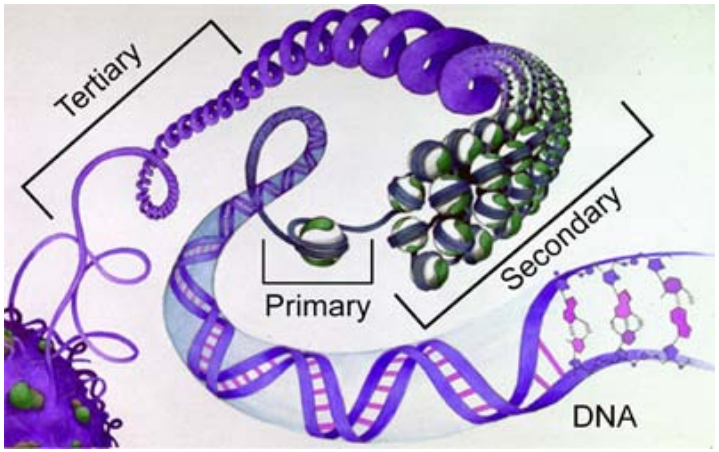
Computational epigenomics: from ChIP-seq data to histone modification patterns

Jianrong Wang

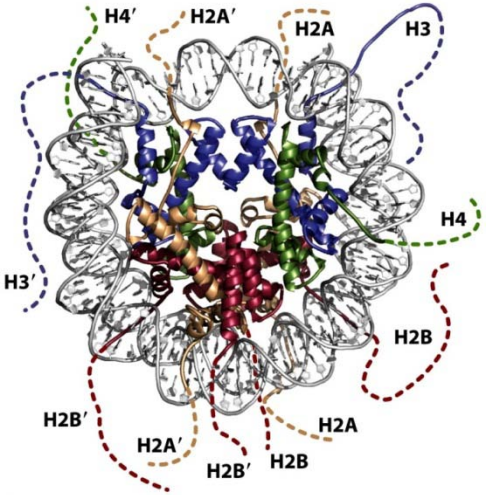
Ph.D Candidate of Bioinformatics
Georgia Institute of Technology

Epigenetics: "changes in gene expression or cellular phenotype caused by mechanisms other than changes in the underlying DNA sequences".

E.g. DNA methylation, histone modifications, nucleosome positioning.



Chromatin Structure



Nucleosome

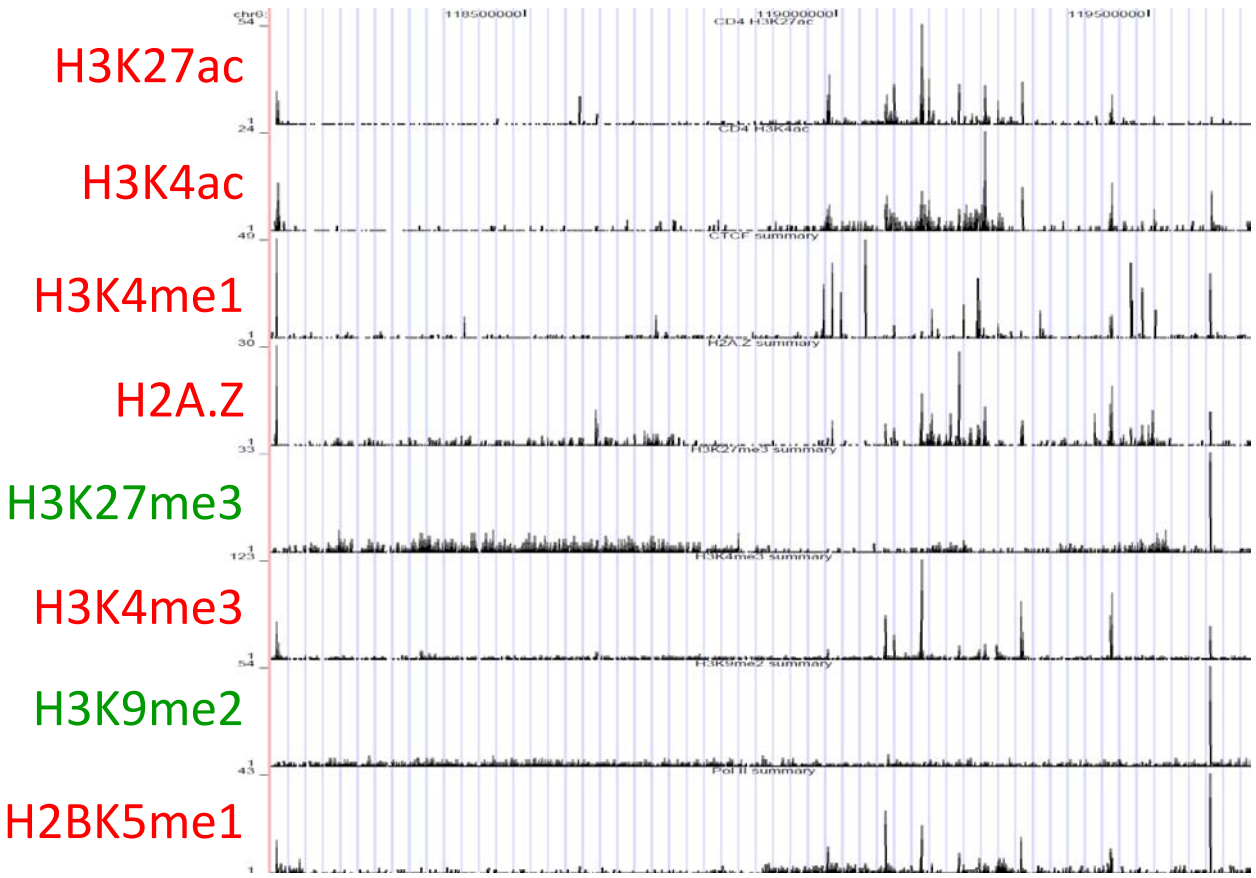
(methylation of H3 Lysine 4)
H3K4me1

H3K27me2
(di-methylation of H3 Lysine 27)

Histone Modifications

Genome-wide ChIP-seq histone modification maps enable systematic characterizations of interesting patterns.

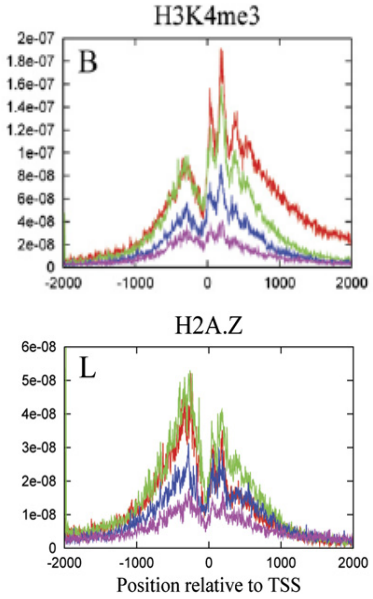
Human $CD4^+$ T cells: 38 histone modifications and 1 histone variant.



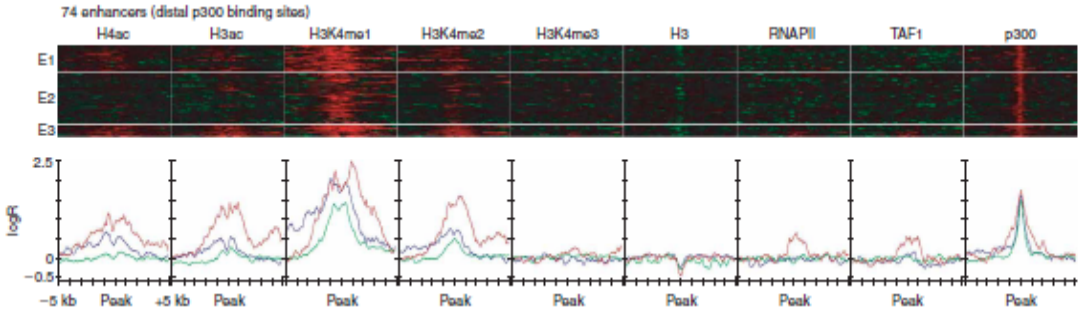
Epigenetic patterns and their significance

