Analiza danych z wysokoprzepustowego sekwencjonowania

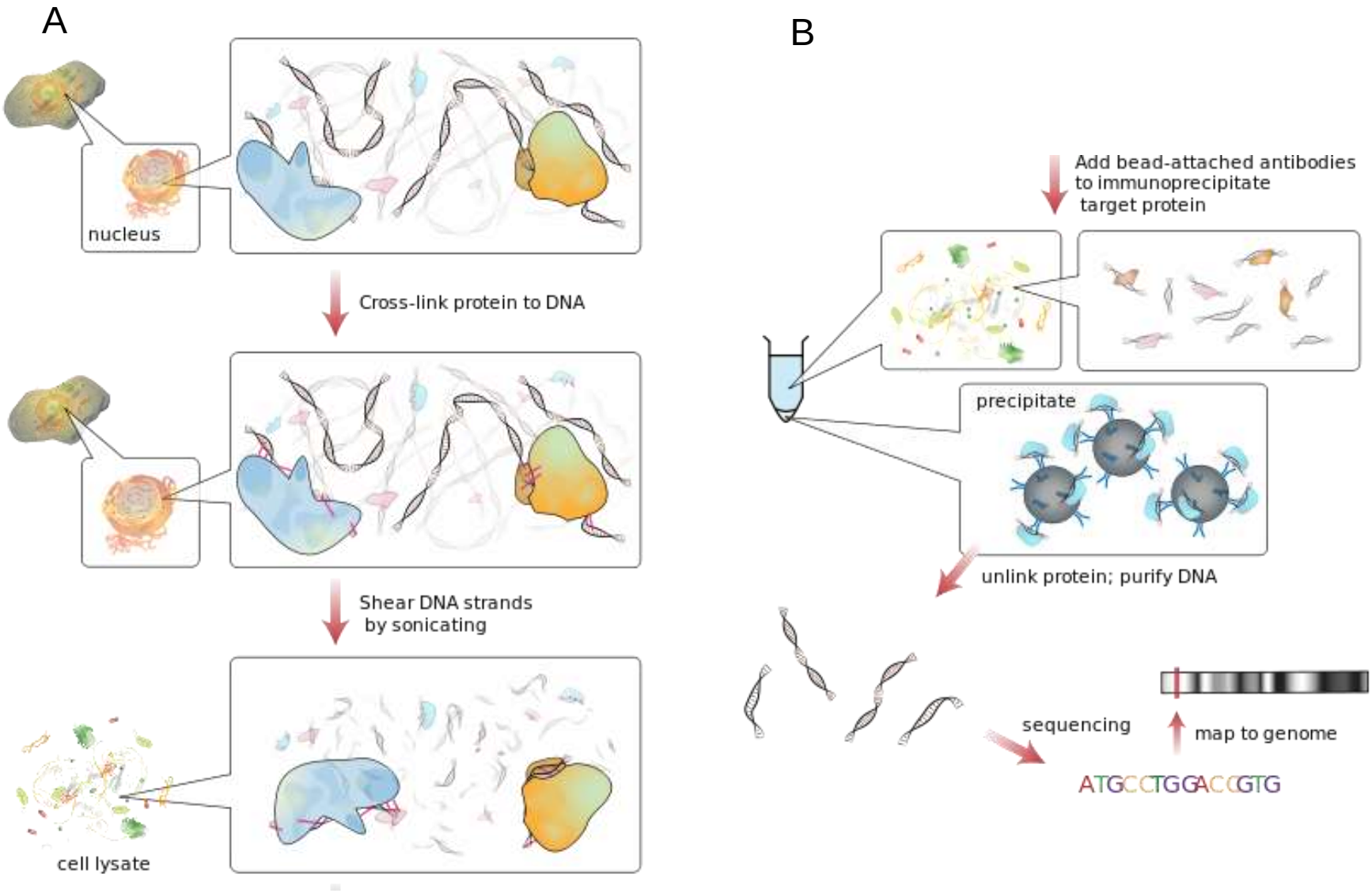
W3: Metody oparte o sekwencjonowanie

Chip-Seq

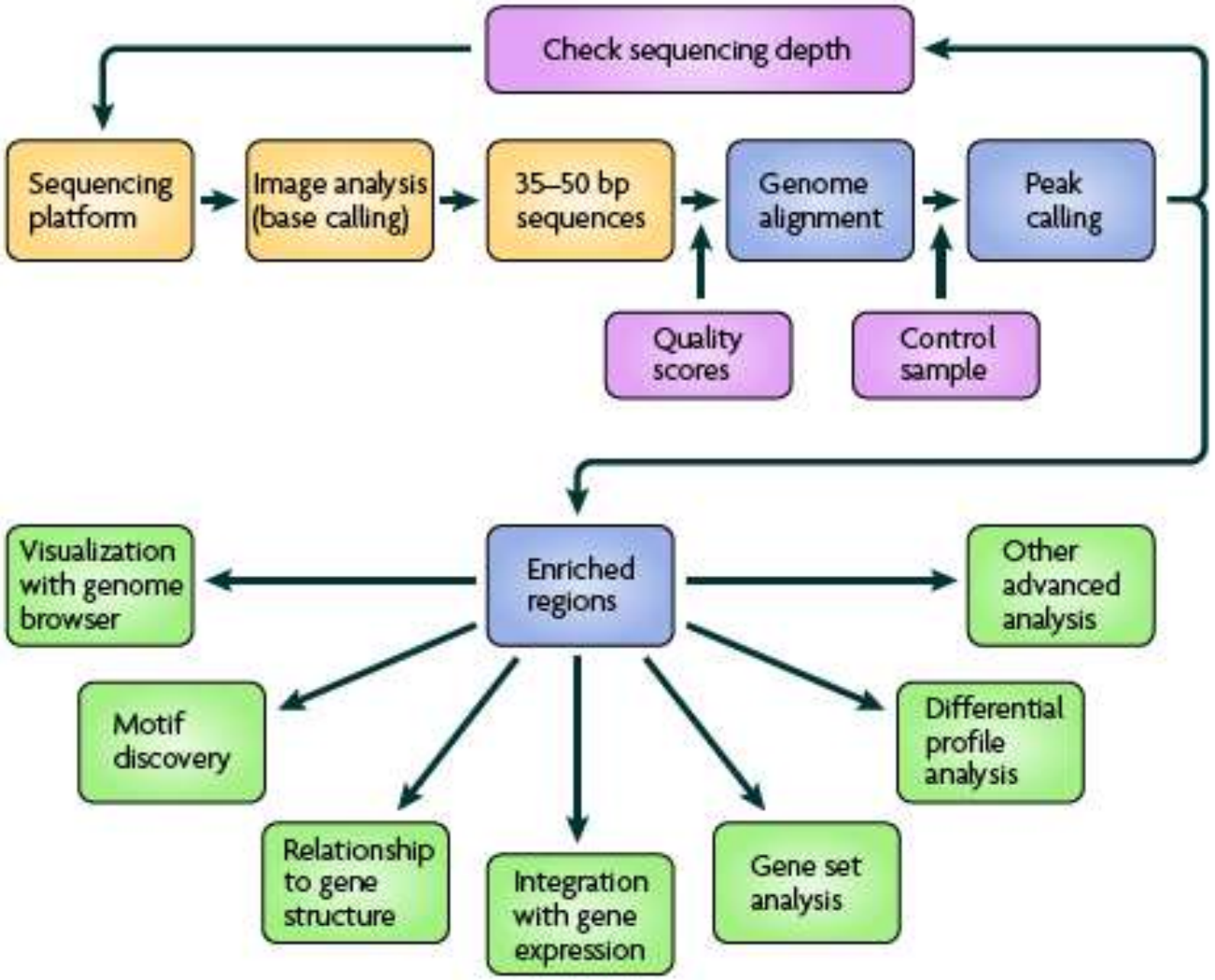
Chromatin immunoprecipitation and sequencing

- Identyfikacja miejsc wiązania białek do DNA:
    - czynniki transkrypcyjne
    - białka regulatorowe
- Identyfikacja modyfikacji histonów:
    - identyfikacja miejsc aktywnych transkrypcyjnie
- ...

Chip-Seq

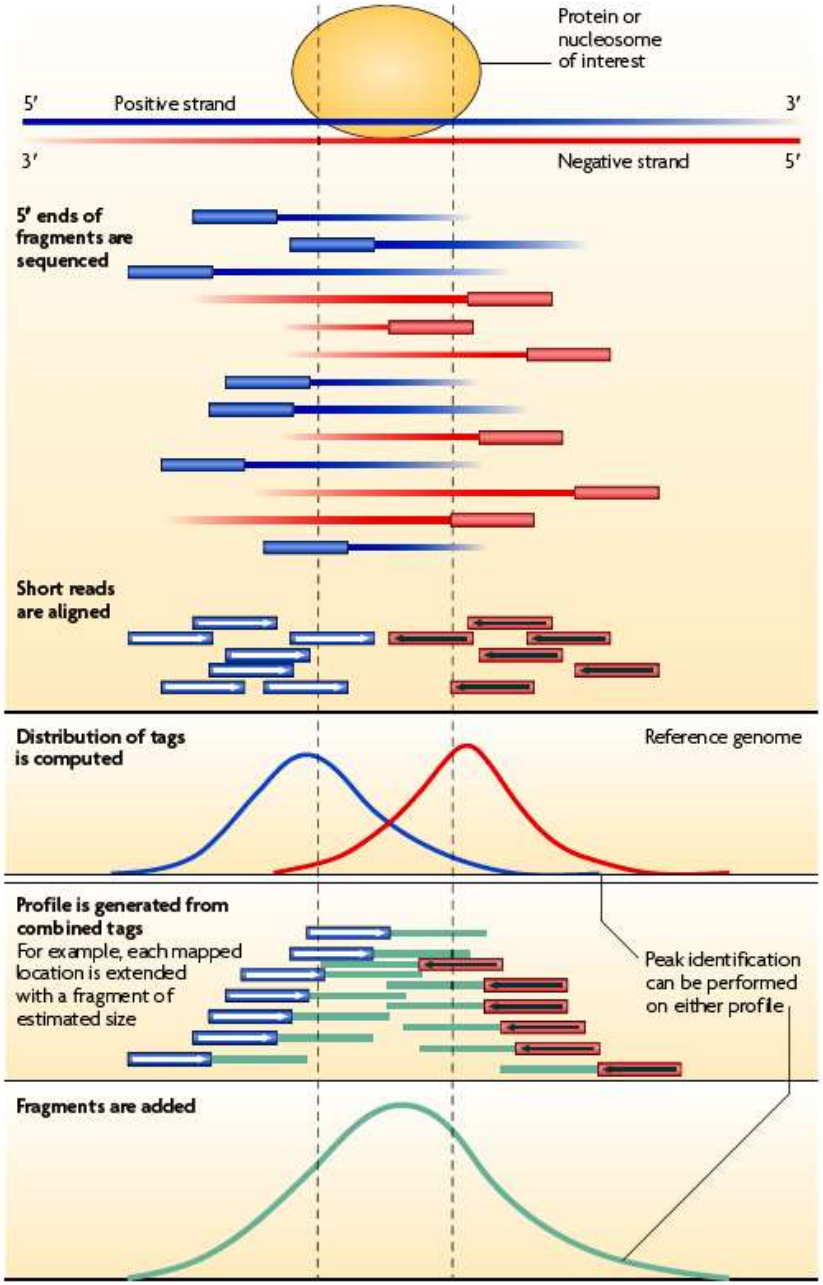

A
nucleus
Cross-link protein to DNA
Shear DNA strands
by sonicating
cell lysate
B
Add bead-attached antibodies
to immunoprecipitate
target protein
precipitate
unlink protein; purify DNA
sequencing
map to genome
ATGCCTGGACCGTG

Chip-Seq



Nat Rev Genet. 2009 Oct;10(10):669-80.

