Application of Machine Learning Methods to Genome-wide Maps of Histone Methylations

Stefan Bekiranov

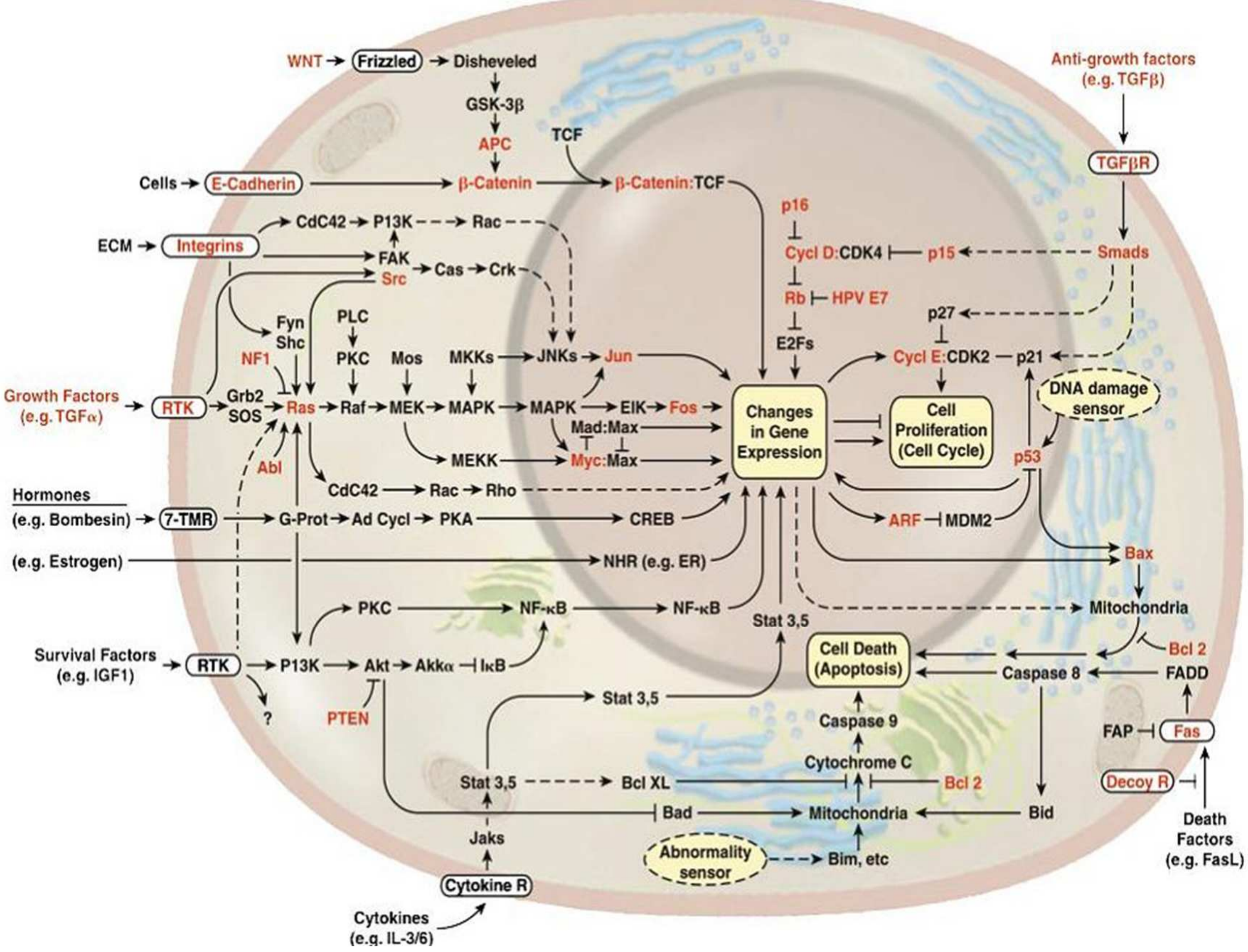
University of Virginia

Outline

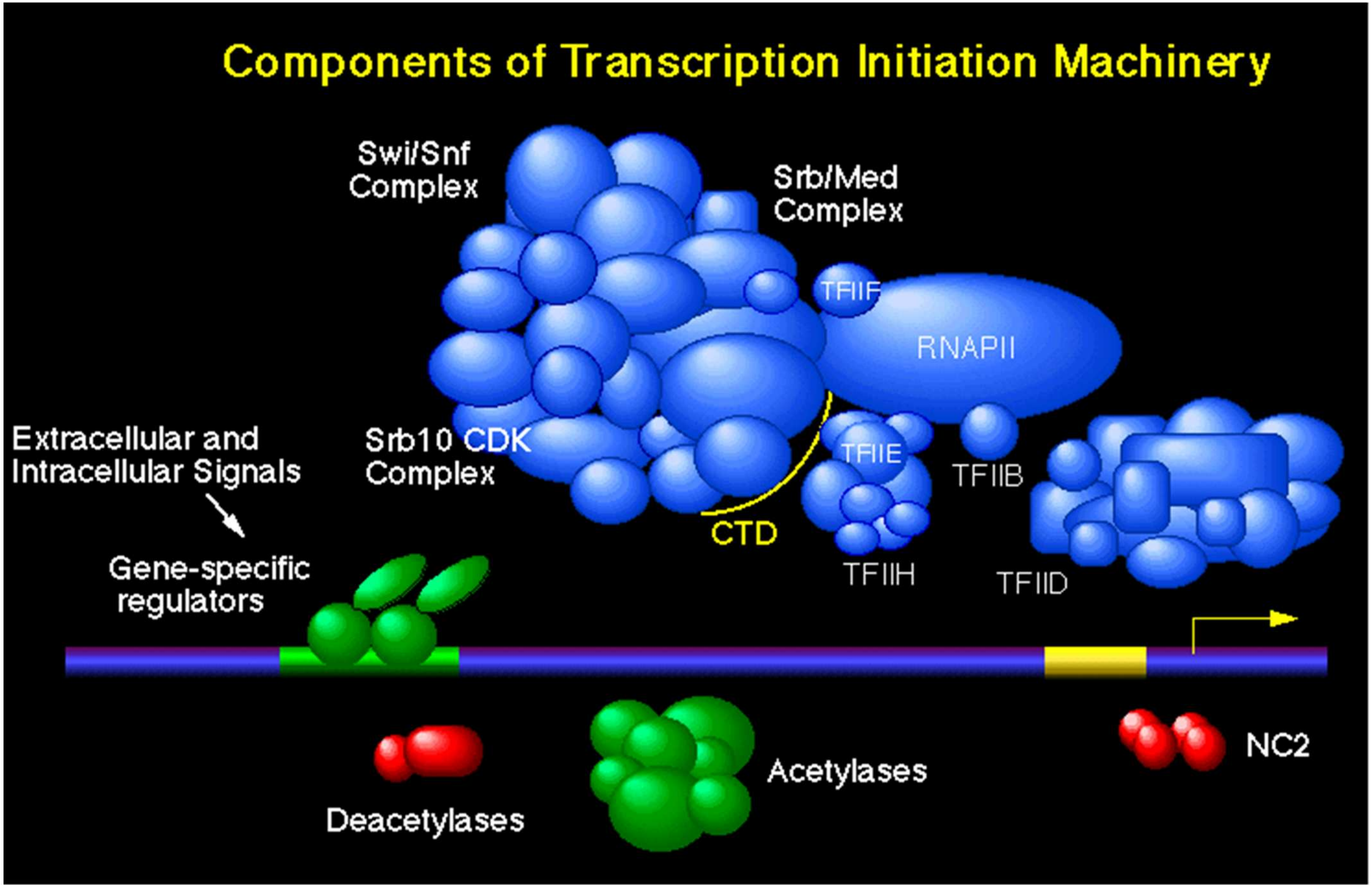
- Molecular Biology Introduction/Overview
- High Throughput Genomics Technologies
  - Gene Expression Microarrays
  - High Throughput Sequencing: ChIP-Seq
- Machine Learning Methods & Our Results
  - Multilinear Regression
  - Multivariate Adaptive Regression Splines (MARS)
  - Prediction: H4R3me2s is globally repressive
- Experimental Studies on H4R3me2s

Molecular Biology Introduction/Overview

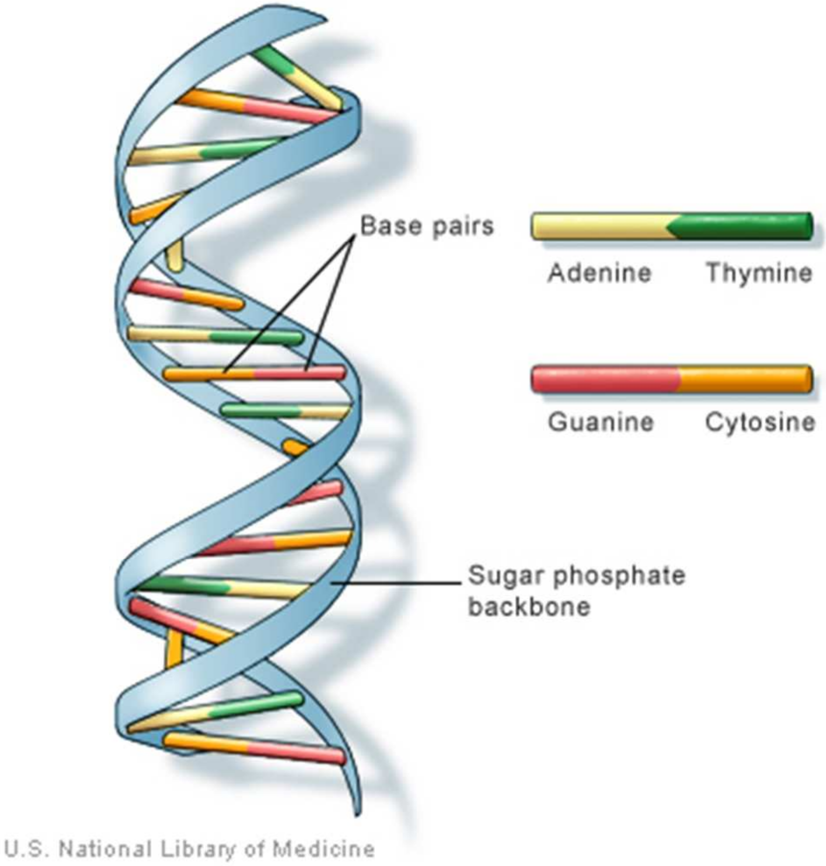
Signal Transduction



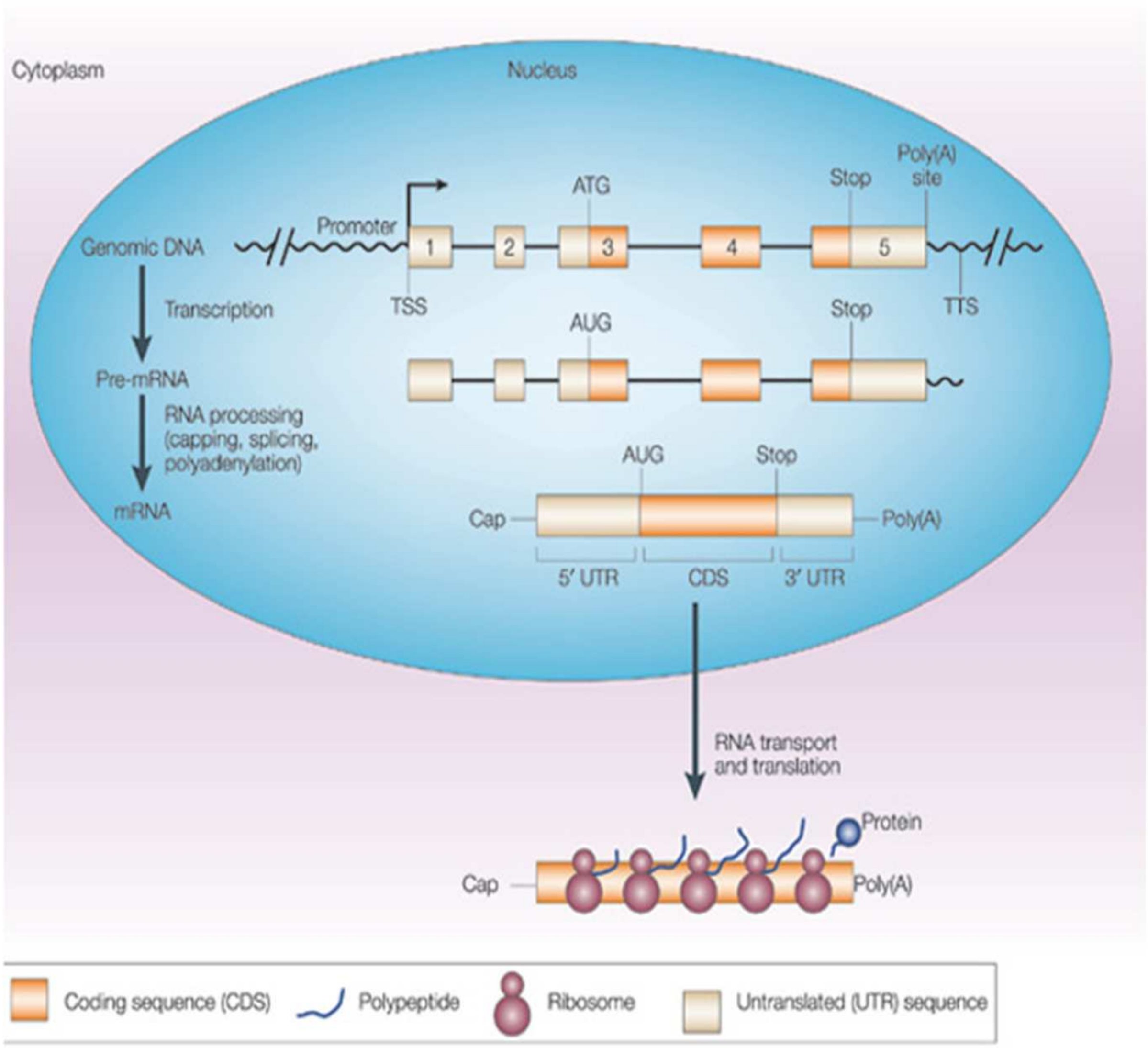
Transcription Initiation: Turn "on" a gene