Histone modification patterns: Chromatin States

Ernst et al. 2007 Nature



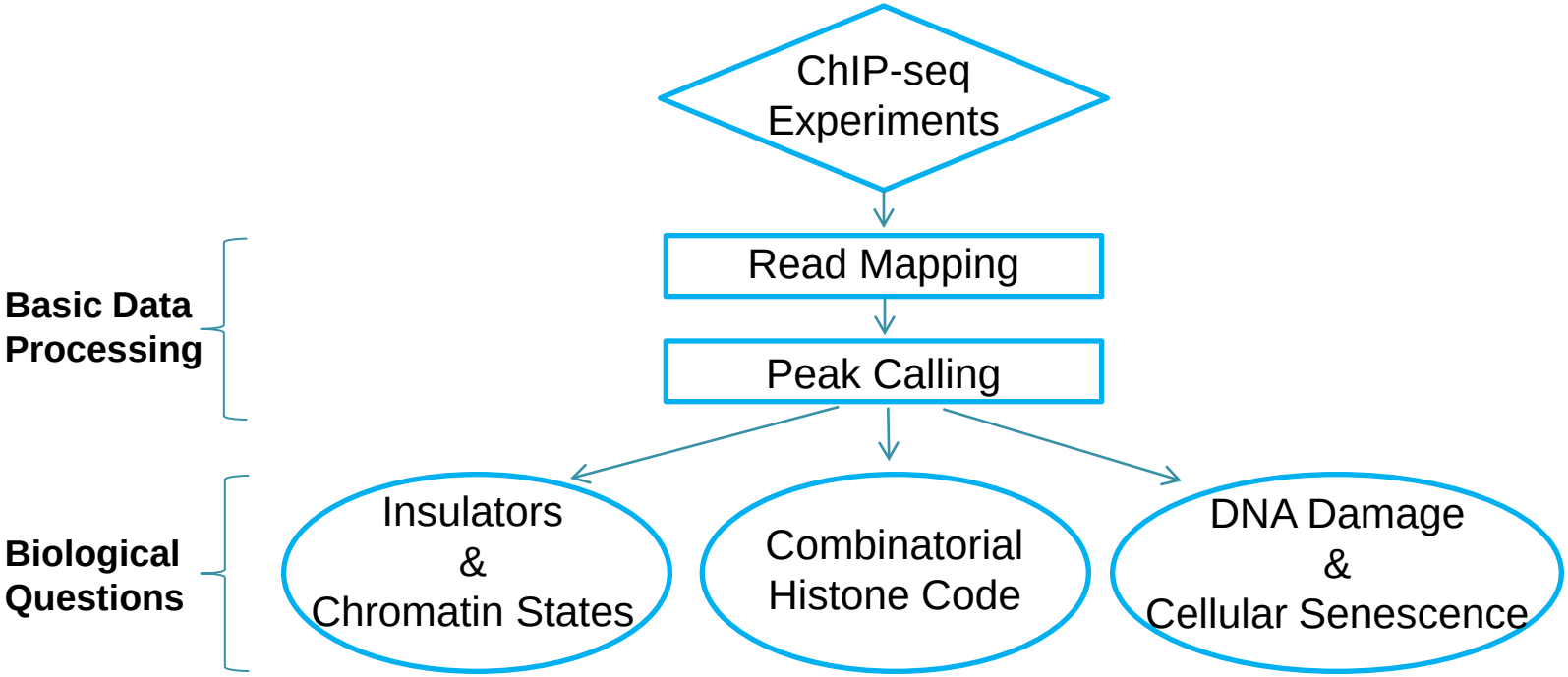
Barski et al. 2007 Cell

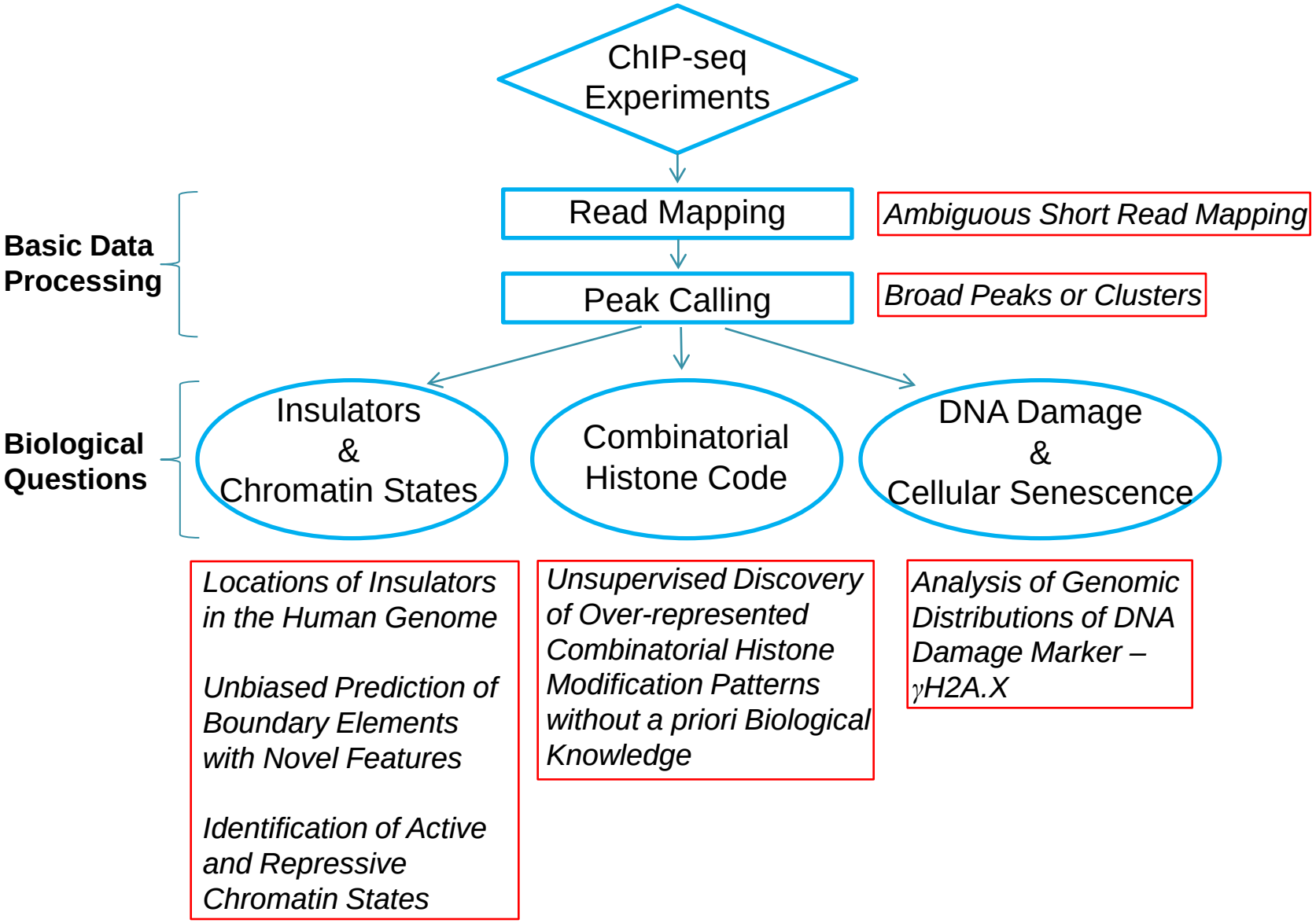


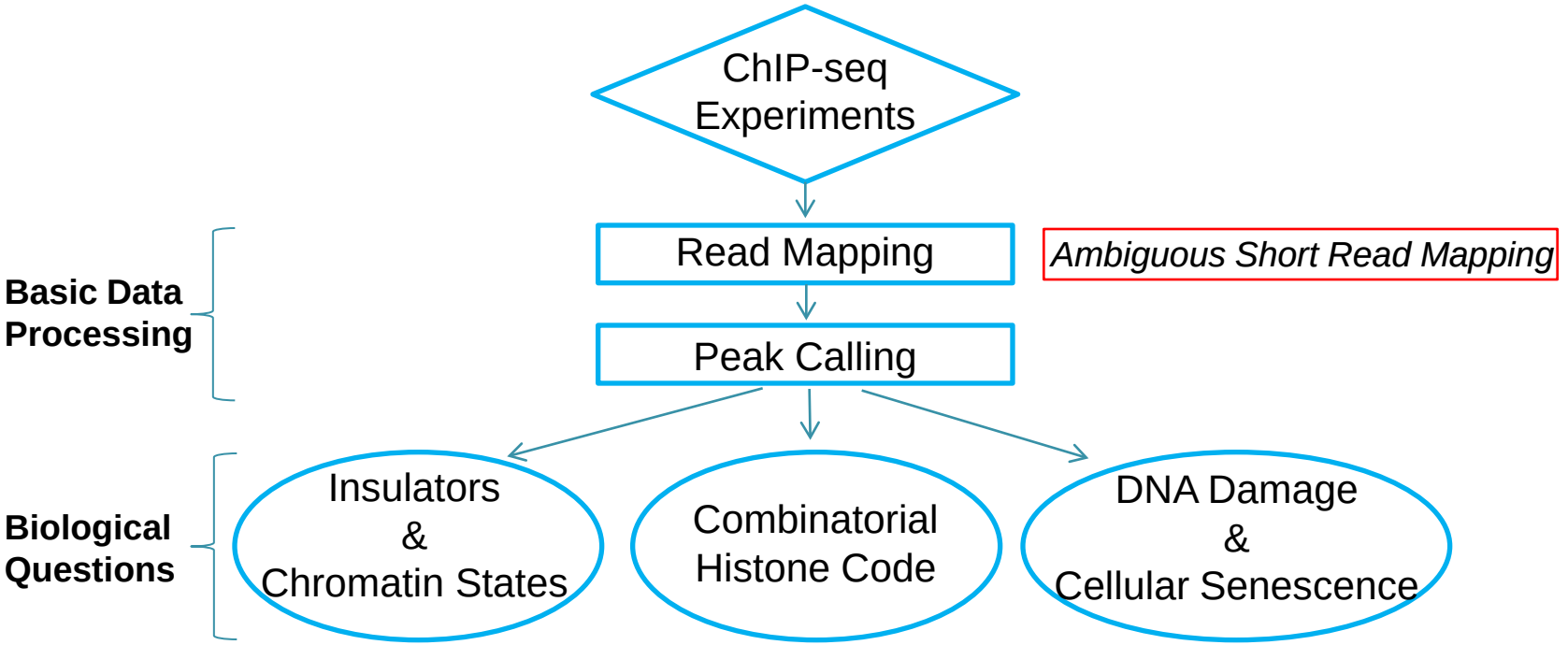
Heintzman et al. 2007
Nature genetics

Histone modification signatures are related with functional features

Computational Questions Related to Epigenetics



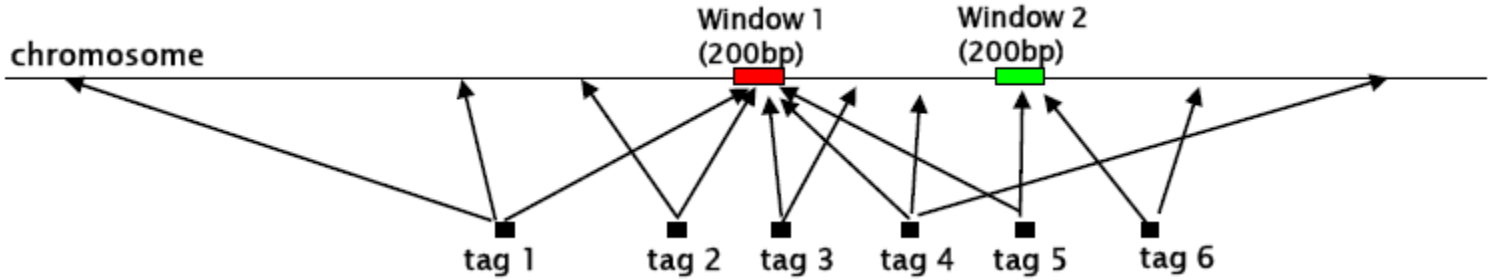

ChIP-seq Experiments
Read Mapping
Ambiguous Short Read Mapping
Peak Calling
Broad Peaks or Clusters
Basic Data Processing
Biological Questions
Insulators & Chromatin States
Combinatorial Histone Code
DNA Damage & Cellular Senescence
Locations of Insulators in the Human Genome
Unbiased Prediction of Boundary Elements with Novel Features
Identification of Active and Repressive Chromatin States
Unsupervised Discovery of Over-represented Combinatorial Histone Modification Patterns without a priori Biological Knowledge
Analysis of Genomic Distributions of DNA Damage Marker – γH2A.X


ChIP-seq
Experiments
Read Mapping
Ambiguous Short Read Mapping
Peak Calling
Basic Data
Processing
Biological
Questions
Insulators
&
Chromatin States
Combinatorial
Histone Code
DNA Damage
&
Cellular Senescence

Ambiguous Short-sequence Read Mapping

1. ChIP-seq reads are short (~30bp), especially for transcription factor binding;

2. Unique reads & Ambiguous reads (reads that can be mapped to >1 genomic locations with similar sequence similarity);



3. Lose Information if simply ignoring ambiguous reads.

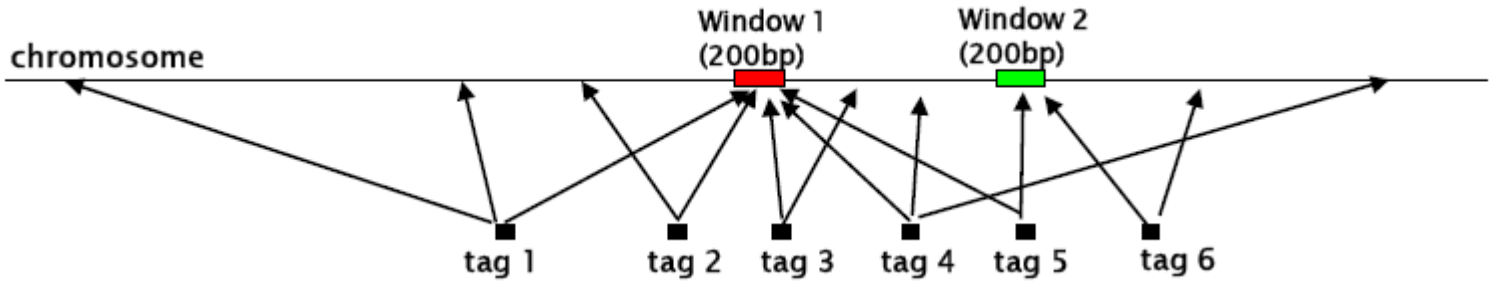
4. Mostly originate from repetitive genomic sequences: transposable elements, simple repeats and segmental duplications.

5. Especially for repeat-rich genomes (*e.g.* human)

Ambiguous Short-sequence Read Mapping

1. Basic Idea: borrow information from co-located reads to estimate the probability that a site is the real one.

2. Genomic sites with more co-located reads are more likely to be the correct sites than genomic sites with only a few co-located reads.



3. Based on the inferred probabilities, ambiguous reads could be stochastically mapped.

4. In order to estimate the probabilities, we need an existing mapping of reads.

5. Gibbs sampling: iteratively map the reads and update the probabilities.

Ambiguous Short-sequence Read Mapping

a_i : ambiguous tag *i;*
$S_i=\{s_{i1},s_{i2}, \ldots, s_{in_i}\}$: the set of totally n_i possible sites for a_i;
s_{ij} : the *j*th possible site for a_i
k_j : the tag count of the *j*th possible site for a_i;
$P_s(k)$: the tag count distribution for real sites;
$P_n(k)$: the tag count distribution for background (non-specific sites);

conditional probability that a_i should be mapped to s_{ij} , given that other reads are already mapped: $P(a_i \sim s_{ij} | M_{[-i]}, D)$

$$P(a_i \sim s_{ij} \mid M_{[-i]}, D) = \frac{P(a_i \sim s_{ij}, M_{[-i]} \mid D)}{P(M_{[-i]} \mid D)} = \frac{\{P_s(k_j+1) \prod_{m \in S_i \setminus j} P_n(k_m)\} \times P(U \setminus S_i)}{\sum_{\tau \in S_i} \left\{ P_s(k_\tau+1) \prod_{m \in S_i \setminus \tau} P_n(k_m) \right\} \times P(U \setminus S_i)} = \frac{\left(\frac{P_s(k_j+1)}{P_n(k_j)}\right)}{\sum_{\tau \in S_i} \frac{P_s(k_\tau+1)}{P_n(k_\tau)}}$$

Predictive formula = Normalized Likelihood Ratio

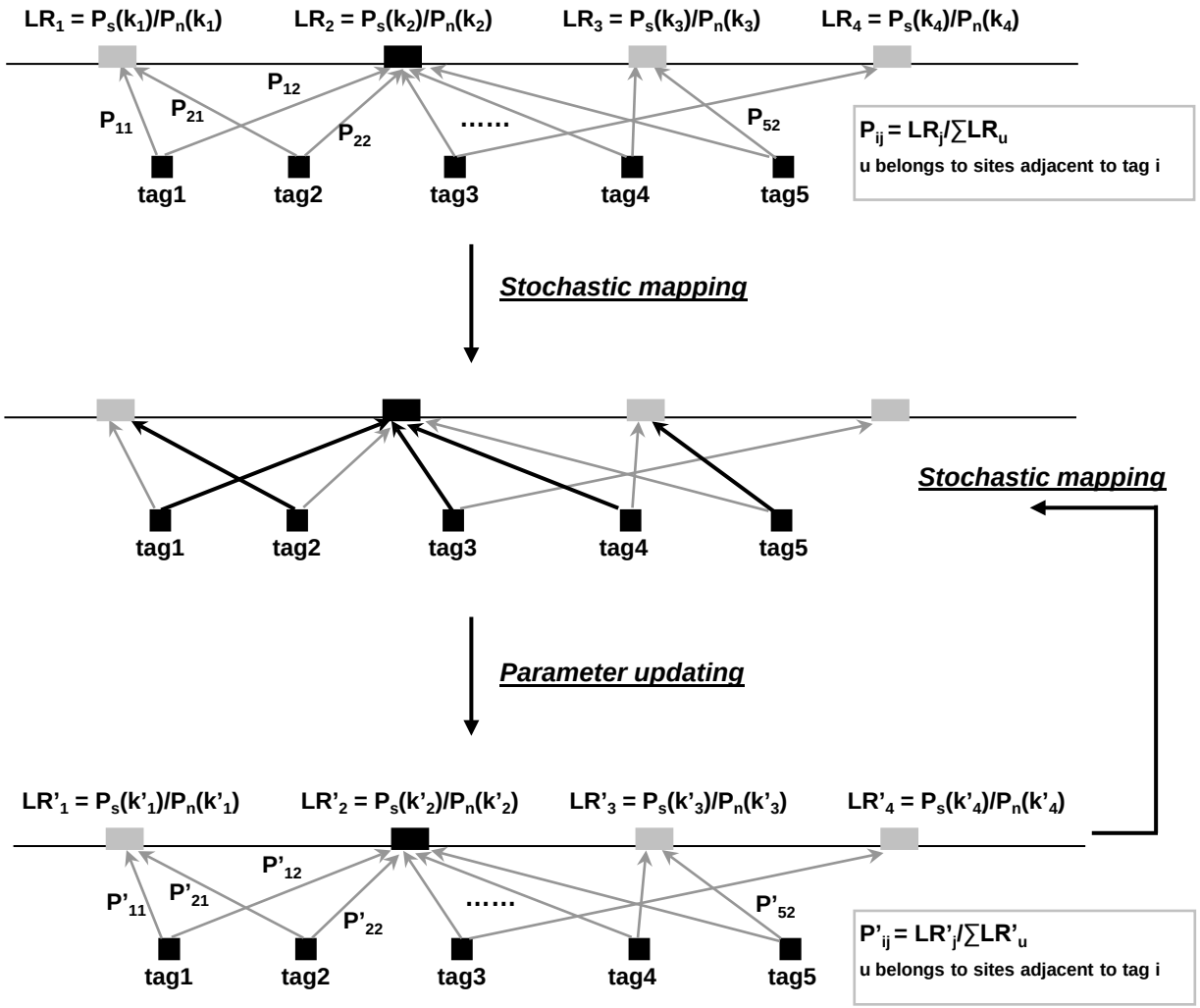
In practice, $P_s(k)$ is assumed to be normal and $P_n(k)$ is assumed to be Poisson:

$$P_s \sim N(\mu, \sigma^2)$$

$$P_n \sim \text{Poisson}(\lambda)$$

Wang et al. 2010 Bioinformatics

Ambiguous Short-sequence Read Mapping



Ambiguous Short-sequence Read Mapping

1. Through Gibbs sampling, we are searching in the space of all possible bipartite graphs to optimize the overall likelihood ratio.

$$\prod_{i \in U} \left(\frac{P_s(k_i)}{P_{n(k_i)}} \right)$$

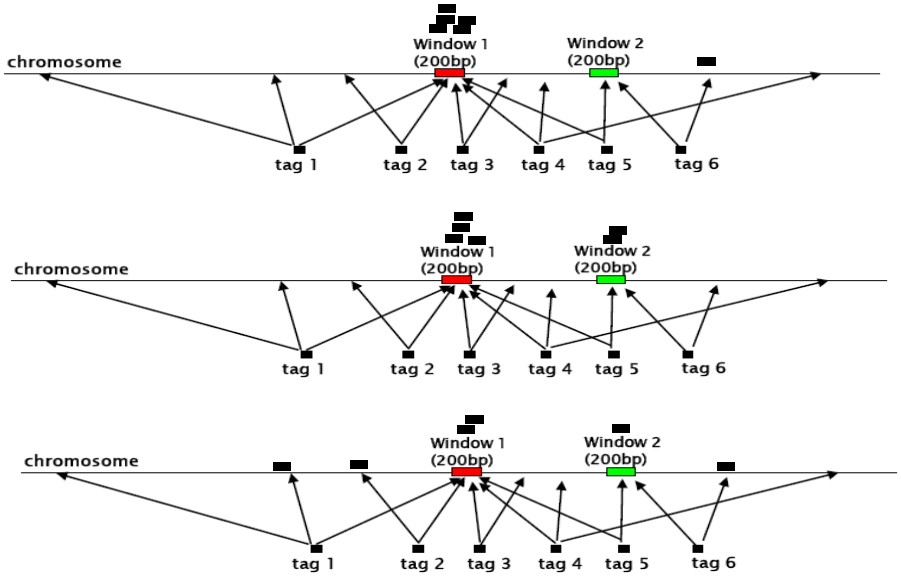
2. Co-located tags are more likely to be grouped than dispersed.