Chip-Seq

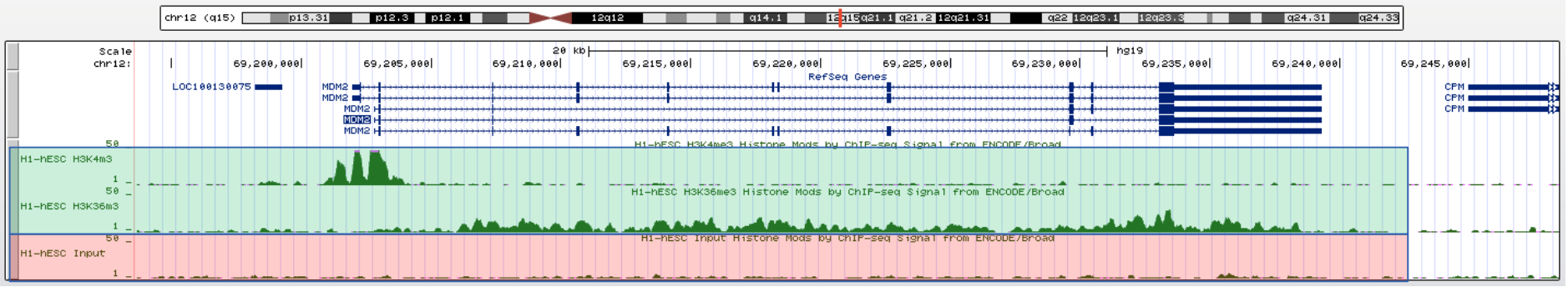


Nat Rev Genet. 2009 Oct;10(10):669-80.

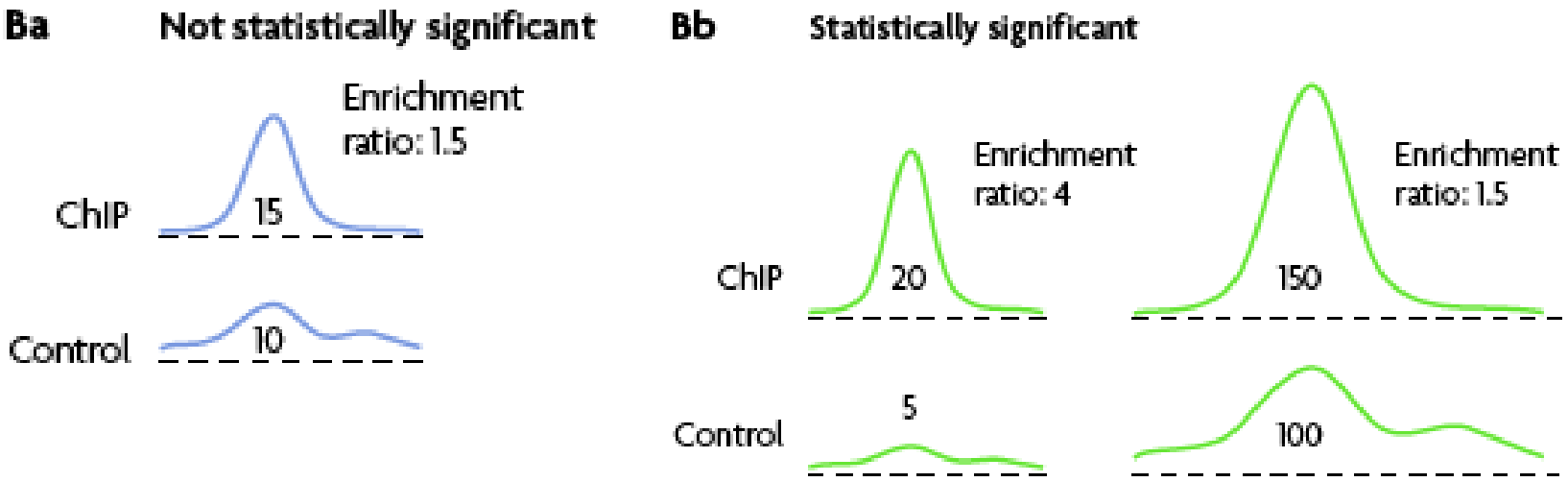
Chip-Seq

Identyfikacja rejonów wiązania:

- Poprzez porównanie sygnału uzyskanego dla danych z próbki z przeciwciałem oraz dla kontroli bez przeciwciała
- Szukamy rejonów o statystycznie istotnym wzbogaceniu

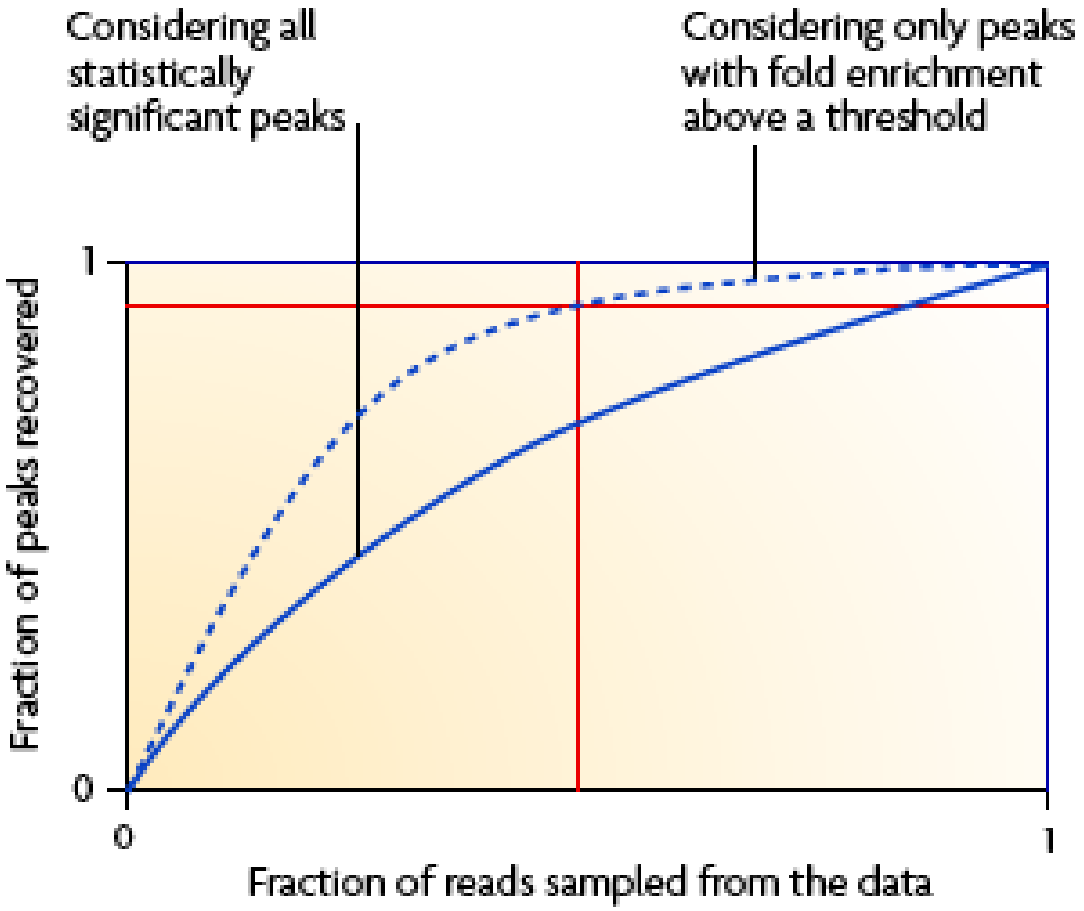


Chip-Seq



Nat Rev Genet. 2009 Oct;10(10):669-80.

Chip-Seq



Nat Rev Genet. 2009 Oct;10(10):669-80.

Chip-Seq

MACS: Model-based Analysis for ChIP-seq

Dwa etapy:

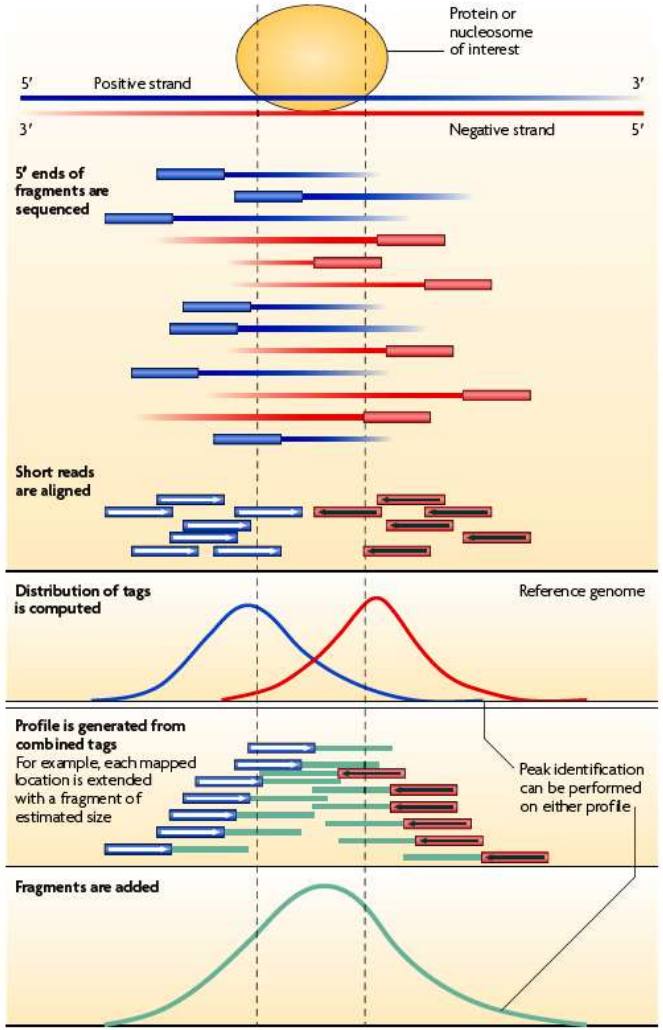
1. Modelowanie przesunięcia sygnału

2. Detekcja wierzchołków

Chip-Seq

Modelowanie przesunięcia sygnału:

Ponieważ prawdopodobieństwo zsekwencjonowania znaczników z nici plus i minus jest takie samo, rzeczywiste miejsca wiązania wykazują wzbogacenie dwumianowe, o charakterystycznym rozkładzie: na nici plus przed miejscem wiązania oraz na nici minus za miejscem wiązania



Nat Rev Genet. 2009 Oct;10(10):669-80.