DNA

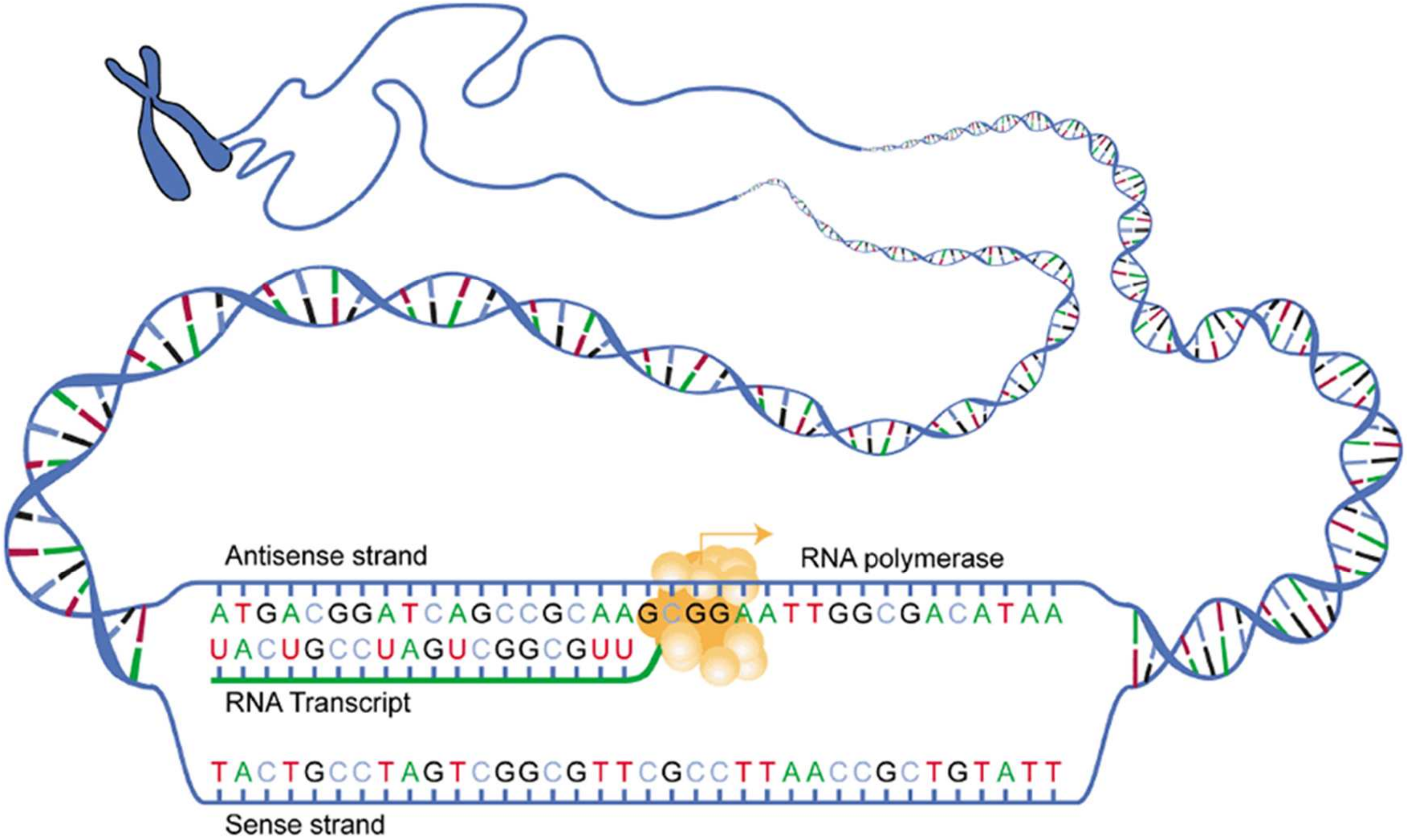

Base pairs
Adenine
Thymine
Guanine
Cytosine
Sugar phosphate backbone
U.S. National Library of Medicine

Gene -> mRNA -> Protein

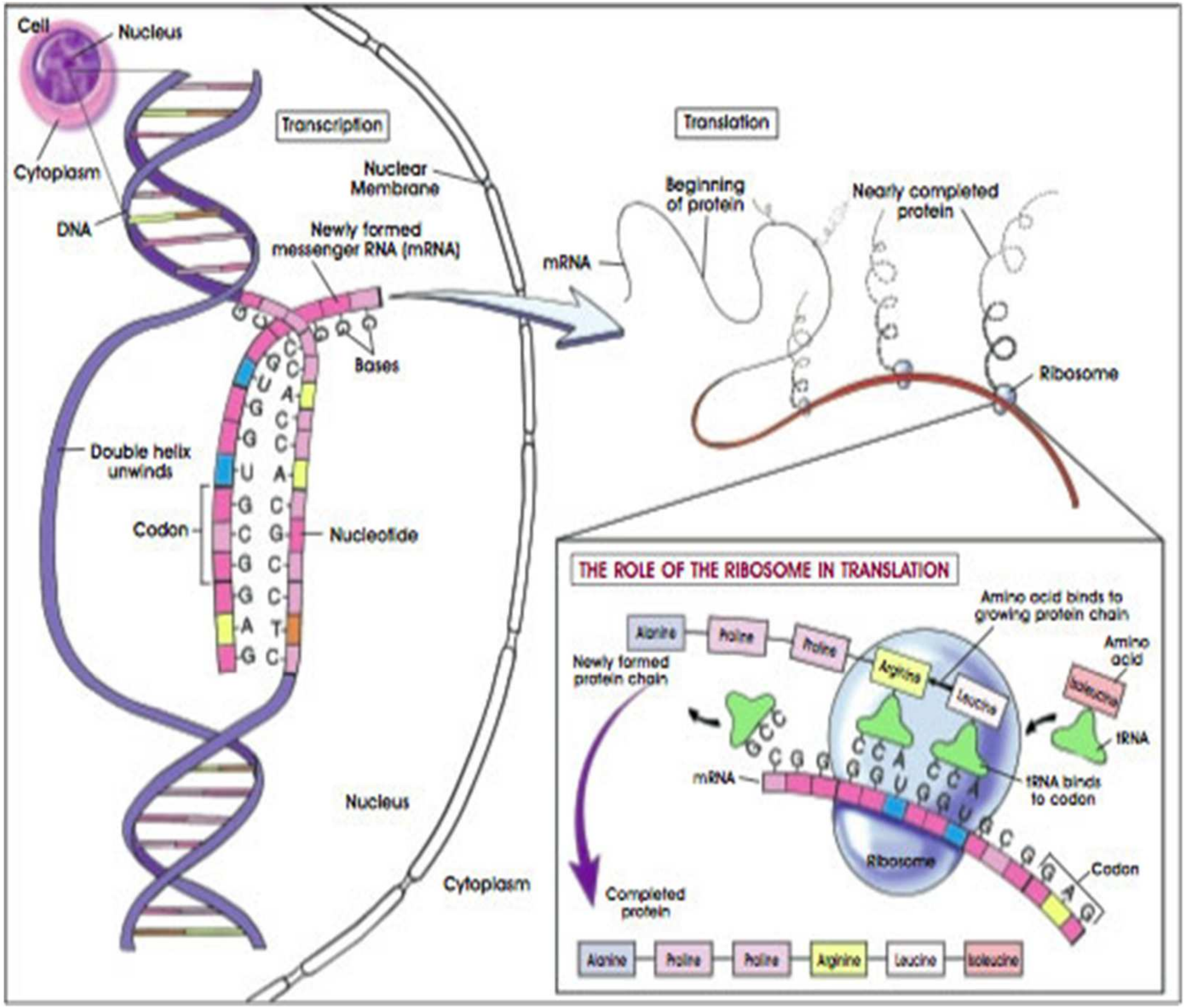


Nature Reviews | Genetics

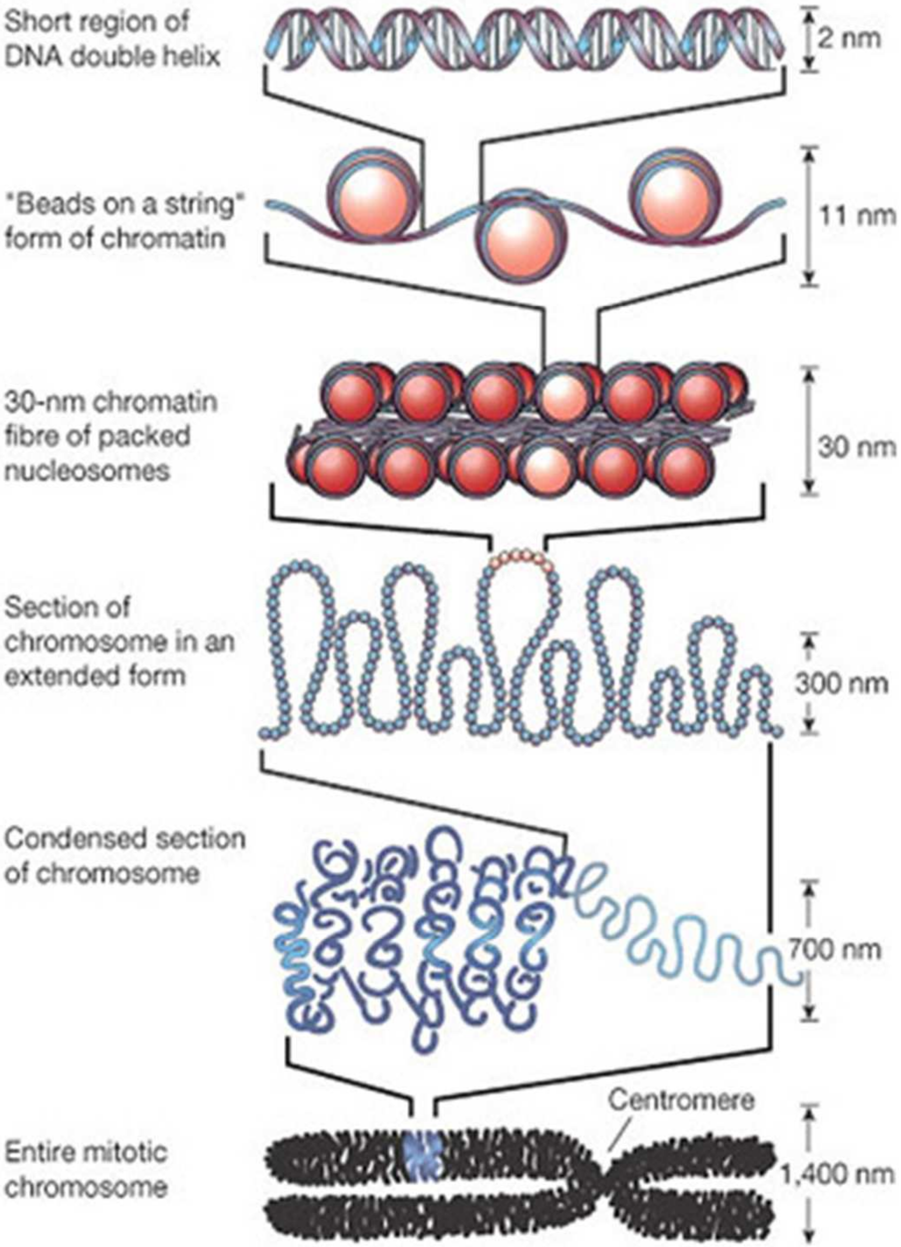
Transcription



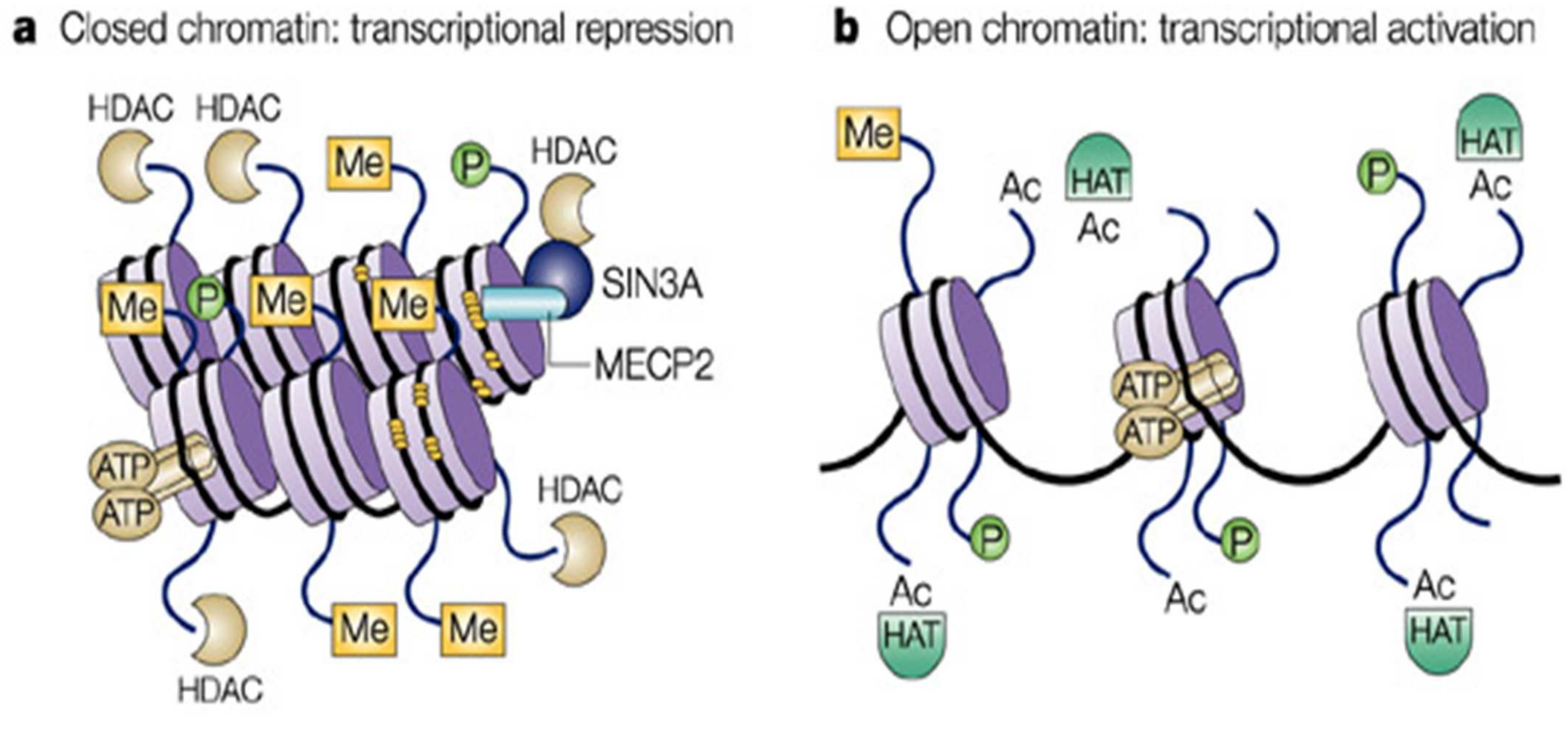
mRNA -> Protein



Chromatin


Short region of DNA double helix
2 nm
"Beads on a string" form of chromatin
11 nm
30-nm chromatin fibre of packed nucleosomes
30 nm
Section of chromosome in an extended form
300 nm
Condensed section of chromosome
700 nm
Centromere
Entire mitotic chromosome
1,400 nm