3. The overall likelihood ratios can be rewritten as $\prod_{\tau \in \sigma} \left(\frac{P_s(\tau)}{P_n(\tau)} \right)^{n(\tau)}$

4. So we are optimizing:

$$\sum_{\tau \in \sigma} \left(\frac{n(\tau)}{Z} \right) \log \left(\frac{P_s(\tau)}{P_n(\tau)} \right)$$

5. The final set σ consists with large numbers.



Ambiguous Short-sequence Read Mapping

Generation of simulated datasets

1. Generated totally 9 libraries of reads from sets of known genomic sites;

2. Parameters:
   a. read length;

   b. ChIP-seq specificity: fractions of reads from random genomic sites (noise);

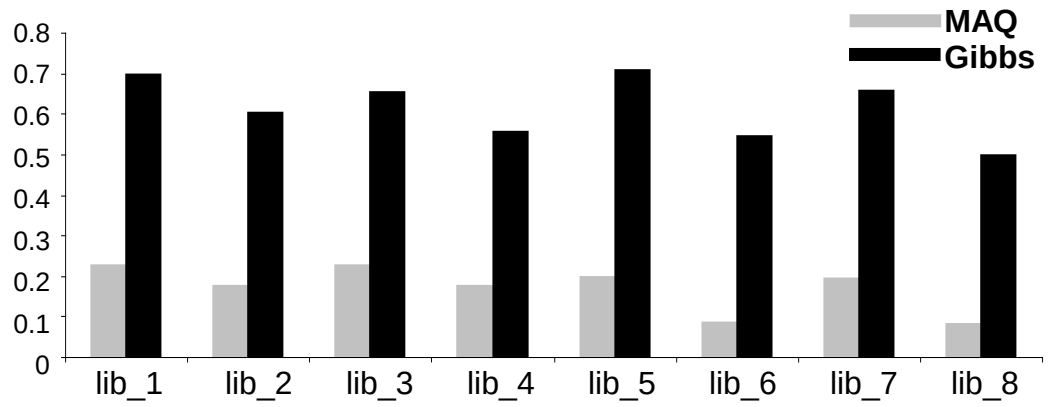
   c. sequencing errors;

   d. library size;

Ambiguous Short-sequence Read Mapping

Performance on simulated datasets

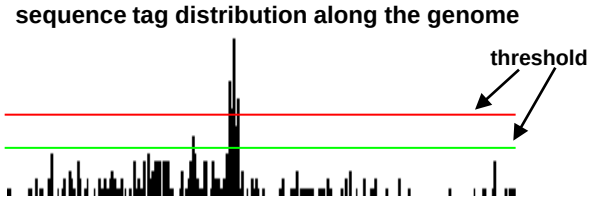
improvements on the fractions of correctly mapped ambiguous reads

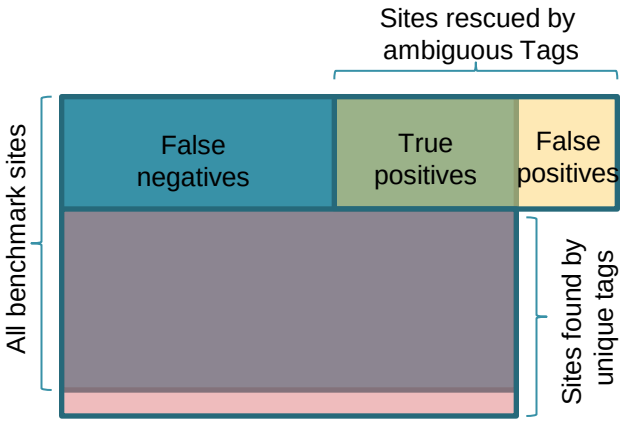


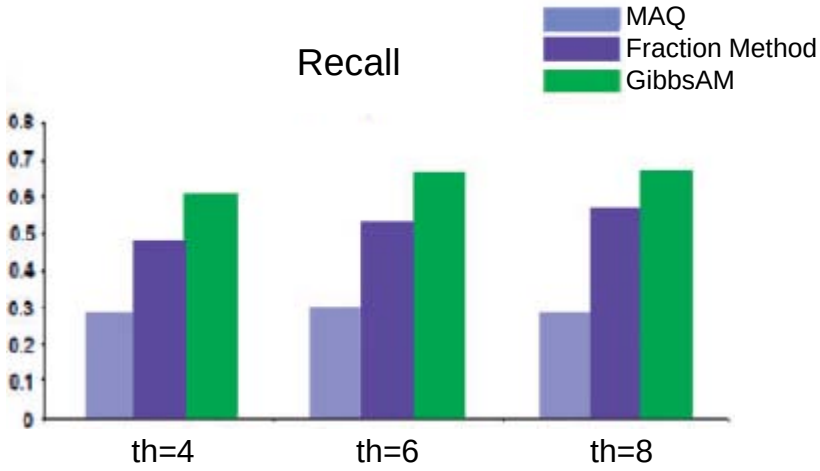
Ambiguous Short-sequence Read Mapping

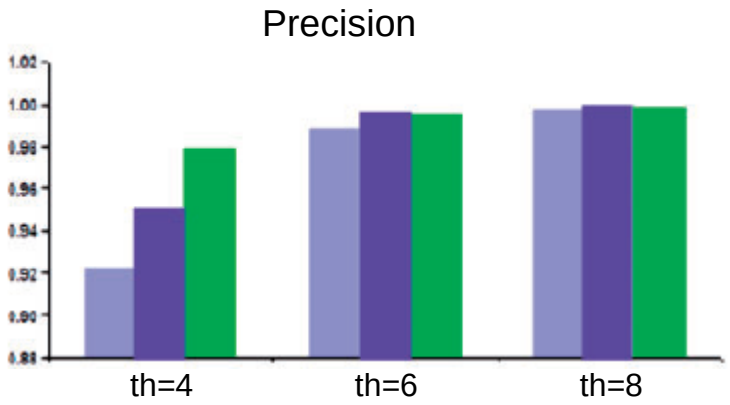
Performance on simulated datasets

improvements on correctly recovered genomic sites
(*F* scores keep the best in all tested datasets)



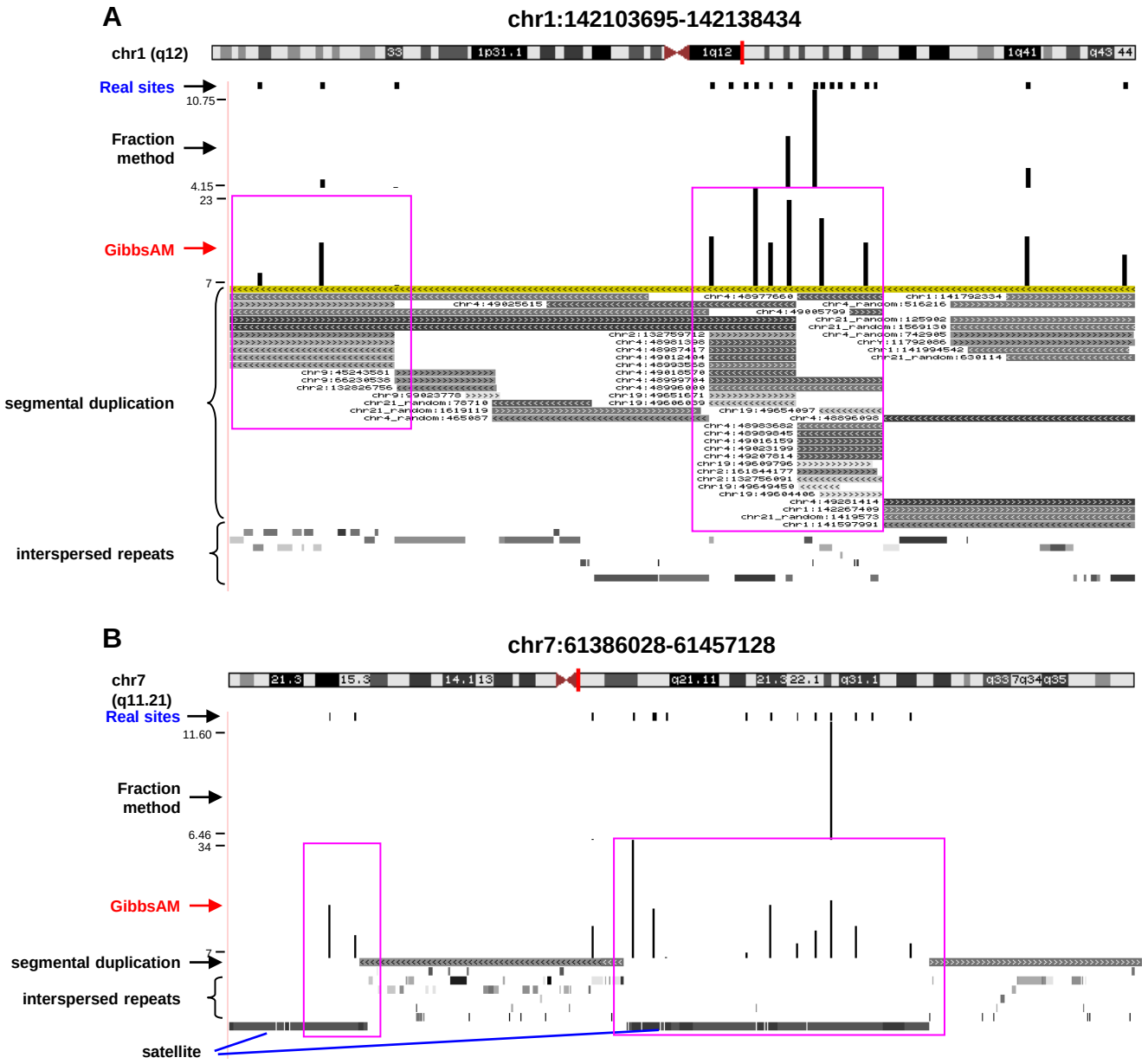






Ambiguous Short-sequence Read Mapping

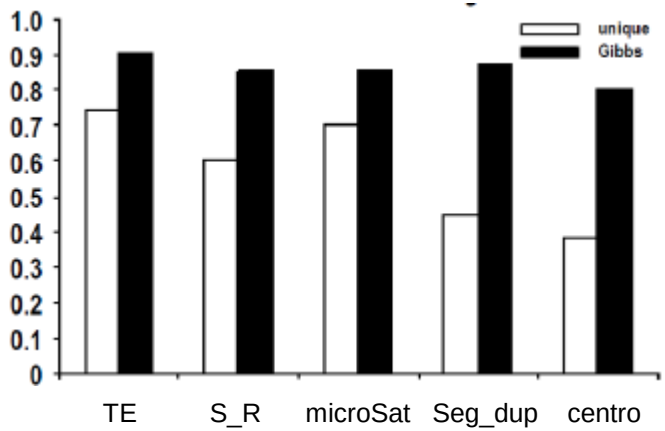
Examples of recovered genomic sites in repetitive regions



Ambiguous Short-sequence Read Mapping

The potential to recover biological information in repetitive genomic regions

Fractions of recovered sites in repetitive genomic regions



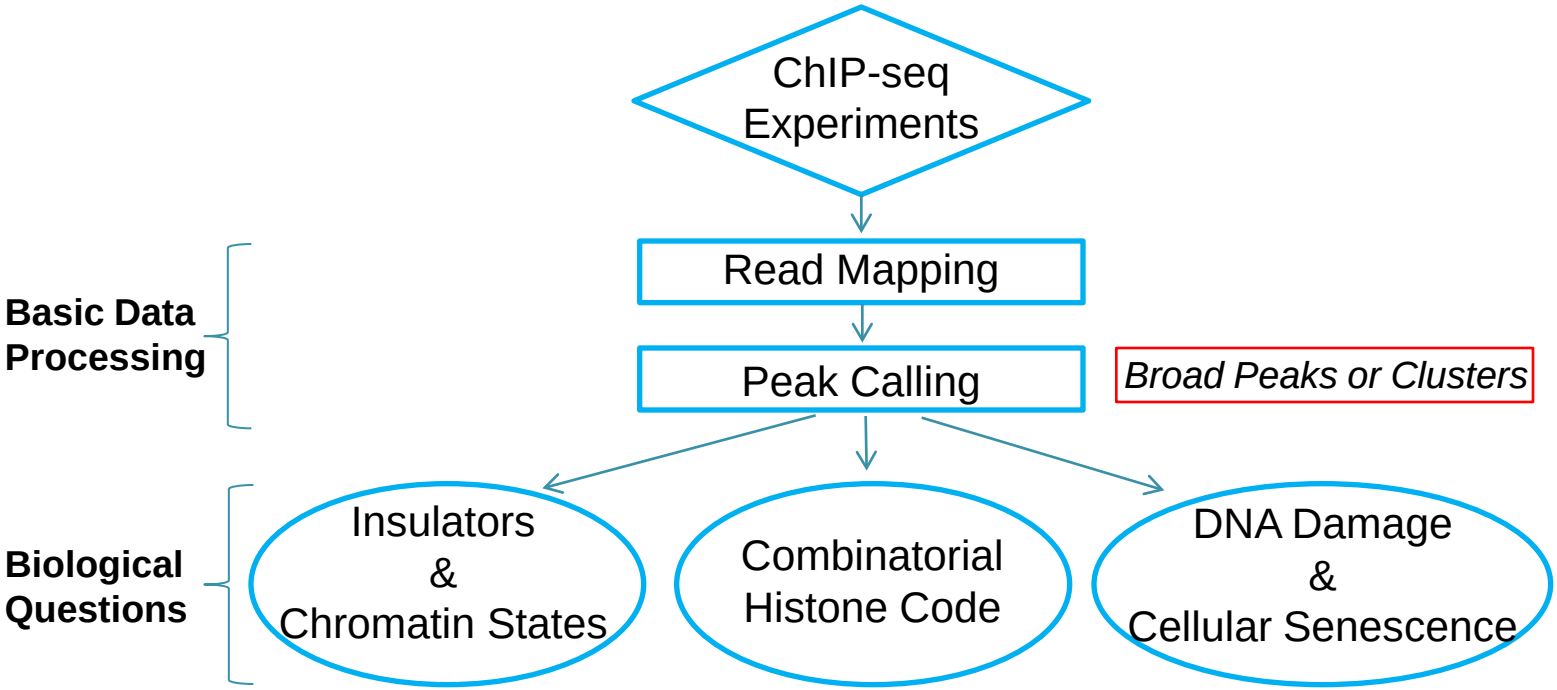
Ambiguous Short-sequence Read Mapping

Summary:

1. Developed a Gibbs sampling algorithm to map ambiguous reads;

2. The performance on simulated datasets is better than previous methods;

3. Provide opportunity to discover more information in repetitive genomic regions.

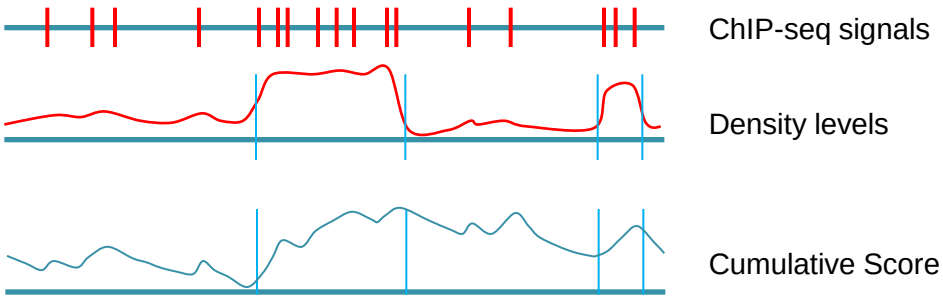

ChIP-seq
Experiments
Read Mapping
Peak Calling
Broad Peaks or Clusters
Basic Data
Processing
Biological
Questions
Insulators
&
Chromatin States
Combinatorial
Histone Code
DNA Damage
&
Cellular Senescence

Broad Peaks: Clustering of Contiguous ChIP-seq signals

1. Peak-calling: transcription factor binding sites, histone modification locations etc.

2. Broad Peaks: large clusters of contiguous ChIP-seq peaks (*e.g.* H3K27me3, H3K36me3, H3K79me3 etc).

3. Segmentation: regions with higher peak densities.



$$S_{i,j} = \sum_{i \le k \le j} x_k$$

4. Maximal Segments: all sub-segments and super-segments have lower scores.

5. Identification of all maximal segments

Ruzzo and Tompa, 1999 Proc. Int. Conf. Intell. Syst. Mol. Biol.

Broad Peaks: Clustering of Contiguous ChIP-seq signals

6. Scoring scheme:

$$s_1 = \ln(\frac{p}{q}) \qquad s_2 = \ln(\frac{1-p}{1-q})$$

p is the density in real peaks and *q* is the background density

7. Maximal segments identified based on this scoring scheme will asymptotically resemble the real peaks.

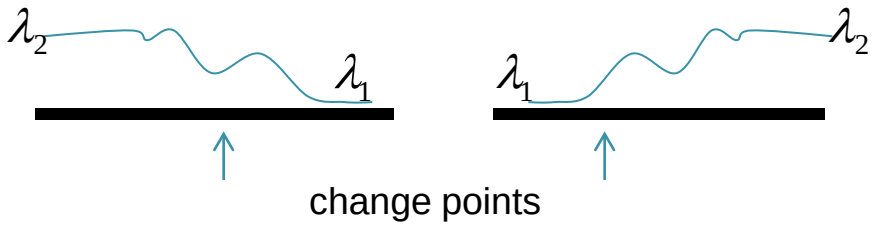
Karlin and Altschul, 1990 PNAS

8. Supervised estimation of *p*: use canonical regions with broad peaks.

e.g. peri-centromeric regions for H3K9me3 and highly expressed genes for H3K36me3.

9. Unsupervised estimation: sliding window to heuristically identify regions with higher and lower peak densities.

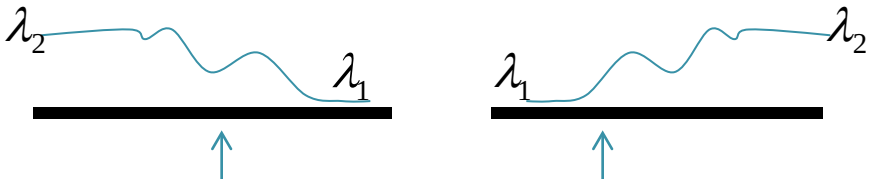
Problem: low accuracy due to the uncertainty of change-point locations



Gibbs sampling on those regions to infer: change point locations, λ_1, λ_2.

9. Unsupervised estimation (continue):

Sampled regions for estimation produced by initial sliding window detection



Non-homogeneous Poisson process: λ_1 & λ_2:

$$P(x \mid \lambda_1) = \frac{\lambda_1^x}{x!} e^{-\lambda_1}$$ for background

$$P(x \mid \lambda_2) = \frac{\lambda_2^x}{x!} e^{-\lambda_2}$$ for peaks

Conjugate prior distributions:

$$\lambda_1 \propto \lambda_1^{\alpha_1 - 1} e^{-\beta_1 \lambda_1} \qquad \lambda_2 \propto \lambda_2^{\alpha_2 - 1} e^{-\beta_2 \lambda_2}$$

Prior distributions for β_1 and β_2 :

$$\beta_1 \propto \beta_1^{\sigma_1 - 1} e^{-\varepsilon_1 \beta_1} \qquad \beta_2 \propto \beta_2^{\sigma_2 - 1} e^{-\varepsilon_2 \beta_2}$$

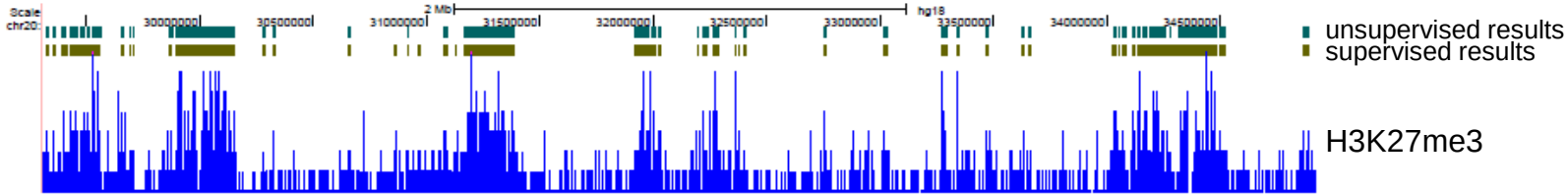
$$P(\lambda_1 \mid x, k, \alpha_1, \beta_1) \propto \lambda_1^{\alpha_1 + \sum_{i=1}^{k} x_i - 1} e^{-(\beta_1 + k)\lambda_1}$$

$$P(\lambda_2 \mid x, k, \alpha_2, \beta_2) \propto \lambda_2^{\alpha_2 + \sum_{i=k}^{N} x_i - 1} e^{-(\beta_2 + N - k)\lambda_2}$$

$$P(\beta_1 \mid x, k, \lambda_1, \alpha_1, \sigma_1, \varepsilon_1) \propto \beta_1^{\alpha_1 + \sigma_1 - 1} e^{-(\lambda_1 + \varepsilon_1)\beta_1}$$

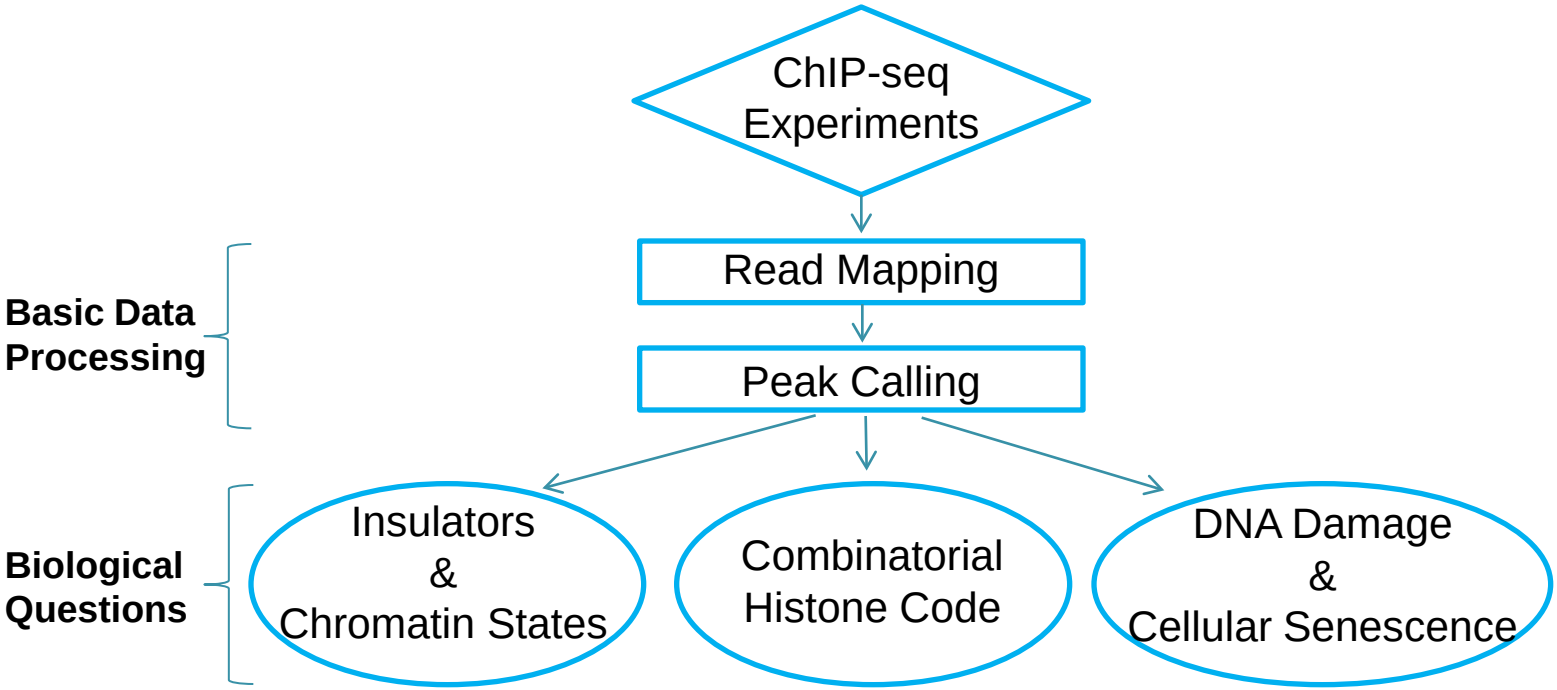
$$P(\beta_2 \mid x, k, \lambda_2, \alpha_2, \sigma_2, \varepsilon_2) \propto \beta_2^{\alpha_2 + \sigma_2 - 1} e^{-(\lambda_2 + \varepsilon_2)\beta_2}$$

$$P(k \mid x, \lambda_1, \lambda_2) \propto \frac{P(x \mid k, \lambda_1, \lambda_2)}{\sum_{c=1}^{N} P(x \mid c, \lambda_1, \lambda_2)}$$



10. Has been applied to other projects to find chromatin domains and γH2A.X clusters.

Wang et al. in preparation



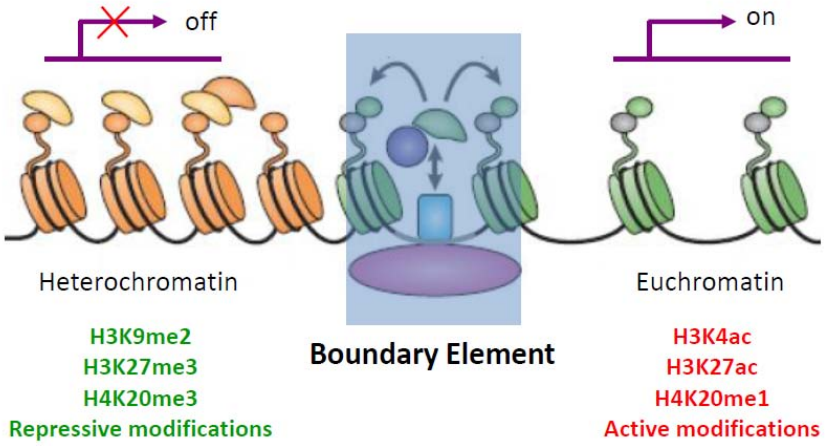
Locations of Insulators in the Human Genome

Unbiased Prediction of Boundary Elements with Novel Features

Identification of Active and Repressive Chromatin States

MIR-retrotransposon derived insulators

1. Insulators (boundary elements): specific DNA sequences that can block enhancer-promoter interactions and/or partition chromatin domains;



Gaszner & Felsenfeld 2006 Nature Reviews Genetics

2. Lack of unified features (CTCF, BEAF, Su(Hw), B-box etc.);

3. An emerging sub-group of insulators:
   Pol III machinery related insulators
   (tRNA genes, B-box, SINE/B2)

MIR-retrotransposon derived insulators

Mammalian-wide interspersed repeats (MIRs):

a. a large number of intergenic MIRs are highly conserved;

Silva et al. 2003 Genet Res

b. have regulatory potentials: *e.g.* c-MYC binding sites;

Wang et al. 2009 Mol. Biosys.