Chip-Seq

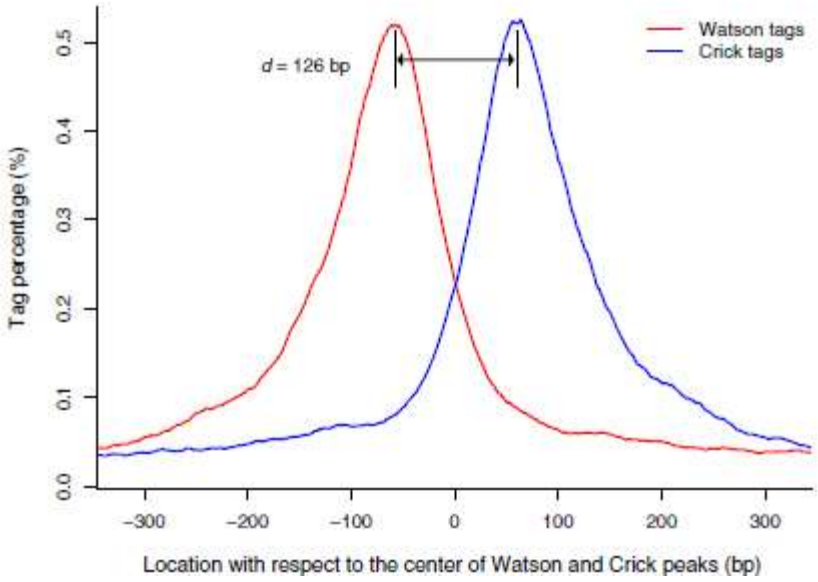
Modelowanie przesunięcia sygnału:

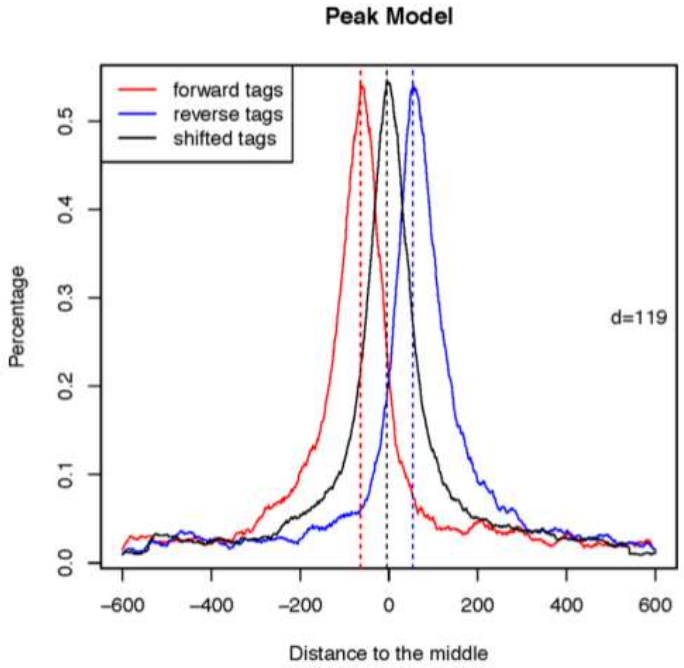
Informacje wejściowe:

- wielkość fragmentów DNA użytych do konstrukcji biblioteki
- minimalne wzbogacenie (krotność sygnału w porównaniu z tłem - mfold)

MACS wybiera losowe 1000 rejonów o wysokim wzbogaceniu i na ich podstawie oznacza przesunięcie sygnału:

- d – odległość pomiędzy sygnałami
- d/2 – wartość przesunięcia sygnału





Nat Rev Genet. 2009 Oct;10(10):669-80.

Chip-Seq

Detekcja miejsc wiązania:

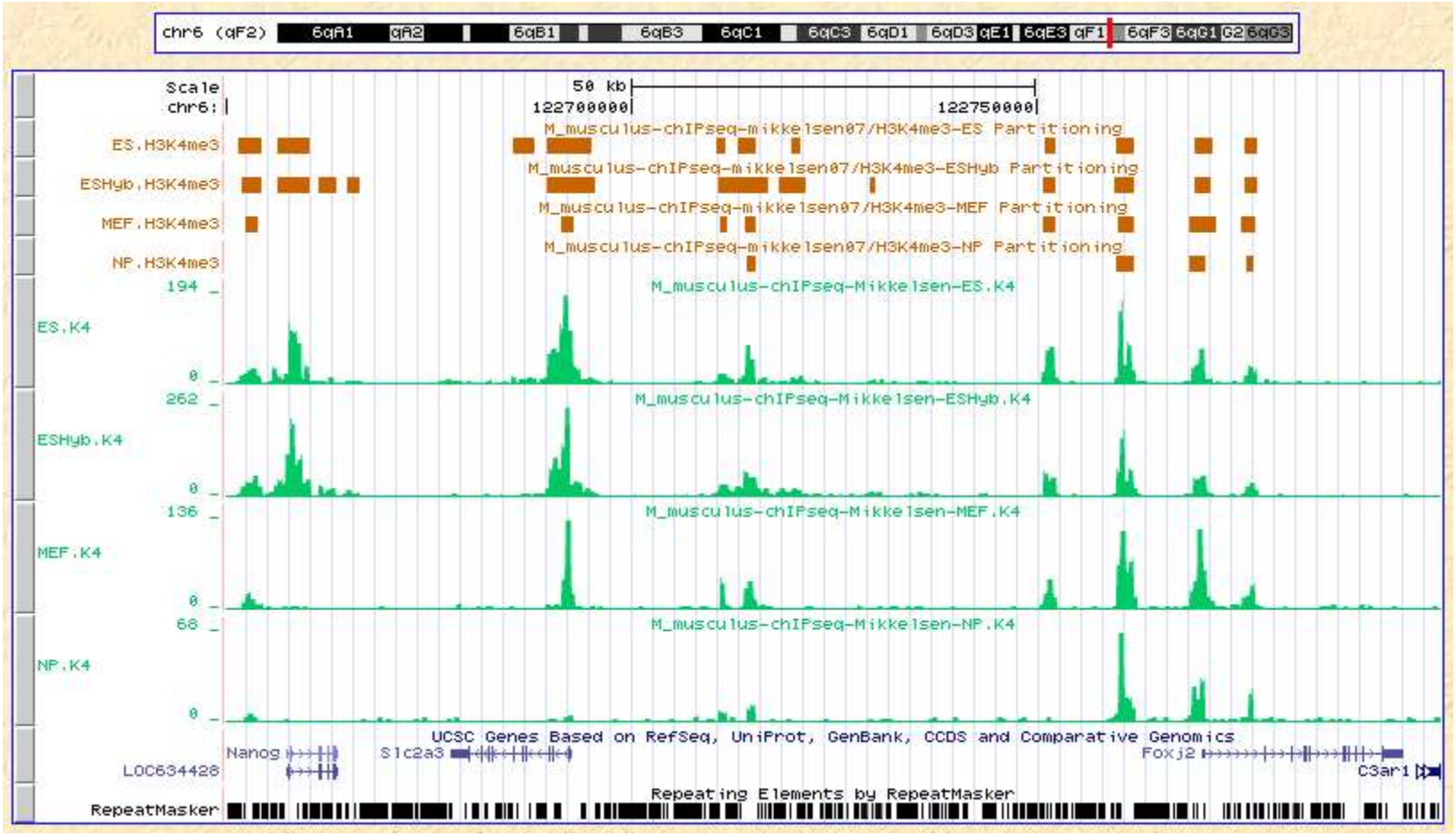
Odczyty zduplikowane są usuwane

Dystrybucja odczytów na genomie może być modelowana za pomocą rozkładu Poisson

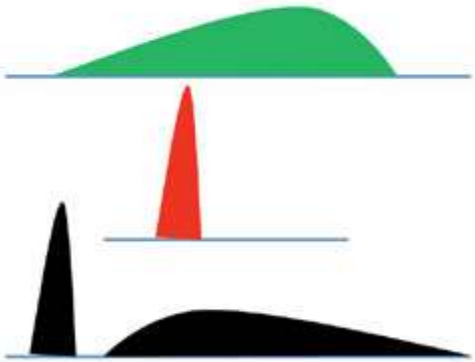
Po przesunięciu sygnału o d/2 MACS skanuje genom z użyciem okna o wielkości 2d i wyszukuje rejony o statystycznie znaczącym wzbogaceniu względem kontroli (wartość $p > 0.05$)

Nat Rev Genet. 2009 Oct;10(10):669-80.

Chip-Seq


chr6 (qF2)
6qA1 6qA2 6qB1 6qB3 6qC1 6qC3 6qD1 6qD3 6qE1 6qE3 6qF1 6qF3 6qG1 6qG2 6qG3
Scale
chr6:
50 kb
122700000
122750000
M_musculus-chIPseq-mikkelsen07/H3K4me3-ES Partitioning
M_musculus-chIPseq-mikkelsen07/H3K4me3-ESHyb Partitioning
M_musculus-chIPseq-mikkelsen07/H3K4me3-MEF Partitioning
M_musculus-chIPseq-mikkelsen07/H3K4me3-NP Partitioning
ES.H3K4me3
ESHyb.H3K4me3
MEF.H3K4me3
NP.H3K4me3
M_musculus-chIPseq-Mikkelsen-ES.K4
194
ES.K4
0
M_musculus-chIPseq-Mikkelsen-ESHyb.K4
262
ESHyb.K4
0
M_musculus-chIPseq-Mikkelsen-MEF.K4
136
MEF.K4
0
M_musculus-chIPseq-Mikkelsen-NP.K4
68
NP.K4
0
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Nanog
Slc2a3
Foxj2
LOC634428
C3ar1
Repeating Elements by RepeatMasker
RepeatMasker

Chip-Seq



Peak calling (choose the right tool)

Type of peak	Example	Representative tools
Broad	H3K27me3	CCAT, SICER
Sharp	CTCF	MACS
Sharp & broad	Pol II	ZINBA

Kidder BL et al (2011) Nat. Immun 12:918

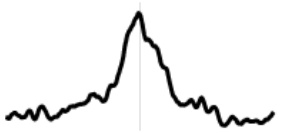
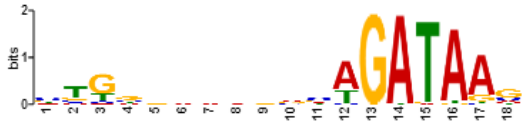
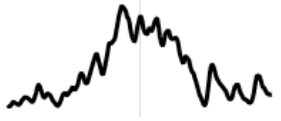
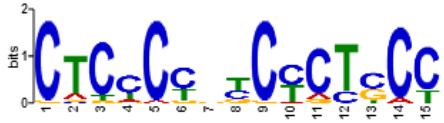
Chip-Seq

chr15:83139085-83139185	-	23	8.93e-06	TGGAGGCCAG AGGGTGTGGTG GATAAAACCC
chr15:91082416-91082516	-	46	8.93e-06	CTCTGCTGGG AGGGTGTGGTG CCTTTTATCT
chr7:139915433-139915533	+	48	8.93e-06	GAGCGCTGTC AGGGCGGGGCT GTTTTGTAAC
chr16:49777000-49777100	-	89	8.93e-06	G AGGGTGTGGTG GCAACTATTA
chr12:70394726-70394826	-	55	8.93e-06	CAGATAGGAG TGGGTGGGGCA GAGGAAAGTT
chr7:103853792-103853892	+	88	8.93e-06	AATGACAGTA TGGGCGTGGCA AG
chr8:109927244-109927344	+	55	9.76e-06	CGCTAGGGGA AGGGTGGGGCA GCAAAACGGA
chr8:107823858-107823958	+	31	9.76e-06	TGATAAGTTC AGGGTGGGGCA GGTAGGGGAG
chr1:135693776-135693876	-	5	9.76e-06	TGAAGAGGAT AGGGTGGGGCA GGTA
chr1:169322898-169322998	-	60	9.76e-06	CCAGAACCAC AGGGTGGGGCA CCTGCAGGGT
chr5:65203207-65203307	+	42	1.02e-05	TCCGAAGTAC TGGGTGTGGGC AGAATCTTAT
chr3:79383670-79383770	-	50	1.02e-05	CACATTTATT TGGGTGTGGGC CTTCAGCACG
chr11:68709695-68709795	-	50	1.02e-05	CAGTGTCTGT TGGGTGTGGAC CTAGGTCTAC
chr5:96888622-96888722	+	28	1.17e-05	TGTGATGGCA TGGGTGTGGAT GGAGGAGGGA
chr11:104842983-104843083	+	83	1.17e-05	NNNTGCTCTC AGGGTGTGGAC ATTTAGT
chr16:14166506-14166606	+	28	1.17e-05	AGAGGGTAAA AGGGTGTGGGC CTGNNNNNNN
chr17:24289205-24289305	-	69	1.17e-05	GCCTGTGGTA TGGGTGTGGAT TGGGGGACAG
chr7:149571622-149571722	-	38	1.41e-05	CCACCTGACT GGGGTGTGGTT GGTGAGGGAA
chr19:6236070-6236170	+	42	1.41e-05	CCTGCTGGGA GGGGTGTGGTT CACTAGAGGC
chr14:56154883-56154983	-	41	1.41e-05	ACCAGGGGGC AGGGTGTGACC TGCAATGTCC
chr7:99891353-99891453	-	21	1.41e-05	ATTCCTAAAG GGGGTGTGGTT GACCCAAAGT
chr11:53839888-53839988	-	60	1.41e-05	GTAGACAGCT AGGGTGTGGGT CCCACCTGCA
chr7:148354574-148354674	+	44	1.41e-05	TGCCACGAAA TGGGTGTGACT GTGATGAGAG

Chip-Seq

MOTIFS

The motifs found by the programs MEME, DREME and CentriMo; clustered by similarity and ordered by E-value.