Histone Modifications

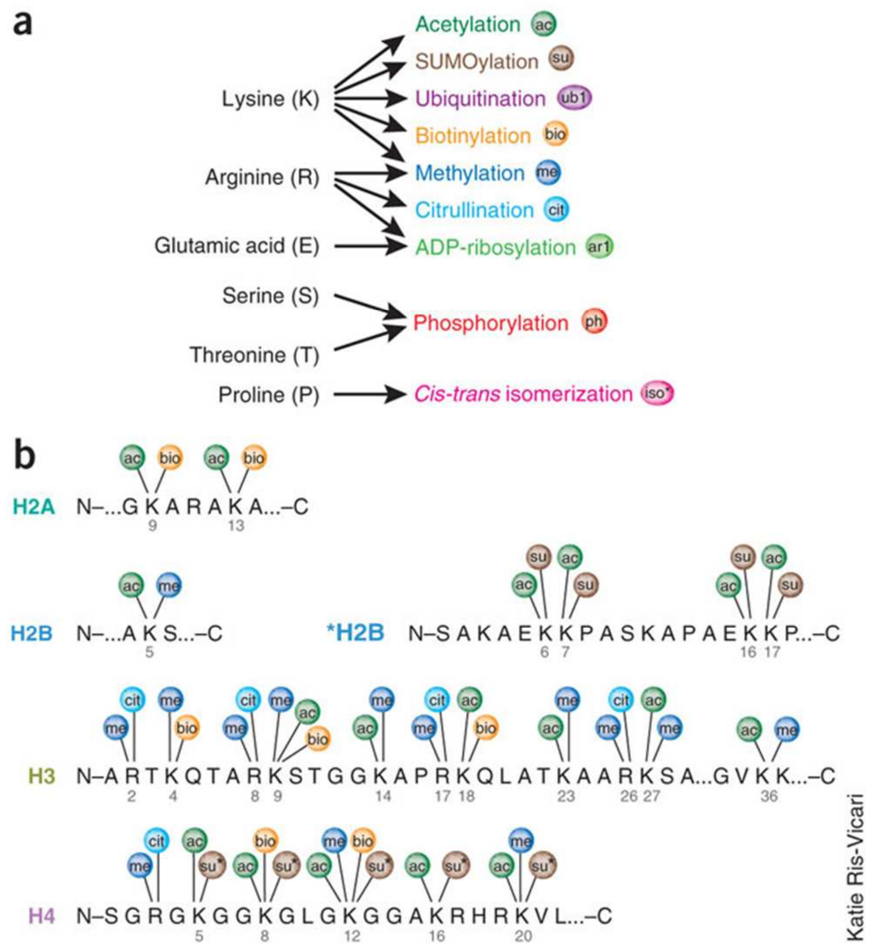


a Closed chromatin: transcriptional repression

b Open chromatin: transcriptional activation

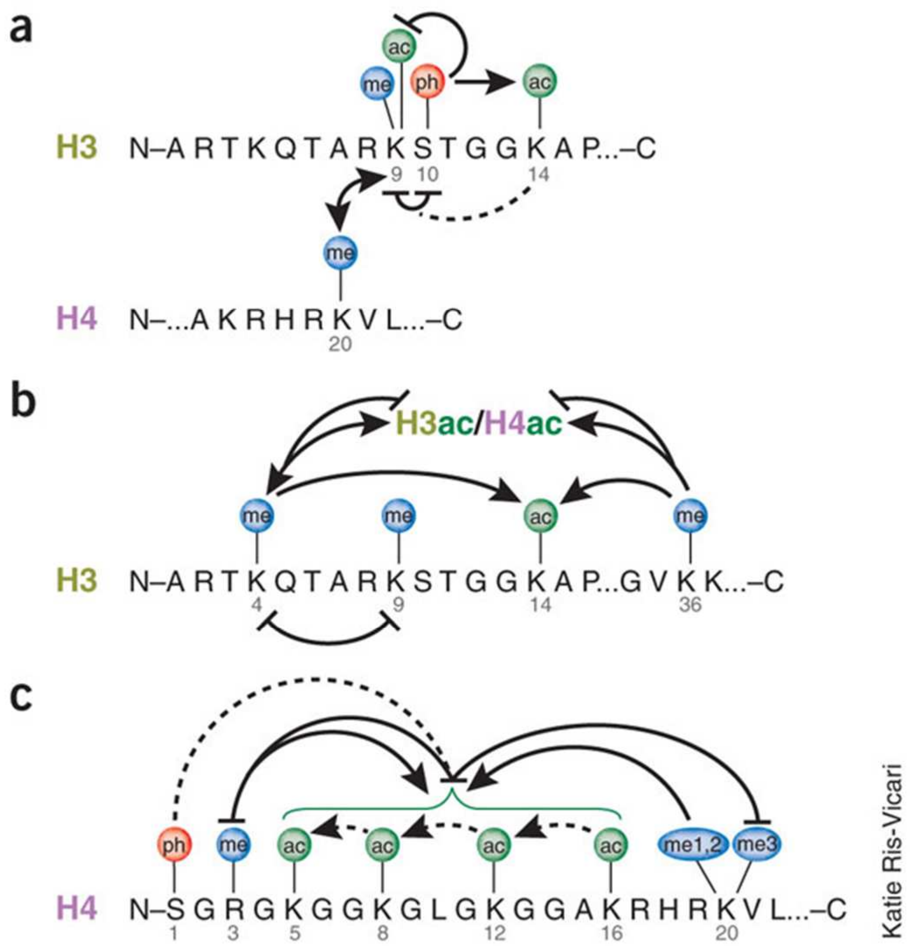
Nature Reviews | Drug Discovery

Histone Modifications



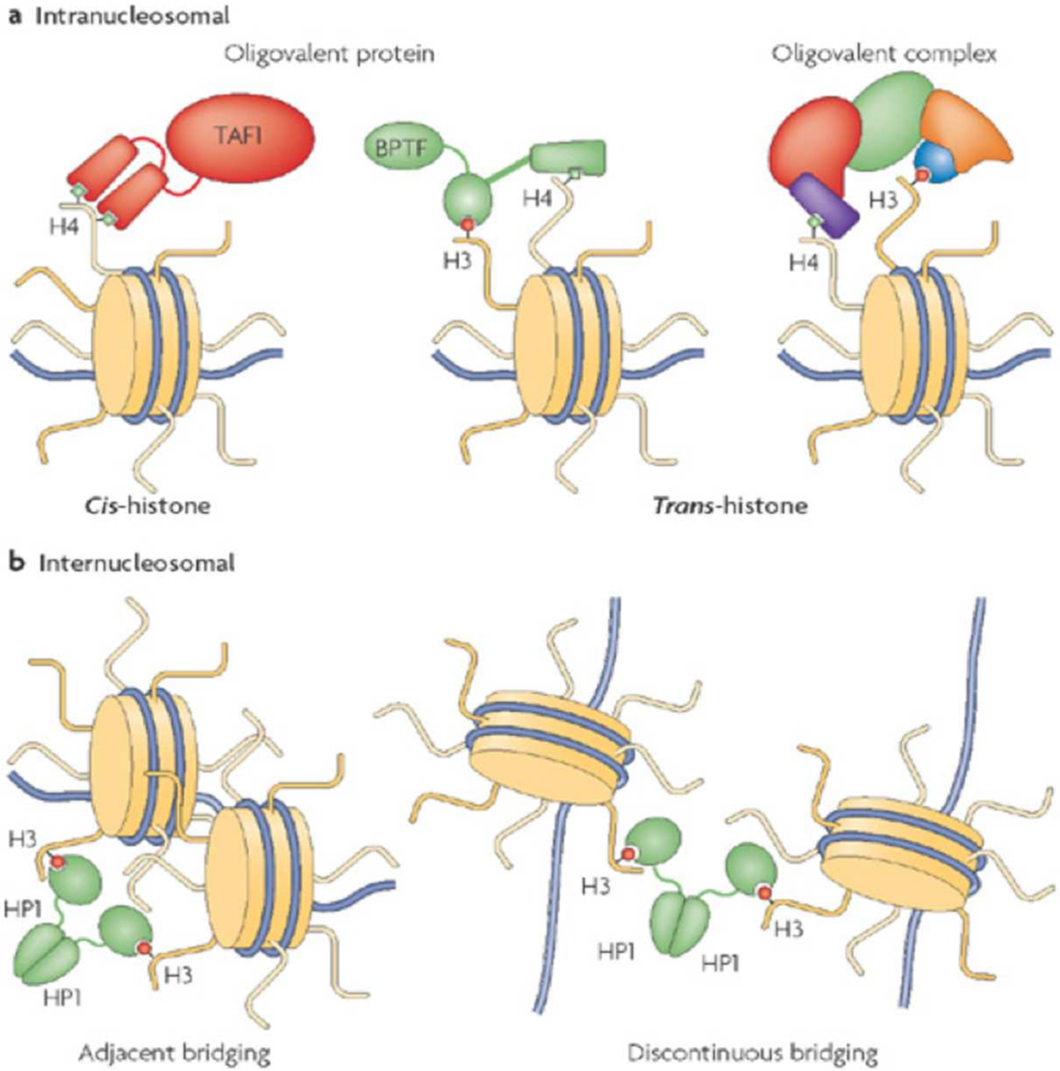
Katie Ris-Vicari

Cross-regulation of Histone Modifications


a
ac
me
ph
ac
H3
N–A R T K Q T A R K S T G G K A P...–C
9
10
14
me
H4
N–...A K R H R K V L...–C
20
b
H3ac/H4ac
me
me
ac
me
H3
N–A R T K Q T A R K S T G G K A P...G V K K...–C
4
9
14
36
c
ph
me
ac
ac
ac
ac
me1,2
me3
H4
N–S G R G K G G K G L G K G G A K R H R K V L...–C
1
3
5
8
12
16
20

Katie Ris-Vicari

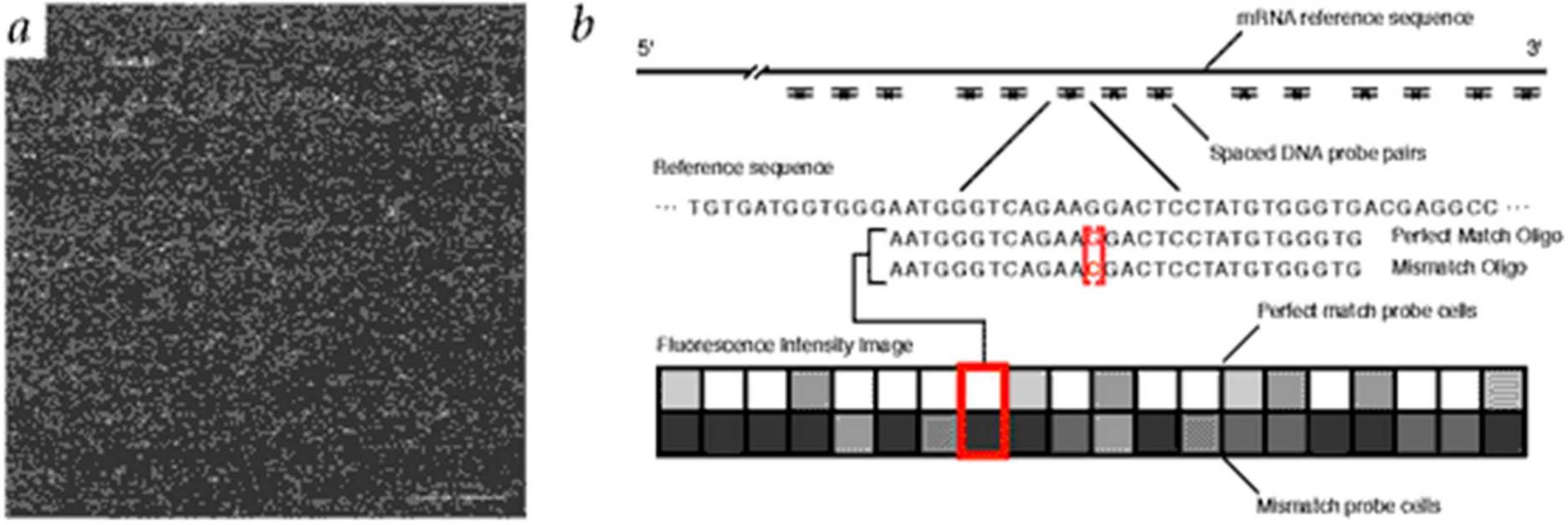
Multivalent Chromatin Engagement



Nature Reviews | Molecular Cell Biology

High Throughput Genomics Technologies

Affymetrix Probe Set: Gene Expression Arrays (Human U133 Array ~2002)


a
b
mRNA reference sequence
5'
3'
Spaced DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC ···
AATGGGTCAGAAGGACTCCTATGTGGGTG
Perfect Match Oligo
AATGGGTCAGAACGACTCCTATGTGGGTG
Mismatch Oligo
Perfect match probe cells
Fluorescence Intensity Image
Mismatch probe cells

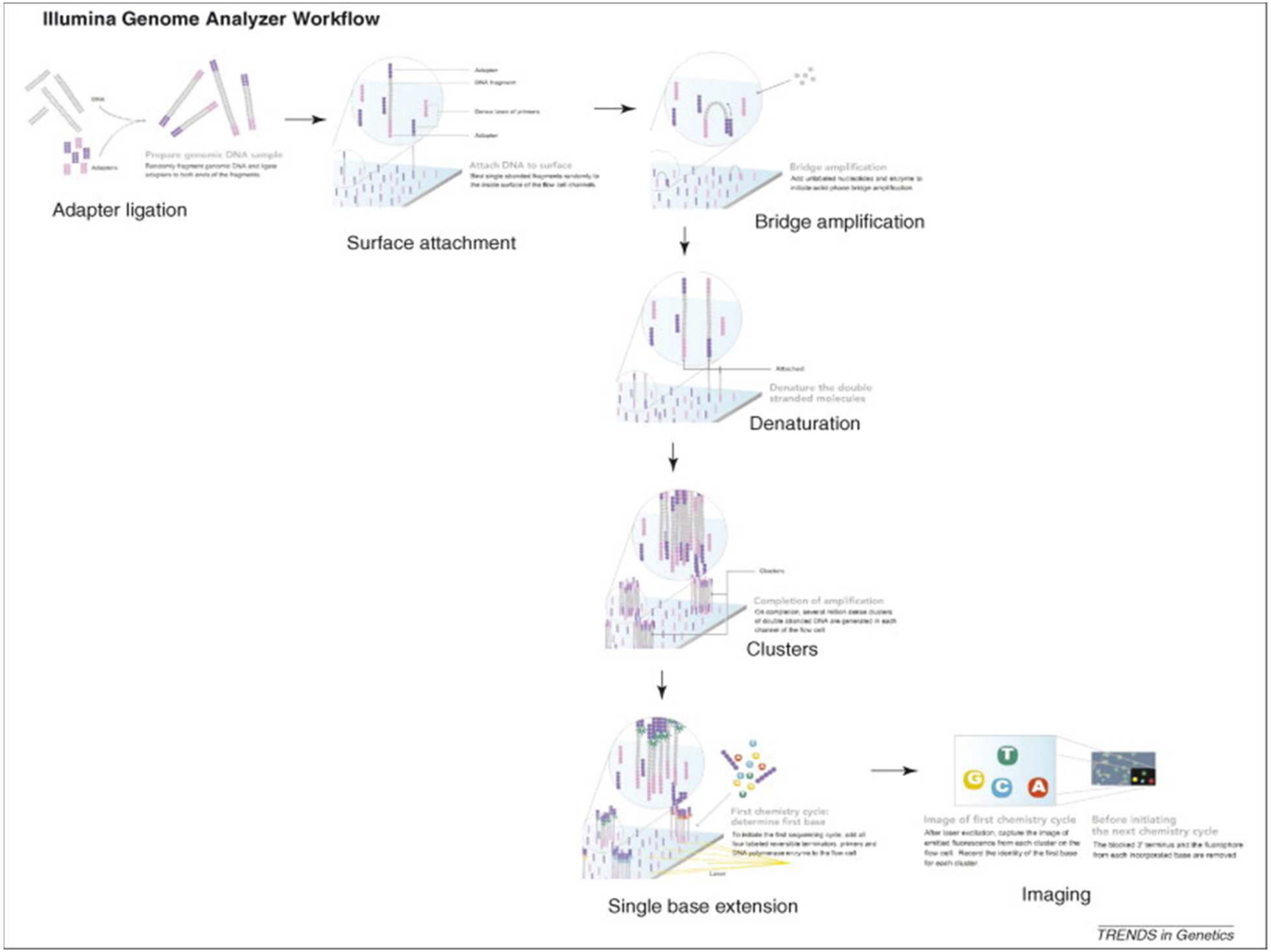
Disruptive Technology: High Throughput Sequencing