c. enriched with open chromatin regions;

Marino-Ramirez & Jordan 2006 Biol Direct

d. evolved from *tRNA* and contain B-box for Pol III machinery;

Jurka et al. 1995 Nucleic Acids Res
Smit & Riggs 1995 Nucleic Acids Res

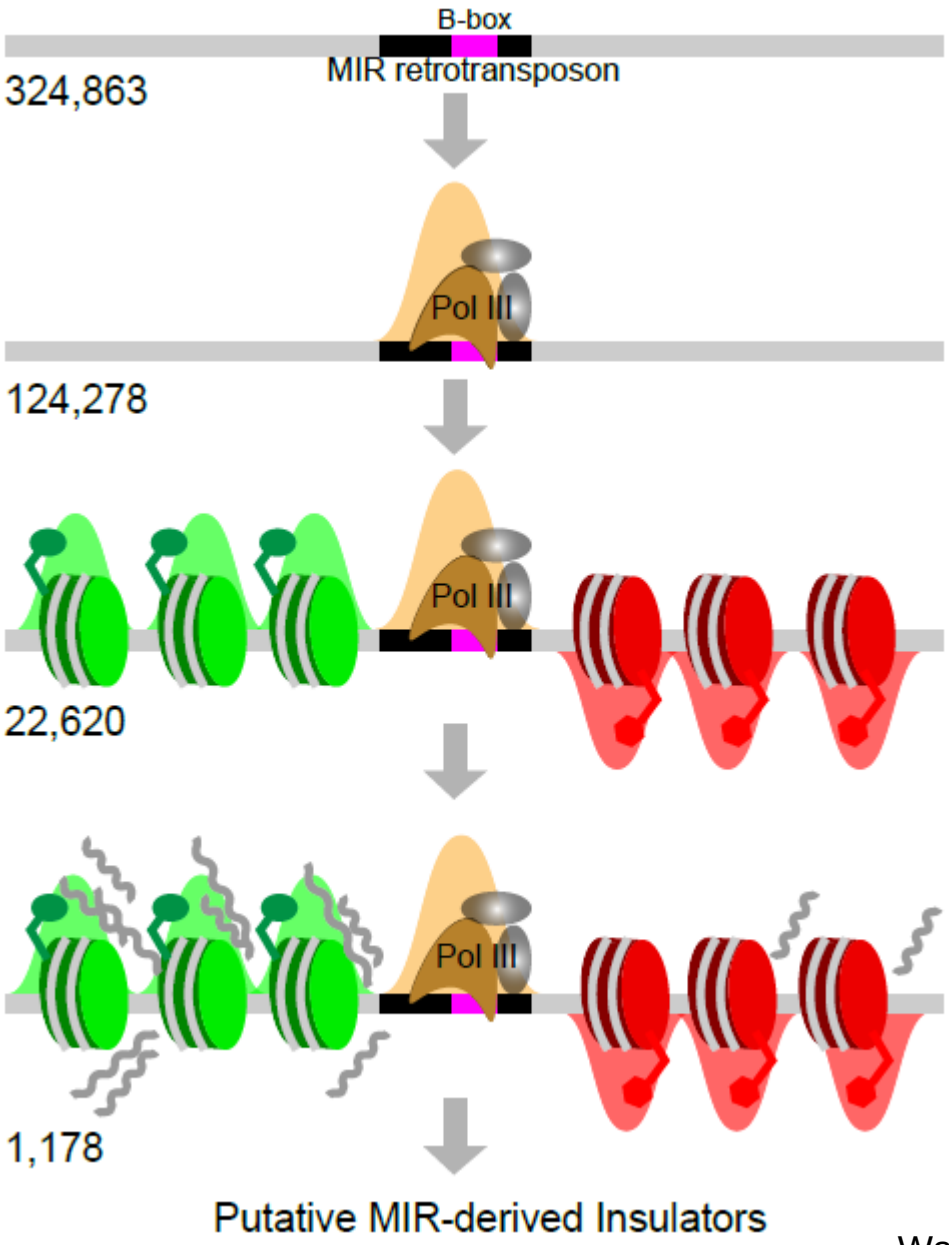
Integrate diverse biological datasets:

a. MIR sequences from RepeatMasker;
b. sequence motif (B-box);
c. Pol III binding (ChIP-seq data);
d. partition between active and repressive histone modifications;
e. located between transcription active and repressive regions (RNA-seq);

All datasets are from human $CD4^+$ T cells.

Wang et al. in preparation.

MIR-retrotransposon derived insulators



Compare with the consensus sequence from Repbase

Poisson(λ)

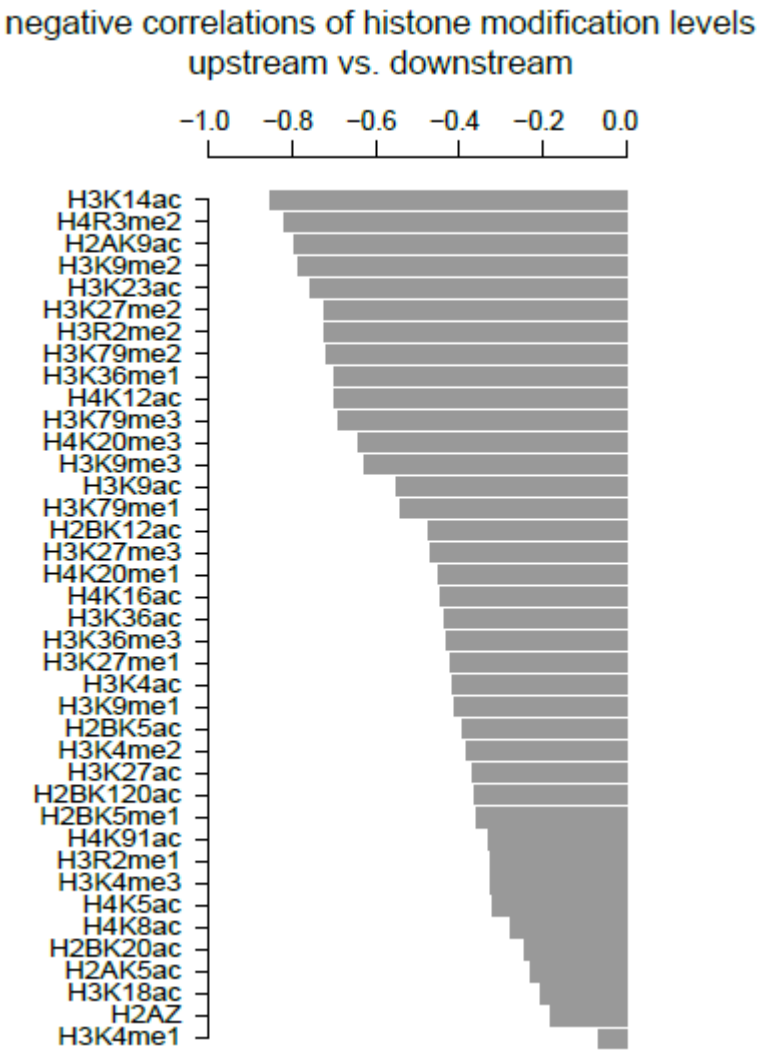
Located between broad peaks of active and repressive marks

Poisson(μ)

Wang et al. in preparation.

MIR-retrotransposon derived insulators

individual histone modifications are blocked to single sides of MIR-insulators

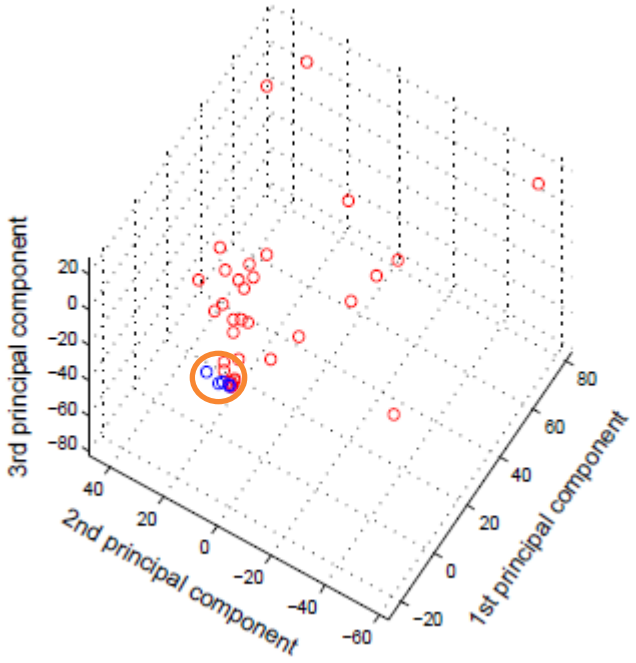


MIR-retrotransposon derived insulators

active and repressive histone modifications are grouped to different clusters

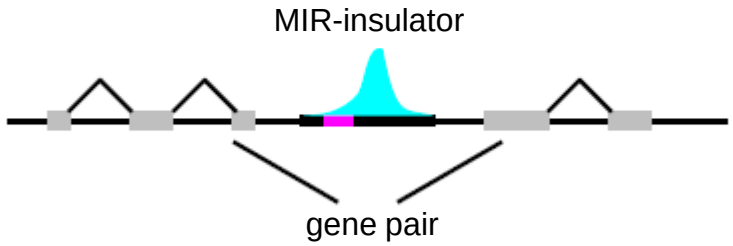
MIR-retrotransposon derived insulators

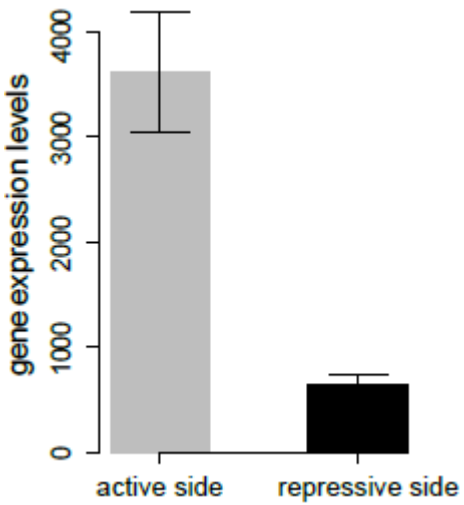
active and repressive histone modifications are grouped to different clusters

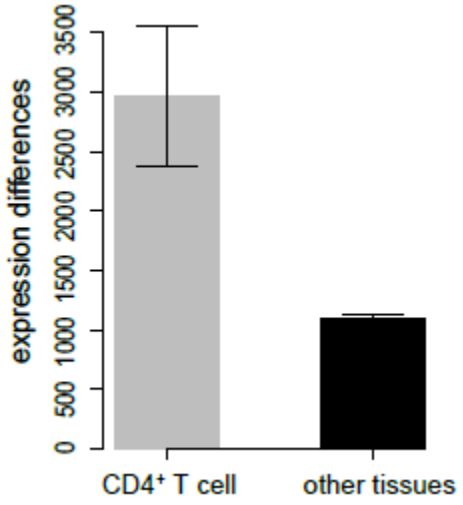


MIR-retrotransposon derived insulators

CD4⁺ T cell specific regulation on gene expressions

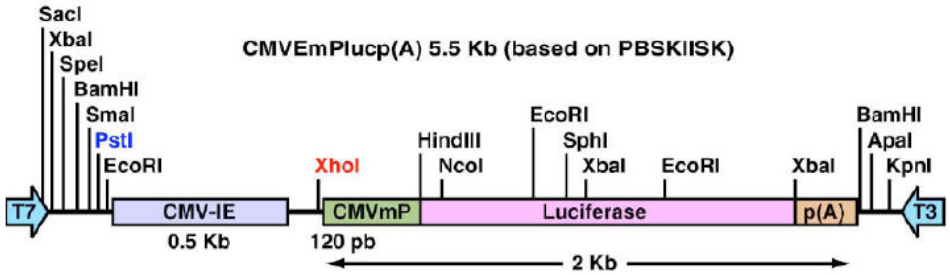


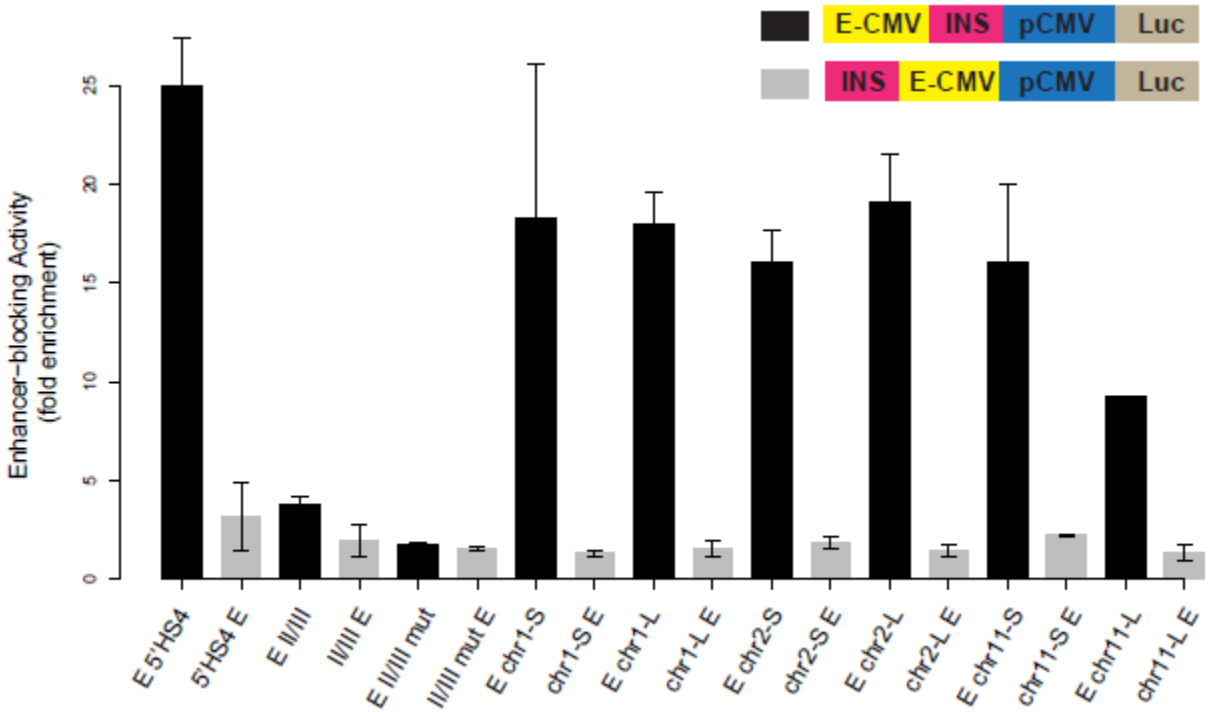




MIR-retrotransposon derived insulators

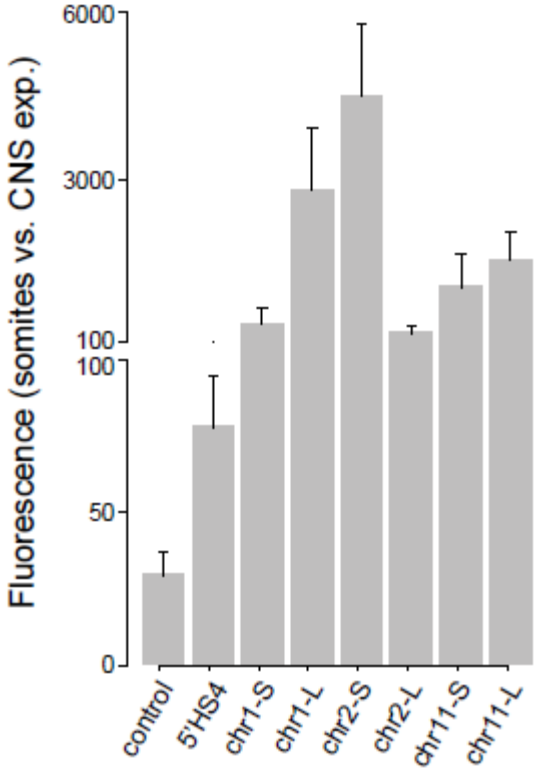
in vitro enhancer-blocking assay in human HEK293 cell lines

