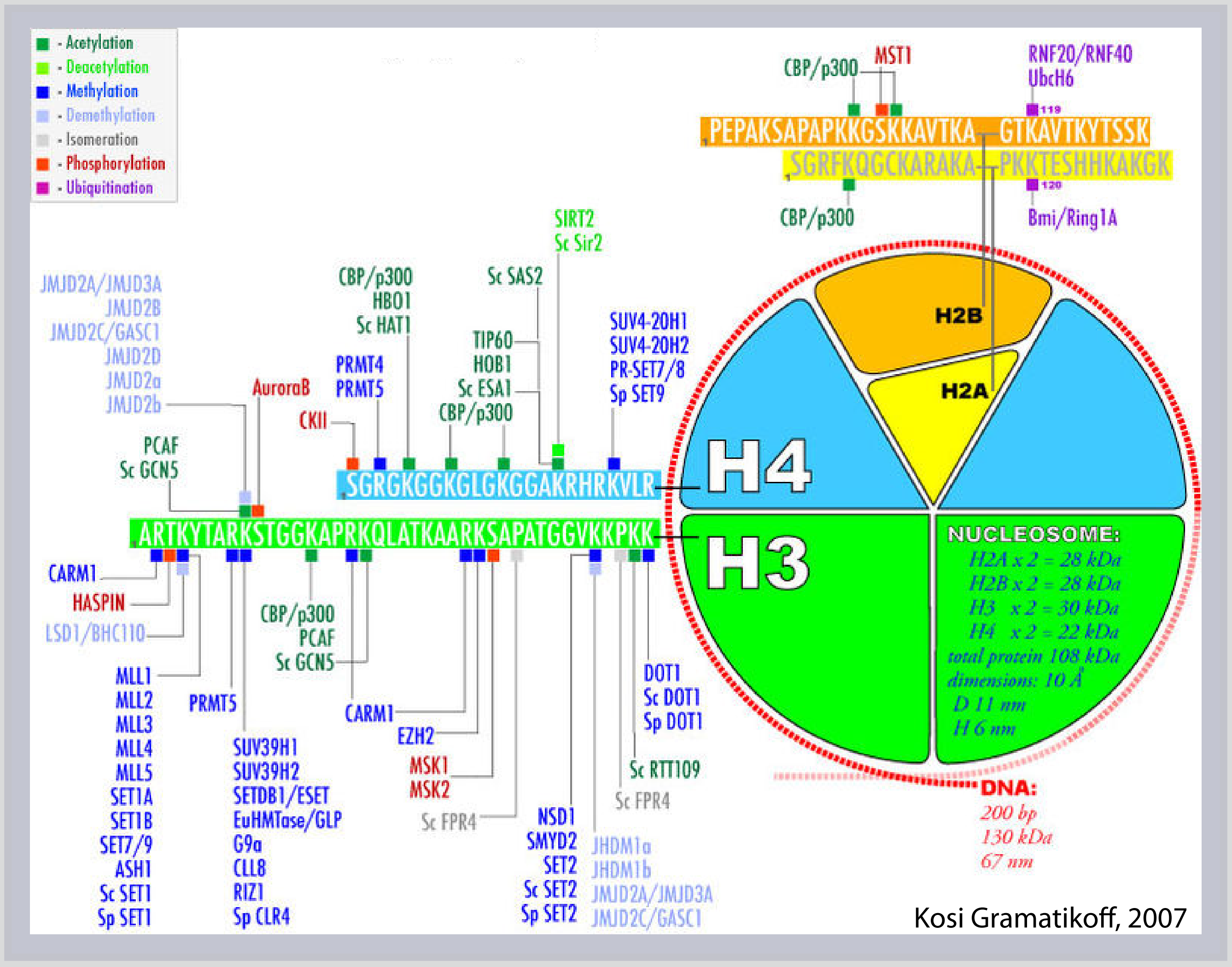
calculate background-normalized binding density

call peaks (punctate marks) or regions (broad marks)

histone modifications

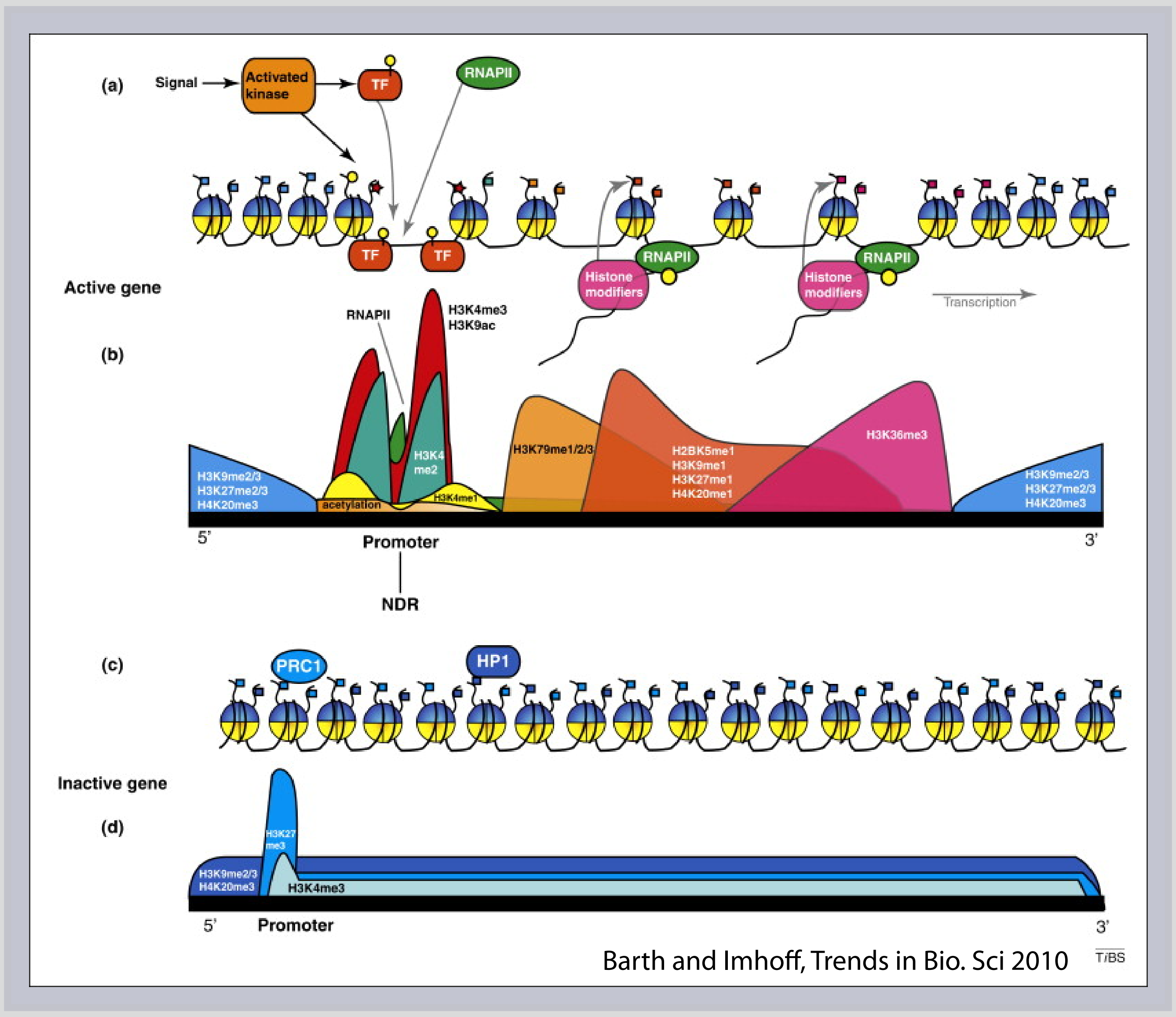
BMI206 November 3, 2014

catalogue



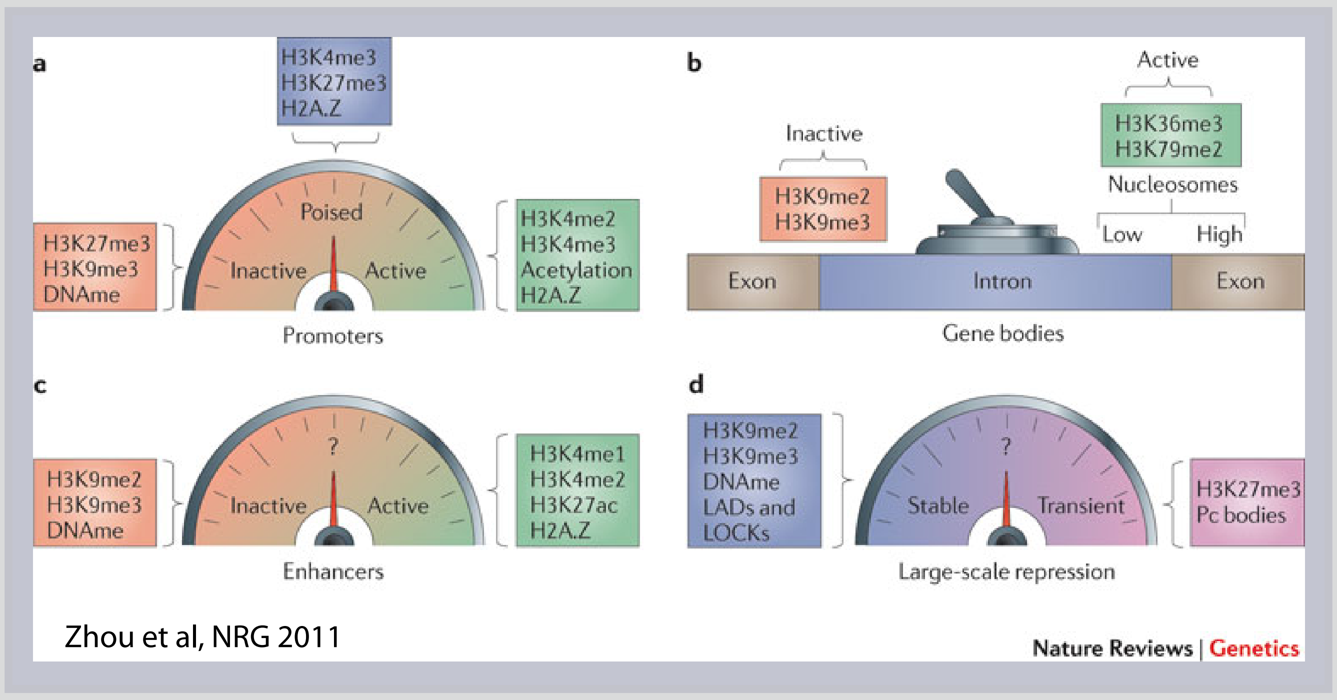
Kosi Gramatikoff, 2007

genomic distribution



Barth and Imhoff, Trends in Bio. Sci 2010

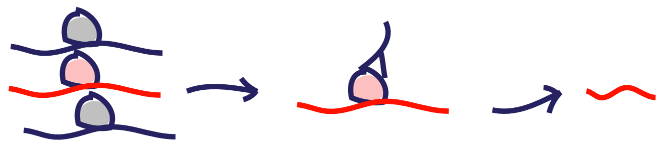
functional association



Zhou et al, NRG 2011

methodology

immunoprecipitation

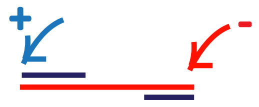


library construction

sequencing

mapping to genome reference (BWA or bowtie2)

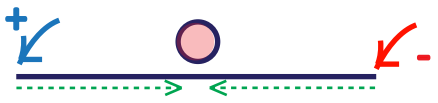
identify tag locations



calculate median insert size (single end)

determine fragment midpoint (paired end)