Prepare gDNA Library
3 hours hands-on, 6 hours total time

Generate Clusters
<1 hour hands-on time, 5 hours total time

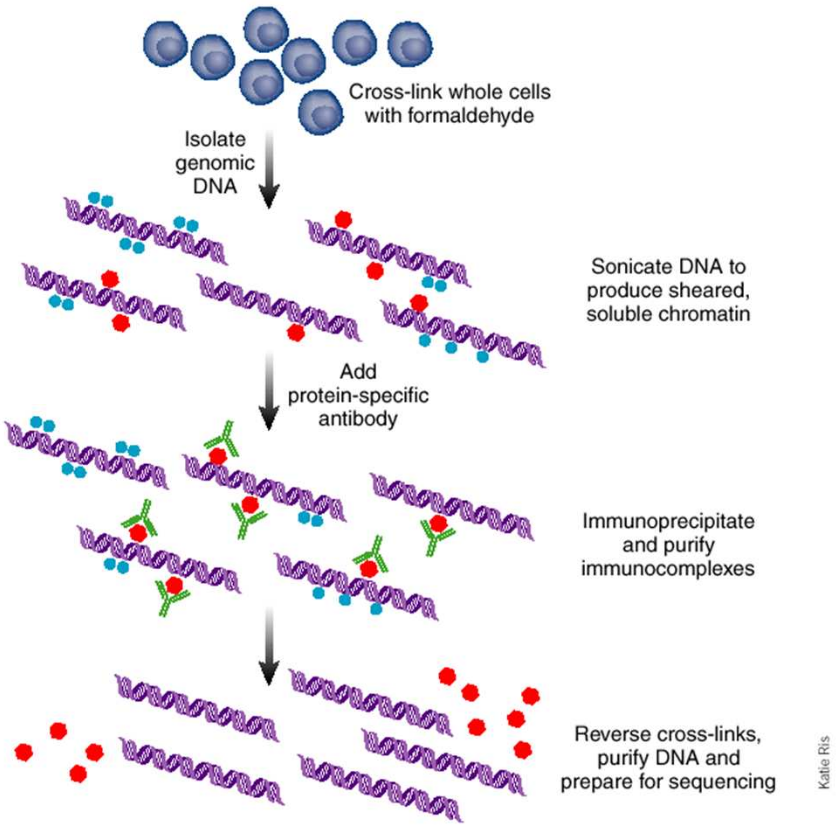
Sequence Clusters
2.5 days single read (36 bases)

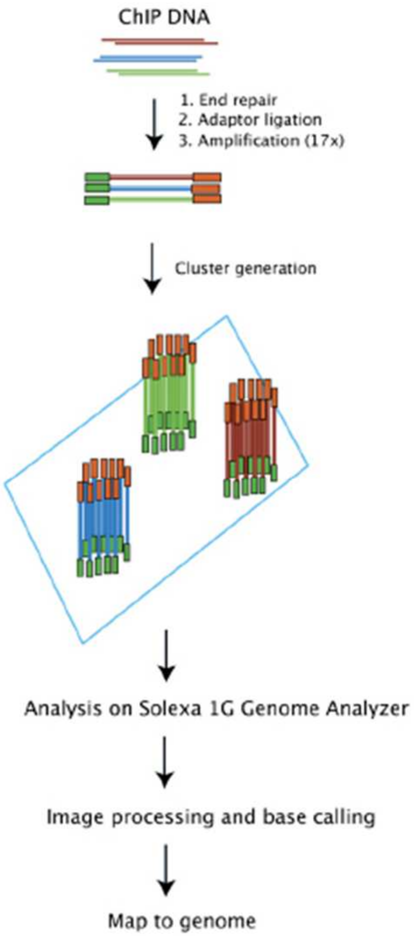
Illumina (Solexa) Workflow (~2006)


Illumina Genome Analyzer Workflow
DNA
Adapters
Prepare genomic DNA sample
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments
Adapter ligation
Adapter
DNA fragment
Dense lawn of primers
Adapter
Attach DNA to surface
Bind single stranded fragments randomly to the inside surface of the flow cell channels
Surface attachment
Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid phase bridge amplification
Bridge amplification
Attached
Denature the double stranded molecules
Denaturation
Clusters
Completion of amplification
On completion, several million dense clusters of double stranded DNA are generated in each channel of the flow cell
Clusters
First chemistry cycle: determine first base
To initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell
Laser
Single base extension
T
G
C
A
Image of first chemistry cycle
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster
Before initiating the next chemistry cycle
The blocked 3' terminus and the fluorophore from each incorporated base are removed
Imaging

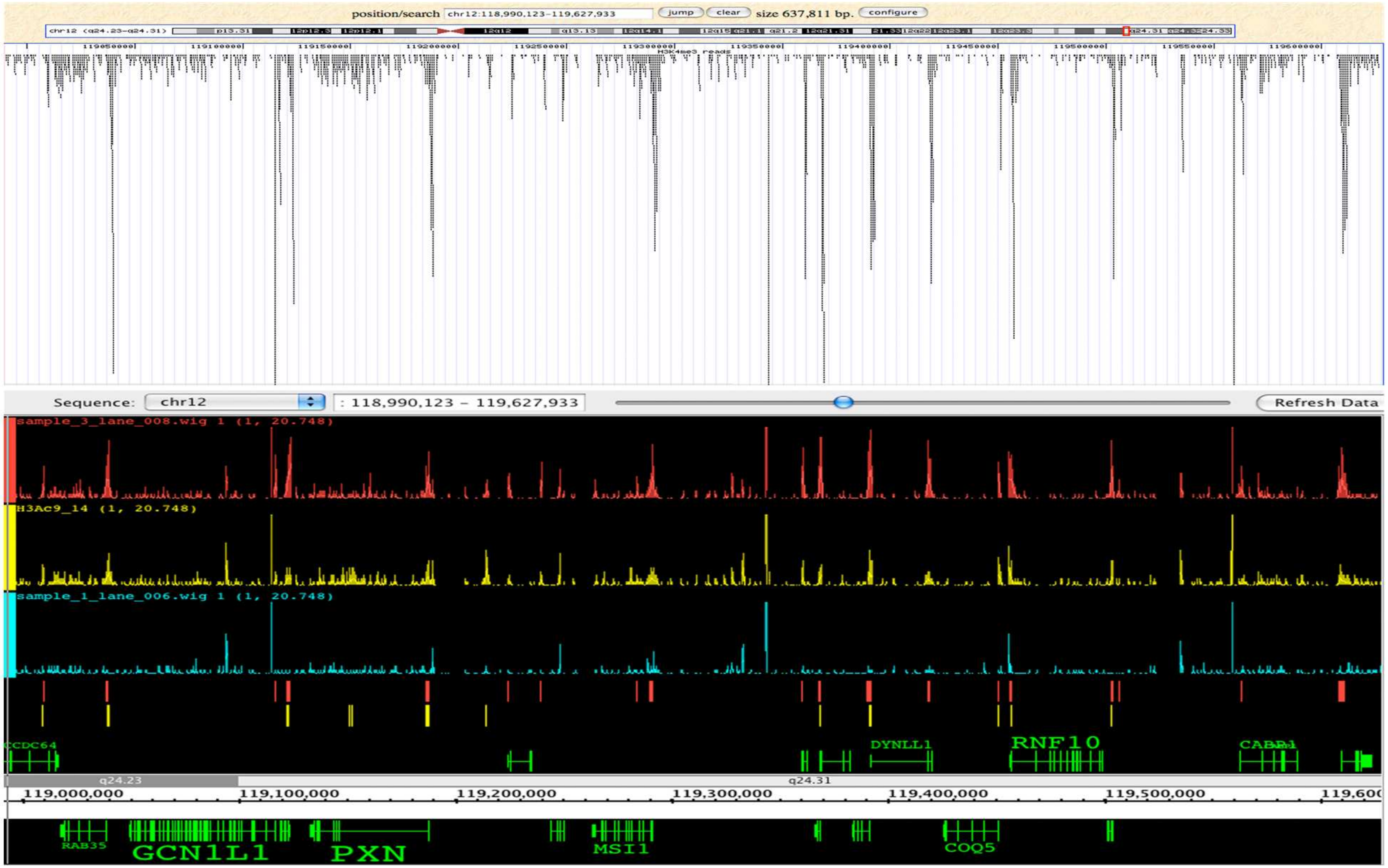
TRENDS in Genetics

ChIP-Seq


Cross-link whole cells with formaldehyde
Isolate genomic DNA
Sonicate DNA to produce sheared, soluble chromatin
Add protein-specific antibody
Immunoprecipitate and purify immunocomplexes
Reverse cross-links, purify DNA and prepare for sequencing
Katie Ris


ChIP DNA
1. End repair
2. Adaptor ligation
3. Amplification (17x)
Cluster generation
Analysis on Solexa 1G Genome Analyzer
Image processing and base calling
Map to genome

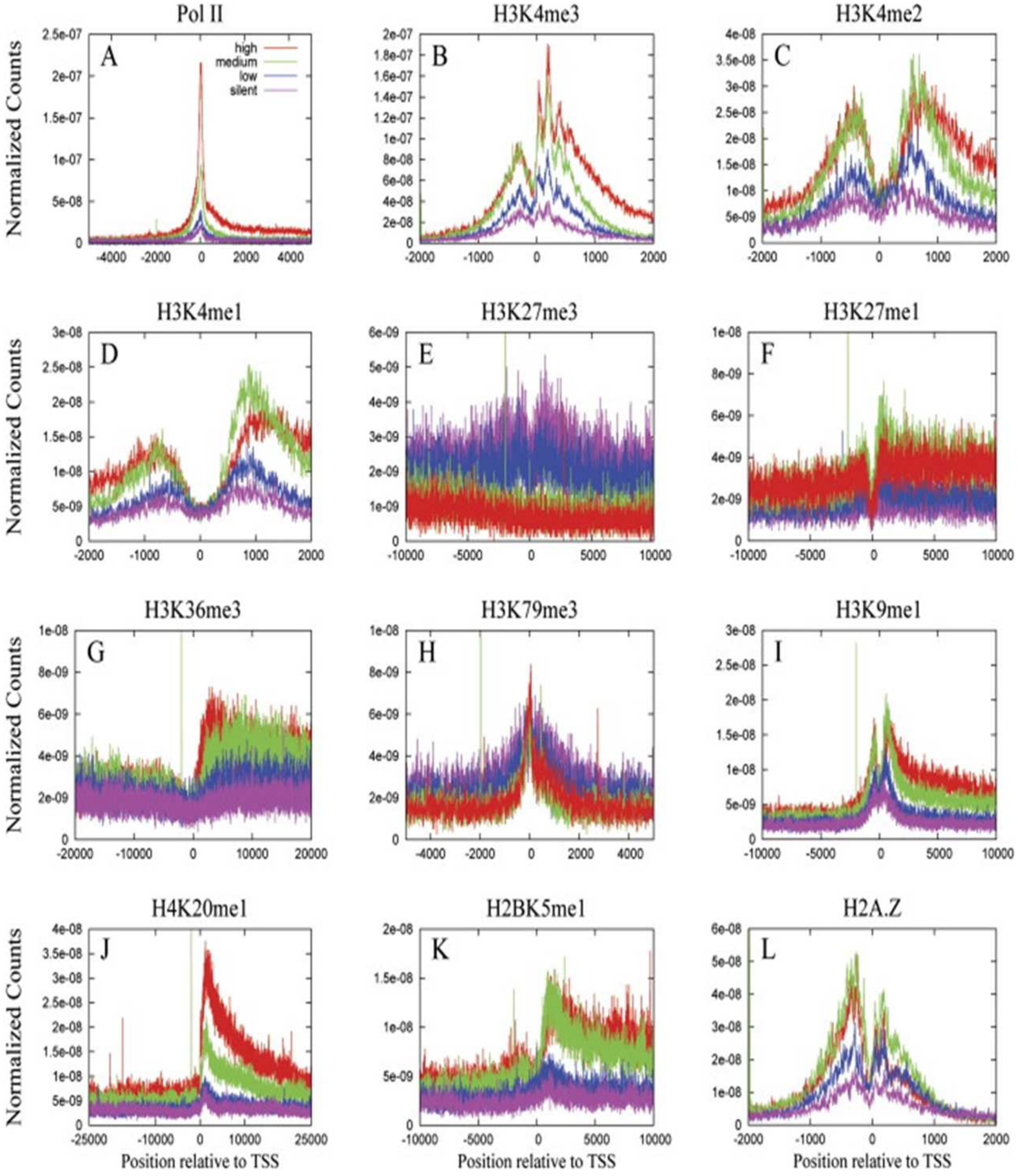
Map Reads; Calculate Enrichment Profile



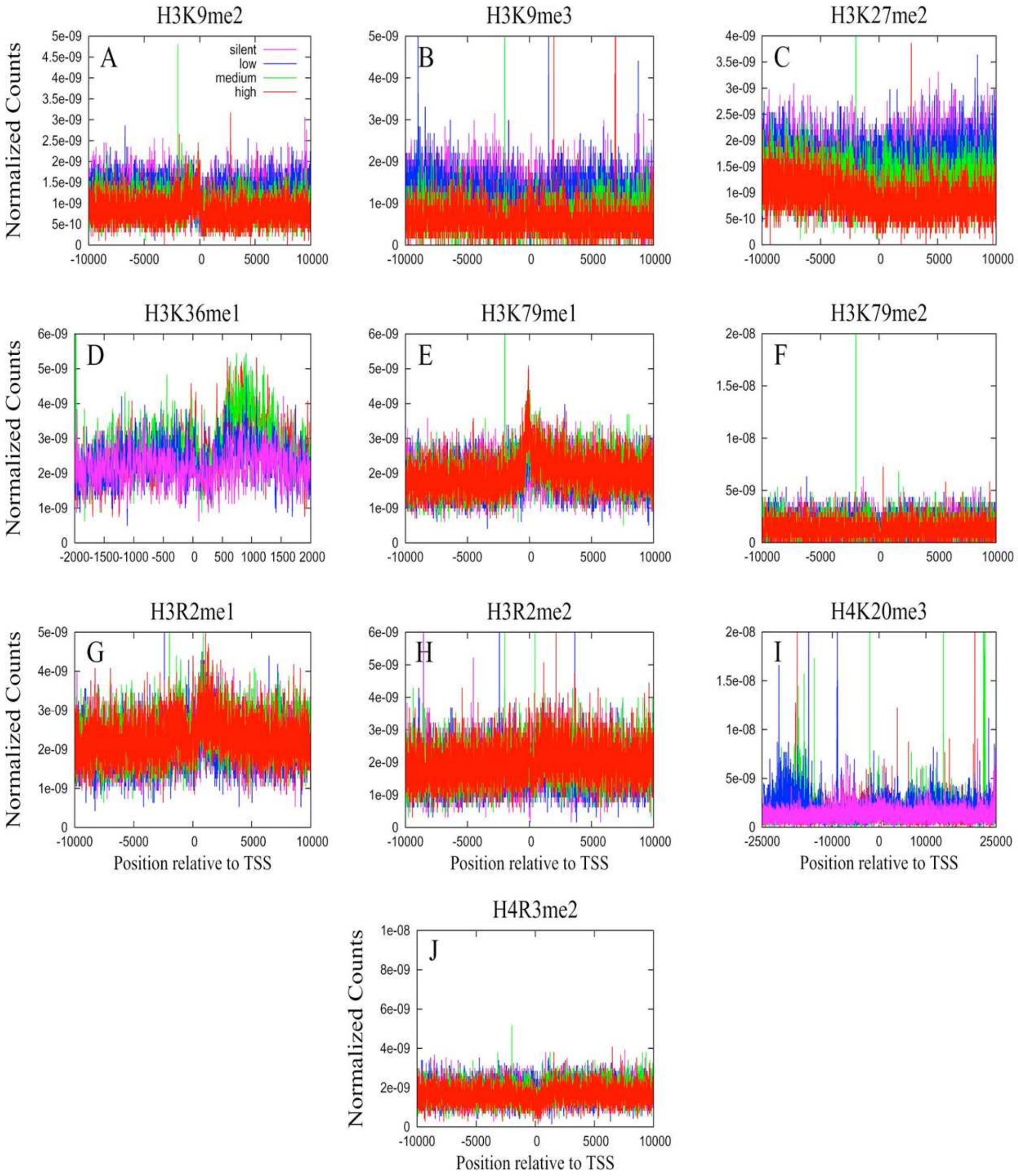
ChIP-Seq Data Sets

- In CD4$^+$ T cells, Keji Zhao Lab (NIH) Generated Maps of:
  - 20 Histone Lysine and Arginine Methylations
  - 18 Histone Acetylations
  - Pol II, CTCF and H2A.Z
  - Nucleosome Positions
  - Effector Proteins
- Epithelial to Mesenchymal Transition (EMT)
  - ~15-20 histone modifications/variants
  - ~6 Master-Switch Transcription Factors including Twist, Snail, Slug, SIP1, SMAD, NF-κB