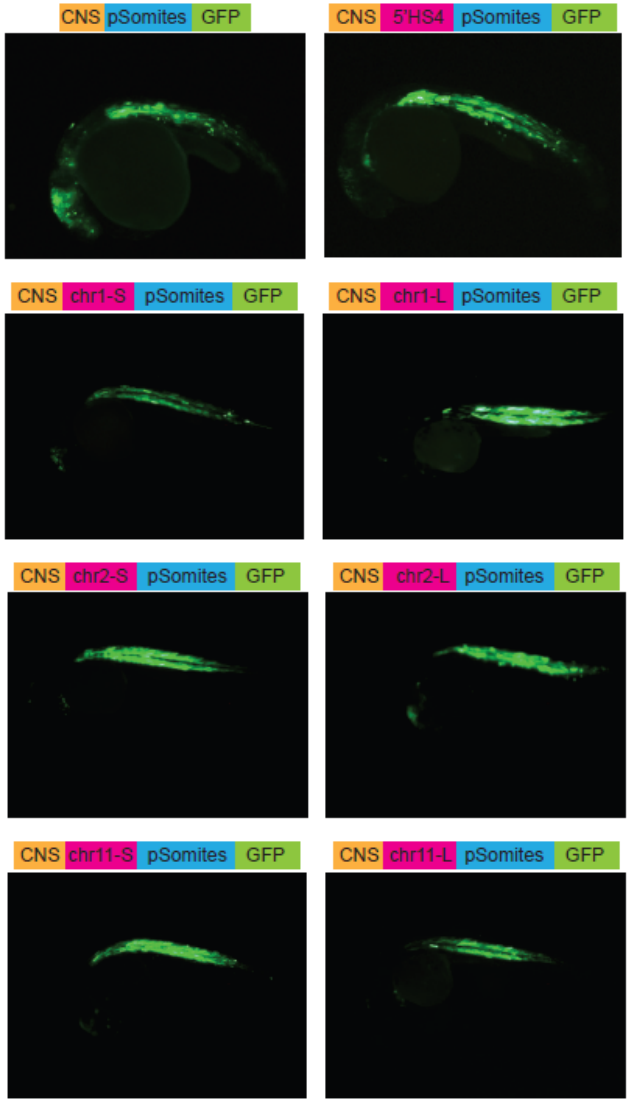




MIR-retrotransposon derived insulators

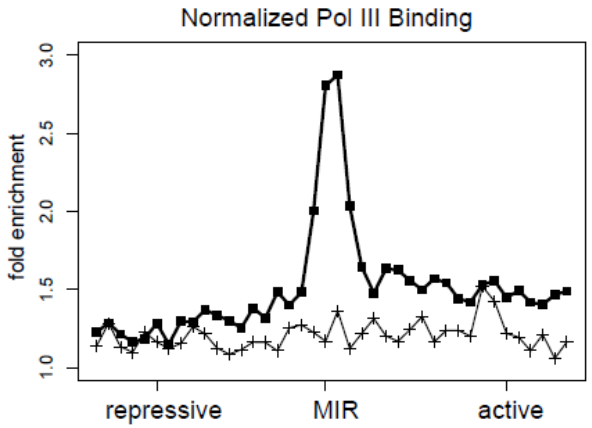
in vivo enhancer-blocking assay in zebrafish embryos

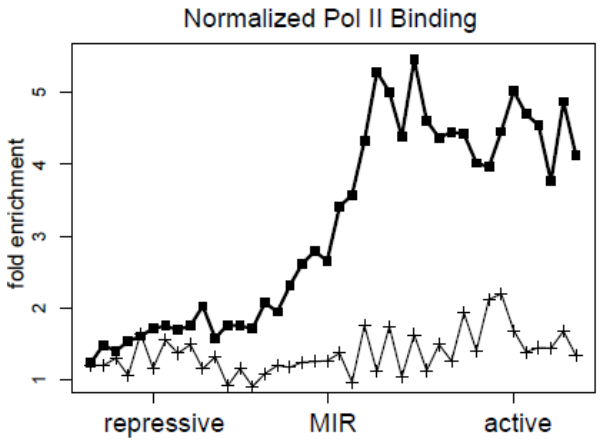


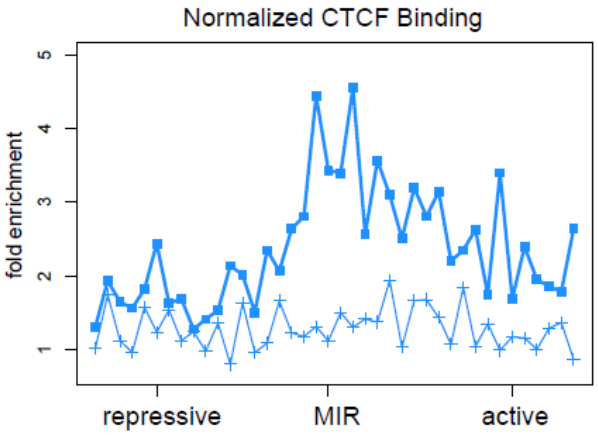


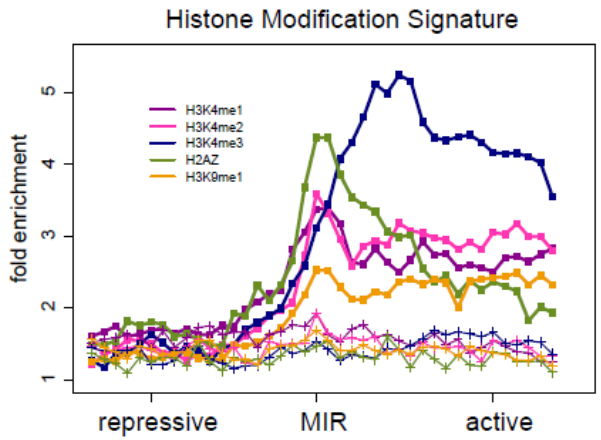
MIR-retrotransposon derived insulators

chromatin signatures of putative MIR-derived boundaries



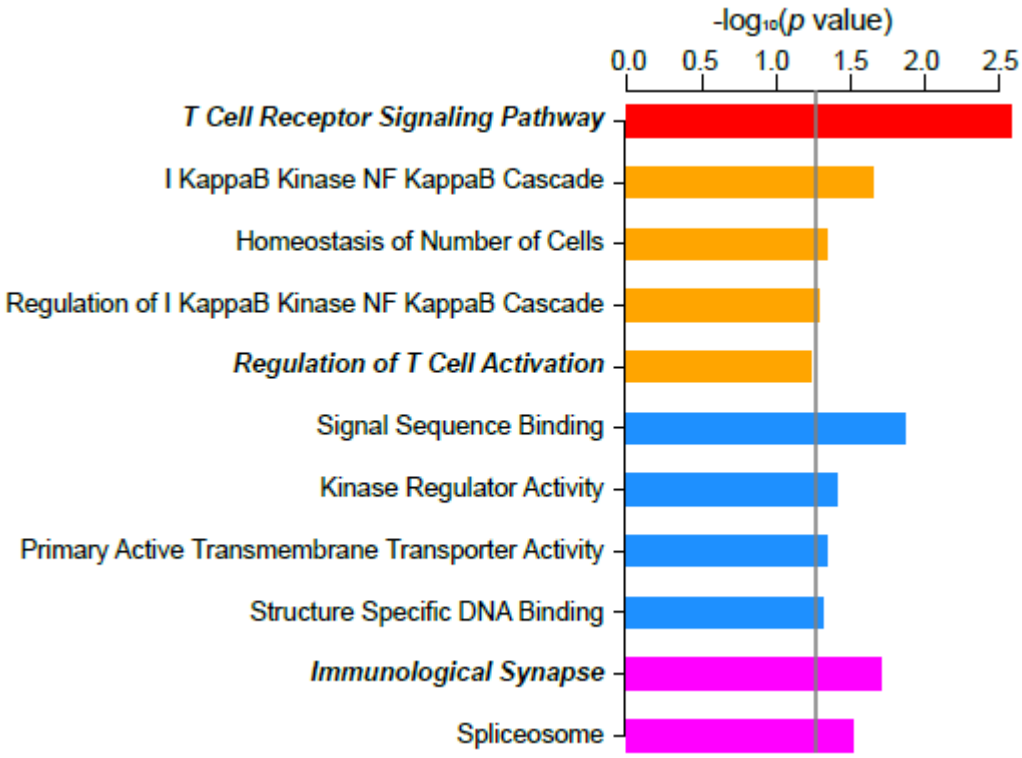






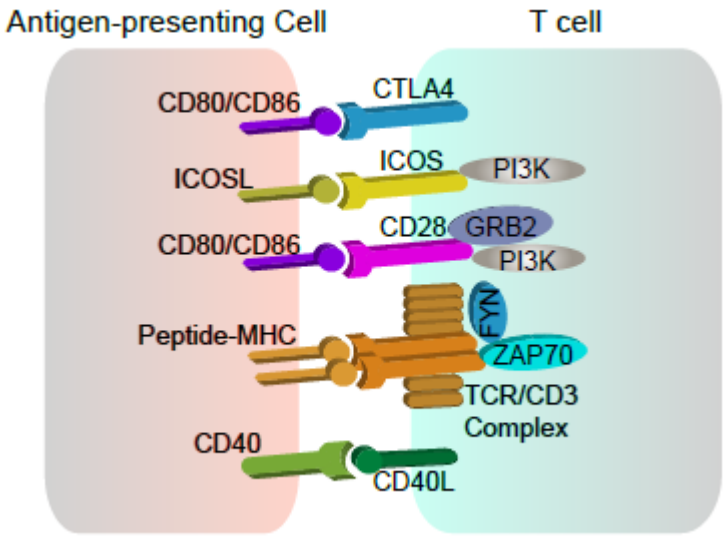
MIR-retrotransposon derived insulators

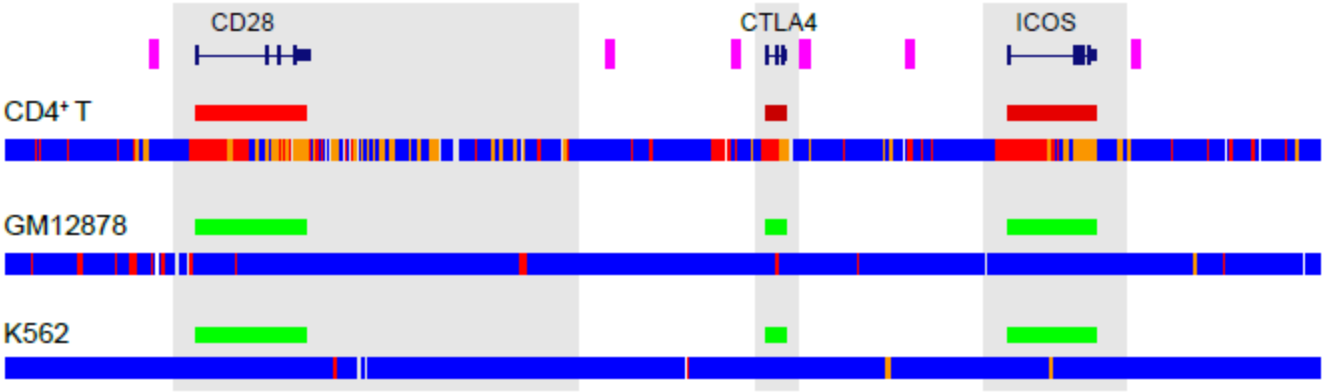
Functional Annotations of Genes Proximal to Putative MIR-insulators