estimate actual binding location per tag

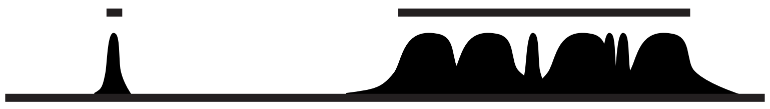


calculate genomic density of shifted tags

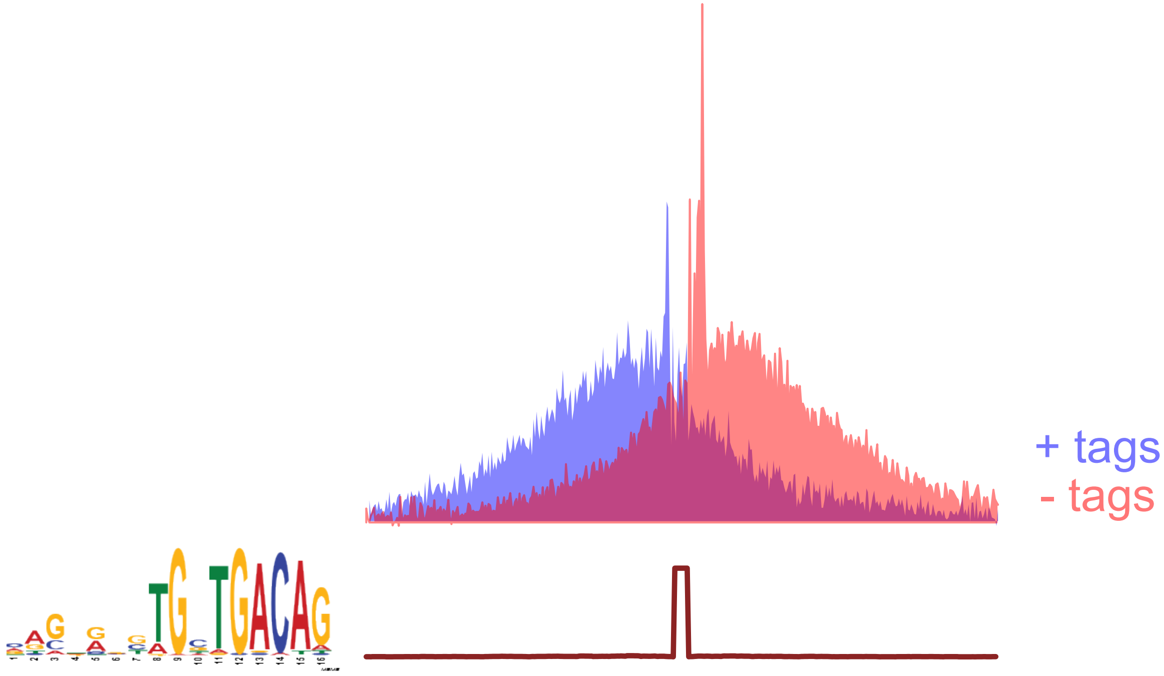
measure background density

calculate background-normalized binding density

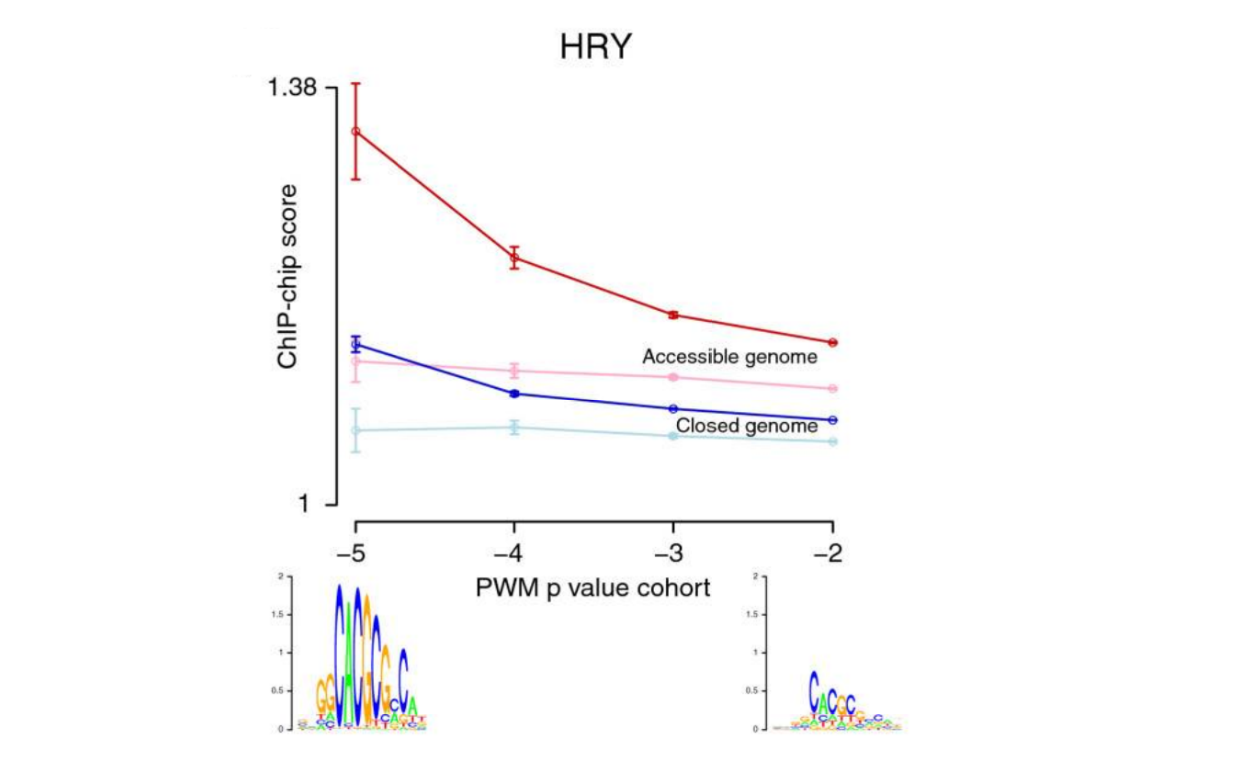
call peaks (punctate marks) or regions (broad marks)



sequence-specific factors

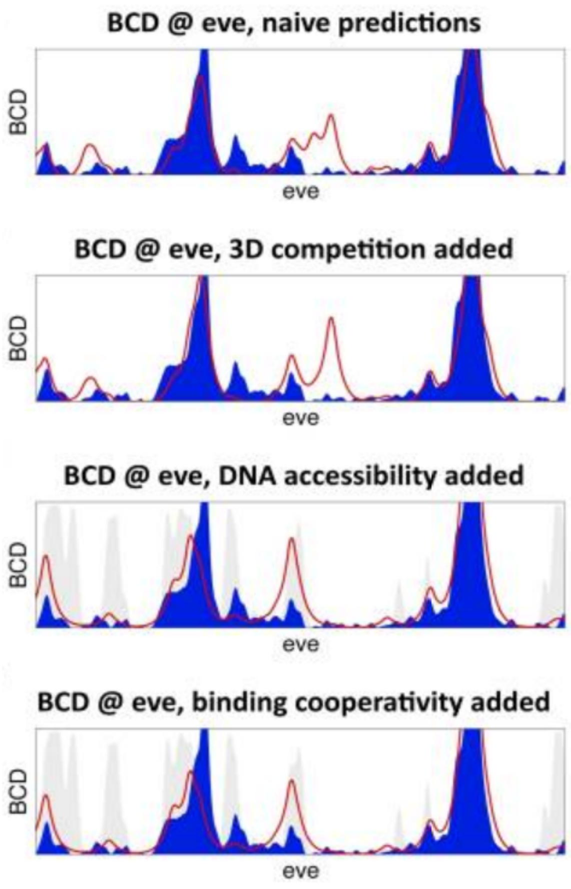


sequence-specific factors



Li et al Genome Biology, 2011

sequence-specific factors



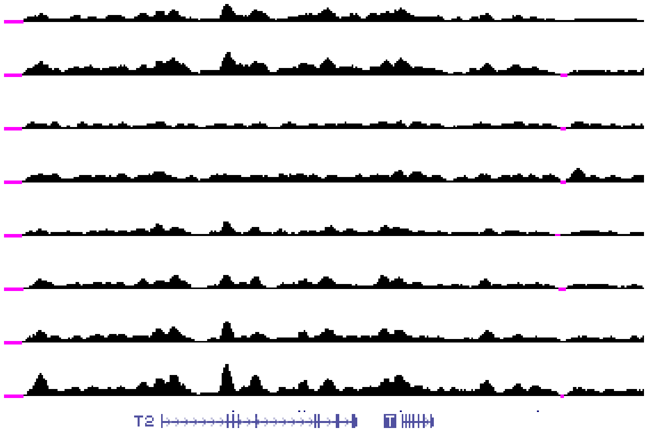
━ modeled ChIP tag density

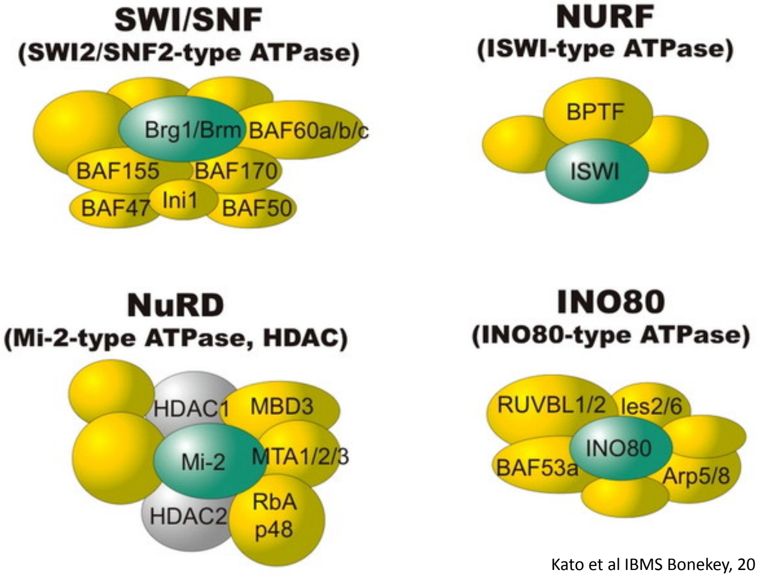
■ actual ChIP tag density

Kaplan et al PLoS Genetics, 2011

chromatin-associated factors

chromatin remodellers



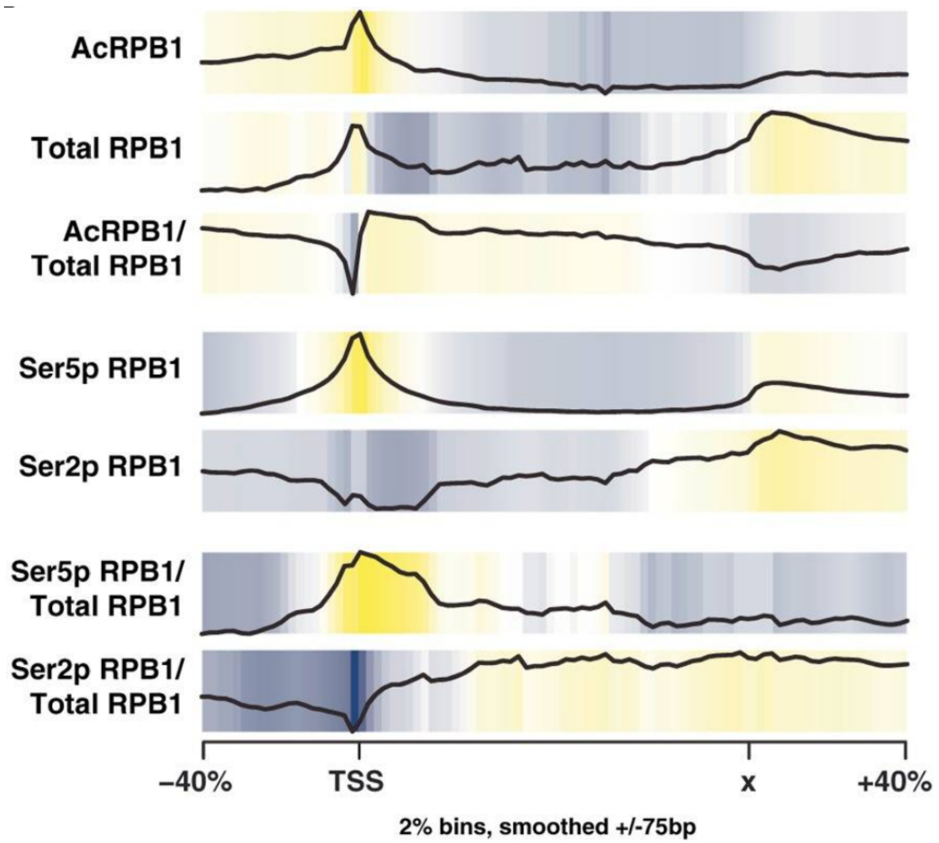


Kato et al IBMS Bonekey, 2010

transcription machinery

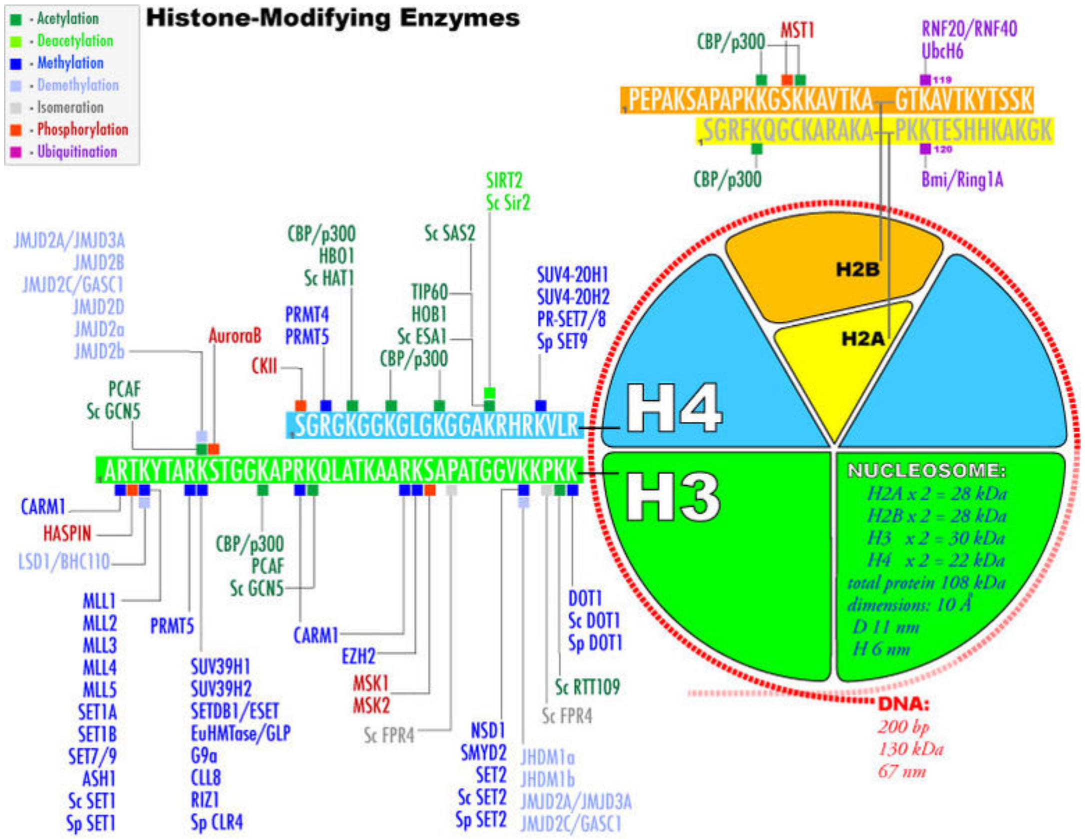
e.g. RNA polymerase II

(and various post-translational modifications)



Schroder et al Mol Cell 2012

histone modifications



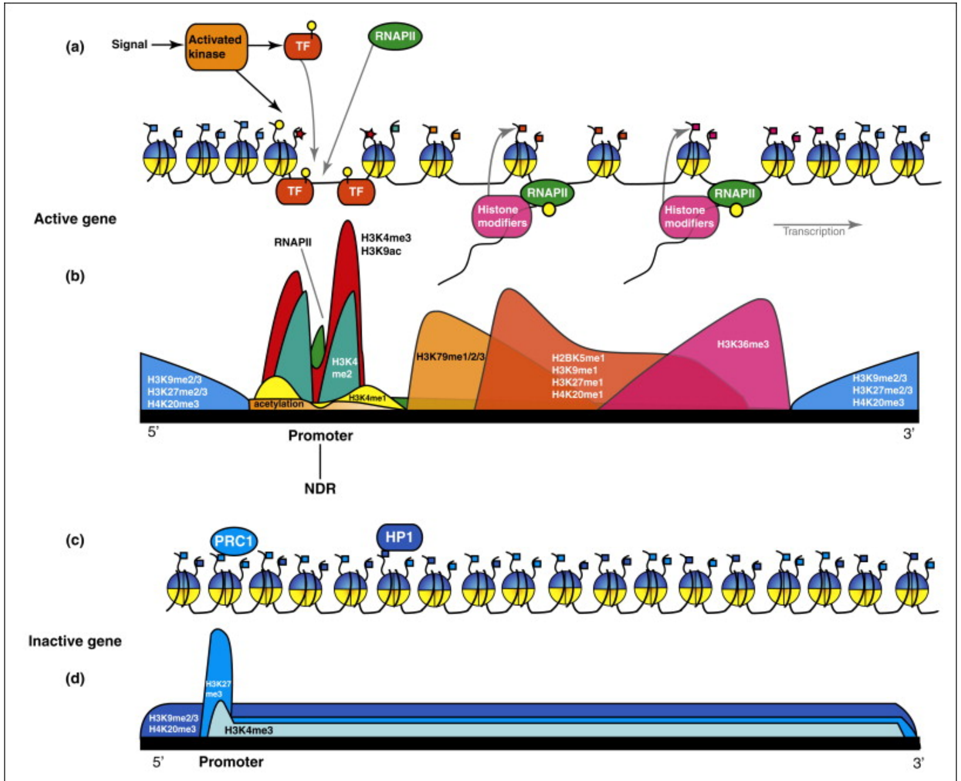
Kosi Gramatikoff, 2007

most commonly studied:

H3K4me3
H3K27ac
H3K27me3

histone modifications

wide range of binding patterns

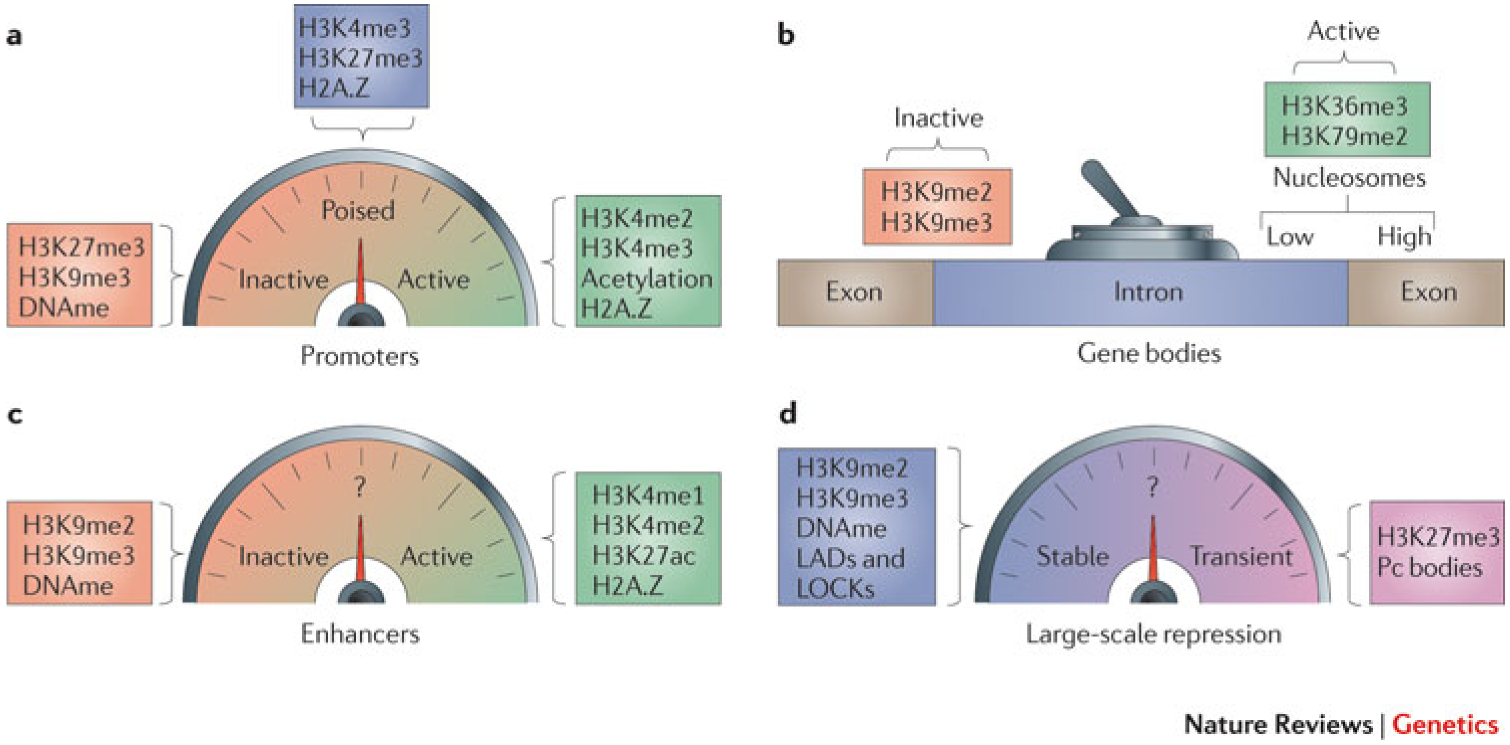


Barth and Imhoff, Trends in Bio. Sci 2010

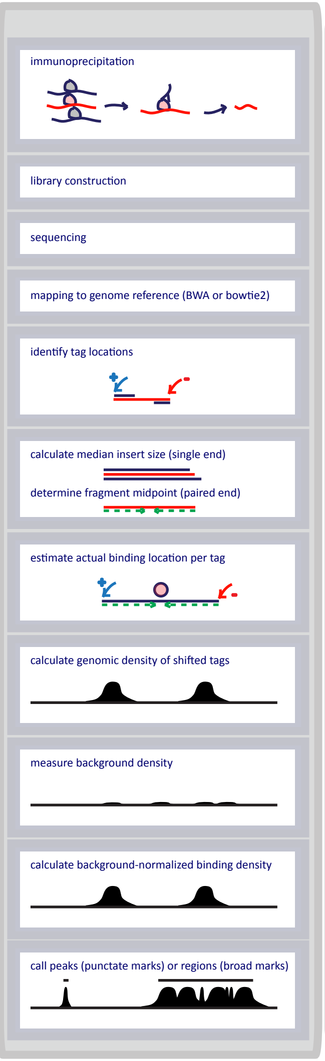
H3K4me3 - transcription start sites
H3K27ac - active enhancers and TSS
H3K27me3 - domain around TSS
H3K36me3 - active gene bodies

histone modifications

wide range of functional associations



Zhou et al, NRG 2011



immunoprecipitation

library construction

sequencing

mapping to genome reference (BWA or bowtie2)

identify tag locations

calculate median insert size (single end)

determine fragment midpoint (paired end)

estimate actual binding location per tag

calculate genomic density of shifted tags

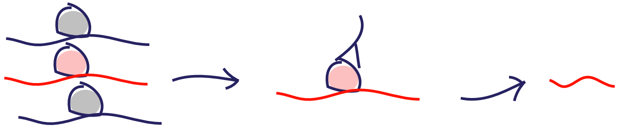
measure background density

calculate background-normalized binding density

call peaks (punctate marks) or regions (broad marks)

methodology

immunoprecipitation



library construction

sequencing

mapping to genome reference (BWA or bowtie2)

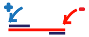
immunoprecipitation

library construction

sequencing

mapping to genome reference (BWA or bowtie2)

identify tag locations



calculate median insert size (single end)

determine fragment midpoint (paired end)

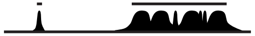
estimate actual binding location per tag

calculate genomic density of shifted tags

measure background density

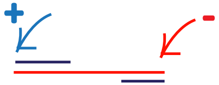
calculate background-normalized binding density

call peaks (punctate marks) or regions (broad marks)

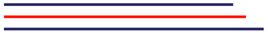


methodology

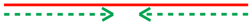
identify tag locations



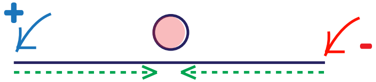
calculate median insert size (single end)



determine fragment midpoint (paired end)



estimate actual binding location per tag



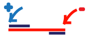
immunoprecipitation

library construction

sequencing

mapping to genome reference (BWA or bowtie2)

identify tag locations



calculate median insert size (single end)

determine fragment midpoint (paired end)

estimate actual binding location per tag

calculate genomic density of shifted tags

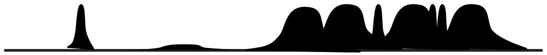
measure background density

calculate background-normalized binding density

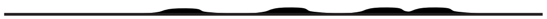
call peaks (punctate marks) or regions (broad marks)

methodology

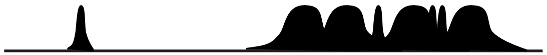
calculate genomic density of shifted tags



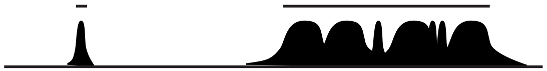
measure background density



calculate background-normalized binding density



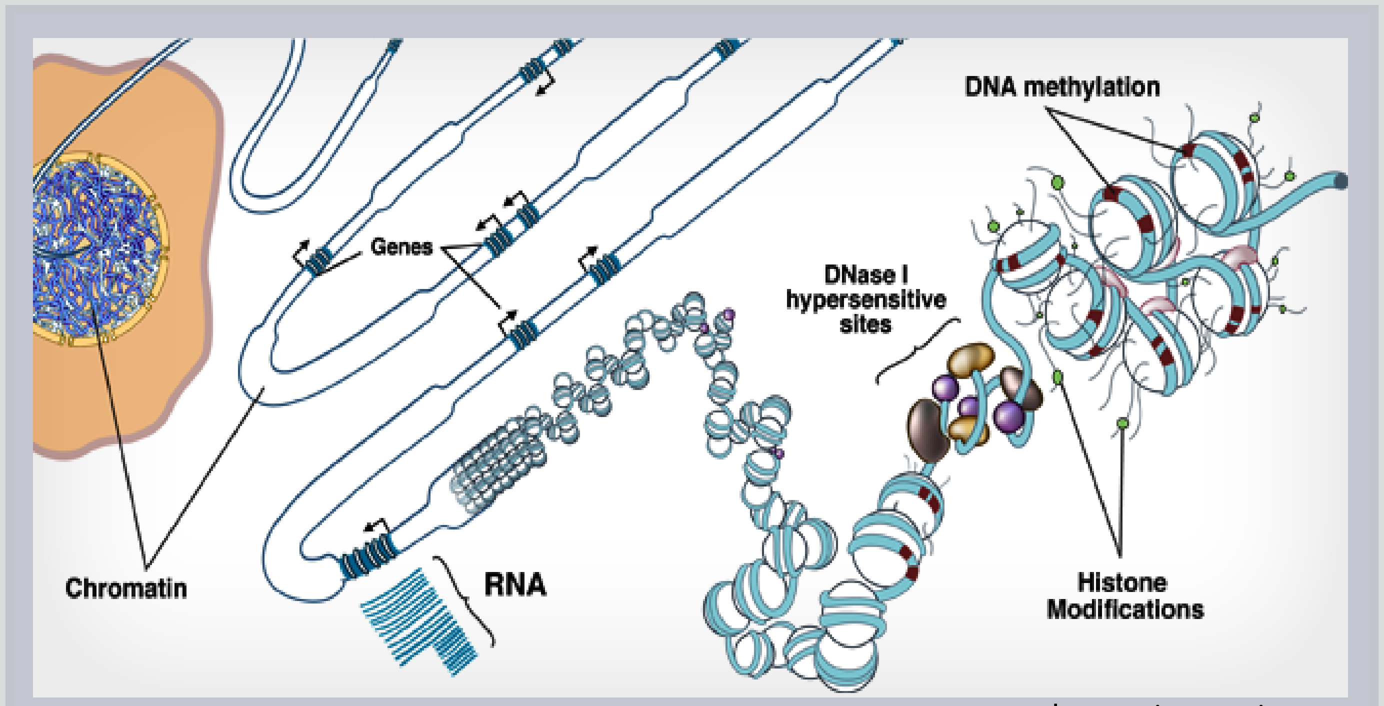
call peaks (punctate marks) or regions (broad marks)



chromatin accessibility

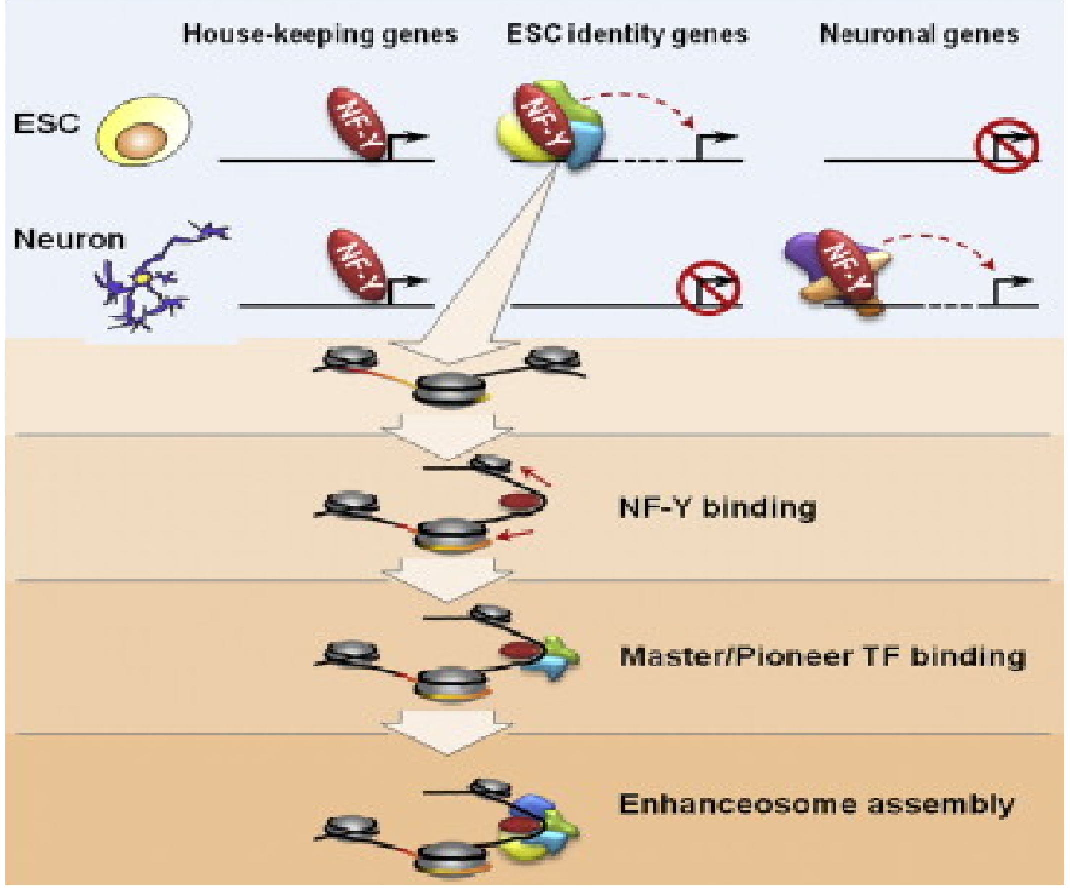
BMI206 November 3, 2014

functional regulatory elements are accessible



roadmapepigenomics.org

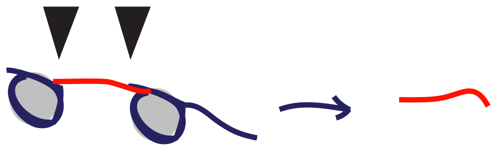
formation of accessible chromatin



Oldfield et al, Mol Cell 2014

methodology (digestion or transposase)

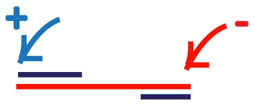
DNase I or Tn5 release fragments of open chromatin



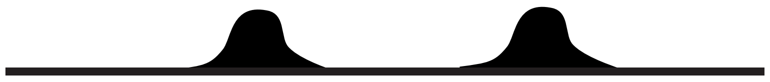
library construction and sequencing

mapping to genome reference (BWA or bowtie2)

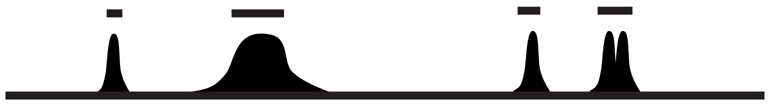
identify cleavage/insertion location



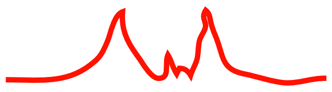
calculate density of cleavage/insertion events