Read Enrichments Stratified By Gene Expression Levels



Rogues Gallery

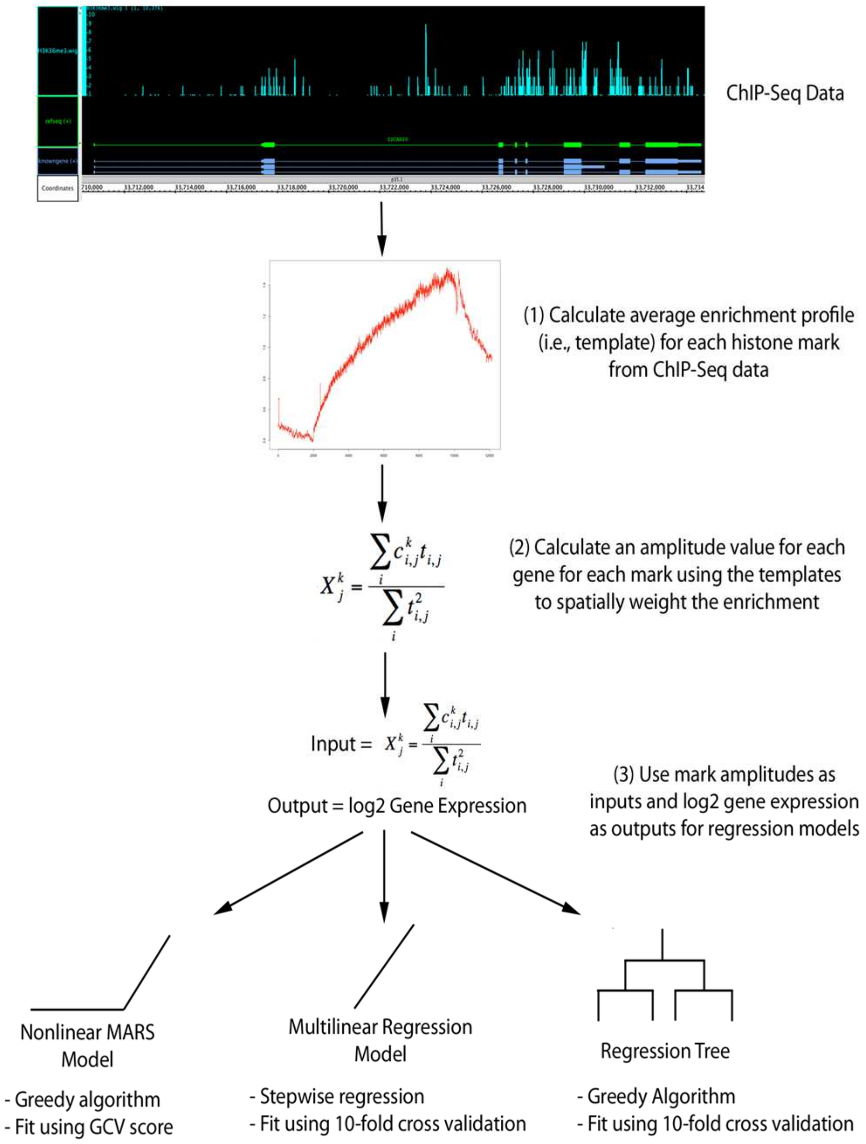


Machine Learning Methods and Results

Questions

- Why does the cell require ~100 histone modifications to maintain two states (open or closed chromatin)?
  - Histone Code Hypothesis (David Allis)
- Are two-body, three-body,… interaction terms strong, dominant given evidence of multivalent readers and writers of histone modifications?
- Can predictive modeling recapitulate known and uncover unknown dependencies among histone modifications?

(Machine) Learning the Histone Code



ChIP-Seq Data

(1) Calculate average enrichment profile (i.e., template) for each histone mark from ChIP-Seq data

$$X_j^k = \frac{\sum_i c_{i,j}^k t_{i,j}}{\sum_i t_{i,j}^2}$$

(2) Calculate an amplitude value for each gene for each mark using the templates to spatially weight the enrichment

Input = $X_j^k = \frac{\sum_i c_{i,j}^k t_{i,j}}{\sum_i t_{i,j}^2}$

Output = log2 Gene Expression

(3) Use mark amplitudes as inputs and log2 gene expression as outputs for regression models

Nonlinear MARS Model

- Greedy algorithm
- Fit using GCV score

Multilinear Regression Model

- Stepwise regression
- Fit using 10-fold cross validation

Regression Tree

- Greedy Algorithm
- Fit using 10-fold cross validation

Estimating Amplitudes

- Normalize Templates: $\frac{1}{N}\sum_{i=1}^{N} t_{i,j} = 1$

- Least Squares Sum: $Q_j^k = \sum_i \left(c_{i,j}^k - X_j^k t_{i,j} \right)^2$

- Least Squares Solution: $X_j^k = \frac{\sum_i c_{i,j}^k t_{i,j}}{\sum_i t_{i,j}^2}$

- Solution for Uniform Template: $X_j^k = \frac{1}{N}\sum_i c_{i,j}^k$

Stepwise Linear Regression

Model: $Y^k = \beta_0 + \sum_{j} \beta_j X_j^k + \sum_{j<l} \beta_{j,l} X_j^k X_l^k + \sum_{j<l<m} \beta_{j,l,m} X_j^k X_l^k X_m^k + \varepsilon^k$

1. Fit the initial (seed) model
2. Use F-test to calculate p-values of potential additional term from pool of candidate terms.
   - If min(p) < 0.05, add term.
   - Repeat until no terms with p < 0.05.
   - If min(p) > 0.05, go to step 3.
3. If any terms in model have p > 0.05, remove term with max p-value and go to step 2; otherwise, end.

Multilinear Model Terms

- Ran stepwisefit 100 times with randomly seeded models.
- Required terms to appear in 35% of the models: 167 core terms.
- 10-fold cross validation with seed model including 167 core + 60 random terms.
- Ran each fold 10 times and selected the model with best test MSE.
- Pruned model using F-test on test data.
- Required a term to appear in ≥ 5 folds: 24 terms total in final model.

Term	B	Z-score	p-value	Impact
H3K79me1	6.74	18.23	0	1.33
H3K36me3	4.09	17.80	0	0.92
H3K79me3	3.08	23.91	0	0.60
H4K20me1	0.977	21.44	0	0.45
H3K4me2-H3R2me1	18.27	7.85	1.66e-15	0.44
H3K4me2*-H3K9me1*	-1.58	-3.34	3.5e-4	-0.12
H4R3me2	-11.12	-13.23	0	-0.30
H3K27me2*-H3K36me3	-31.77	-8.91	0	-0.31
H3K27me2*-H3K79me1	-56.54	-9.53	0	-0.45
H3R2me1	-11.94	-16.50	0	-0.60

Multivariate Adaptive Regression Splines

$$Y^k = c_0 + \sum_{i=1}^{n} c_i b_i(\vec{X}_i^k)$$

$b_i(\cdot)$ is a basis function that is made up of either one or a product of two or more hinge functions. Hinge functions are splines that take on the form $h(X_j^k) = \max\left(0, X_j^k - X_j^{k^*}\right)$ or $h(X_j^k) = \max\left(0, X_j^{k^*} - X_j^k\right)$ where $X_j^{k^*}$ is a special constant known as a *knot*.

Model Built in One Forward and One Reverse Pass:

Forward Pass:

(1) Add Intercept = mean of response variable Y^k

(2) Add 1-body terms which maximally reduce RSS at every step

(3) Add 2-body terms (constrained to include one of the 1-body terms) which maximally reduce the RSS at every step

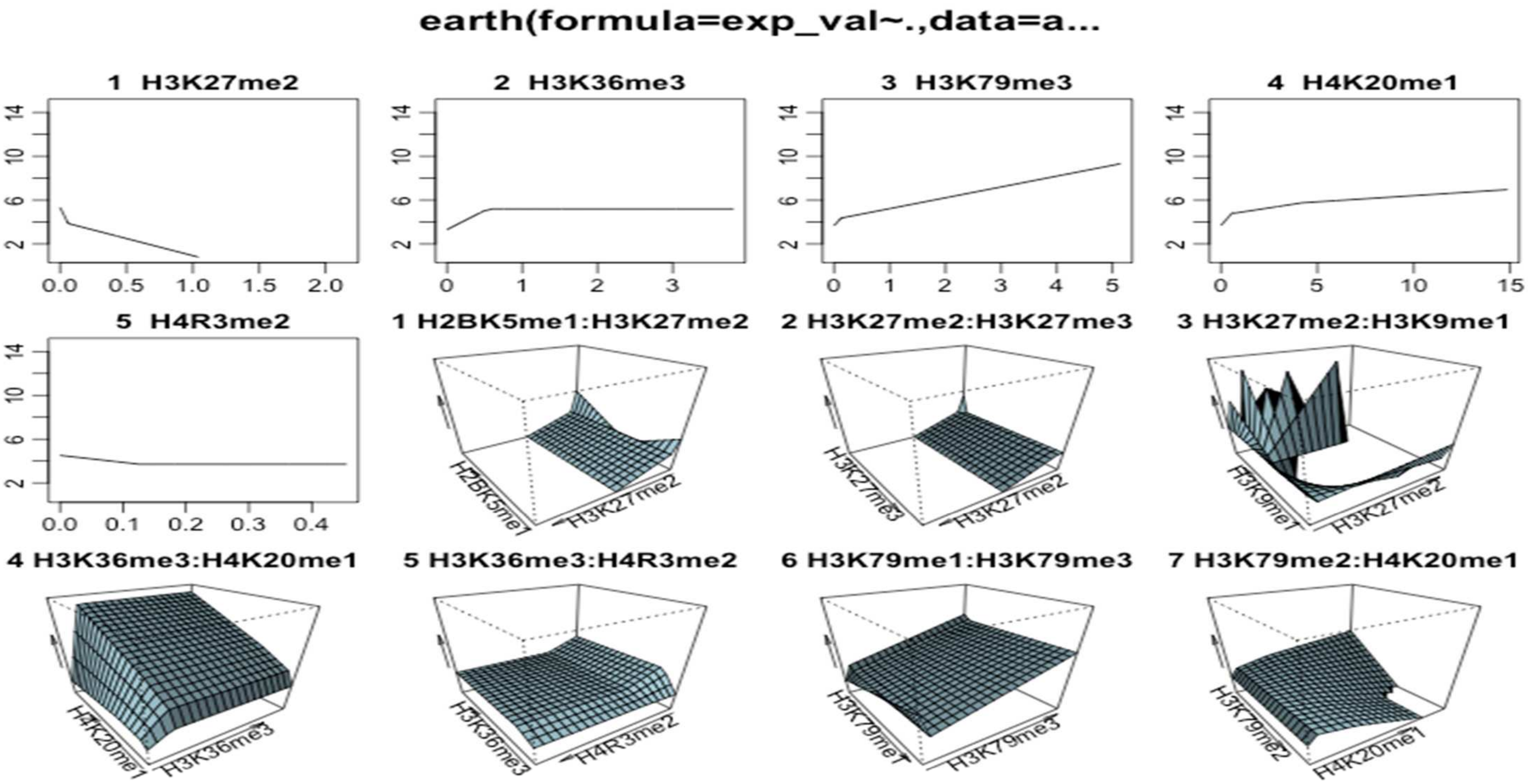
(4) Add 3-body terms (constrained to include one of the 2-body terms) which maximally reduce the RSS at every step

Reverse Pass:

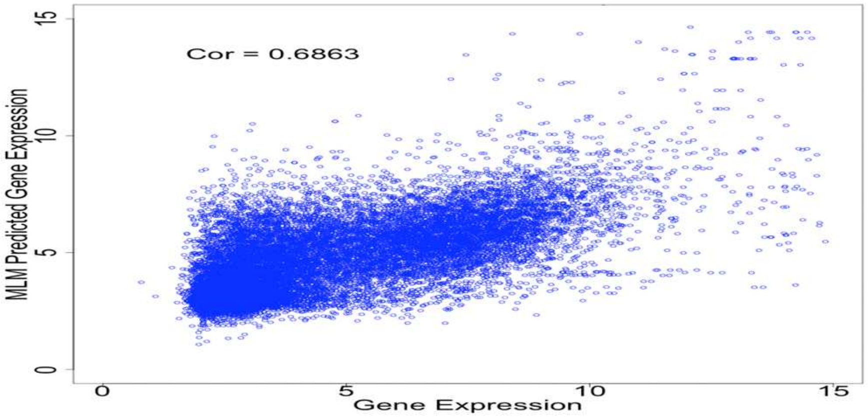
(1) Remove terms to optimize a Generalized Cross Validation Score which penalizes model complexity by dividing the RSS by the effectve number of degrees of freedom in the model

Final Model Contains 24 terms.