MIR-retrotransposon derived insulators

Functional Annotations of Genes Proximal to Putative MIR-insulators





MIR-retrotransposon derived insulators

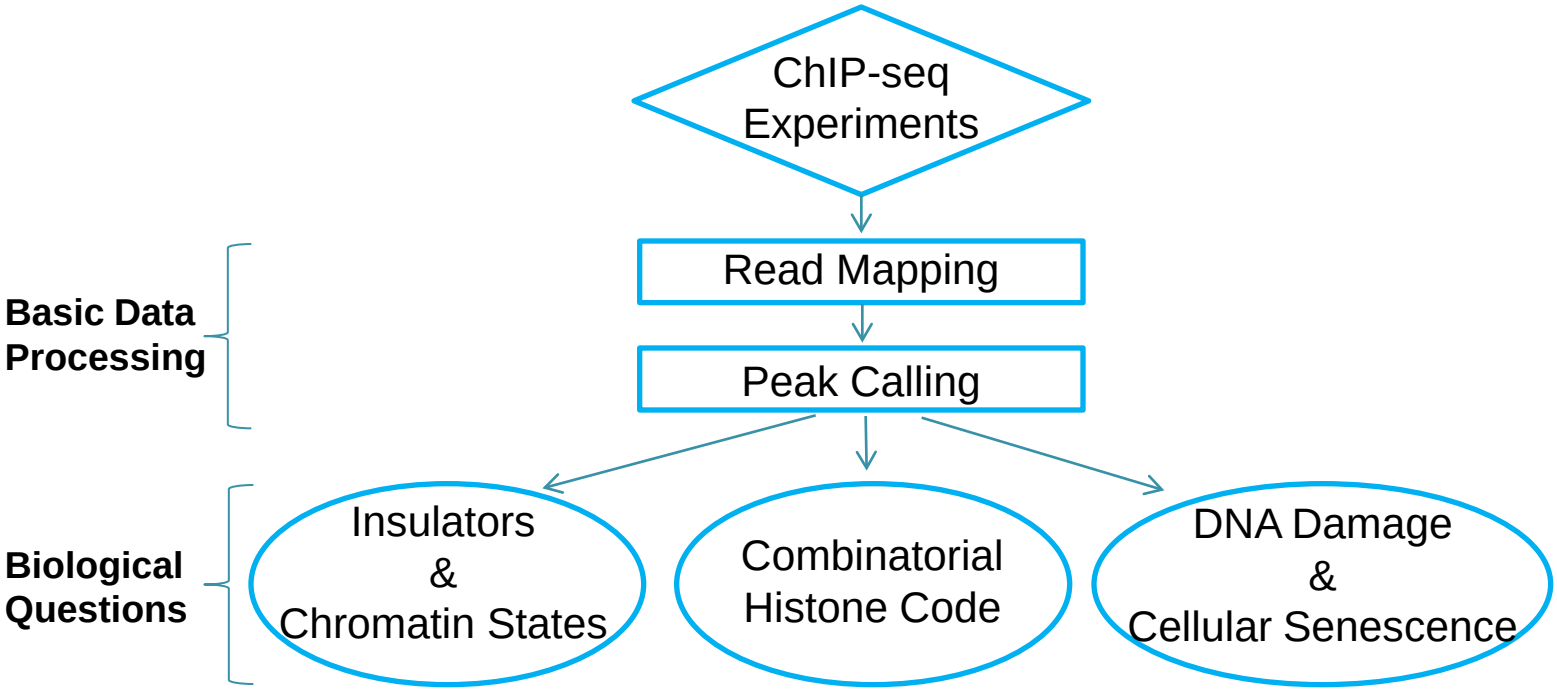
Summary:

1. Discovered a set of putative MIR-derived insulators by integrating specific features;

2. Selected MIR-derived insulators are experimentally tested *in vitro* and *in vivo*;

3. Specific histone modifications are enriched around MIR-derived insulators;

4. MIR-derived insulator functions might be cell type specific.



Locations of Insulators in the Human Genome

Unbiased Prediction of Boundary Elements with Novel Features

Identification of Active and Repressive Chromatin States

Chromatin State and Boundary Elements

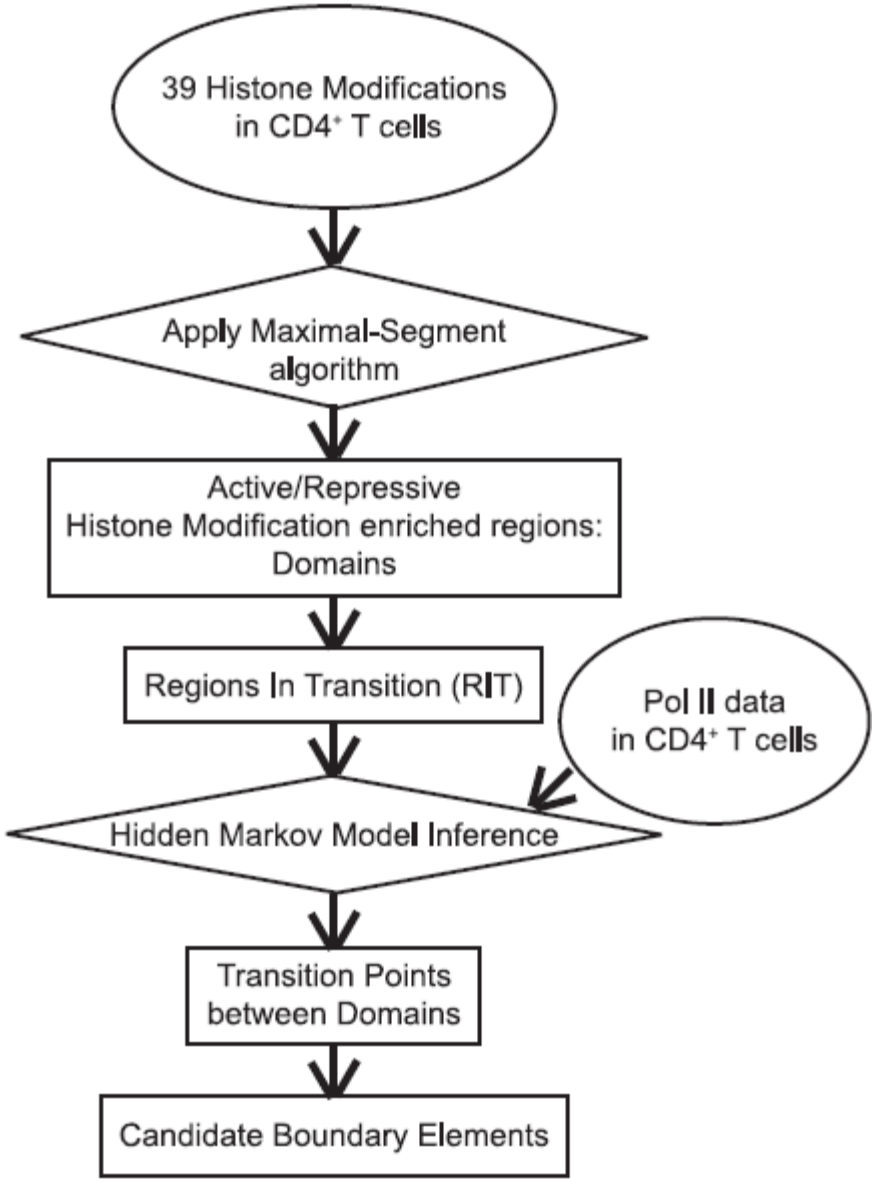
1. Using a specific set of features (*e.g.* CTCF or B-box) of boundary elements can't find elements with different mechanisms.

2. Looking for novel features (or mechanisms) is more important than searching for locations of boundary elements.

Need to design a method to search boundary elements without using any known features.

Chromatin State and Boundary Elements

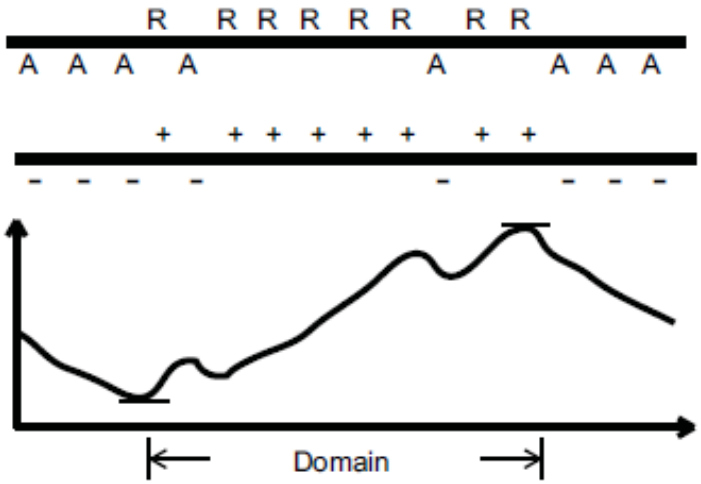
1. Consider boundary elements as transition points between active and repressive domains.
2. Start from larger scales – chromatin domains.
3. Then focus on regions in transition (RIT).
4. HMM: transcriptional active state and repressive state.



Wang et al. 2012 Nucleic Acids Res

Chromatin State and Boundary Elements

chromatin states: broad peaks of active (or repressive) histone modifications



$s_1 = ln(p/q)$ and $s_2 = ln((1-p)/(1-q))$

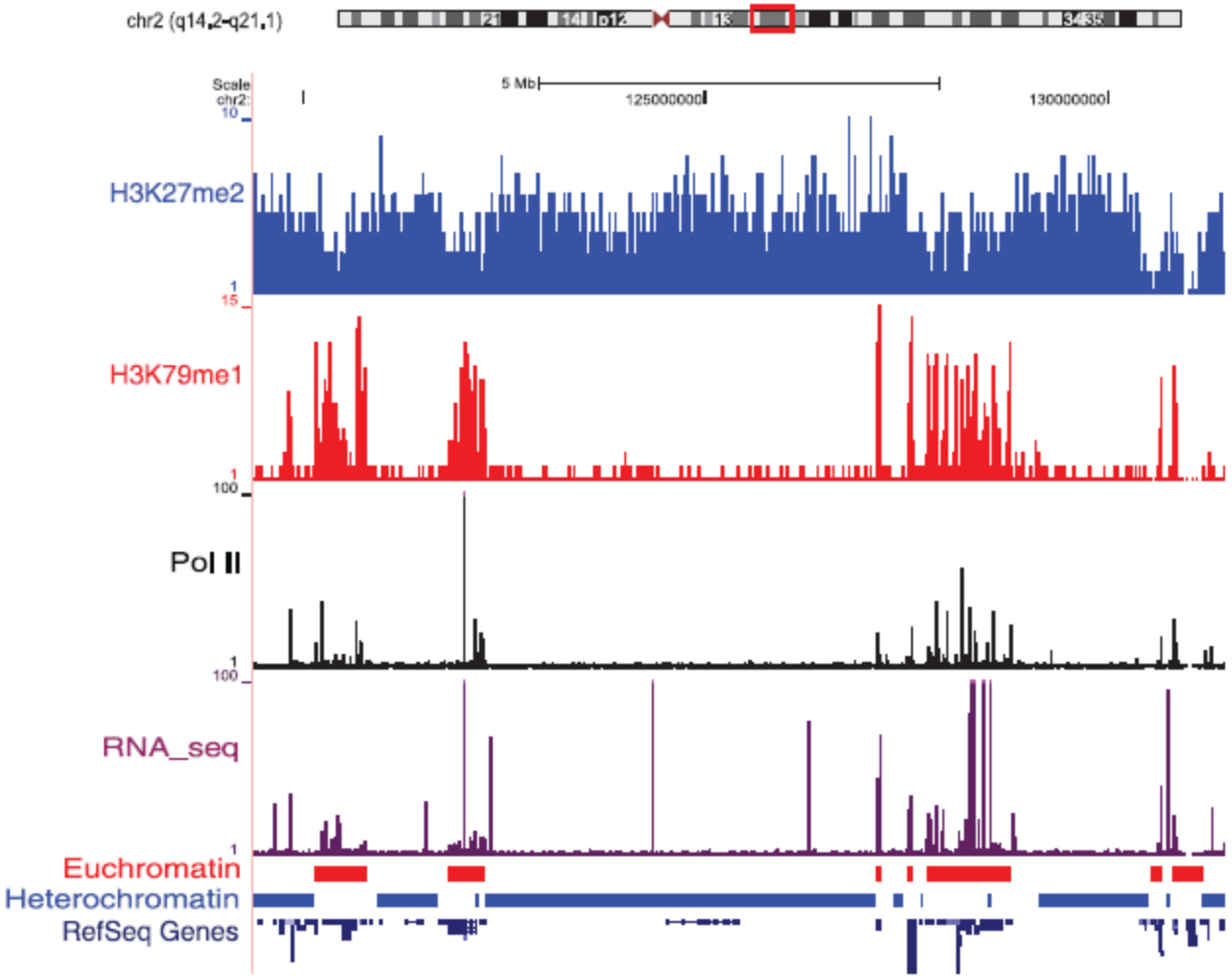
For repressive state inferences, *p* is the density of genomic sites with repressive modifications in real repressive domains and *q* is the corresponding background density.

For active state inferences, *p* is the density of genomic sites with active modifications in real active domains and *q* is the corresponding background density.

These parameters are estimated from canonical repressive or active regions (*e.g.* peri-centromere, highly expressed genes).