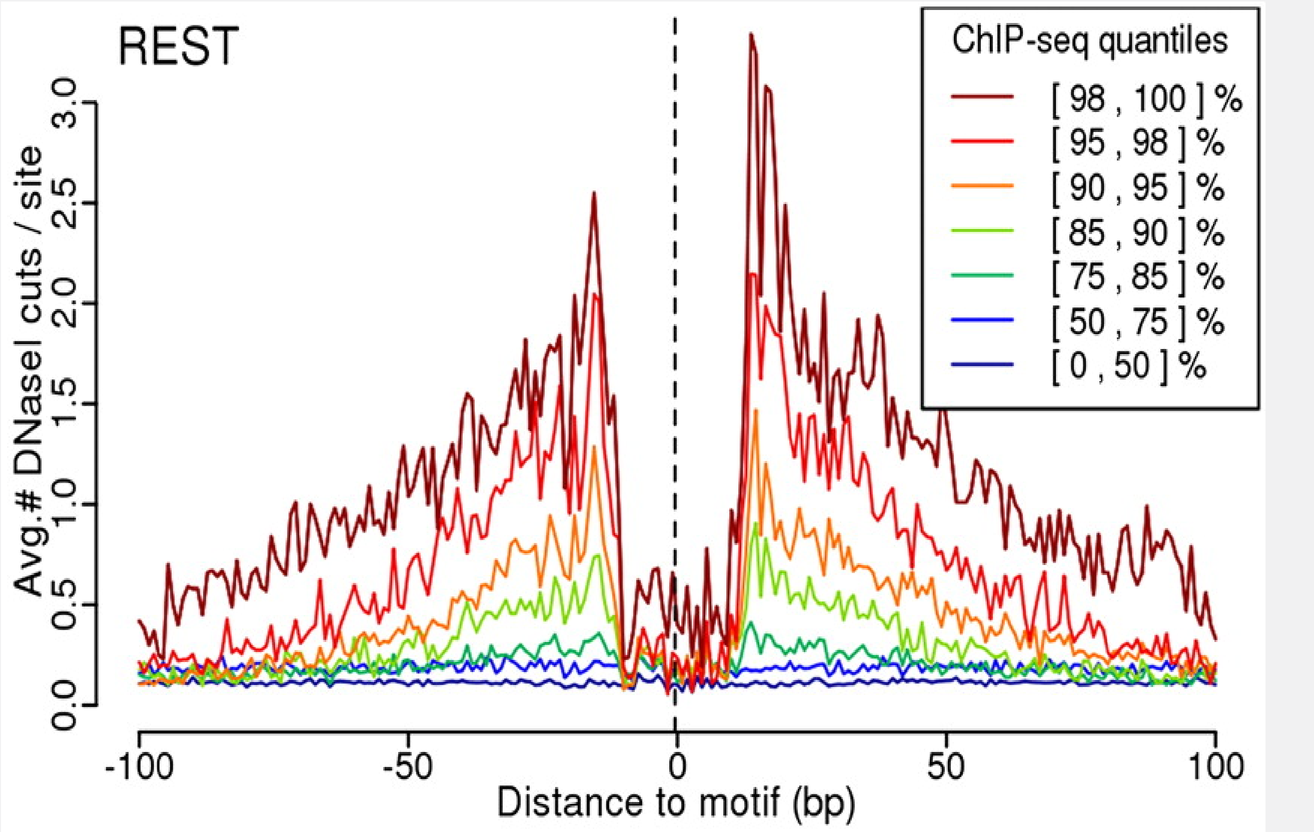
call regions of accessibility



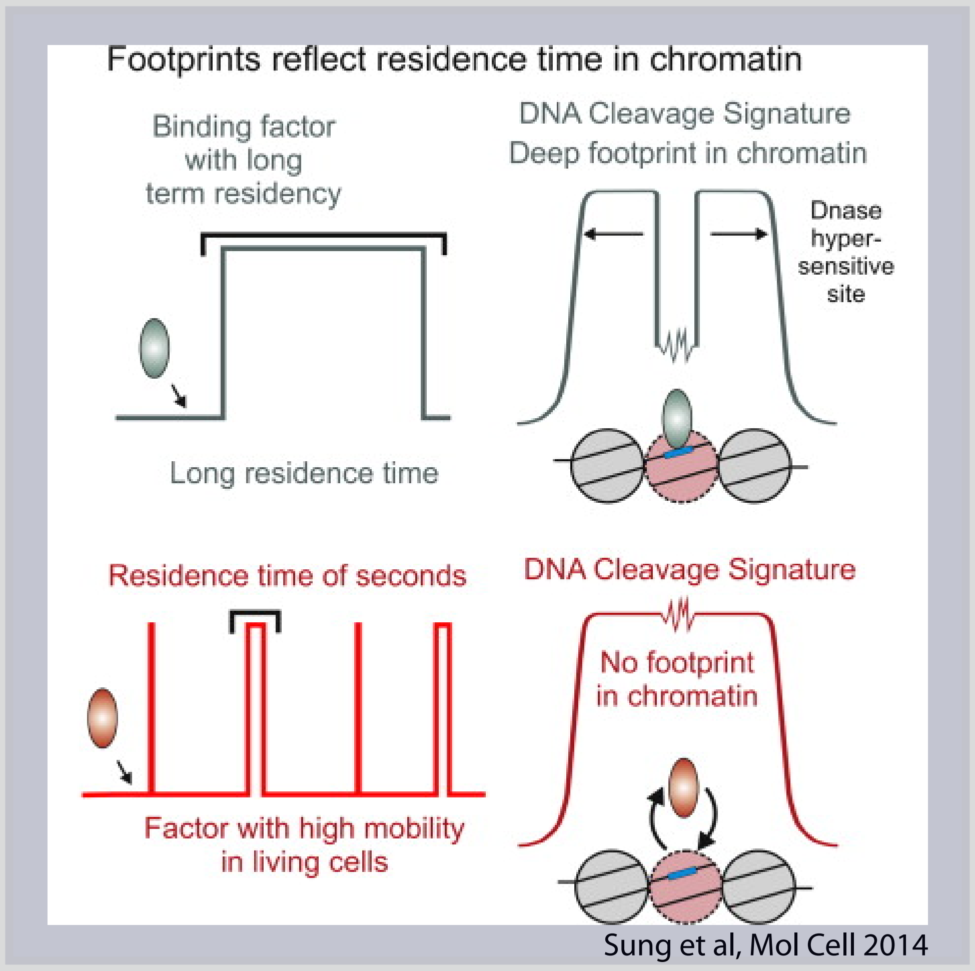
call footprints within accessible chromatin



footprinting

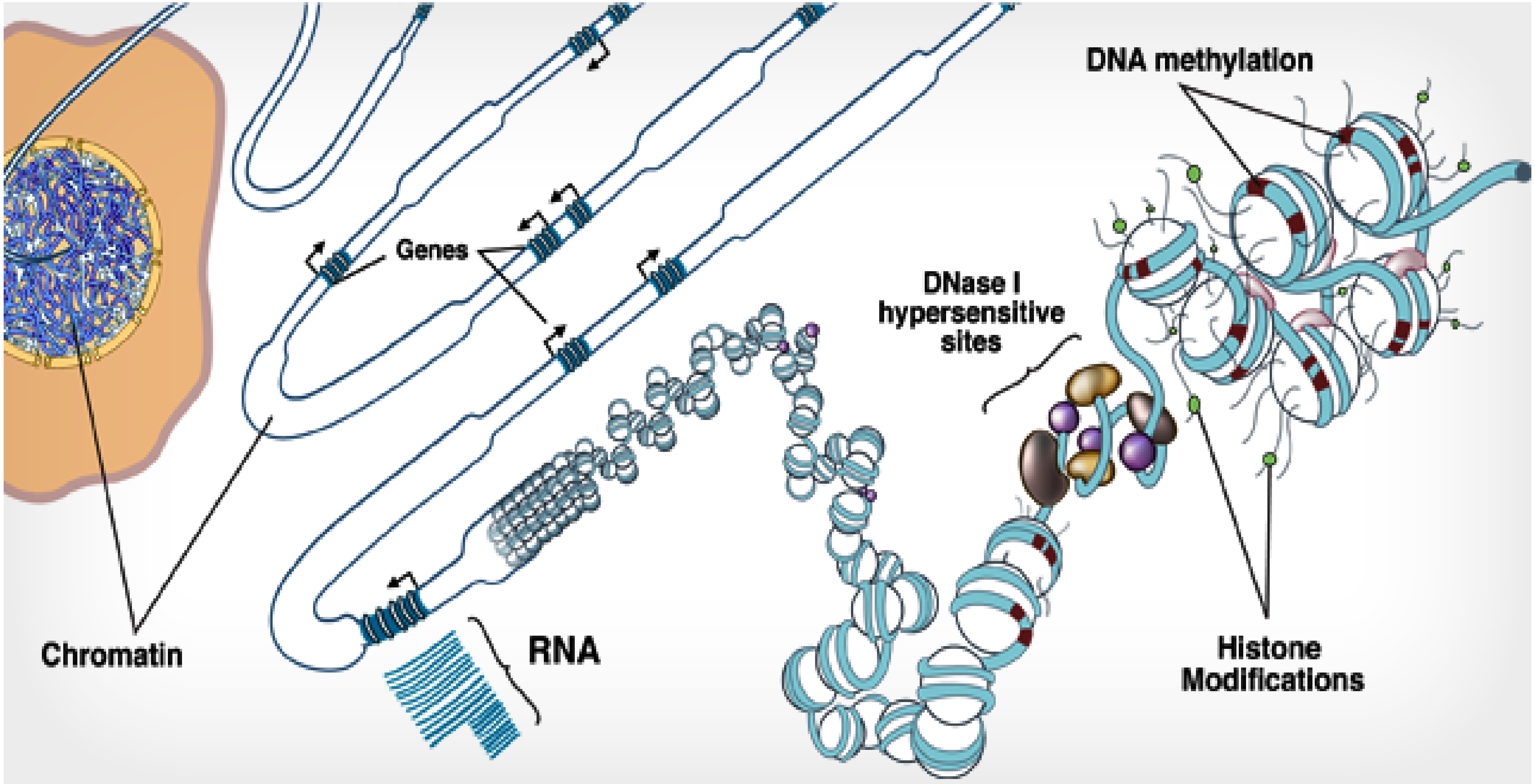


Pique-Regi et al Genome Research 2011


Footprints reflect residence time in chromatin
Binding factor with long term residency
DNA Cleavage Signature
Deep footprint in chromatin
Dnase hyper-sensitive site
Long residence time
Residence time of seconds
DNA Cleavage Signature
No footprint in chromatin
Factor with high mobility in living cells

Sung et al, Mol Cell 2014

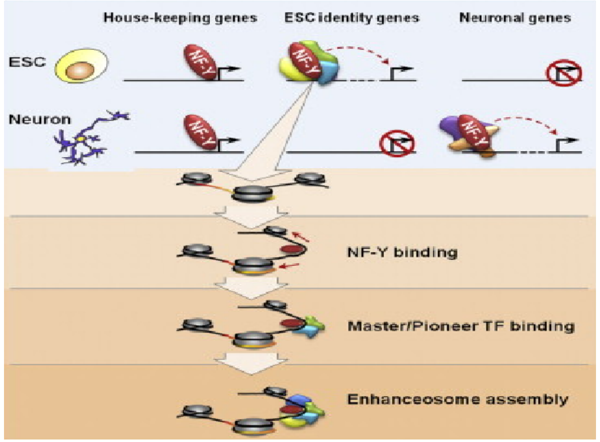
functional regulatory elements are accessible



roadmapepigenomics.org

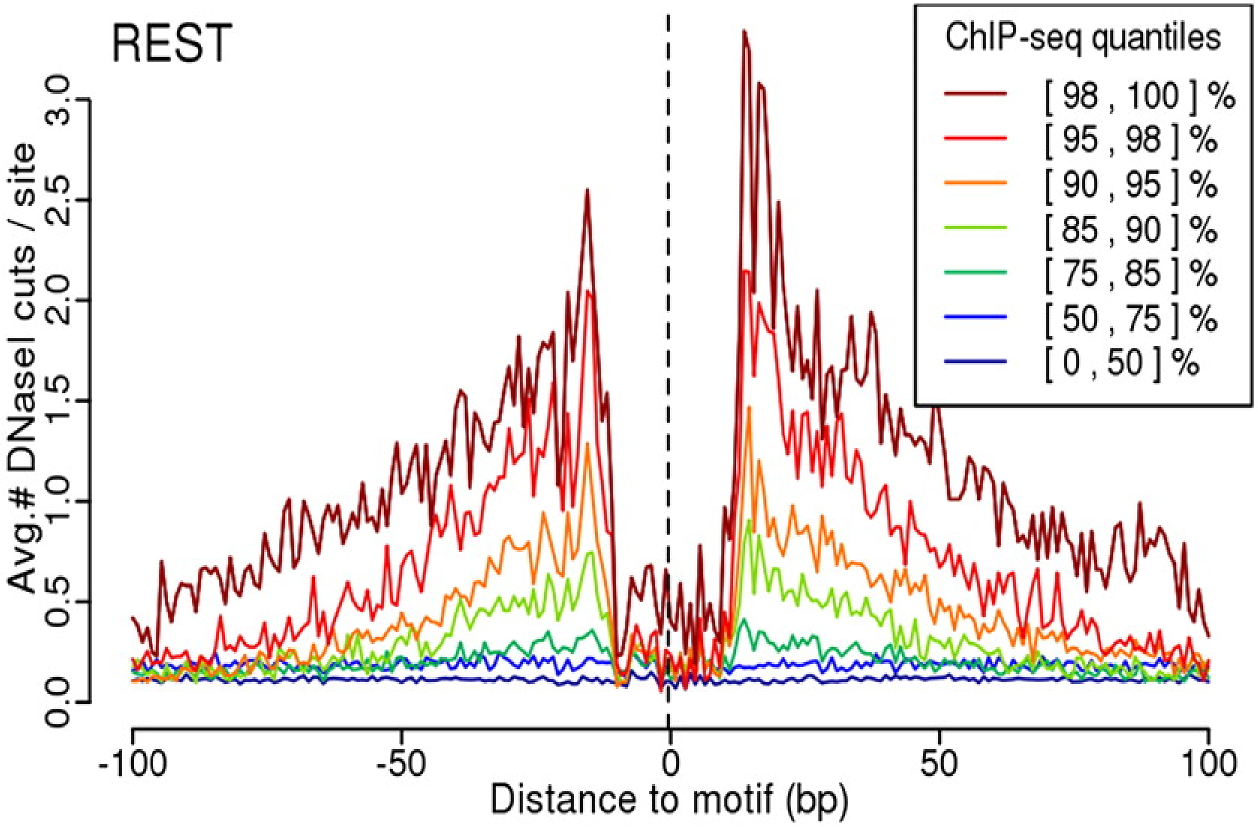
formation of accessible chromatin

pioneering factors make important genes accessible during development



Oldfield et al, Mol Cell 2014

footprinting

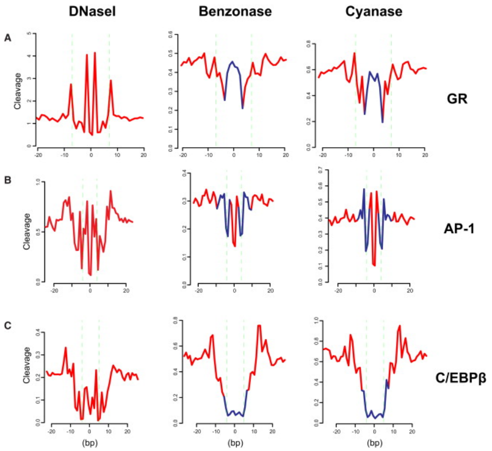


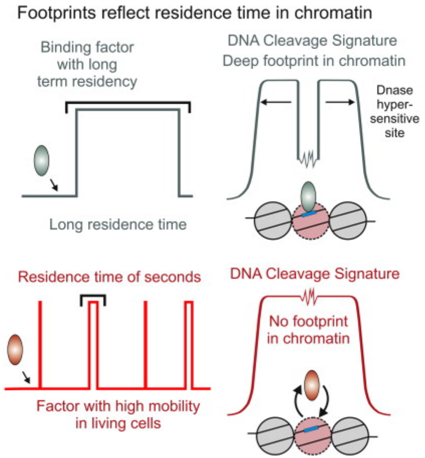
Pique-Regi et al Genome Research 2011

footprinting

distinct for
different factors

not consistently
present for all factors




Footprints reflect residence time in chromatin
Binding factor with long term residency
DNA Cleavage Signature
Deep footprint in chromatin
Dnase hyper-sensitive site
Long residence time
Residence time of seconds
DNA Cleavage Signature
No footprint in chromatin
Factor with high mobility in living cells