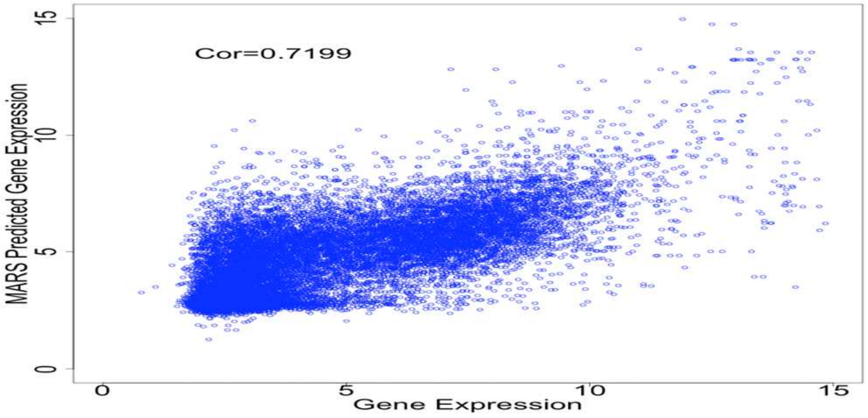
MARS Model Terms



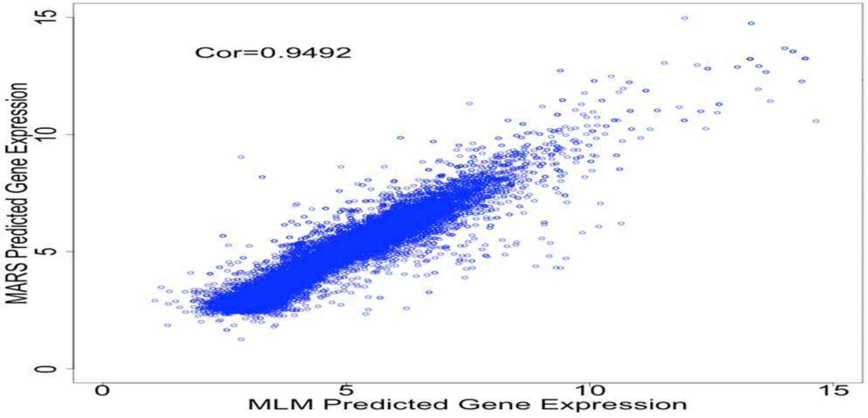
Scatter Plots



Multilinear Model vs Gene Expression
(24 terms in model)

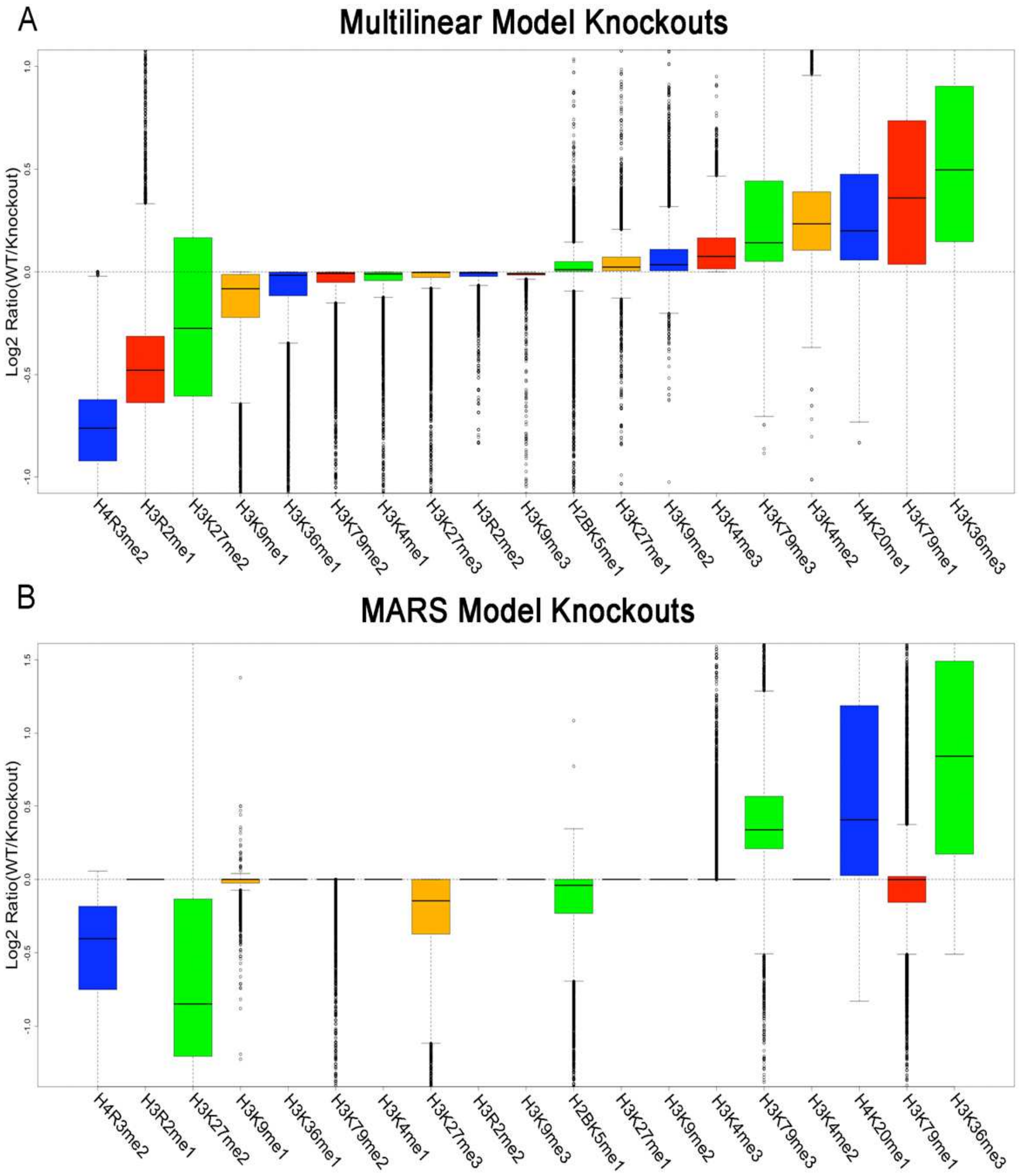


MARS Model vs Gene Expression
(24 terms in model)



MARS vs Multilinear Model

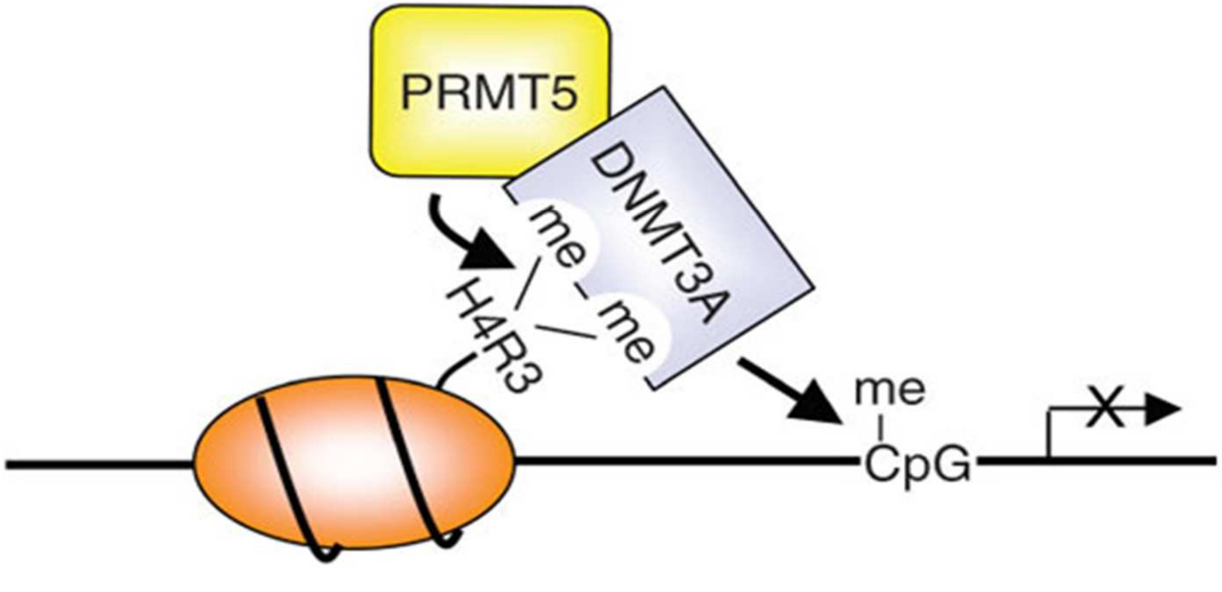
In-silico Knockout Predictions



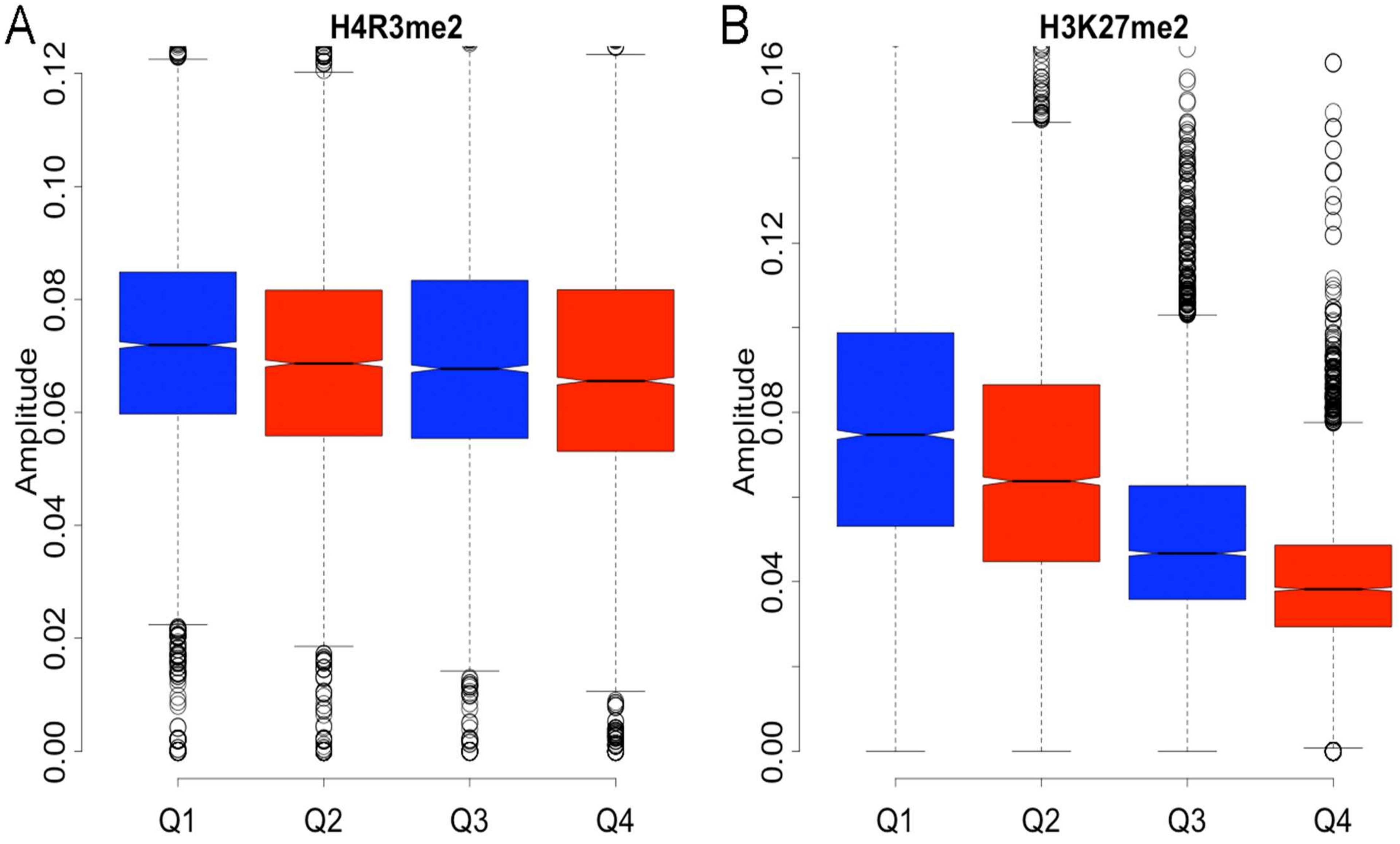
PRMT5-mediated methylation of histone H4R3 recruits DNMT3A, coupling histone and DNA methylation in gene silencing

Quan Zhao[1,2,8], Gerhard Rank[1,8], Yuen T Tan[1], Haitao Li[3], Robert L Moritz[4], Richard J Simpson[4], Loretta Cerruti[1], David J Curtis[1], Dinshaw J Patel[3], C David Allis[5], John M Cunningham[6] & Stephen M Jane[1,7]

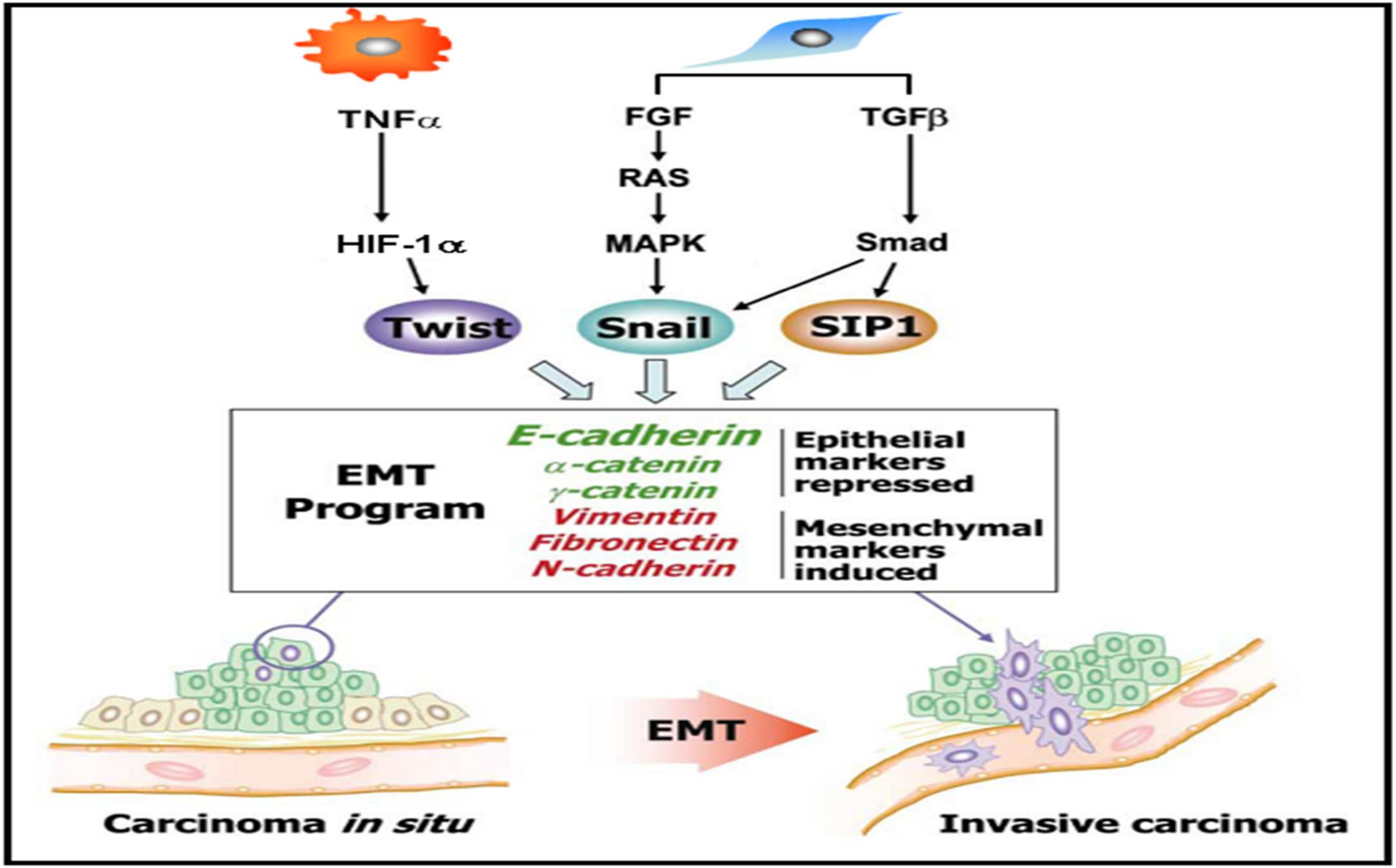
Mammalian gene silencing is established through methylation of histones and DNA, although the order in which these modifications occur remains contentious. Using the human β-globin locus as a model, we demonstrate that symmetric methylation of histone H4 arginine 3 (H4R3me2s) by the protein arginine methyltransferase PRMT5 is required for subsequent DNA methylation. H4R3me2s serves as a direct binding target for the DNA methyltransferase DNMT3A, which interacts through the ADD domain containing the PHD motif. Loss of the H4R3me2s mark through short hairpin RNA–mediated knockdown of PRMT5 leads to reduced DNMT3A binding, loss of DNA methylation and gene activation. In primary erythroid progenitors from adult bone marrow, H4R3me2s marks the inactive methylated globin genes coincident with localization of PRMT5. Our findings define DNMT3A as both a reader and a writer of repressive epigenetic marks, thereby directly linking histone and DNA methylation in gene silencing.