Motif Found	Discovery/Enrichment Program	E-value	Known or Similar Motifs	Distribution
	MEME	8.1e-229	Klf4 (MA0039.2) AFT2 (MA0270.1) SP1 (MA0079.1)	
Reverse Complement ⇆ Show 10 More Ⅰ CentriMo Group ↷				
	CentriMo	1.3e-045	Tal1::Gata1 (MA0140.1)	
Reverse Complement ⇆ Show 12 More Ⅰ CentriMo Group ↷				
	DREME	8.2e-003	TAL1::TCF3 (MA0091.1)	
Reverse Complement ⇆				
	MEME	1.7e-001	SP1 (MA0079.2) Pax4 (MA0068.1) SP1 (MA0079.1)	
Reverse Complement ⇆				

Chip-Seq

Zalety:

- Identyfikacja rejonów wiązania białek do DNA w skali genomowej
- Możliwość uzyskania informacji o motywie sekwencyjnym odpowiedzialnym za wiązanie białka

Wady:

- Czułość metody zależy od siły wiązania białka bądź efektywności cross-link
- Rozdzielczość zależy od analizowanego białka (wysoka dla czynników transkrypcyjnych, niska dla histonów

Chip-Seq

Na co zwracać uwagę:

- Spodziewany kształt pików zależny jest od charakterystyki wiązania DNA przez badane białko
- Czy mamy dostępną kontrolę (eksperyment bez przeciwciała)

Dobór programu do rodzaju pików:

Chip-Seq

Software tool	Version	Availability	Point- source (peaks)	Broad regions (domains)
BayesPeak [88]	1.10.0	http://bioconductor.org/packages/release/bioc/html/BayesPeak.html	Yes	
BEADS§ [84]	1.1	http://beads.sourceforge.net/	Yes	Yes
CCAT [91]	3.0	http://cmb.gis.a-star.edu.sg/ChIPSeq/paperCCAT.htm		Yes
CisGenome [56]	2.0	http://www.biostat.jhsph.edu/~hji/cisgenome/	Yes	
CSAR [85]	1.10.0	http://bioconductor.org/packages/release/bioc/html/CSAR.html	Yes	
dPeak	0.9.9	http://www.stat.wisc.edu/~chungdon/dpeak/	Yes	
GPS/GEM [67,18]	1.3	http://cgs.csail.mit.edu/gps/	Yes	
HPeak [87]	2.1	http://www.sph.umich.edu/csg/qin/HPeak/	Yes	
MACS [17]	**2.0.10**	**https://github.com/taoliu/MACS/**	**Yes**	**Yes**
NarrowPeaks§	1.4.0	http://bioconductor.org/packages/release/bioc/html/NarrowPeaks.html	Yes	
PeakAnalyzer/ PeakSplitter§ [89]	1.4	http://www.bioinformatics.org/peakanalyzer	Yes	
PeakRanger [93]	1.16	http://ranger.sourceforge.net/	Yes	Yes
PeakSeq [24]	1.1	http://info.gersteinlab.org/PeakSeq	Yes	
polyaPeak§	0.1	http://web1.sph.emory.edu/users/hwu30/polyaPeak.html	Yes	
RSEG [92]	0.6	http://smithlab.usc.edu/histone/rseg/		Yes
SICER [90]	1.1	http://home.gwu.edu/~wpeng/Software.htm		Yes
SIPeS [21]	2.0	http://gmdd.shgmo.org/Computational-Biology/ChIP-Seq/download/SIPeS	Yes	
SISSRs [19]	1.4	http://sissrs.rajajothi.com/	Yes	
SPP [9]	1.1	http://compbio.med.harvard.edu/Supplements/ChIP-seq/	Yes	Yes
USeq [97]	8.5.1	http://sourceforge.net/projects/useq/	Yes	
ZINBA [86]	2.02.03	http://code.google.com/p/zinba/	Yes	Yes

Chip-Seq – metody normalizacji

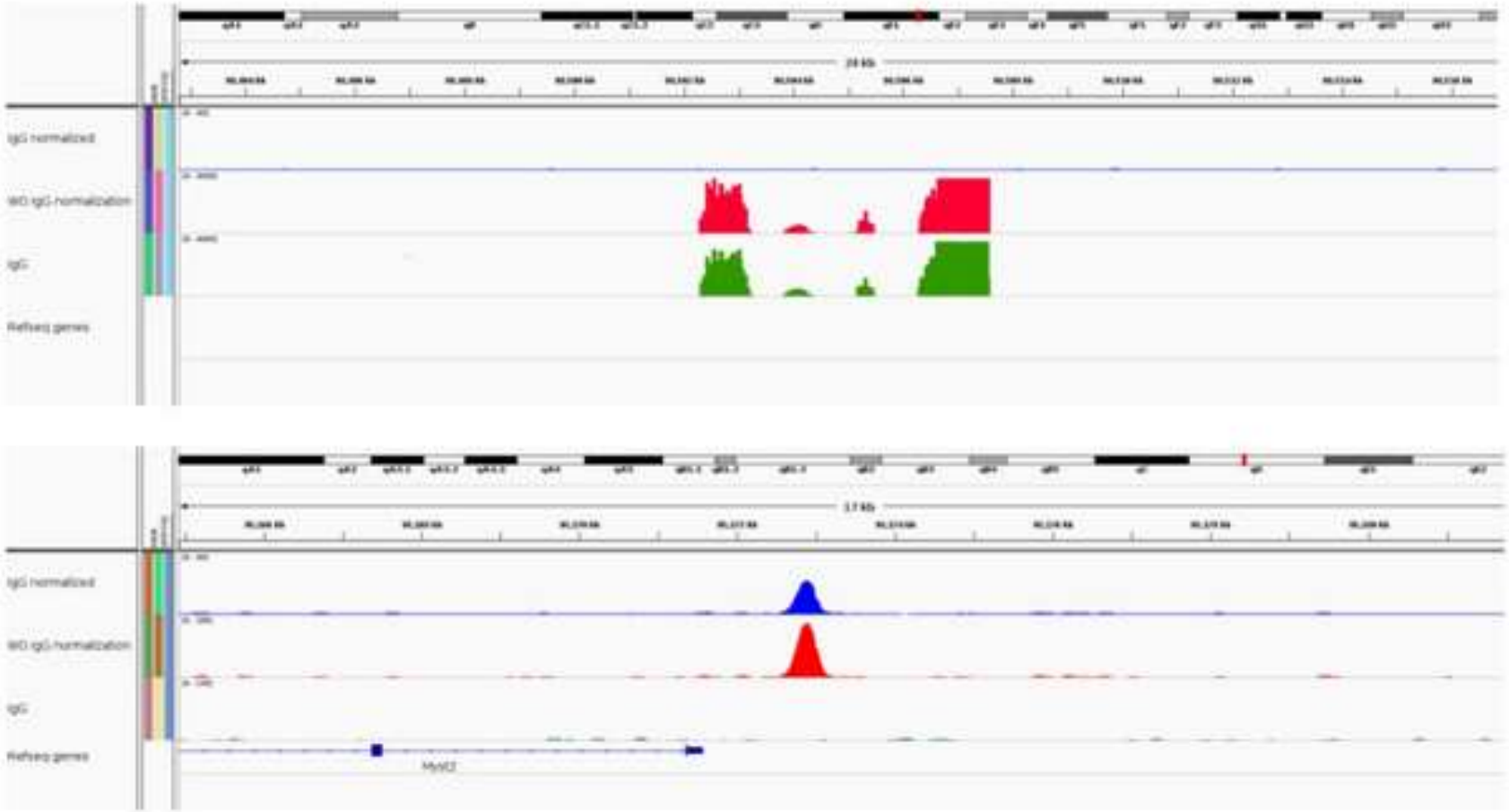
Linear Algorithms	Focus on negative control sample	Publication Time
PeakSeq [24]	Yes	Jan 2009
Cisgenome [56]	Yes	Nov 2008
MACS [17]	Yes	Sept 2008
USeq [97]	Yes	Dec 2008
RPKM [98]	No	May 2008

Non-linear Algorithms	Focus on negative control sample	Publication Time
Locally weighted regression with respect to mean and variance [28]	No	Jan 2009
MAnorm [34]	No	March 2012
POLYPHEMUS [37] (Quantile and locally weighted regression)	No	Feb 2012