(effect seen in naked DNA for many factors)

Sung et al, Mol Cell 2014

methodology (digestion or transposase)

library construction and sequencing

mapping to genome reference (BWA or bowtie2)

identify cleavage/insertion location

calculate density of cleavage/insertion events

call regions of accessibility

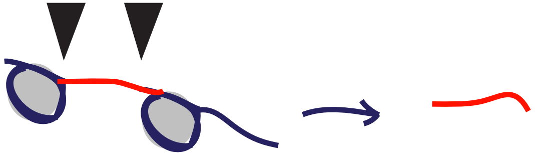
call footprints within accessible chromatin

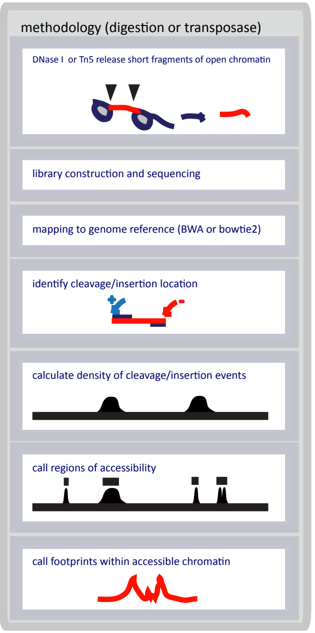
methodology

DNase I digestion
(DNase-seq)

transposon integration
(ATAC-seq)

DNase I or Tn5 release short fragments of open chromatin





methodology (digestion or transposase)

DNase I or Tn5 release short fragments of open chromatin

library construction and sequencing

mapping to genome reference (BWA or bowtie2)

identify cleavage/insertion location

calculate density of cleavage/insertion events

call regions of accessibility

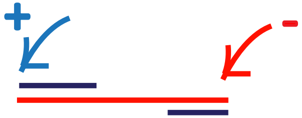
call footprints within accessible chromatin

methodology

library construction and sequencing

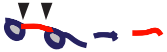
mapping to genome reference (BWA or bowtie2)

identify cleavage/insertion location



methodology (digestion or transposase)

DNase I or Tn5 release short fragments of open chromatin



library construction and sequencing

mapping to genome reference (BWA or bowtie2)

identify cleavage/insertion location

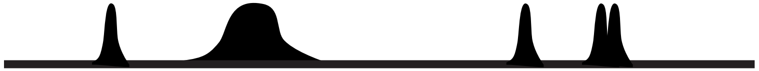
calculate density of cleavage/insertion events

call regions of accessibility

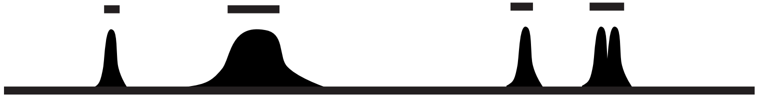
call footprints within accessible chromatin

methodology

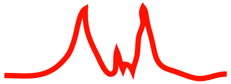
calculate density of cleavage/insertion events



call regions of accessibility

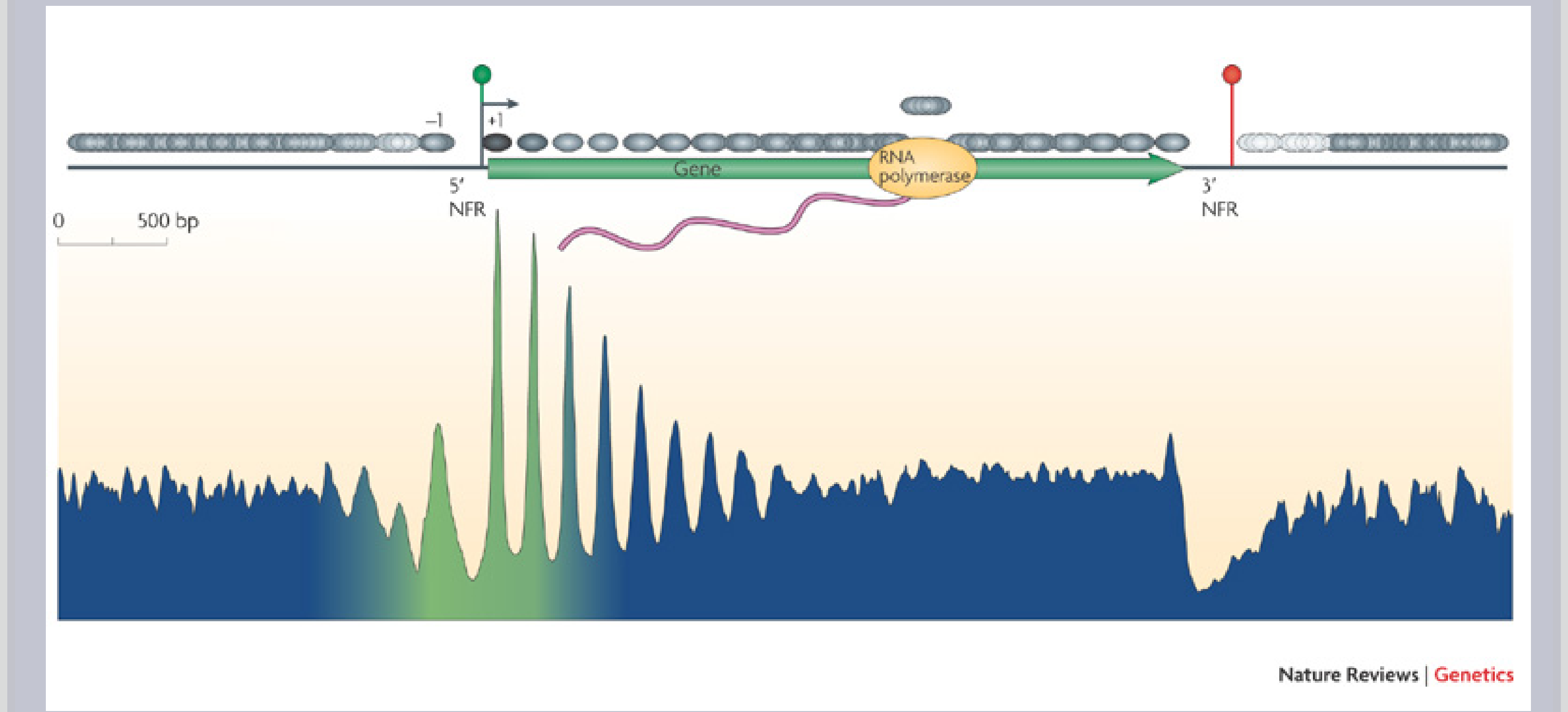


call footprints within accessible chromatin



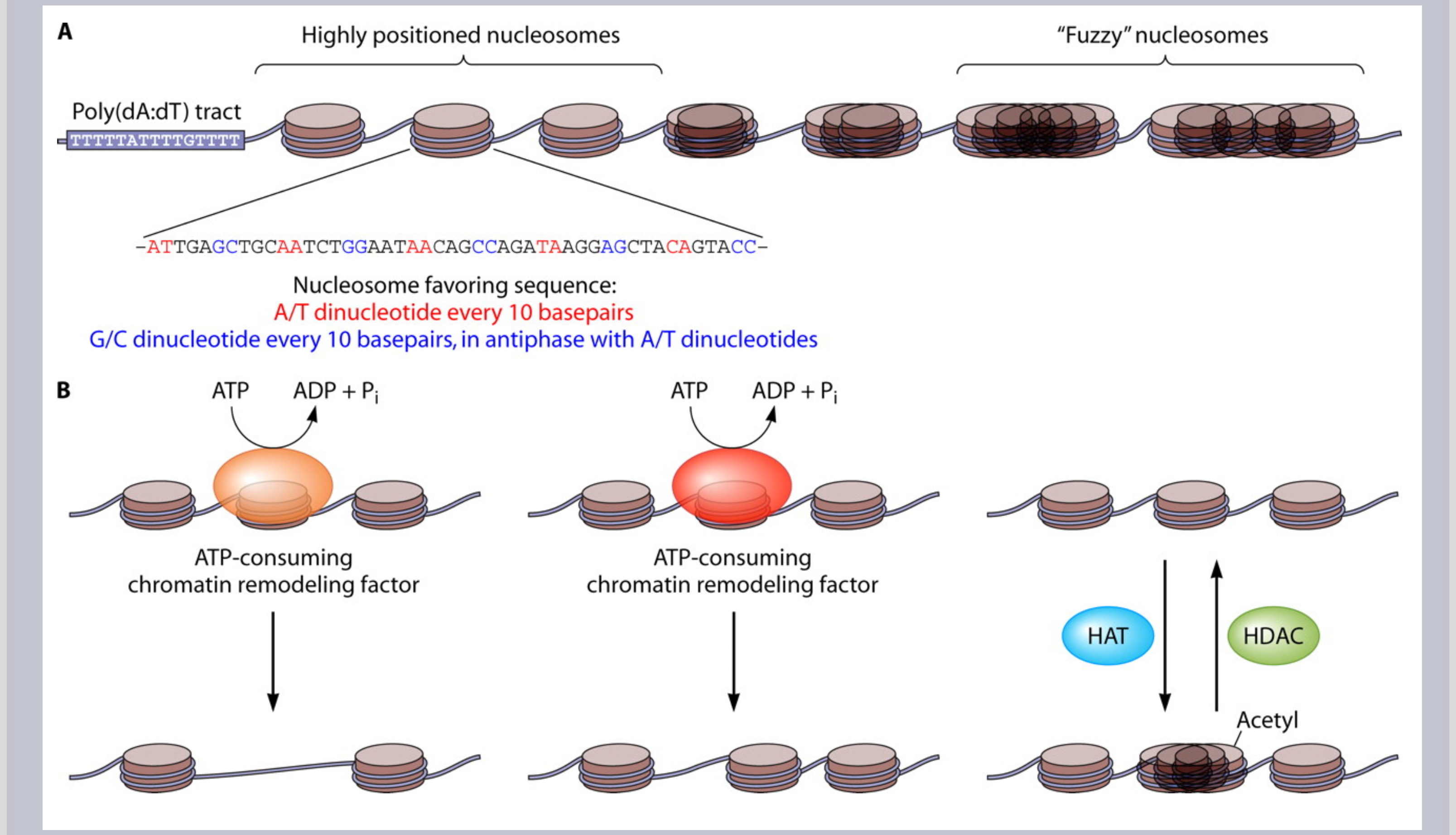
nucleosome position

nucleosome positioning during transcription



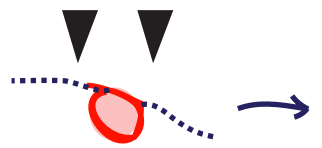
Jiang et al NRG 2009

determinants of positioning stability

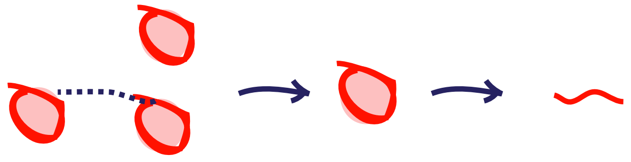

A
Highly positioned nucleosomes
"Fuzzy" nucleosomes
Poly(dA:dT) tract
TTTTTATTTTGTTTT
–ATTGAGCTGCAATCTGGAATAACAGCCAGATAAGGAGCTACAGTACC–
Nucleosome favoring sequence:
A/T dinucleotide every 10 basepairs
G/C dinucleotide every 10 basepairs, in antiphase with A/T dinucleotides
B
ATP
ADP + P$_i$
ATP-consuming chromatin remodeling factor
ATP
ADP + P$_i$
ATP-consuming chromatin remodeling factor
HAT
HDAC
Acetyl

methodology (MNase digestion)

MNase I digests linker DNA, releasing multisomes



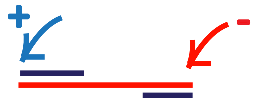
purify mononucleosome-bound DNA



library construction and sequencing

mapping to genome reference (BWA or bowtie2)

identify cleavage location

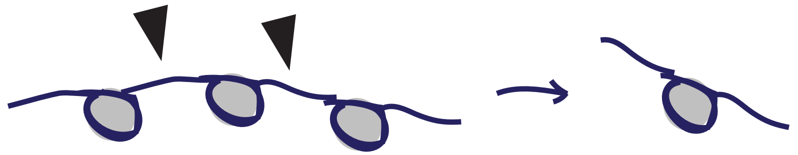


calculate nucleosomal density

call positioned nucleosomes

methodology (transposon insertion)

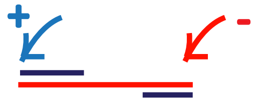
Tn5 integrates into accessible chromatin, and releases multisomes