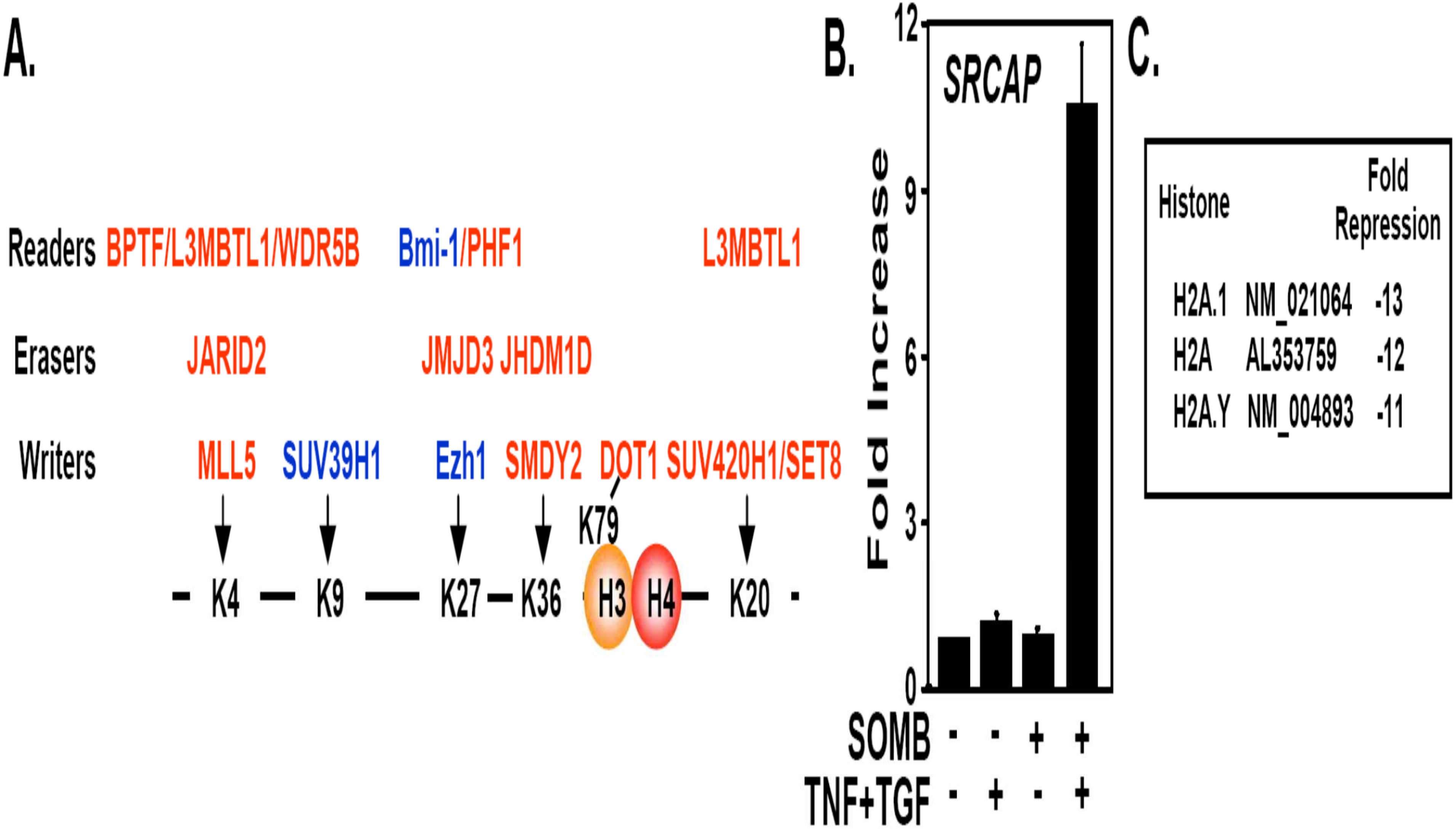
Amplitude vs Gene Expression Levels



Epithelial to Mesenchymal Transition



~5,000 Genes Differentially Expressed during EMT including Multiple Readers, Writers and Erasers of Histone Modifications



Histone		Fold Repression
H2A.1	NM_021064	-13
H2A	AL353759	-12
H2A.Y	NM_004893	-11

Summary

- Multilinear, MARS and Regression Tree Models recapitulate known/univariate trends
  - H3K36me3, H3K79me3, H3K4me3, H4K20me1 activating
  - H3K27me2,3 repressive
- They predict novel synergies:
  - H3K36me3-H3K79me1,3-H4K20me1
- They reveal activating/repressive potential of marks not observed in univariate analysis
  - H3K79me1 activating
  - H4R3me2 repressive
- H4R3me2s (PRMT5 writer) shown to be globally repressive from a number of experimental studies
  - Required for DNMT3A mediated DNA methylation and gene silencing
  - Knockdown of PRMT5 yield more de-repressed than repressed genes

Acknowledgements

Bekiranov Lab

Xiaojiang Xu
Stephen Hoang
Kunal Poorey
Veena Valsakumar

Computational Biologists

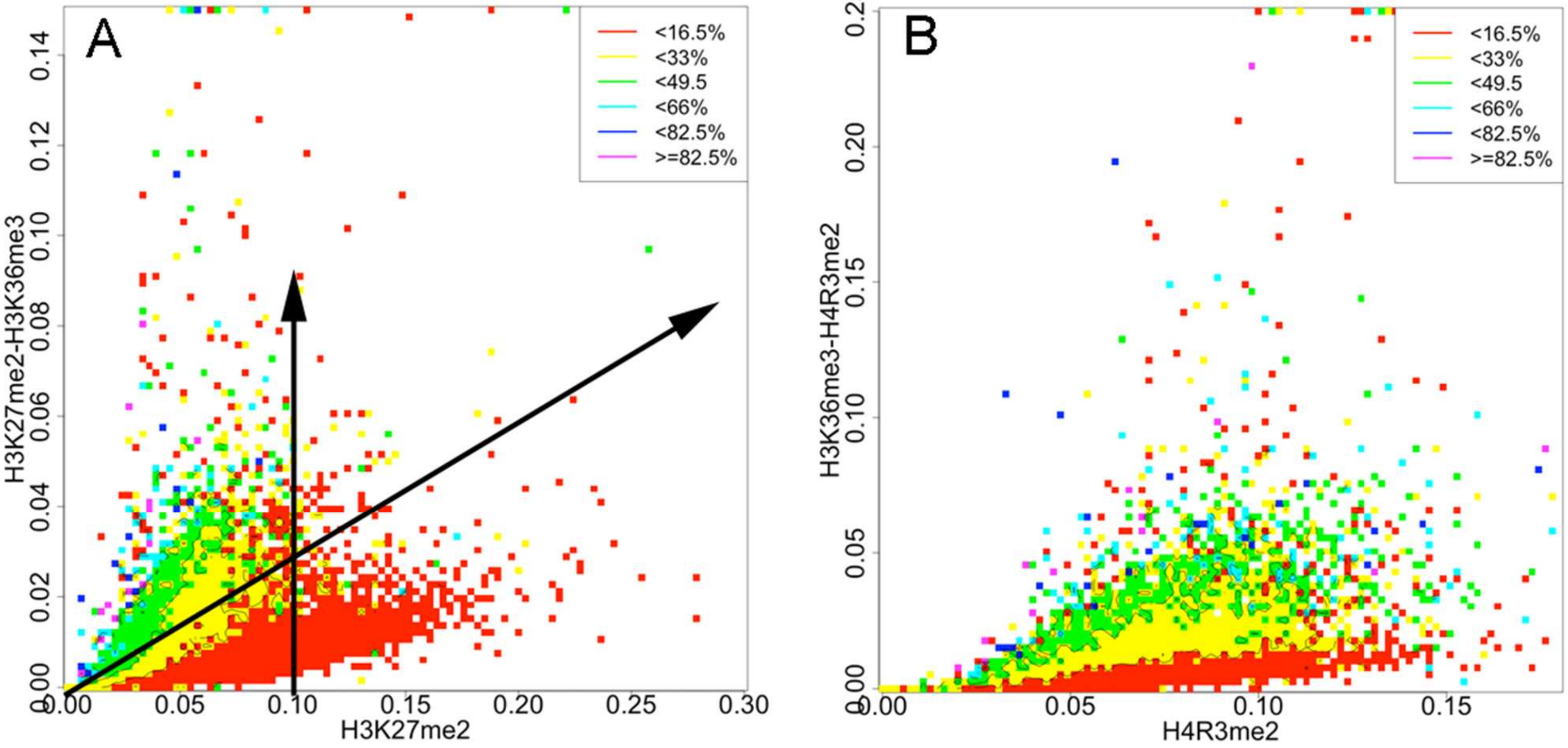
Bill Pearson
Aaron Quinlan
Royden Clark

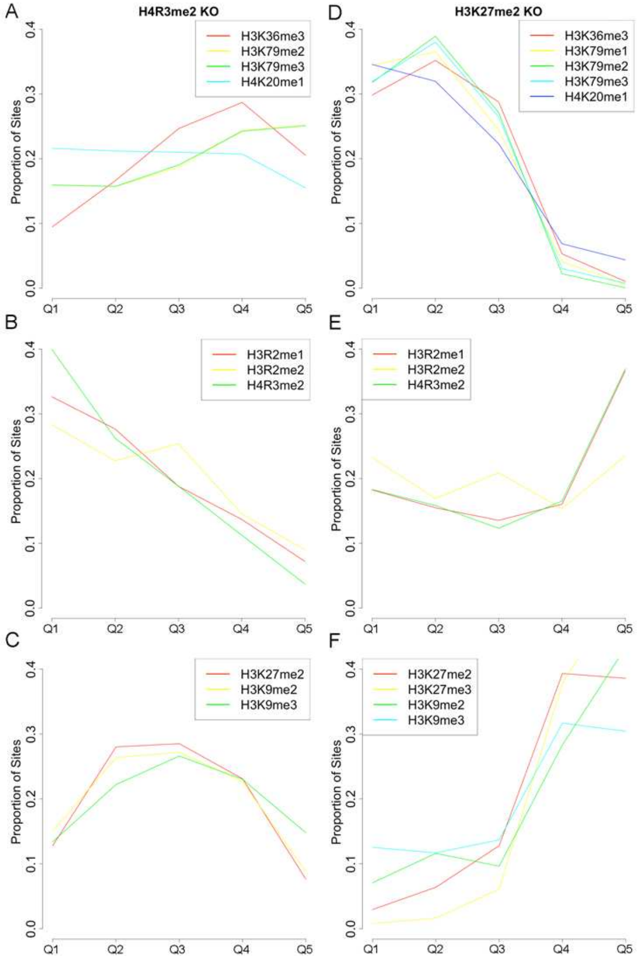
Collaborators

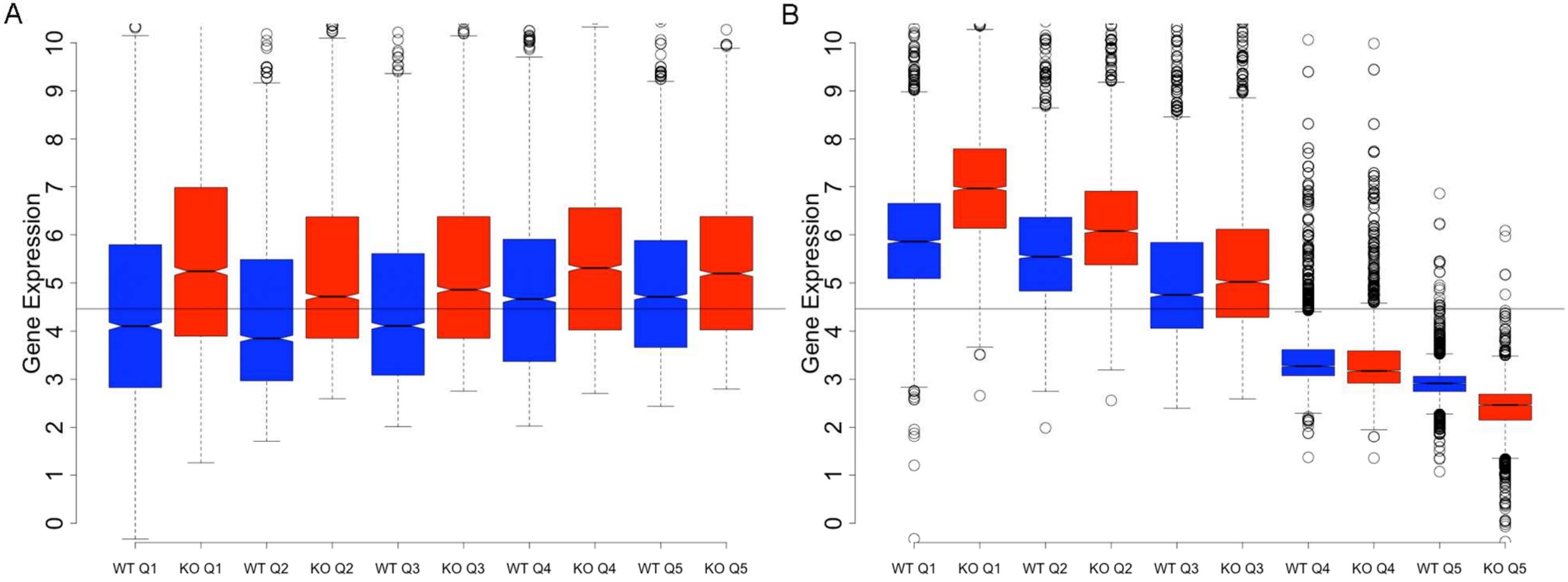
Marty Mayo
Manish Kumar
Natalya N. Baranova
Mayo Lab
Patrick Grant
Mitch Smith
David Auble
Jeff Smith