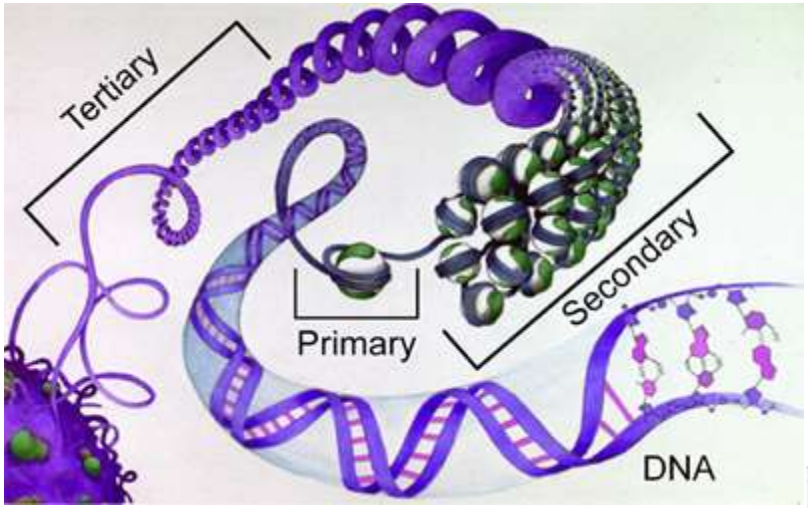
Chip-Seq – różnicowe wiązanie

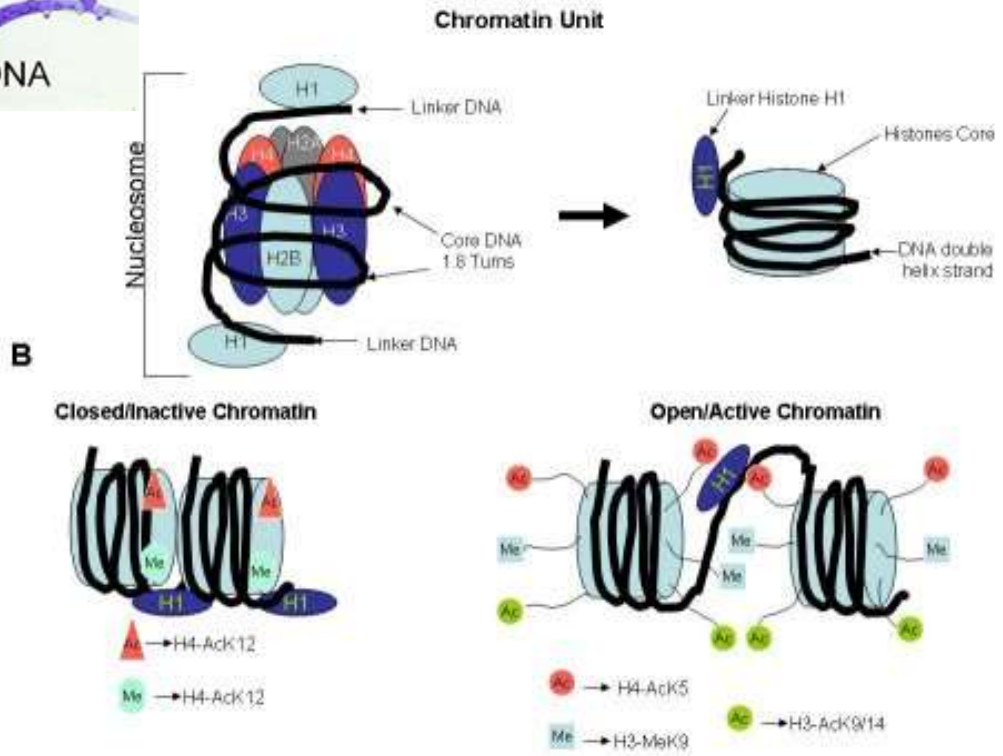
Software tool	Availability	Notes
ChIPDiff [36]	http://cmb.gis.a-star.edu.sg/ChIPSeq/paperChIPDiff.htm	Differential histone modification sites using a hidden Markov model
Comparative ChIP-seq [25]	http://www.starklab.org/data/bardet_natprotoc_2011/	Fold change ratio between normalized peak heights
DBChIP [33]	http://pages.cs.wisc.edu/~kliang/DBChIP/	Assigns uncertainty measures in a test of non-differential binding (uses edgeR)
DESeq§ [31]	http://www.bioconductor.org/packages/release/bioc/html/DESeq.html	Test based on a model using the negative binomial distribution
DiffBind	http://www.bioconductor.org/packages/release/bioc/html/DiffBind.html	Differential binding affinity analysis (uses edgeR and DESeq)
DIME [35]	http://cran.r-project.org/web/packages/DIME/	Differential identification using mixtures ensemble
edgeR§ [32]	http://www.bioconductor.org/packages/release/bioc/html/edgeR.html	Empirical Bayes estimation and exact tests based on the negative binomial distribution
MACS [17] (version 2)	https://github.com/taoliu/MACS/	Differential peak detection based on paired four bedGraph files
MAnorm [34]	http://bcb.dfci.harvard.edu/~gcyuan/MAnorm/MAnorm.htm	Robust regression to derive a linear model
MMDiff	http://bioconductor.org/packages/release/bioc/html/MMDiff.html	Differences in shape using Kernel methods
NarrowPeaks	http://bioconductor.org/packages/release/bioc/html/NarrowPeaks.html	Shape-based analysis of variation using functional PCA
POLYPHEMUS [37]	http://cran.r-project.org/web/packages/polyphemus/	Non-linear normalization on RNA Pol II profiling

Chip-Seq – wpływ kontroli

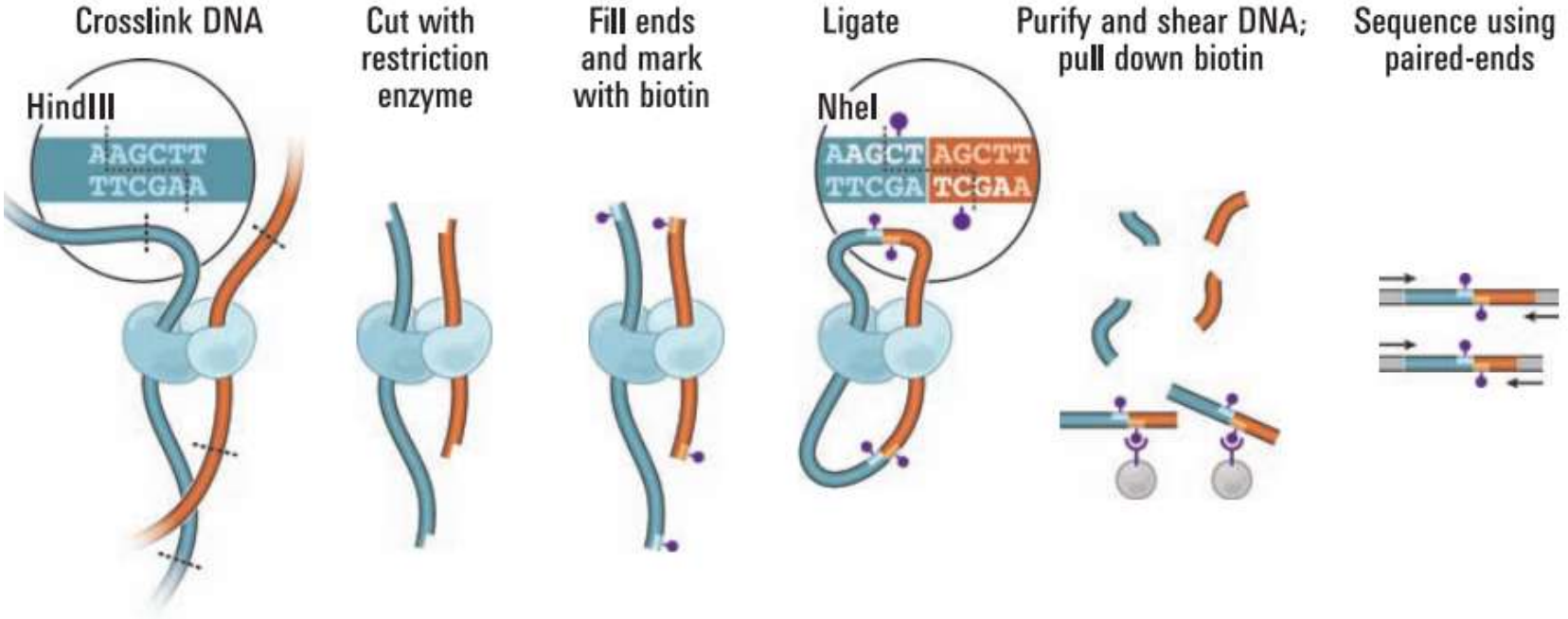


Struktura DNA



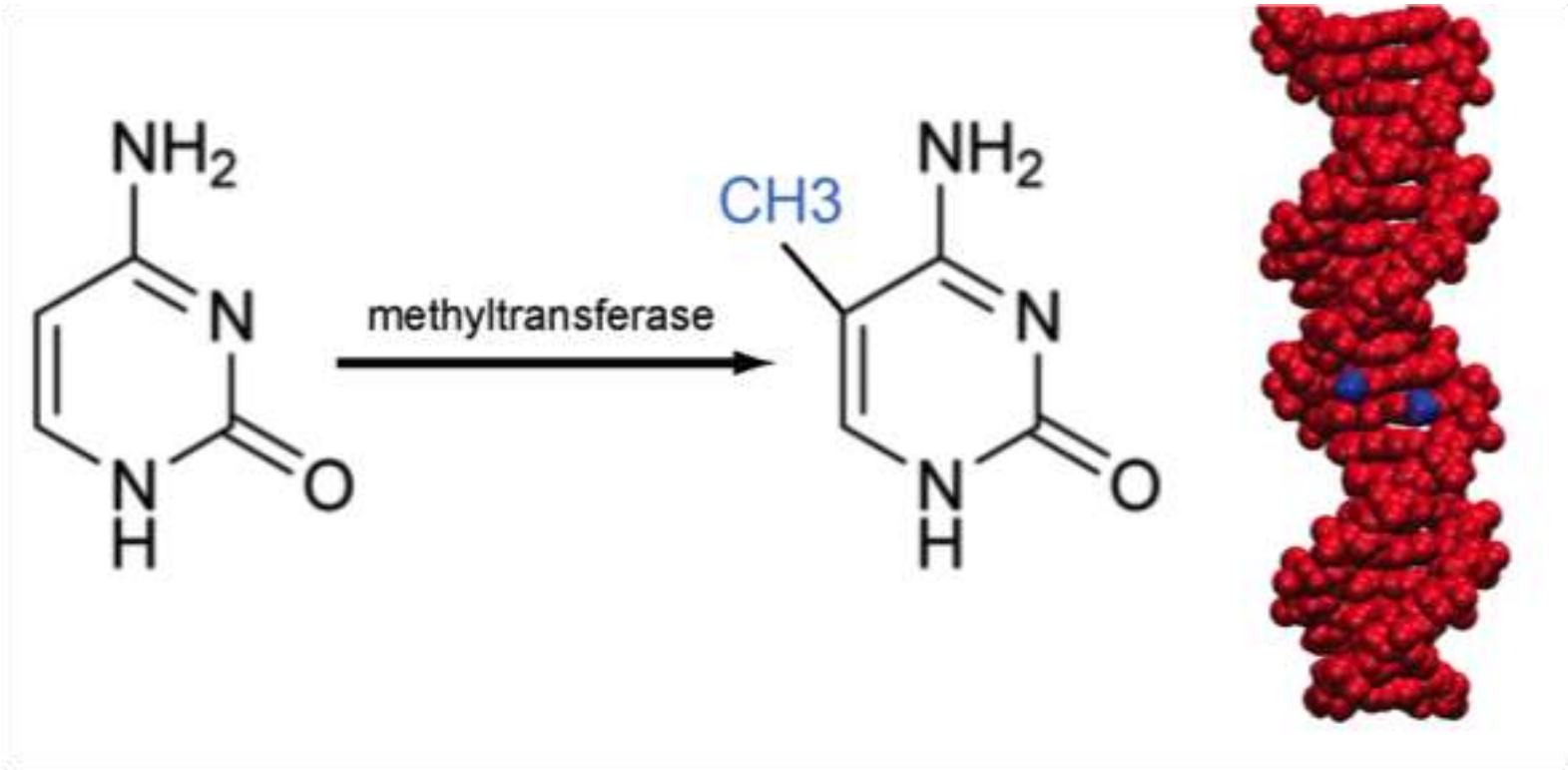


Chromatin interactions mapping (Hi-C)