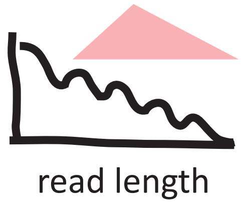
library construction and sequencing

mapping to genome reference (BWA or bowtie2)

identify cleavage location



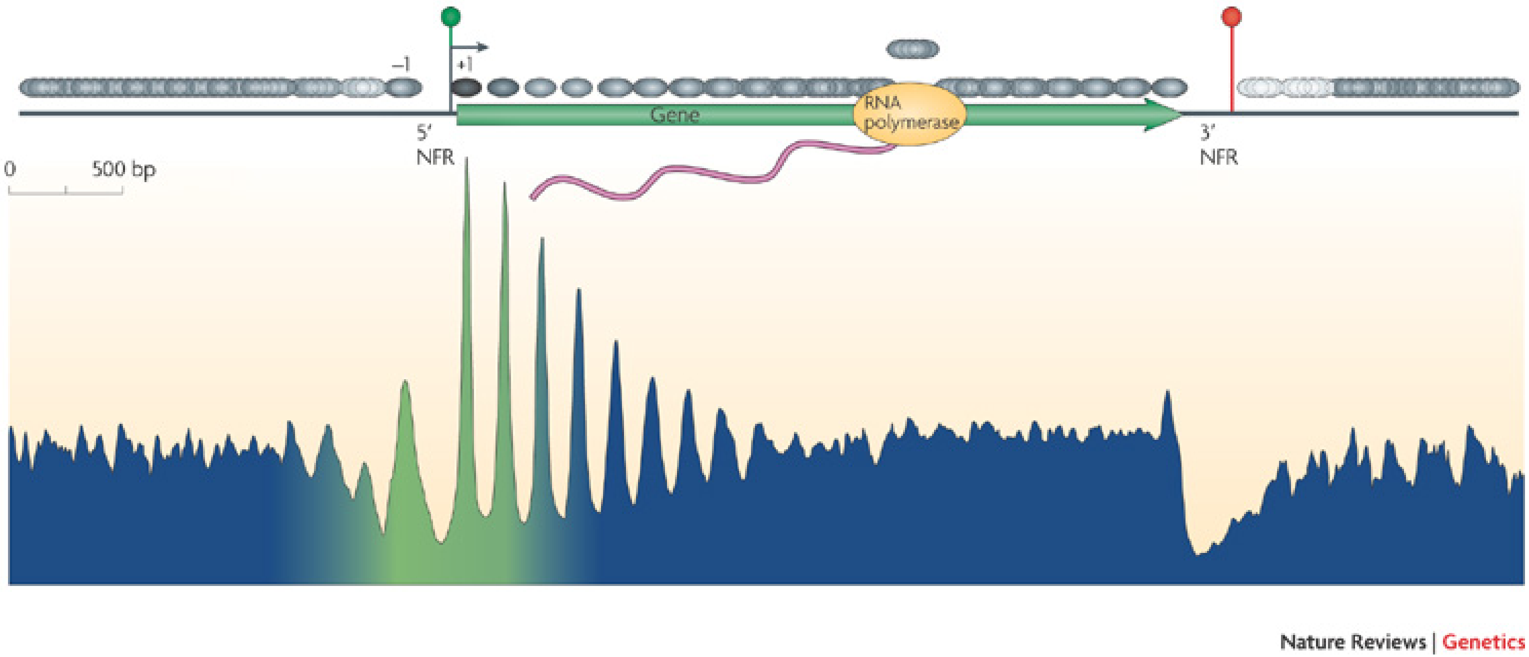
identify reads that span at least one nucleosome



calculate density of tags / number of insertion events

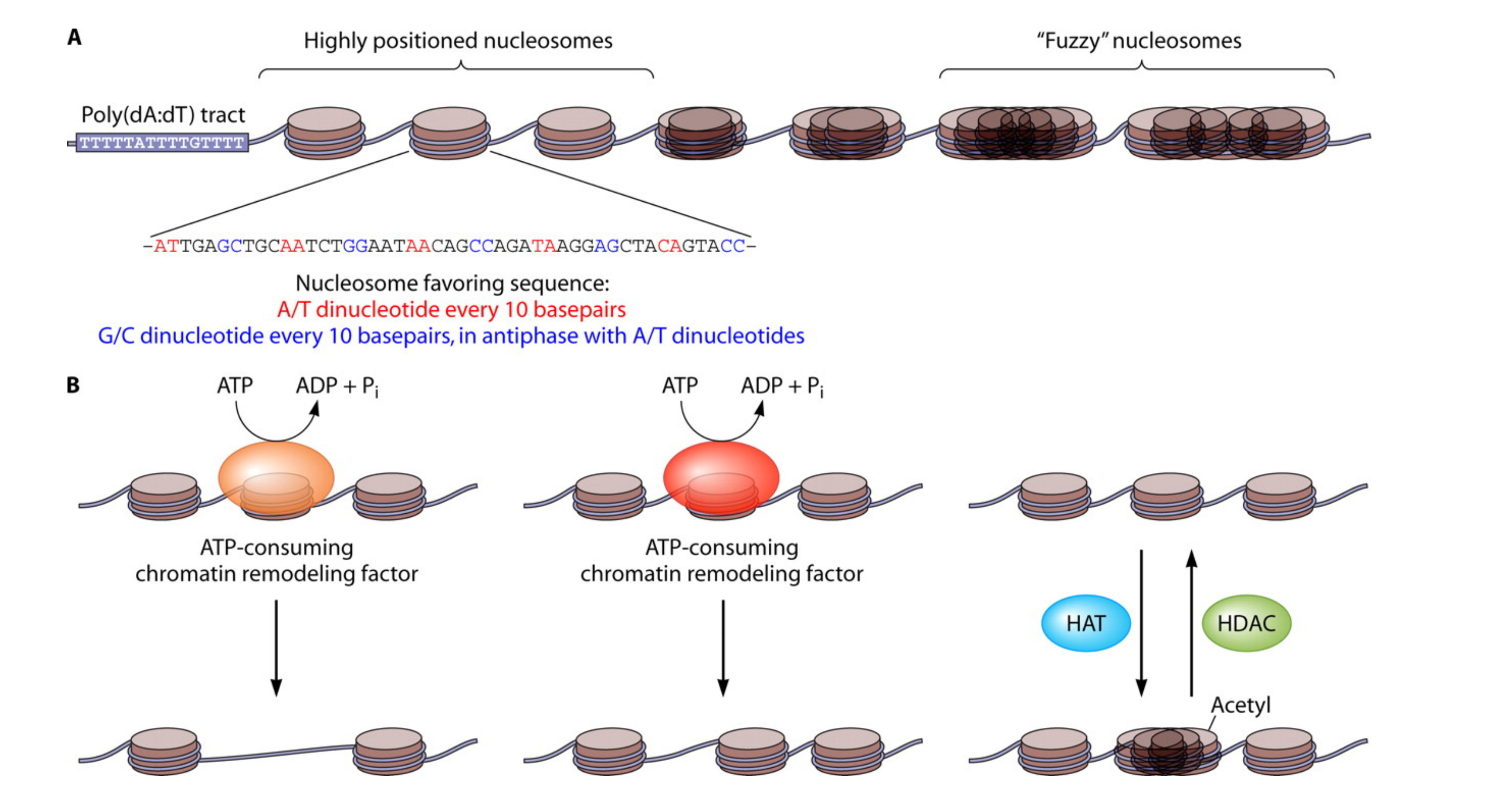
call positioned nucleosomes

nucleosome positioning during transcription

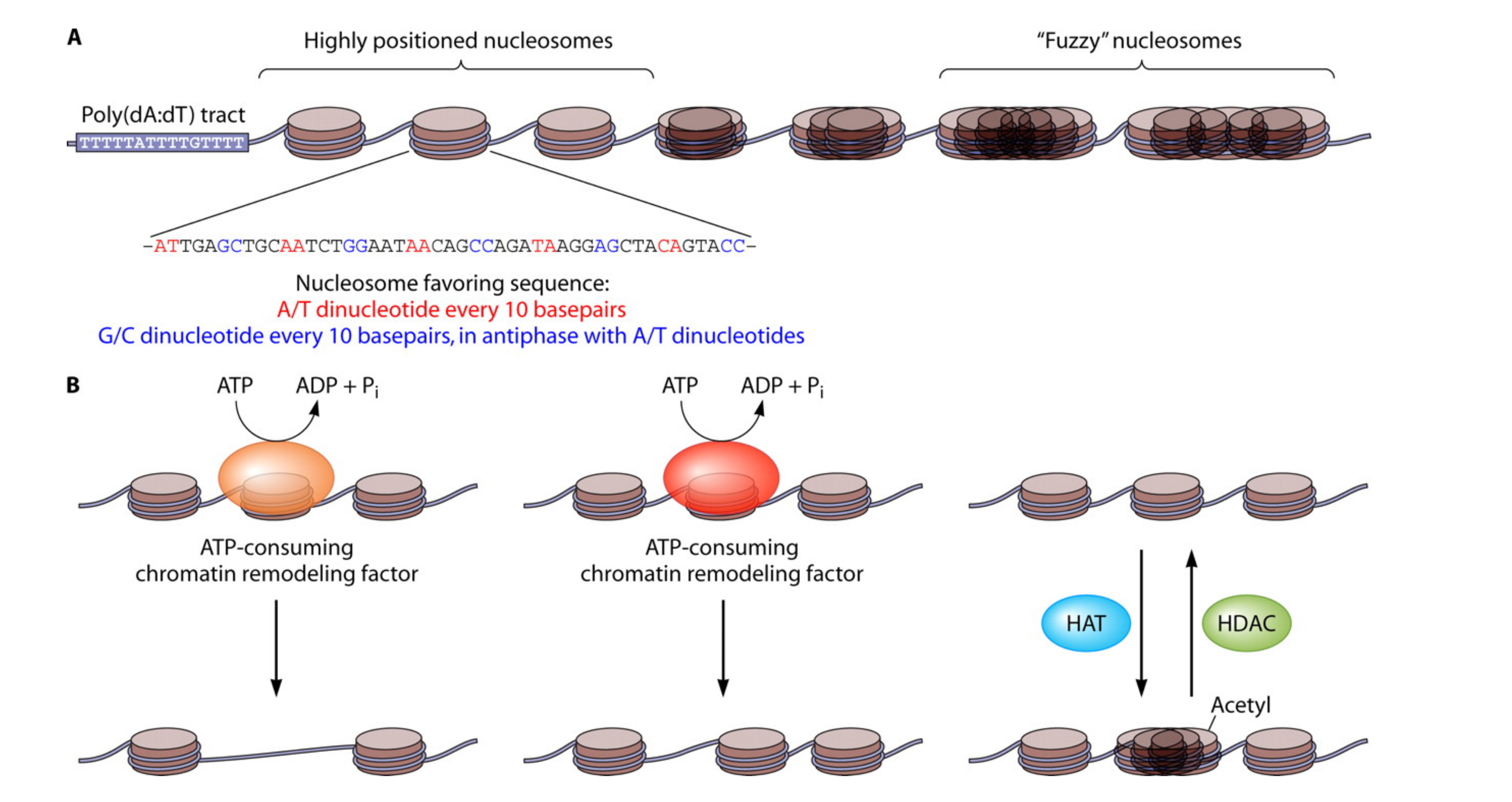


Jiang et al NRG 2009

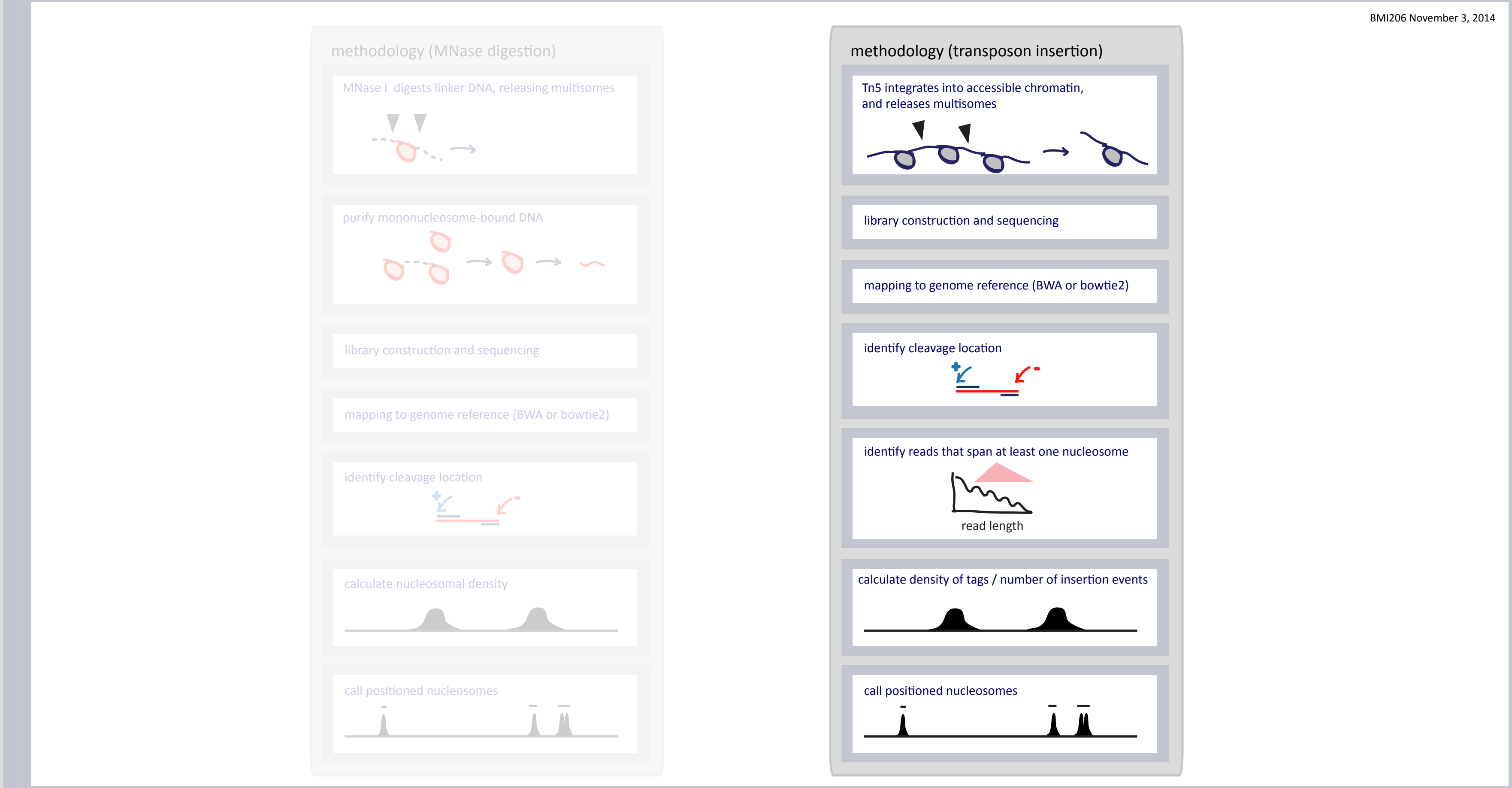
positioning stability


A
Highly positioned nucleosomes
"Fuzzy" nucleosomes
Poly(dA:dT) tract
TTTTTATTTTGTTTT
–ATTGAGCTGCAATCTGGAATAACAGCCAGATAAGGAGCTACAGTACC–
Nucleosome favoring sequence:
A/T dinucleotide every 10 basepairs
G/C dinucleotide every 10 basepairs, in antiphase with A/T dinucleotides
B
ATP
ADP + Pi
ATP-consuming chromatin remodeling factor
ATP
ADP + Pi
ATP-consuming chromatin remodeling factor
HAT
HDAC
Acetyl

positioning stability


A
Highly positioned nucleosomes
"Fuzzy" nucleosomes
Poly(dA:dT) tract
TTTTTATTTTGTTTT
-ATTGAGCTGCAATCTGGAATAACAGCCAGATAAGGAGCTACAGTACC-
Nucleosome favoring sequence:
A/T dinucleotide every 10 basepairs
G/C dinucleotide every 10 basepairs, in antiphase with A/T dinucleotides
B
ATP
ADP + P$_i$
ATP
ADP + P$_i$
ATP-consuming
chromatin remodeling factor
ATP-consuming
chromatin remodeling factor
HAT
HDAC
Acetyl

two common methodologies (MNase, ATAC-seq)