Consultation

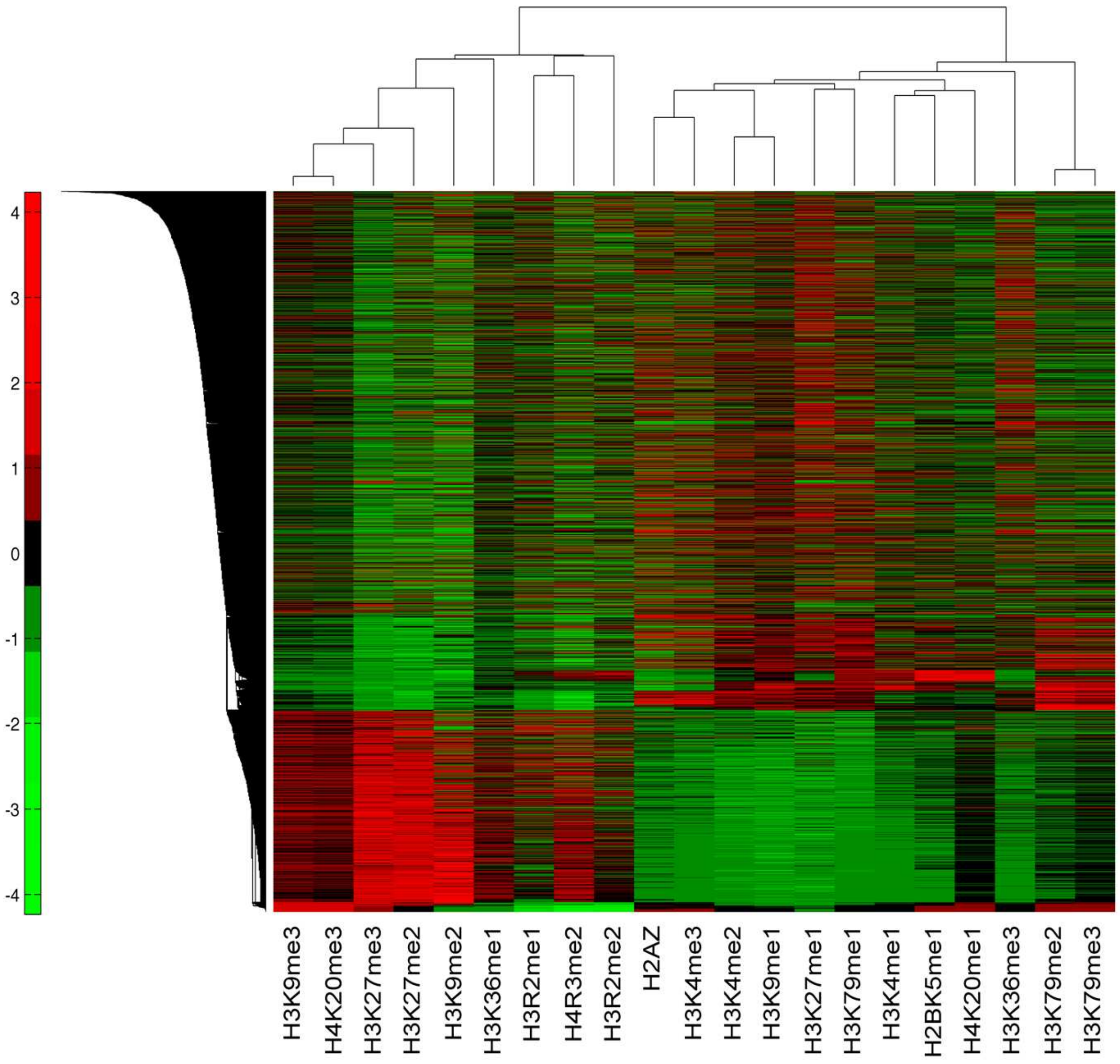
Sepideh Khorasanizadeh

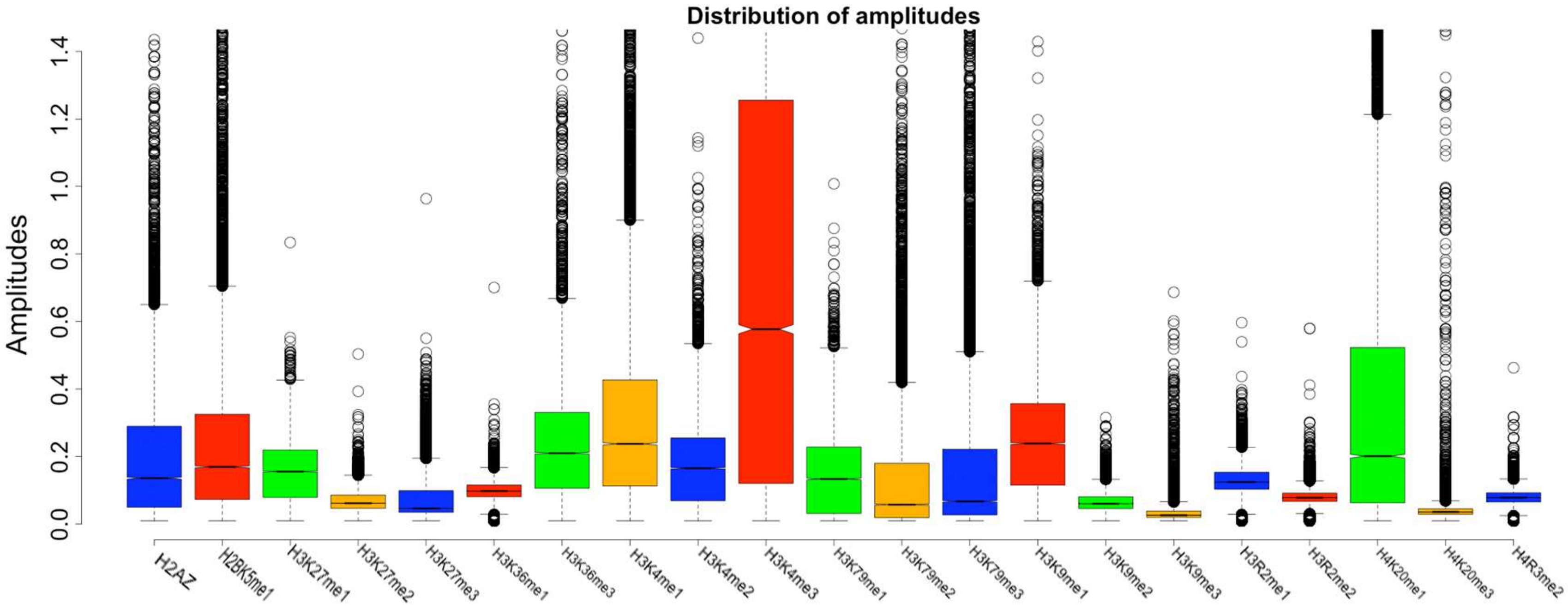

A
H3K27me2-H3K36me3
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.00
0.05
0.10
0.15
0.20
0.25
0.30
H3K27me2
<16.5%
<33%
<49.5
<66%
<82.5%
>=82.5%
B
H3K36me3-H4R3me2
0.00
0.05
0.10
0.15
0.20
0.25
0.00
0.05
0.10
0.15
H4R3me2
<16.5%
<33%
<49.5
<66%
<82.5%
>=82.5%


A
H4R3me2 KO
D
H3K27me2 KO
B
E
C
F
Proportion of Sites
0.0
0.1
0.2
0.3
0.4
Q1
Q2
Q3
Q4
Q5
H3K36me3
H3K79me2
H3K79me3
H4K20me1
H3K36me3
H3K79me1
H3K79me2
H3K79me3
H4K20me1
H3R2me1
H3R2me2
H4R3me2
H3R2me1
H3R2me2
H4R3me2
H3K27me2
H3K9me2
H3K9me3
H3K27me2
H3K27me3
H3K9me2
H3K9me3


A
B
Gene Expression
0
1
2
3
4
5
6
7
8
9
10
WT Q1
KO Q1
WT Q2
KO Q2
WT Q3
KO Q3
WT Q4
KO Q4
WT Q5
KO Q5

Cluster Amplitudes: Genes x Marks




Distribution of amplitudes
Amplitudes
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
H2AZ
H2BK5me1
H3K27me1
H3K27me2
H3K27me3
H3K36me1
H3K36me3
H3K4me1
H3K4me2
H3K4me3
H3K79me1
H3K79me2
H3K79me3
H3K9me1
H3K9me2
H3K9me3
H3R2me1
H3R2me2
H4K20me1
H4K20me3
H4R3me2

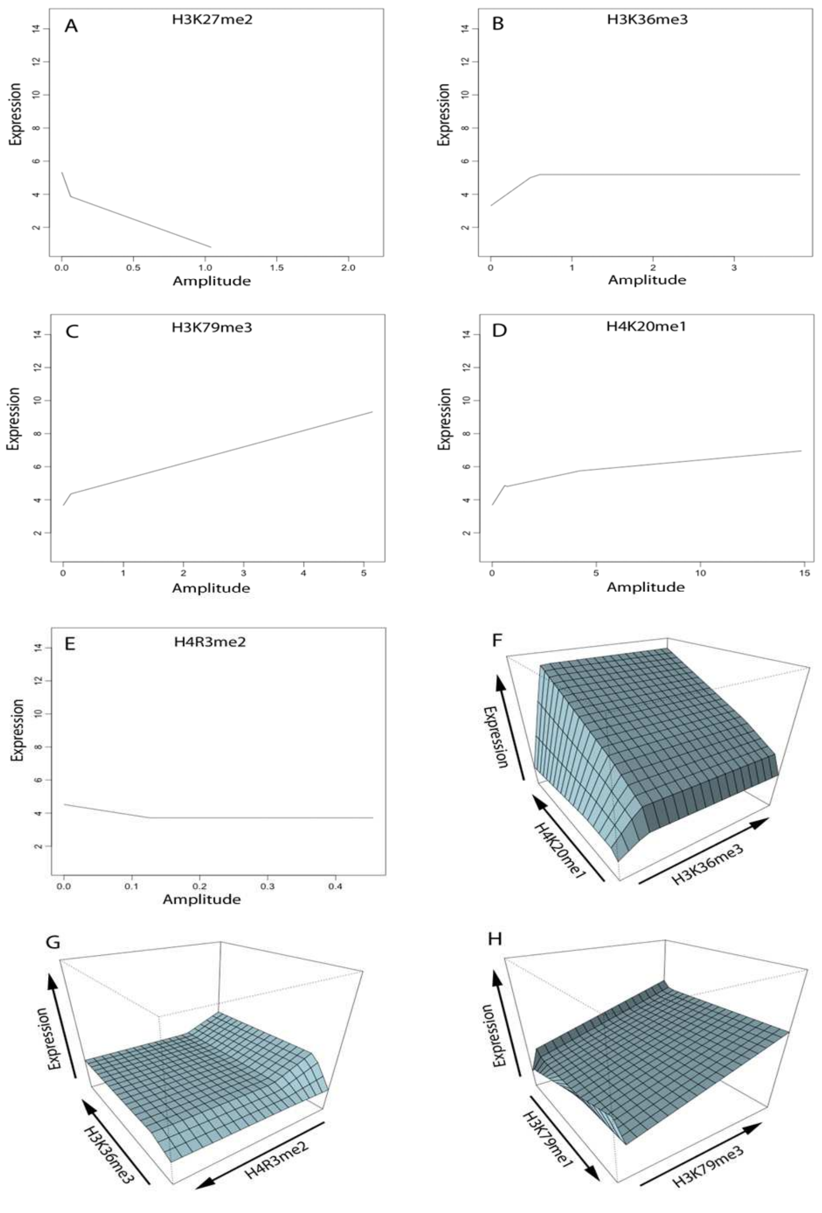
Calculating Enrichment Amplitudes

- Fixed Enrichment Profile Model
  - Average enrichment profile is characteristic of mark/gene boundary
  - Amplitude, not shape, varies across genes
  - Calculate amplitude from shape weighted average of read depth
- Gene Dependant Enrichment Profile Model
  - Use Unsupervised Learning Methods (PCA, Clustering Analysis, SOM) to identify enrichment profiles classes
  - Calculate amplitude from shape weighted average of read depth for each profile class
- Annotation Based Estimate
  - Calculate Average read depth in various functional elements: promoters, 5’ UTR, exons, introns, 3’UTR
  - Use least squares to weight/summarize vector into one amplitude
- Significantly Enriched Sites Approach
  - Identify significantly enriched sites (at 5% False Discovery Rate)
  - Estimate Amplitude using average, median, or max within site
  - Cases below cutoff: Impute positions of sites from relative position of significant sites within genes & estimate amplitude

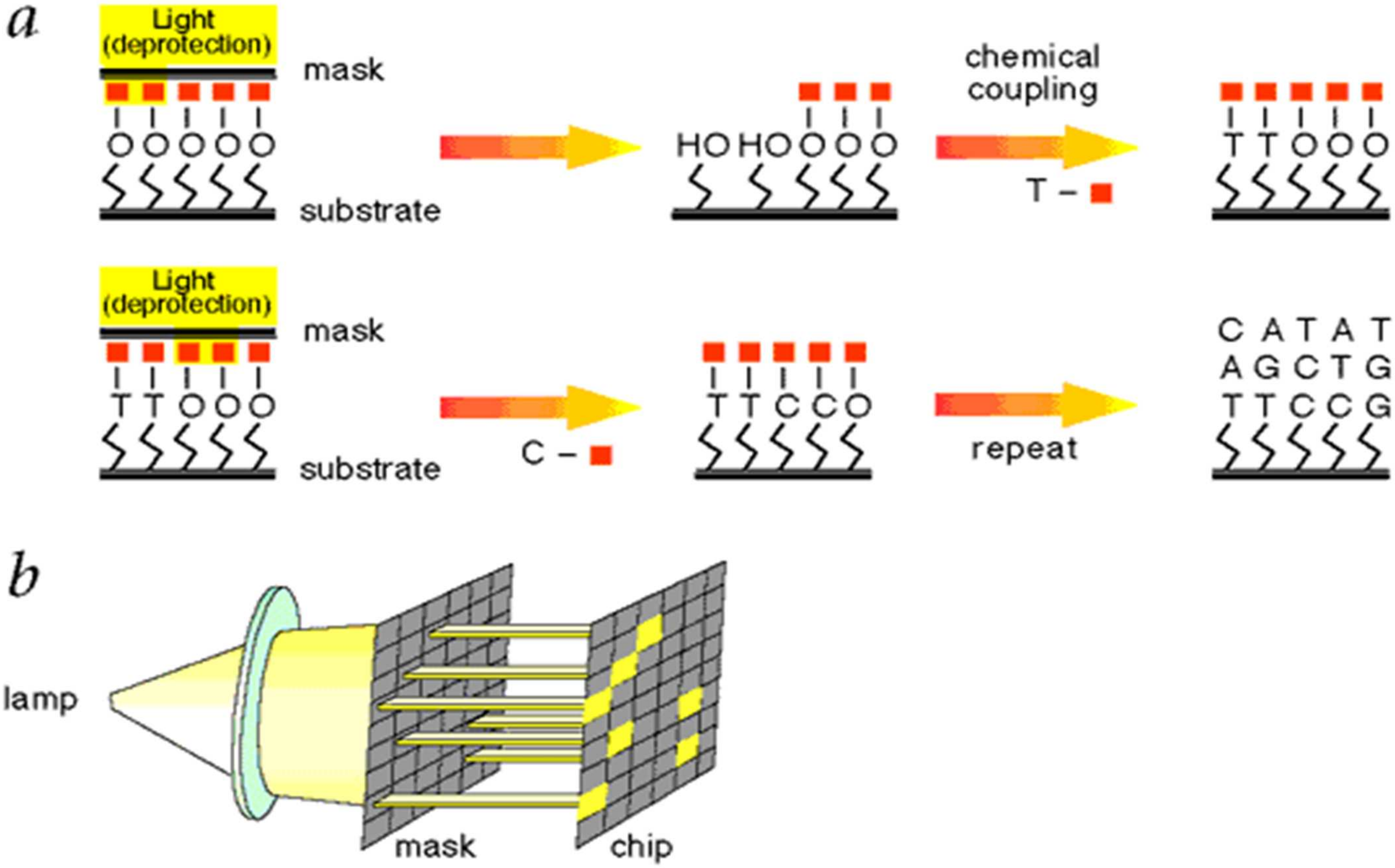
Stepwise Regression

1. Fit the initial model
2. Use F-test to calculate p-values of potential additional term from pool of candidate terms.
   - If min(p) < 0.05, add term.
   - Repeat until no terms with p < 0.05.
   - If min(p) > 0.05, go to step 3.
3. If any terms in model have p > 0.05, remove term with max p-value and go to step 2; otherwise, end.

MARS Model Terms



Synthetic Oligonucleotide Arrays (1991)


a
Light (deprotection)
mask
substrate
HO HO
chemical coupling
T –
T T O O O
Light (deprotection)
mask
T T O O O
substrate
C –
T T C C O
repeat
C A T A T
A G C T G
T T C C G
b
lamp
mask
chip