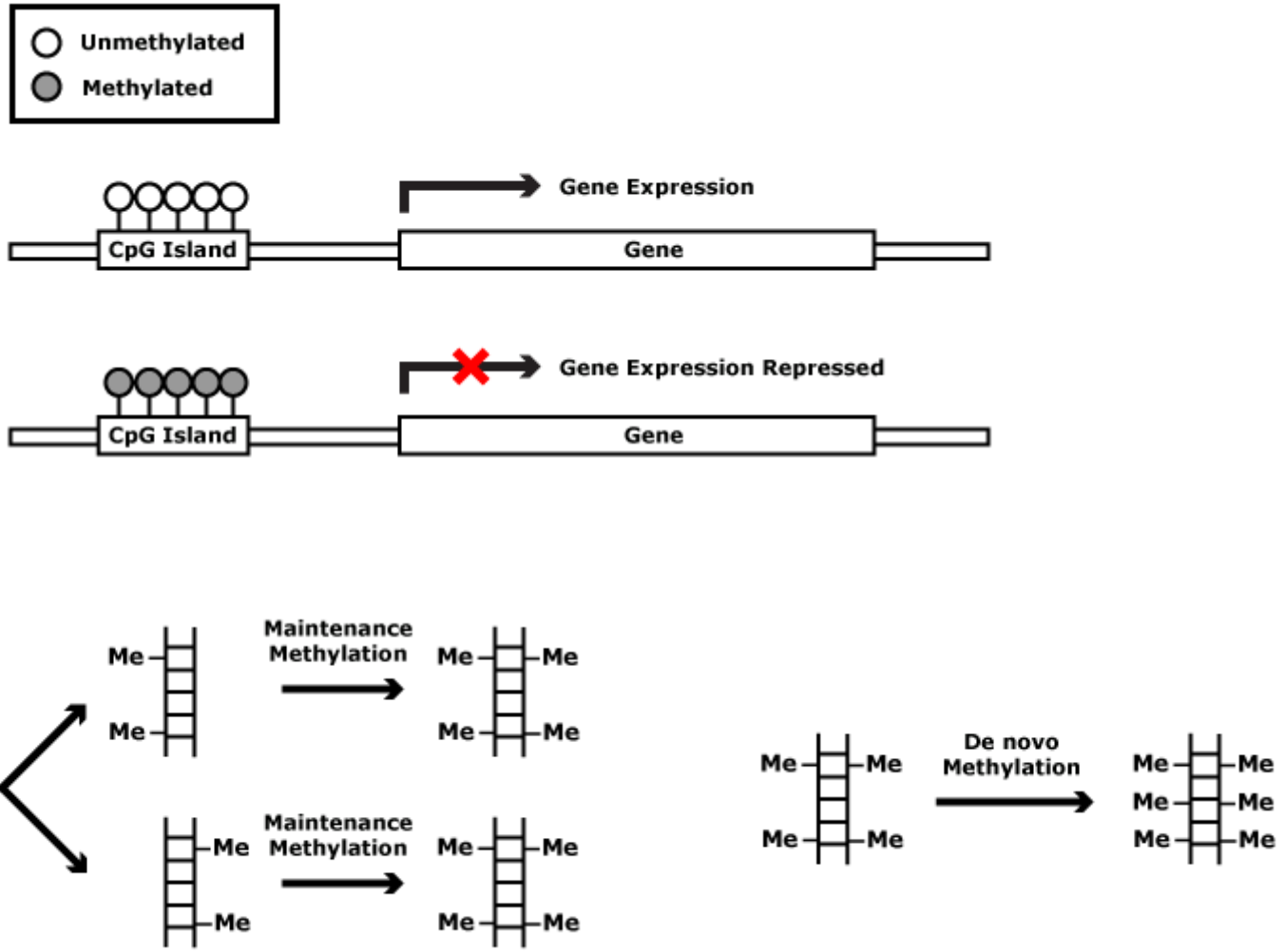


Lieberman-Aiden E (2009) Science 326:289

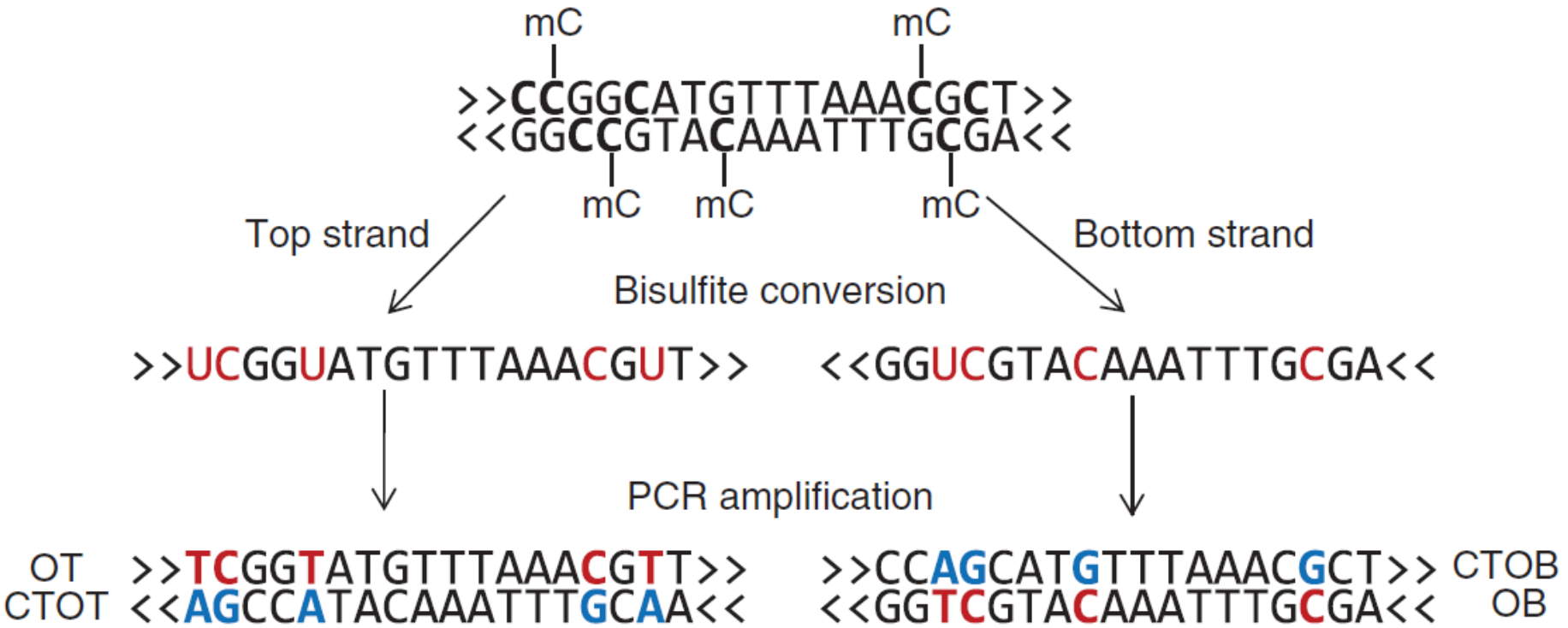
metylacja DNA



metylacja DNA

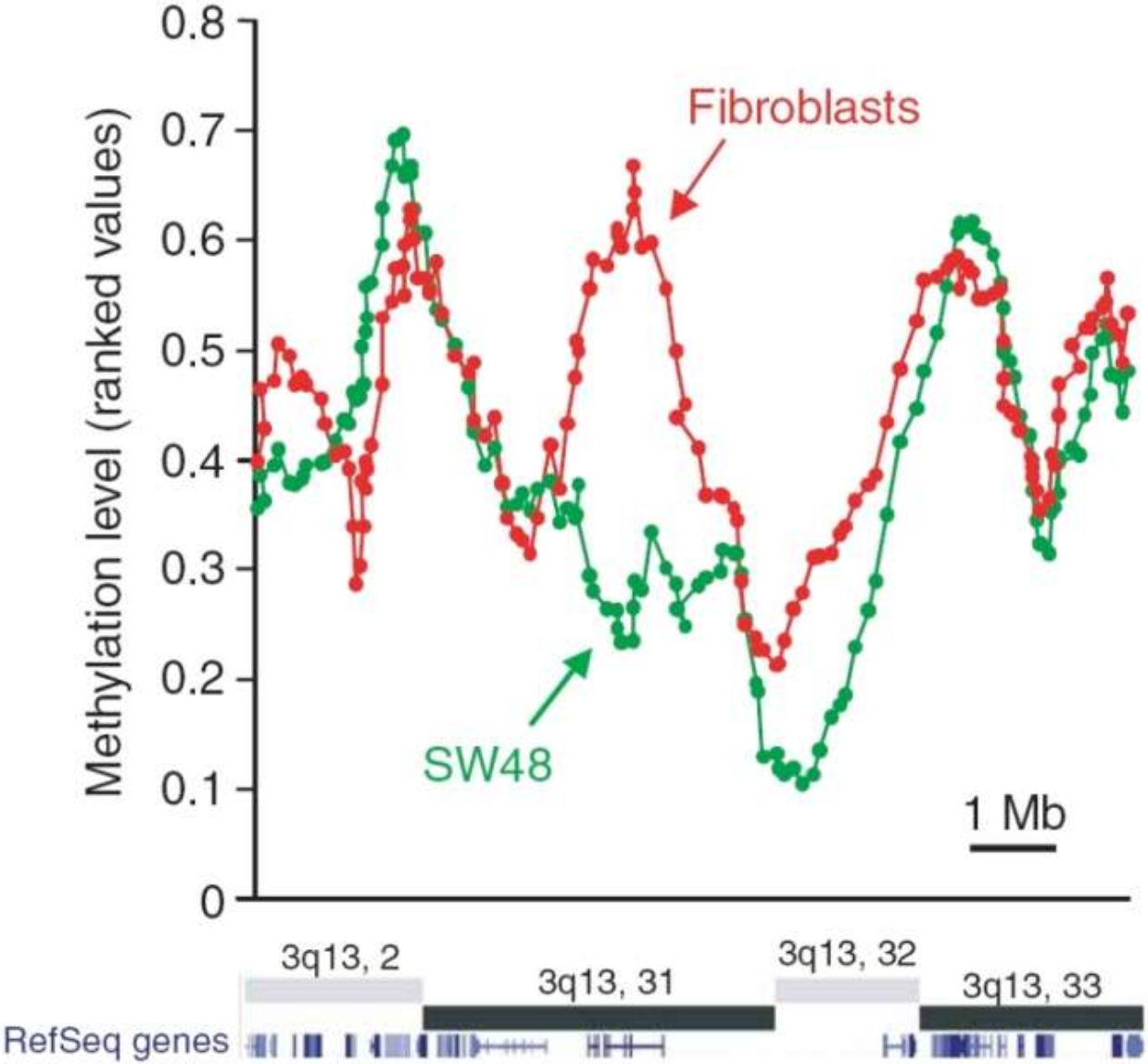


profilowanie metylacji DNA



Krueger F (2012) Nature Meth. 9:145

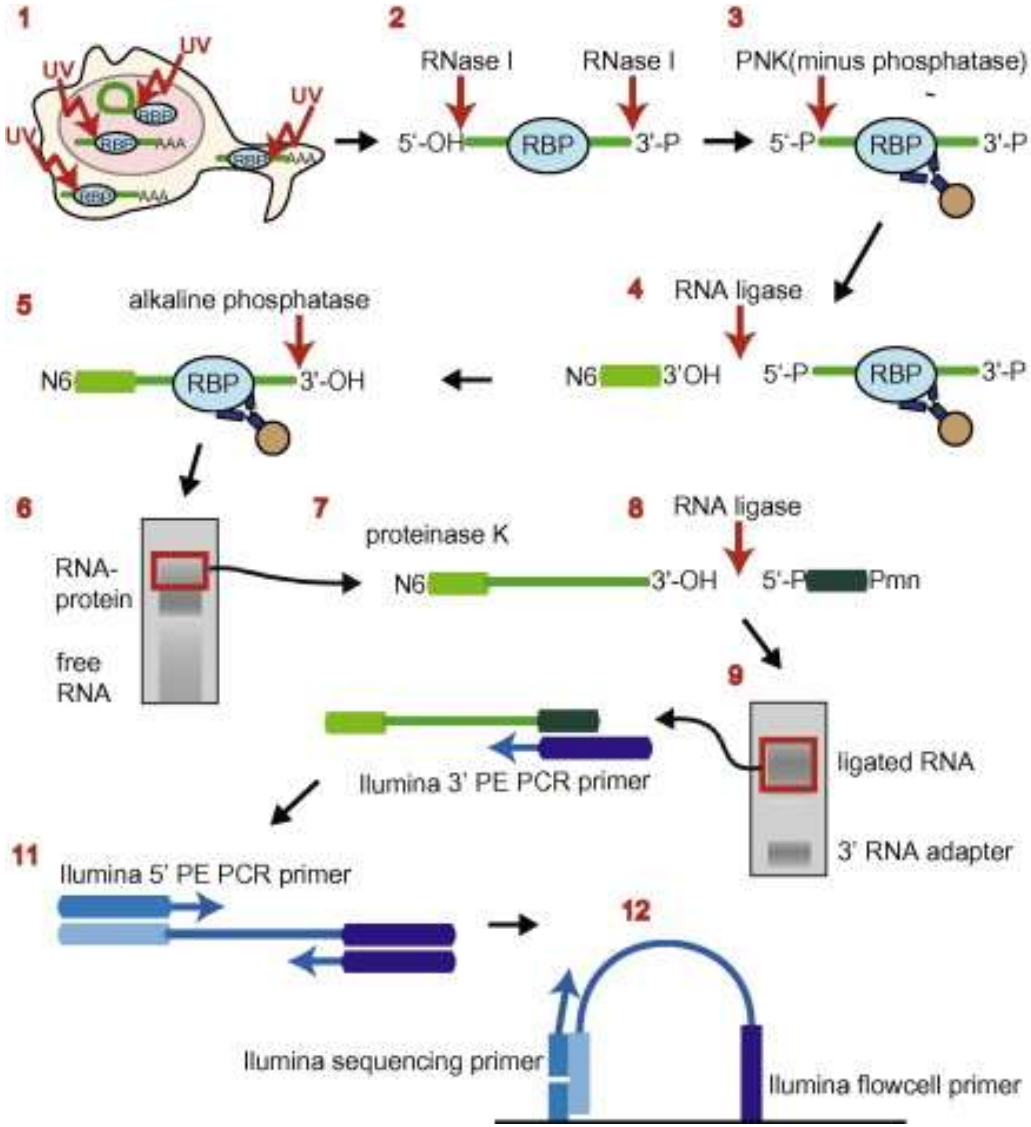
profilowanie metylacji DNA



Weber W et al (2005) Nat. Genet. 37:856

Oddziaływania białek z RNA (CLIP)