BMI206 November 3, 2014

methodology (MNase digestion)

MNase I digests linker DNA, releasing multisomes

purify mononucleosome-bound DNA

library construction and sequencing

mapping to genome reference (BWA or bowtie2)

identify cleavage location

calculate nucleosomal density

call positioned nucleosomes

methodology (transposon insertion)

Tn5 integrates into accessible chromatin, and releases multisomes

library construction and sequencing

mapping to genome reference (BWA or bowtie2)

identify cleavage location

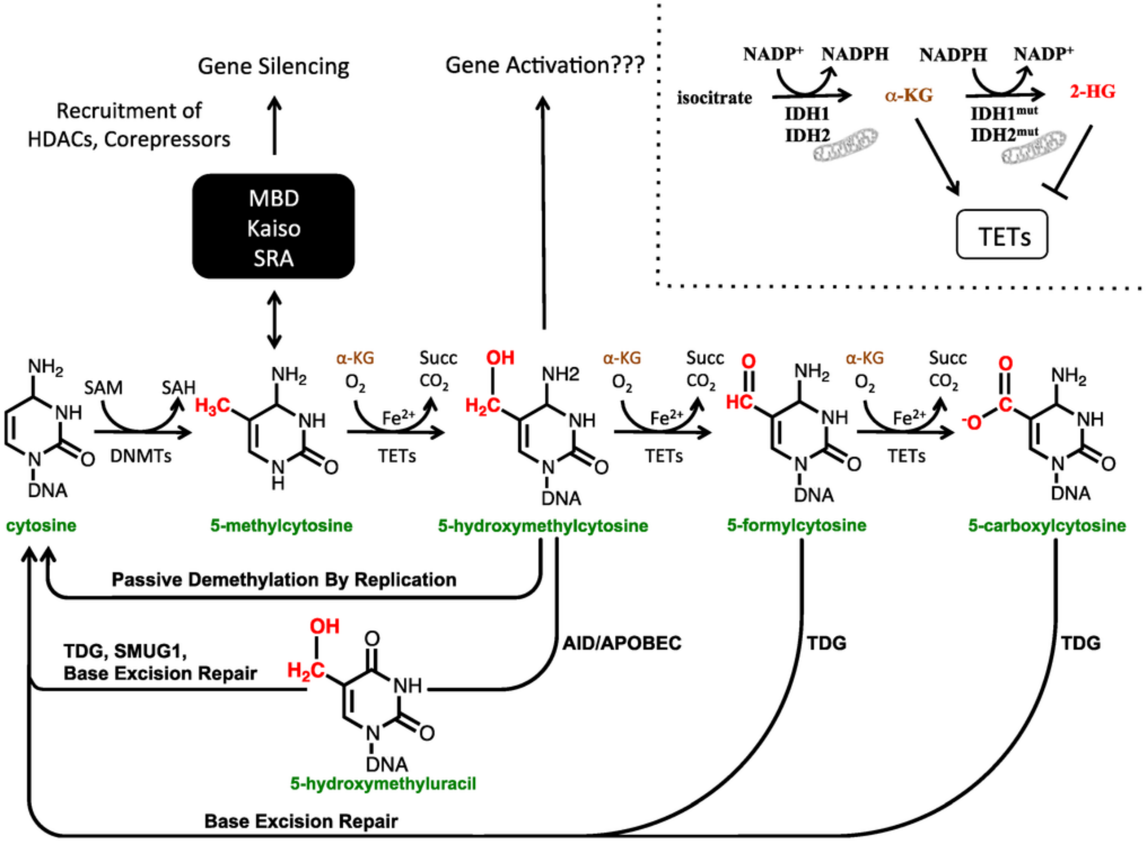
identify reads that span at least one nucleosome

read length

calculate density of tags / number of insertion events

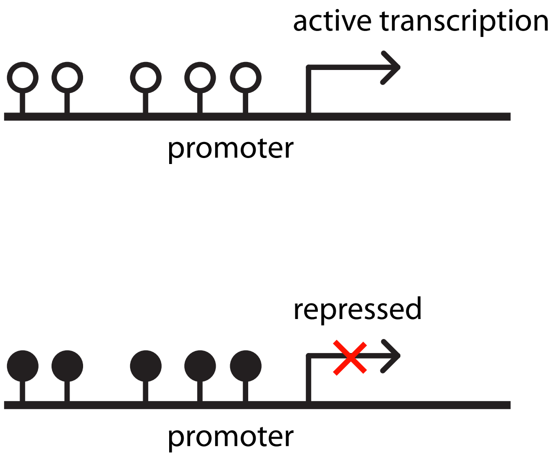
call positioned nucleosomes

catalogue of common base modifications (in mammals)



Mariani et al, Cancers 2013

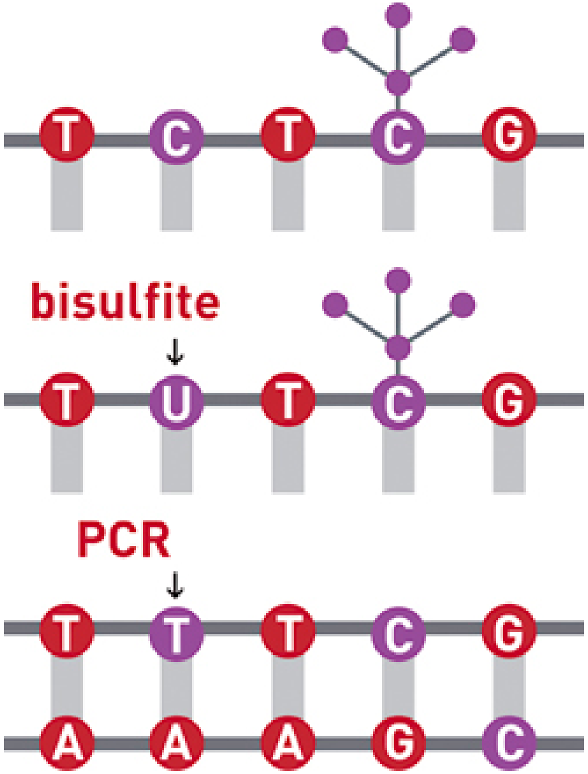
5mC repression of gene expression



● 5mC 5 methyl cytosine

○ C unmodified cytosine

bisulfite conversion / protection



daigenode.com

methodology

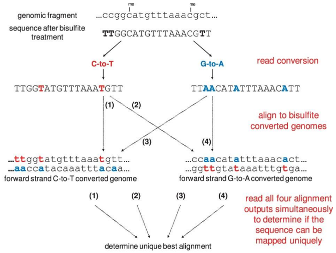
convert DNA with bisulfite

methyl CGTTCG → CGTTUG

fragmentation

library construction and sequencing

mapping


me
me
genomic fragment
...ccggcatgtttaaacgct...
sequence after bisulfite treatment
TTGGCATGTTTAAACGTT
C-to-T
G-to-A
read conversion
TTGGTATGTTTAAATGTT
TTAACATATTTAAACATT
(1)
(2)
align to bisulfite converted genomes
(3)
(4)
...ttggtatgtttaaatgtt...
...aaccatacaaatttacaa...
forward strand C-to-T converted genome
...ccaacatatttaaacact...
...ggttgtataaatttgtga...
forward strand G-to-A converted genome
(1)
(2)
(3)
(4)
read all four alignment outputs simultaneously to determine if the sequence can be mapped uniquely
determine unique best alignment

Krueger et al Bioinformatics 2011

calculate % of tags methylated per genomic position

methodology

convert DNA with bisulfite

methyl CGTTCG → CGTTUG

fragmentation

library construction and sequencing

methodology

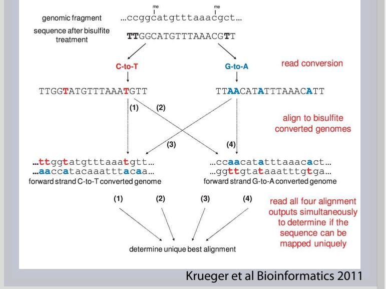
convert DNA with bisulfite

meth
CGTTCG
CGTTUG

fragmentation

library construction and sequencing

mapping

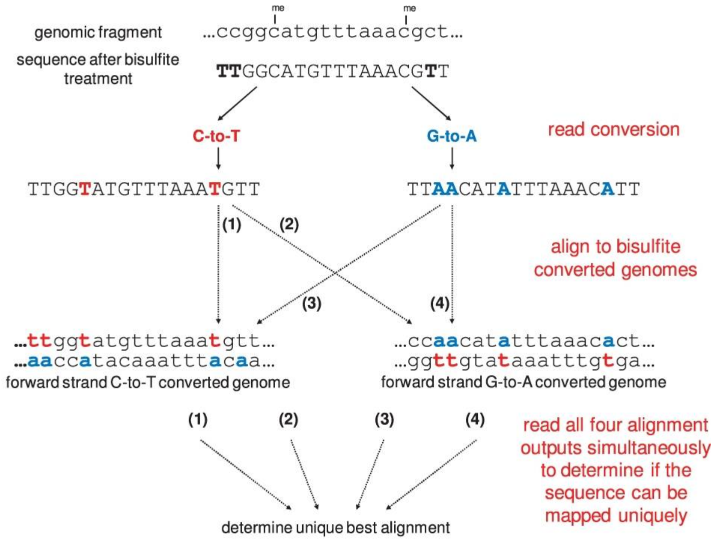


Krueger et al Bioinformatics 2011

calculate % of tags methylated per genomic position

methodology

mapping


me
me
genomic fragment
...ccggcatgtttaaacgct...
sequence after bisulfite treatment
TTGGCATGTTTAAACGTT
C-to-T
G-to-A
read conversion
TTGGTATGTTTAAATGTT
TTAACATATTTAAACATT
(1)
(2)
(3)
(4)
align to bisulfite converted genomes
...ttggtatgtttaaatgtt...
...aaccatacaaatttacaa...
forward strand C-to-T converted genome
...ccaacatatttaaacact...
...ggttgtataaatttgtga...
forward strand G-to-A converted genome
(1)
(2)
(3)
(4)
read all four alignment outputs simultaneously to determine if the sequence can be mapped uniquely
determine unique best alignment

Krueger et al Bioinformatics 2011

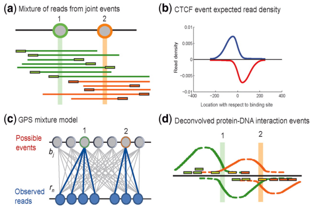
calculate % of tags methylated per genomic position

GEM (Guo et al, PLoS Comp. Bio 2012)

algorithm

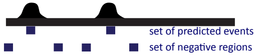
1. predict protein-DNA binding events with sparse prior
2. discover set of enriched kmers at binding event
3. cluster set of enriched k-mers into k-mer classes
4. generate positional prior for discovery in most enriched class
5. predict improved binding event probabilities with pos. prior
6. repeat (2) and (3) to generate improved motif enrichments

1. predict events with sparse prior


(a) Mixture of reads from joint events
1
2
(b) CTCF event expected read density
0.01
0.005
0
−0.005
−0.01
Read density
−400
−200
0
200
400
Location with respect to binding site
(c) GPS mixture model
Possible events
b_j
1
2
Observed reads
r_n
(d) Deconvolved protein-DNA interaction events
1
2

Guo et al, Bioinformatics 2010

2. discover set of enriched kmers



set of predicted events from (1)

set of negative regions (+/-300bp)

	positive region count	negative region count
kmer1	400	34
kmer2	25	42
kmer...	17	25

hypergeometric test

Barash et al WABI 2001

3. cluster kmer classes

K-mer	Offset	Pos Hit	Neg Hit
-----ATGCAAAT	-3	739	30
----TATGCAAA	-4	628	33
------TGCAAATG	-2	460	22
-----ATGCTAAT	-3	382	12
---TTATGCAA	-5	358	13
-----ATGCATAT	-3	320	21
------TGCAAATT	-2	222	18
... ...	...	...	...

Guo et al, PLoS Comp Bio 2012

4. generate positional prior for events in most enriched kmer class

$$p(\pi) \propto \prod_{m=1}^{M} (\pi_m)^{-\alpha_S+\alpha_m}$$

5. predict binding event probabilities

$$\hat{\pi}_m{}^{(i)} = \frac{\max(0, N_m - \alpha_S + \alpha_m)}{\sum_{m'=1}^{M} \max(0, N_{m'} - \alpha_S + \alpha_m)}, \; N_m = \sum_{n=1}^{N} \gamma(z_n = m)$$

6. Redo steps (2) and (3) to improve motif quality

http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002638

GEM (Guo et al, PLoS Comp. Bio 2012)

algorithm

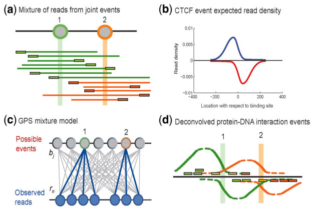
1. predict protein-DNA binding events with sparse prior
2. discover set of enriched kmers at binding event
3. cluster set of enriched k-mers into k-mer classes
4. generate positional prior for discovery in most enriched class
5. predict improved binding event probabilities with pos. prior
6. repeat (2) and (3) to generate improved motif enrichments

GEM (Guo et al, PLoS Comp. Bio 2012)

algorithm

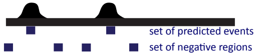
1. predict protein-DNA binding events with sparse prior
2. discover set of enriched kmers at binding event
3. cluster set of enriched k-mers into k-mer classes
4. generate positional prior for discovery in most enriched class
5. predict improved binding event probabilities with pos. prior
6. repeat (2) and (3) to generate improved motif enrichments

1. predict events with sparse prior


(a) Mixture of reads from joint events
(b) CTCF event expected read density
(c) GPS mixture model
(d) Deconvolved protein-DNA interaction events

Guo et al, Bioinformatics 2010

2. discover set of enriched kmers



set of predicted events from (1)

set of negative regions (+/-300bp)

	positive region count	negative region count
kmer1	400	34
kmer2	25	42
kmer...	17	25

hypergeometric test

Barash et al WABI 2001

3. cluster kmer classes

K-mer	Offset	Pos Hit	Neg Hit
-----ATGCAAAT	-3	739	30
----TATGCAAA	-4	628	33
------TGCAAATG	-2	460	22
-----ATGCTAAT	-3	382	12
---TTATGCAA	-5	358	13
-----ATGCATAT	-3	320	21
------TGCAAATT	-2	222	18
... ...	...	...	...