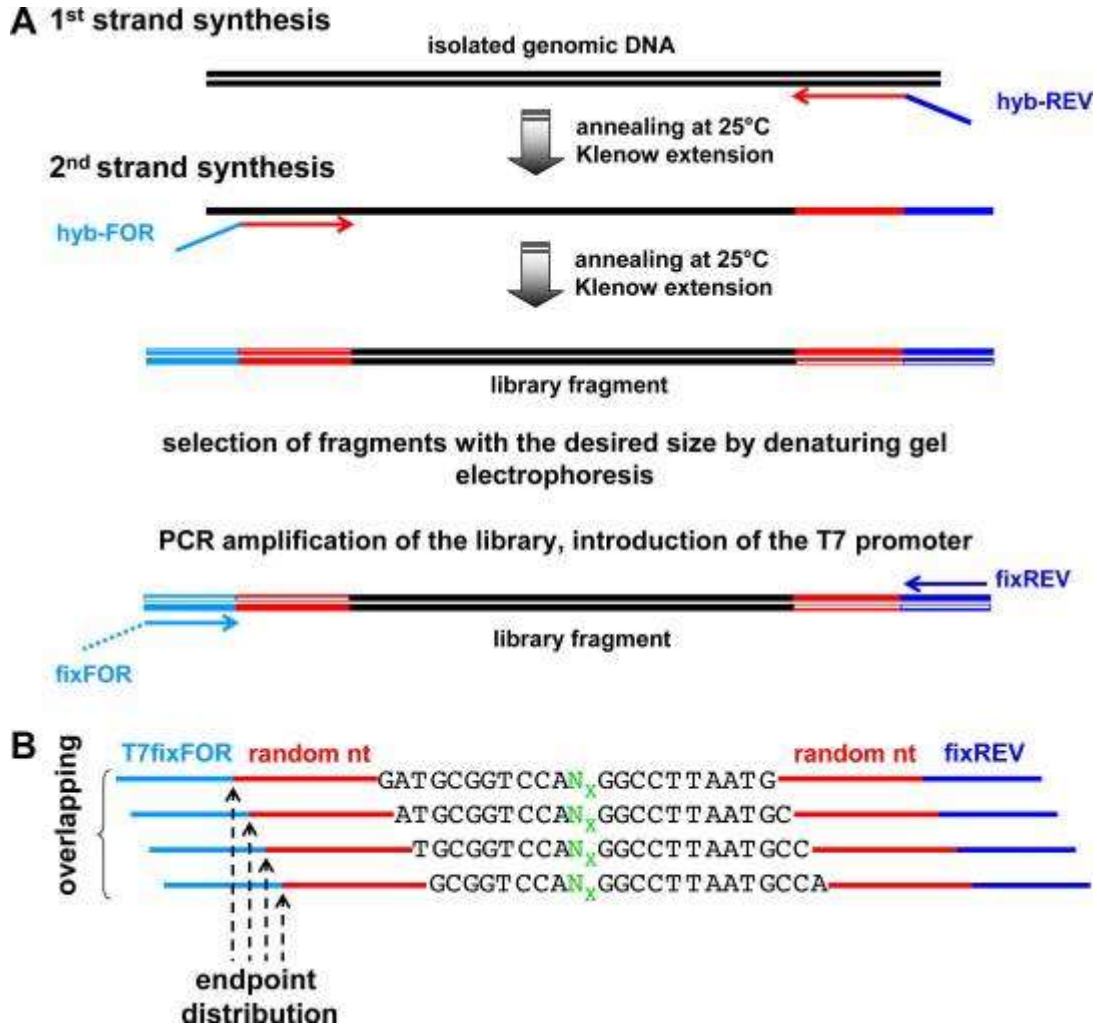
Cross-linking and immunoprecipitation



1. UV crosslink Cells or Tissue.
2. Partial RNA digestion.
3. Immunoprecipitate RBP and phosphorylate RNA 5' end.
4. Ligate the 5' RNA adapter.
5. Dephosphorylate RNA 3' end.
6. Purify RBP-RNA on SDS-PAGE
7. Digest the RBP.
8. Ligate the 3' RNA adapter.
9. Purify RNA on urea-TBE gel.
10. Reverse transcription.
11. PCR.
12. Illumina paired-end sequencing.

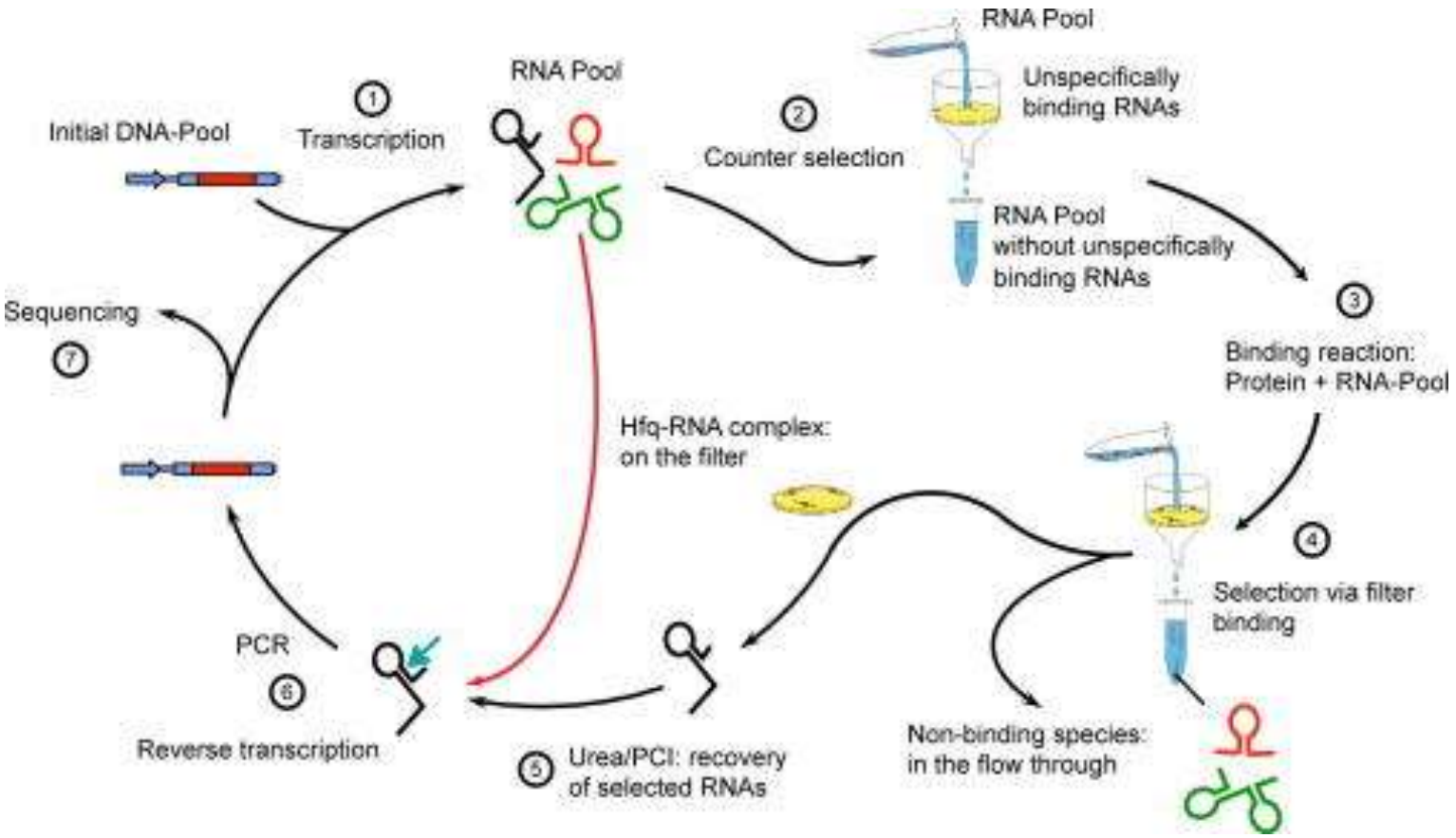
Wang Z et al. (2009) Methods 48:287

Genomic selex



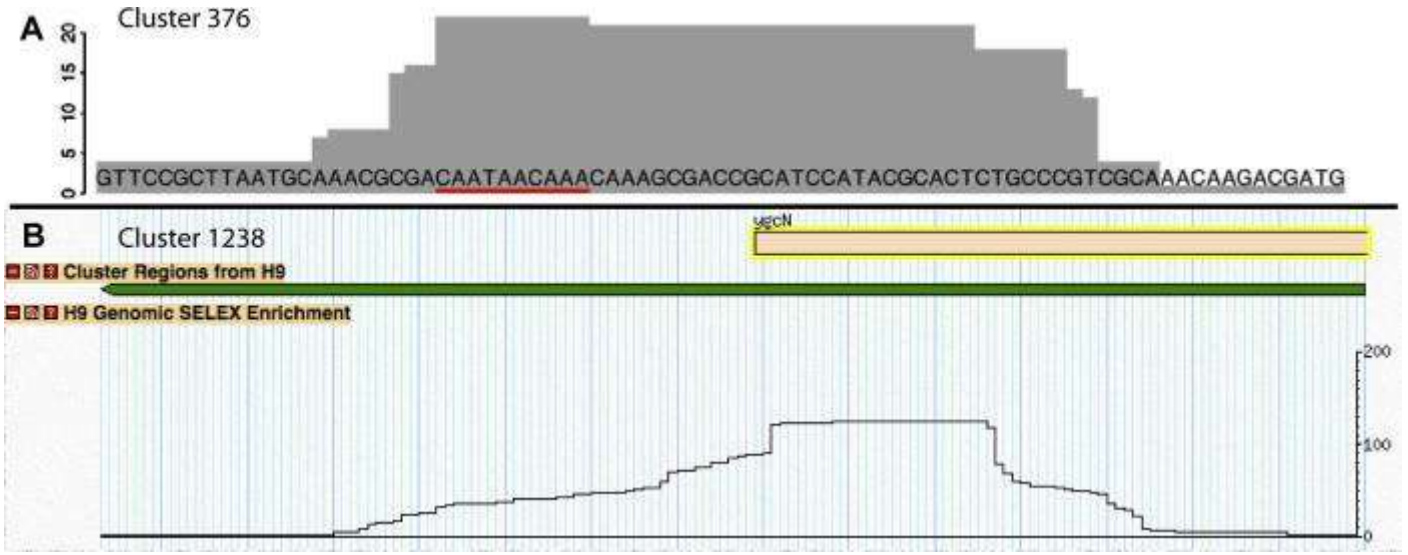
Zimmermann B et al. (2010) Mathods 52;125