Guo et al, PLoS Comp Bio 2012

4. generate positional prior for events in most enriched kmer class

$$p(\pi) \propto \prod_{m=1}^{M} (\pi_m)^{-\alpha_S + \alpha_m}$$

5. predict binding event probabilities

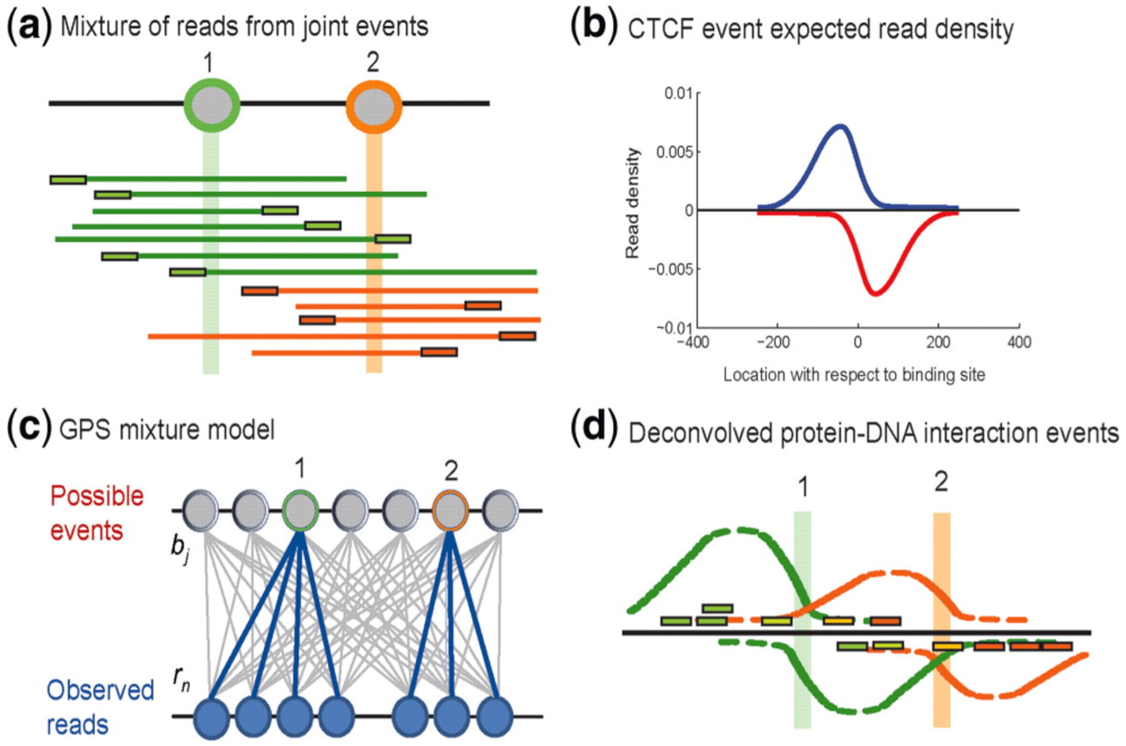
$$\hat{\pi}_m{}^{(i)} = \frac{\max(0, N_m - \alpha_S + \alpha_m)}{\sum_{m'=1}^{M} \max(0, N_{m'} - \alpha_S + \alpha_m)}, N_m = \sum_{n=1}^{N} \gamma(z_n = m)$$

6. Redo steps (2) and (3) to improve motif quality

http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002638

GEM (Guo et al, PLoS Comp. Bio 2012)

1. predict events with sparse prior

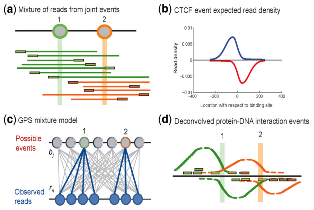

(a) Mixture of reads from joint events
1
2
(b) CTCF event expected read density
0.01
0.005
0
−0.005
−0.01
Read density
−400
−200
0
200
400
Location with respect to binding site
(c) GPS mixture model
1
2
Possible events
b_j
Observed reads
r_n
(d) Deconvolved protein-DNA interaction events
1
2

GEM (Guo et al, PLoS Comp. Bio 2012)

algorithm

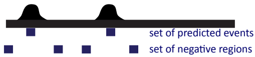
1. predict protein-DNA binding events with sparse prior
2. discover set of enriched kmers at binding event
3. cluster set of enriched k-mers into k-mer classes
4. generate positional prior for discovery in most enriched class
5. predict improved binding event probabilities with pos. prior
6. repeat (2) and (3) to generate improved motif enrichments

1. predict events with sparse prior


(a) Mixture of reads from joint events
1
2
(b) CTCF event expected read density
0.01
0.005
0
−0.005
−0.01
Read density
−400
−200
0
200
400
Location with respect to binding site
(c) GPS mixture model
Possible events
b_j
1
2
Observed reads
r_n
(d) Deconvolved protein-DNA interaction events
1
2

Guo et al, Bioinformatics 2010

2. discover set of enriched kmers



set of predicted events from (1)

set of negative regions (+/-300bp)

	positive region count	negative region count
kmer1	400	34
kmer2	25	42
kmer...	17	25

hypergeometric test

Barash et al WABI 2001

3. cluster kmer classes

K-mer	Offset	Pos Hit	Neg Hit
-----ATGCAAAT	-3	739	30
----TATGCAAA	-4	628	33
------TGCAAATG	-2	460	22
-----ATGCTAAT	-3	382	12
---TTATGCAA	-5	358	13
-----ATGCATAT	-3	320	21
------TGCAAATT	-2	222	18
... ...	...	...	...

Guo et al, PLoS Comp Bio 2012

4. generate positional prior for events in most enriched kmer class

$$p(\pi) \propto \prod_{m=1}^{M} (\pi_m)^{-\alpha_S + \alpha_m}$$

5. predict binding event probabilities

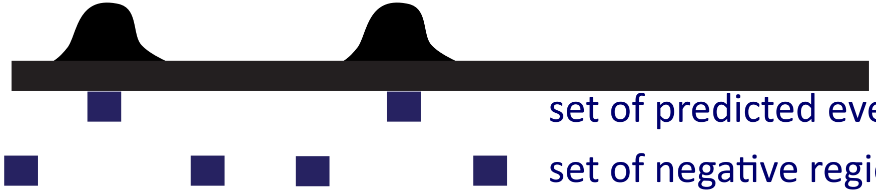
$$\hat{\pi}_m{}^{(i)} = \frac{\max(0, N_m - \alpha_S + \alpha_m)}{\sum_{m'=1}^{M} \max(0, N_{m'} - \alpha_S + \alpha_m)}, N_m = \sum_{n=1}^{N} \gamma(z_n = m)$$

6. Redo steps (2) and (3) to improve motif quality

http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002638

GEM (Guo et al, PLoS Comp. Bio 2012)

2. discover set of enriched kmers



set of predicted events from (1)

set of negative regions (+/-300bp)

	positive region count	negative region count
kmer1	400	34
kmer2	25	42
kmer...	17	25

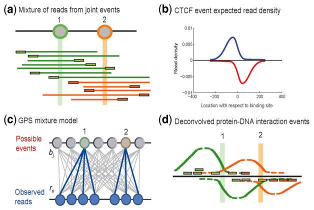
hypergeometric test

GEM (Guo et al, PLoS Comp. Bio 2012)

algorithm

1. predict protein-DNA binding events with sparse prior
2. discover set of enriched kmers at binding event
3. cluster set of enriched k-mers into k-mer classes
4. generate positional prior for discovery in most enriched class
5. predict improved binding event probabilities with pos. prior
6. repeat (2) and (3) to generate improved motif enrichments

1. predict events with sparse prior


(a) Mixture of reads from joint events
1
2
(b) CTCF event expected read density
0.01
0.005
0
−0.005
−0.01
Read density
−400
−200
0
200
400
Location with respect to binding site
(c) GPS mixture model
Possible events
b_j
1
2
Observed reads
r_n
(d) Deconvolved protein-DNA interaction events
1
2

Guo et al, Bioinformatics 2010

2. discover set of enriched kmers

set of predicted events from (1)

set of negative regions (+/-300bp)

	positive region count	negative region count
kmer1	400	34
kmer2	25	42
kmer...	17	25

hypergeometric test

Barash et al WABI 2001

3. cluster kmer classes

K-mer	Offset	Pos Hit	Neg Hit
-----ATGCAAAT	-3	739	30
----TATGCAAA	-4	628	33
------TGCAAATG	-2	460	22
-----ATGCTAAT	-3	382	12
---TTATGCAA	-5	358	13
-----ATGCATAT	-3	320	21
------TGCAAATT	-2	222	18
... ...	...	...	...

Guo et al, PLoS Comp Bio 2012

4. generate positional prior for events in most enriched kmer class

$$p(\pi) \propto \prod_{m=1}^{M} (\pi_m)^{-\alpha_S+\alpha_m}$$

5. predict binding event probabilities

$$\hat{\pi}_m{}^{(i)} = \frac{\max(0, N_m - \alpha_S + \alpha_m)}{\sum_{m'=1}^{M} \max(0, N_{m'} - \alpha_S + \alpha_m)}, \; N_m = \sum_{n=1}^{N} \gamma(z_n = m)$$

6. Redo steps (2) and (3) to improve motif quality

http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002638

GEM (Guo et al, PLoS Comp. Bio 2012)

3. cluster kmer classes

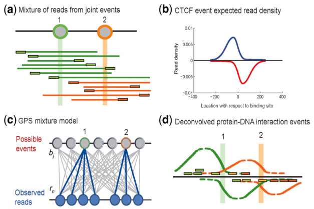
K-mer	Offset	Pos Hit	Neg Hit
-----ATGCAAAT	-3	739	30
----TATGCAAA	-4	628	33
------TGCAAATG	-2	460	22
-----ATGCTAAT	-3	382	12
---TTATGCAA	-5	358	13
-----ATGCATAT	-3	320	21
------TGCAAATT	-2	222	18
... ...	...	...	...

GEM (Guo et al, PLoS Comp. Bio 2012)

algorithm

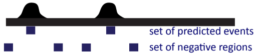
1. predict protein-DNA binding events with sparse prior
2. discover set of enriched kmers at binding event
3. cluster set of enriched k-mers into k-mer classes
4. generate positional prior for discovery in most enriched class
5. predict improved binding event probabilities with pos. prior
6. repeat (2) and (3) to generate improved motif enrichments

1. predict events with sparse prior


(a) Mixture of reads from joint events
(b) CTCF event expected read density
(c) GPS mixture model
(d) Deconvolved protein-DNA interaction events

Guo et al, Bioinformatics 2010

2. discover set of enriched kmers

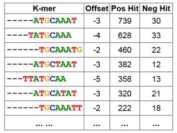

set of predicted events from (1)
set of negative regions (+/-300bp)

	positive region count	negative region count
kmer1	400	34
kmer2	25	42
kmer...	17	25

hypergeometric test

Barash et al WABI 2001

3. cluster kmer classes



K-mer	Offset	Pos Hit	Neg Hit
-----ATGCAAAT	-3	739	30
----TATGCAAA	-4	628	33
------TGCAAATG	-2	460	22
-----ATGCTAAT	-3	382	12
---TTATGCAA	-5	358	13
-----ATGCATAT	-3	320	21
------TGCAAATT	-2	222	18
... ...	...	...	...

Guo et al, PLoS Comp Bio 2012

4. generate positional prior for events in most enriched kmer class

$$p(\pi) \propto \prod_{m=1}^{M} (\pi_m)^{-\alpha_S + \alpha_m}$$

5. predict binding event probabilities

$$\hat{\pi}_m{}^{(i)} = \frac{\max(0, N_m - \alpha_S + \alpha_m)}{\sum_{m'=1}^{M} \max(0, N_{m'} - \alpha_S + \alpha_m)}, N_m = \sum_{n=1}^{N} \gamma(z_n = m)$$

6. Redo steps (2) and (3) to improve motif quality

http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002638

GEM (Guo et al, PLoS Comp. Bio 2012)

4. generate positional prior for events in most enriched kmer class

$$p(\pi) \propto \prod_{m=1}^{M} (\pi_m)^{-\alpha_S + \alpha_m}$$

5. predict binding event probabilities

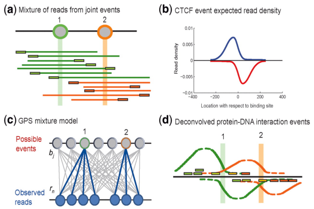
$$\hat{\pi}_m{}^{(i)} = \frac{\max(0, N_m - \alpha_S + \alpha_m)}{\sum_{m'=1}^{M} \max(0, N_{m'} - \alpha_S + \alpha_m)}, N_m = \sum_{n=1}^{N} \gamma(z_n = m)$$

GEM (Guo et al, PLoS Comp. Bio 2012)

algorithm

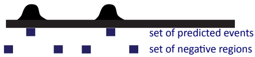
1. predict protein-DNA binding events with sparse prior
2. discover set of enriched kmers at binding event
3. cluster set of enriched k-mers into k-mer classes
4. generate positional prior for discovery in most enriched class
5. predict improved binding event probabilities with pos. prior
6. repeat (2) and (3) to generate improved motif enrichments

1. predict events with sparse prior


(a) Mixture of reads from joint events
1
2
(b) CTCF event expected read density
0.01
0.005
0
−0.005
−0.01
Read density
−400
−200
0
200
400
Location with respect to binding site
(c) GPS mixture model
Possible events
b_j
1
2
Observed reads
r_n
(d) Deconvolved protein-DNA interaction events
1
2

Guo et al, Bioinformatics 2010

2. discover set of enriched kmers



set of predicted events from (1)

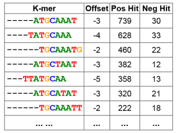
set of negative regions (+/-300bp)

	positive region count	negative region count
kmer1	400	34
kmer2	25	42
kmer...	17	25

hypergeometric test

Barash et al WABI 2001

3. cluster kmer classes



K-mer	Offset	Pos Hit	Neg Hit
-----ATGCAAAT	-3	739	30
----TATGCAAA	-4	628	33
------TGCAAATG	-2	460	22
-----ATGCTAAT	-3	382	12
---TTATGCAA	-5	358	13
-----ATGCATAT	-3	320	21
------TGCAAATT	-2	222	18
... ...	...	...	...

Guo et al, PLoS Comp Bio 2012

4. generate positional prior for events in most enriched kmer class

$$p(\pi) \propto \prod_{m=1}^{M} (\pi_m)^{-\alpha_S + \alpha_m}$$

5. predict binding event probabilities

$$\hat{\pi}_m{}^{(i)} = \frac{\max(0, N_m - \alpha_S + \alpha_m)}{\sum_{m'=1}^{M} \max(0, N_{m'} - \alpha_S + \alpha_m)}, \; N_m = \sum_{n=1}^{N} \gamma(z_n = m)$$

6. Redo steps (2) and (3) to improve motif quality

http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002638

GEM (Guo et al, PLoS Comp. Bio 2012)

6. Redo steps (2) and (3) to improve motif quality