Genomic selex



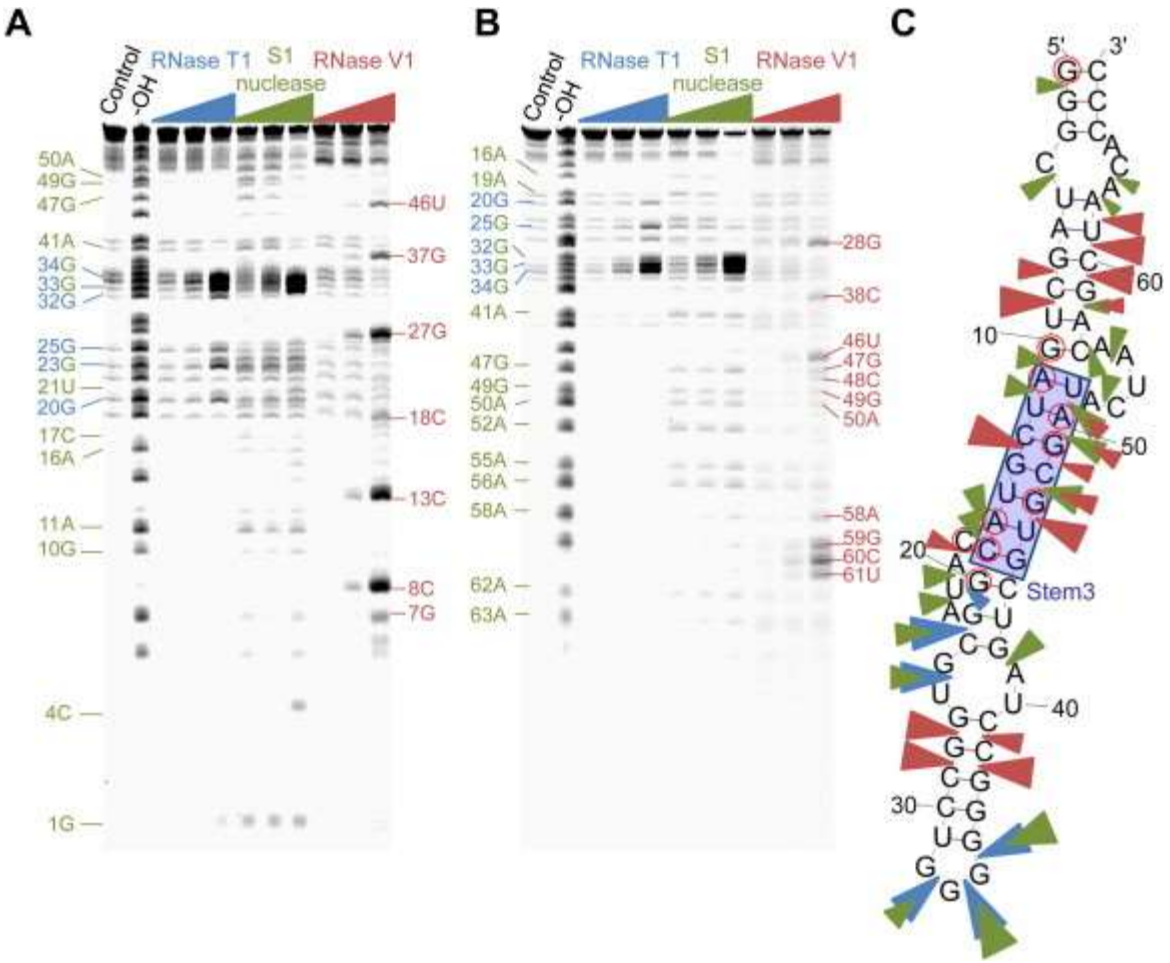
Zimmermann B et al. (2010) Mathods 52;125

Genomic selex



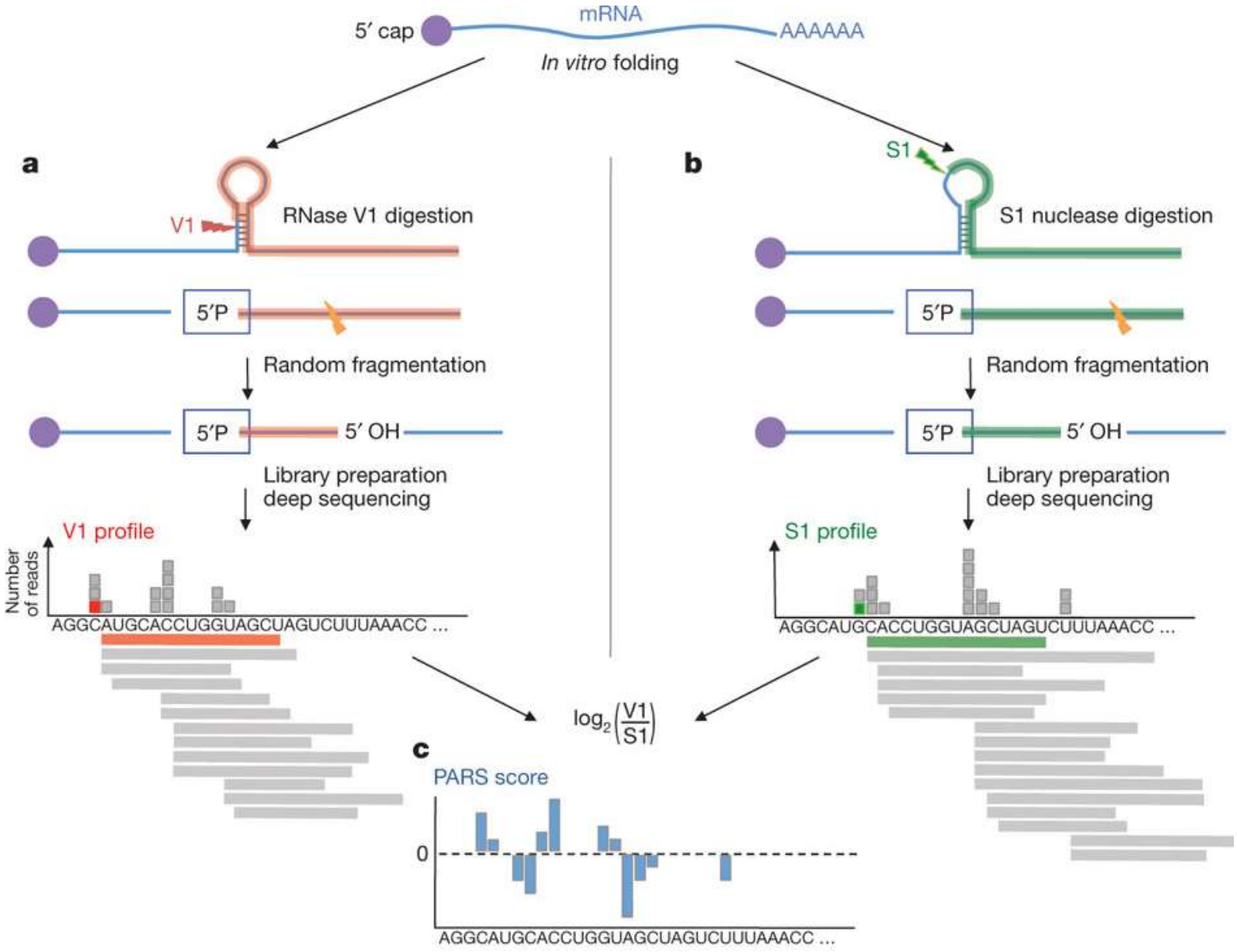
Zimmermann B et al. (2010) Mathods 52;125

badanie struktury RNA



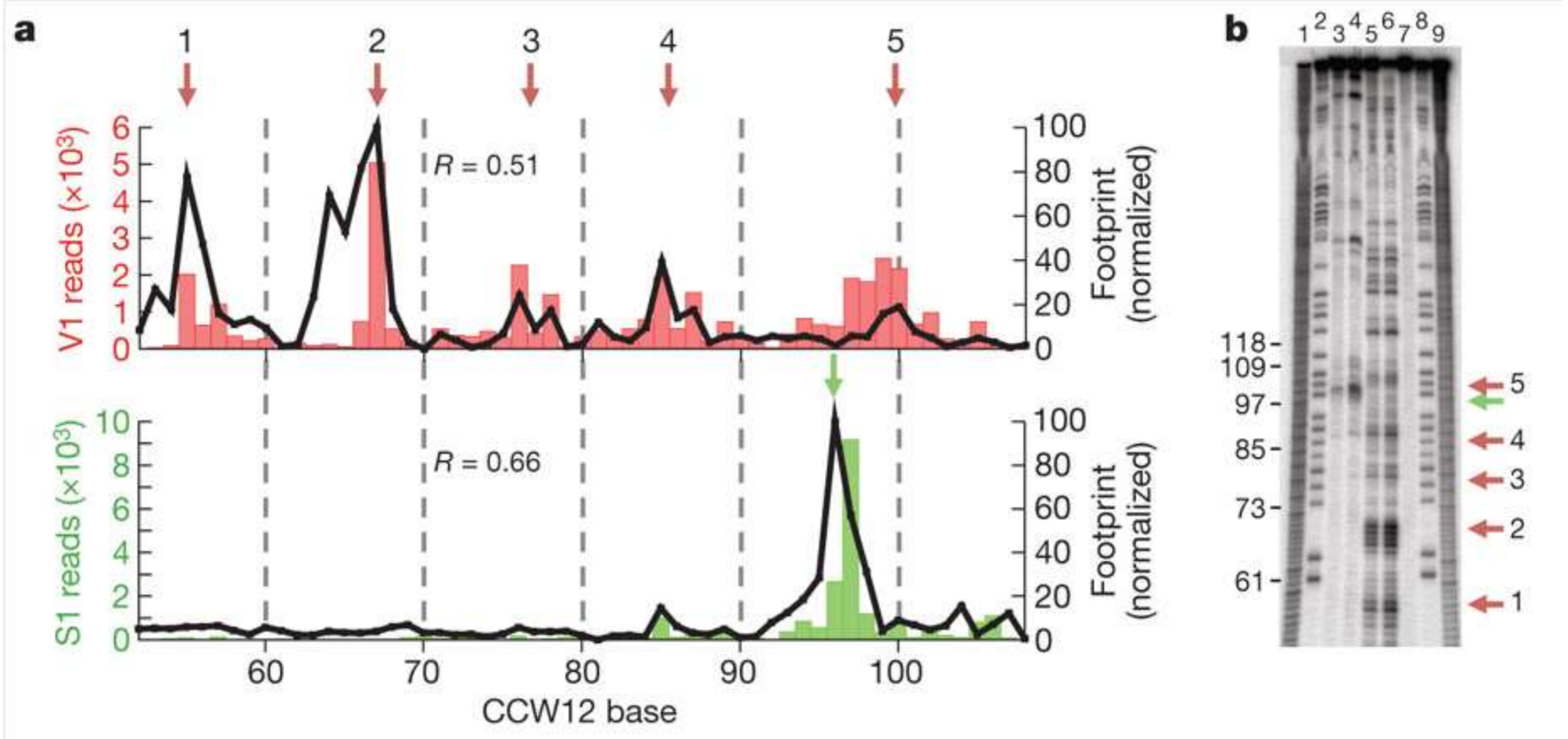
Adachi H et al (2000) Biochimie. 93:1081

badanie struktury RNA



Kertesz M et al (2010) Nature 467:103

badanie struktury RNA



Kertesz M et al (2010) Nature 467:103