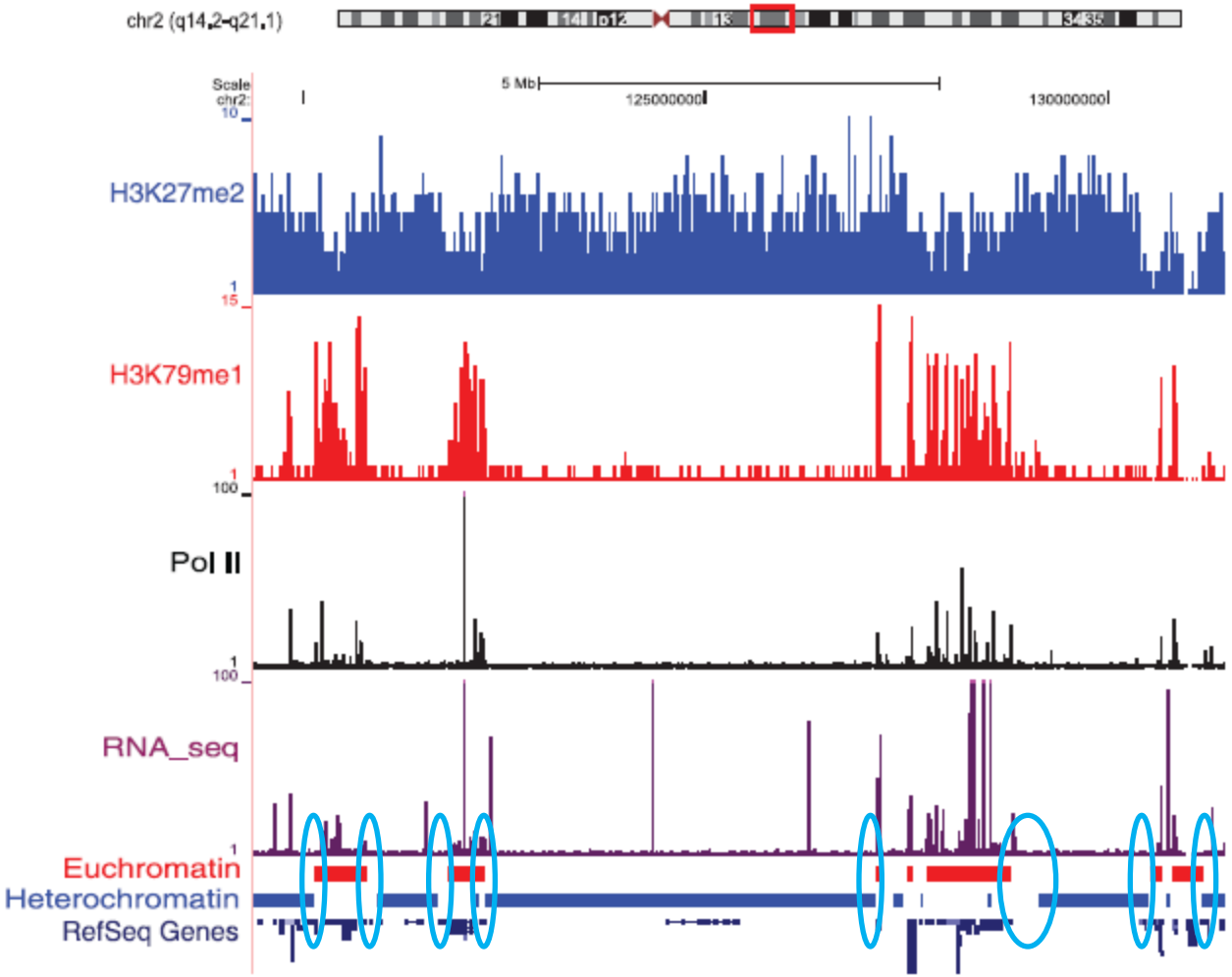
Chromatin State and Boundary Elements

chromatin states: broad peaks of active (or repressive) histone modifications



Chromatin State and Boundary Elements

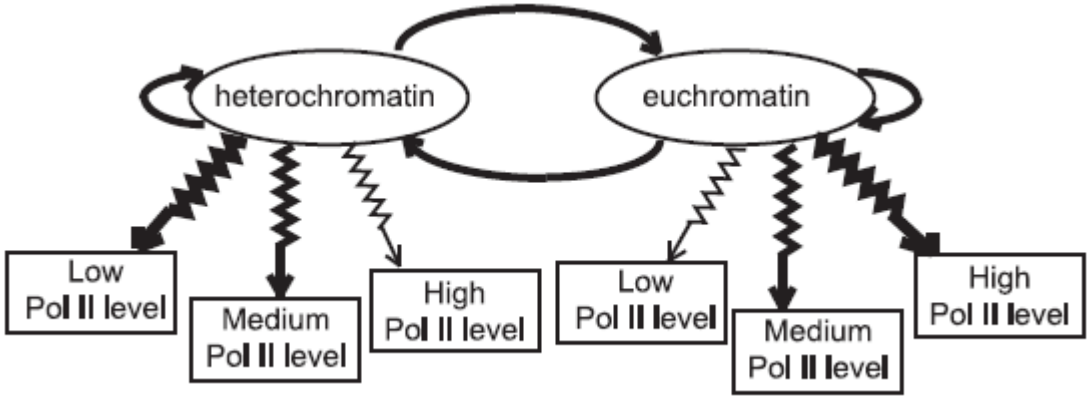
chromatin states: broad peaks of active (or repressive) histone modifications



regions in transition (RIT)

Chromatin State and Boundary Elements

boundary element: transition point between transcriptional active and repressive states



RNA Pol II binding levels are classified into 3 classes:

1 – low Pol II level;
2 – medium Pol II level;
3 – high Pol II level;

{ e(x|s) | s=(H,E), x=(1,2,3) }: emission probability of Pol II level *x* from hidden state *s*.
estimated from canonical repressive and active regions

{ t_{ij} | i=(H,E), j=(H,E) }: transition probability from hidden state *i* to hidden state *j*.
estimated from the basic chromatin state configurations produced by the first stage

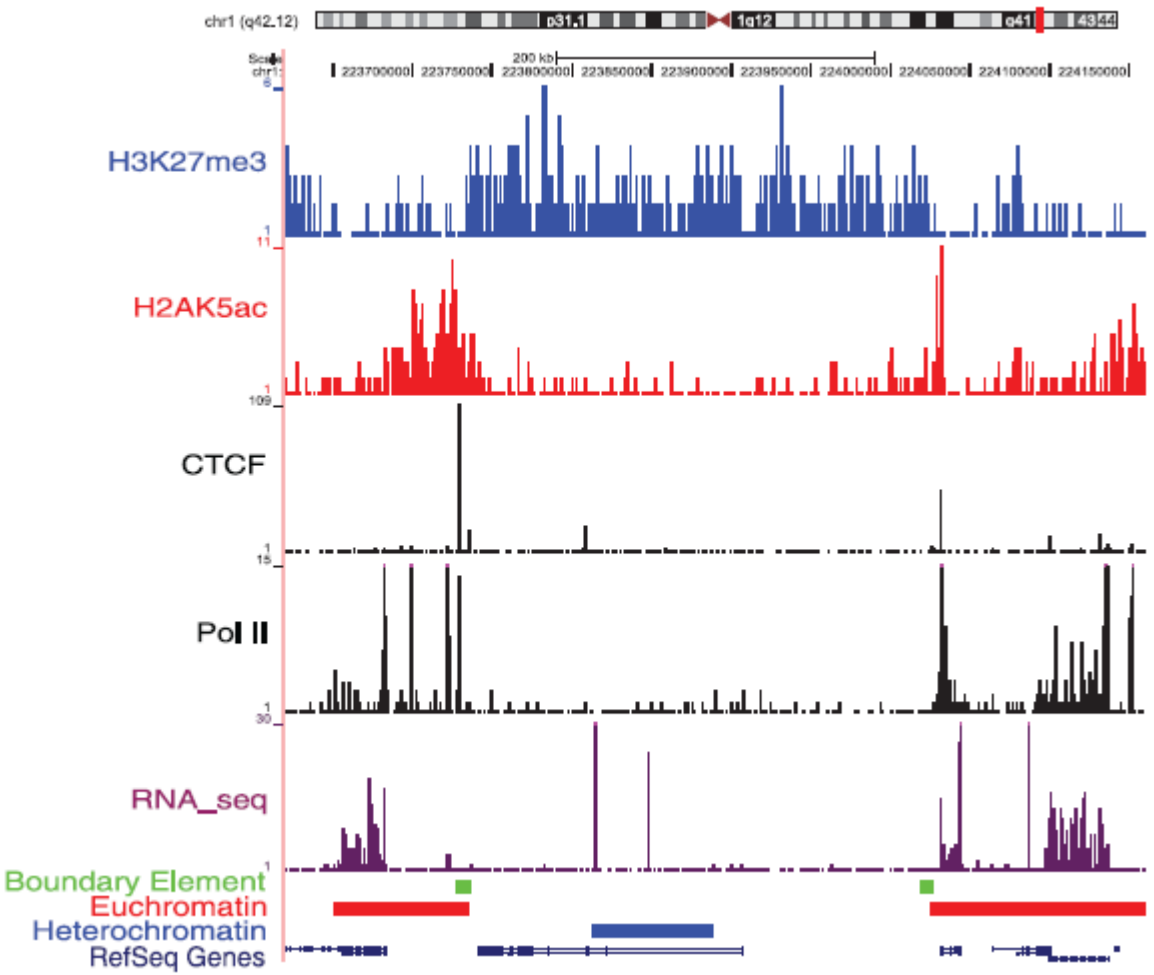
Viterbi algorithm to infer the most probable hidden state path.

$$V(s_k,n) = e(x_n|s_k) \times max_y \{t_{yk} V(s_y,n-1)\}$$

Only focus on regions with a unique transition point.

Chromatin State and Boundary Elements

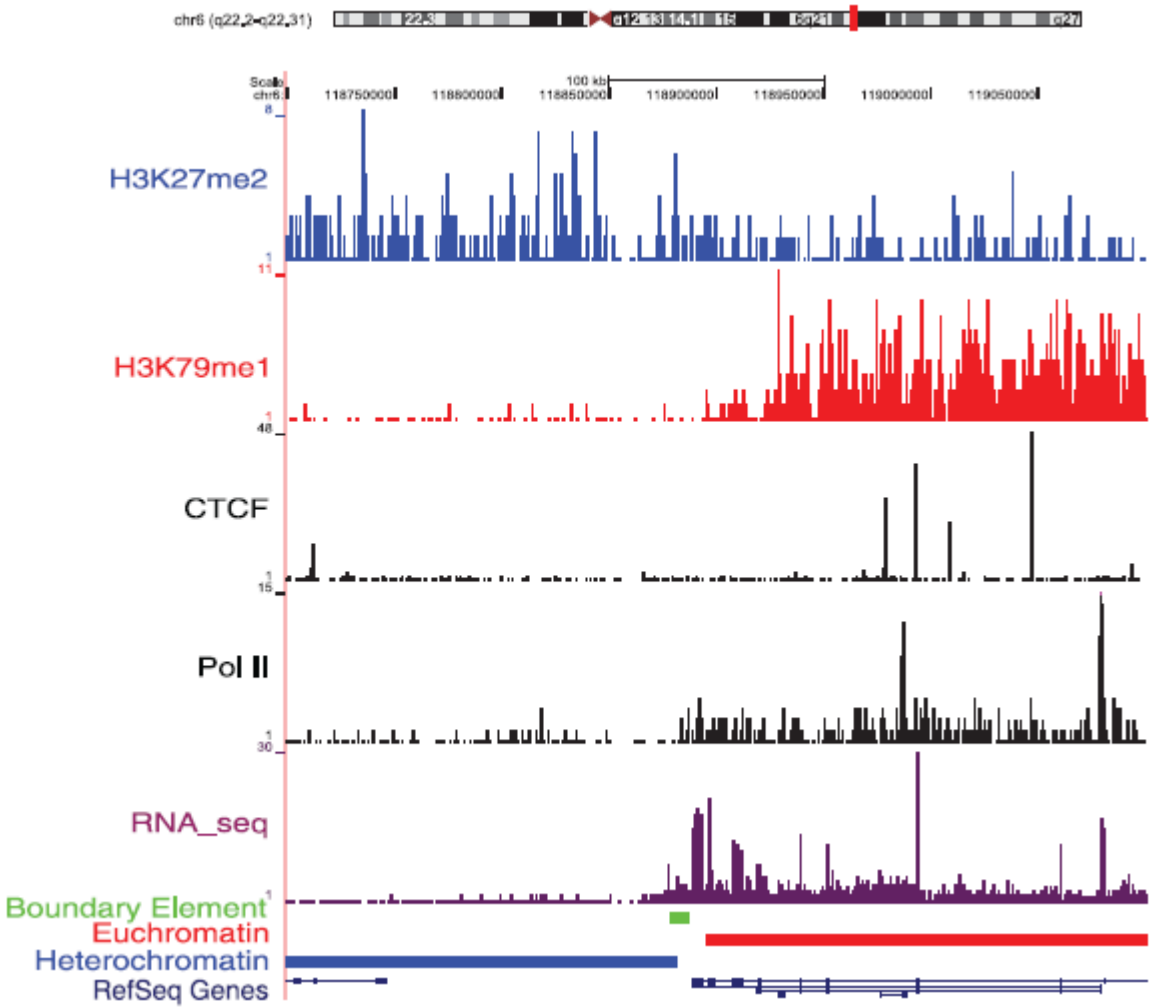
boundary element: transition point between transcriptional active and repressive states



CTCF-related putative boundaries

Chromatin State and Boundary Elements

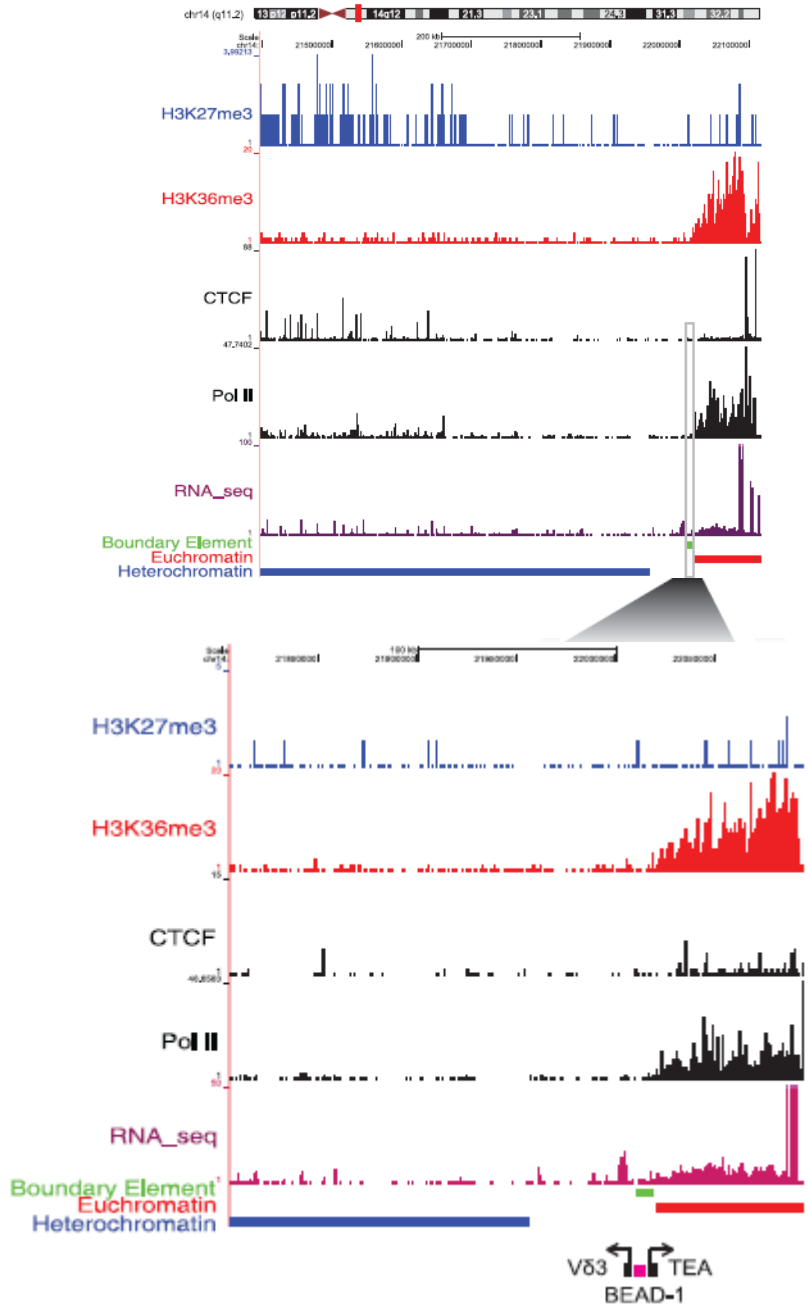
boundary element: transition point between transcriptional active and repressive states



CTCF unrelated putative boundaries

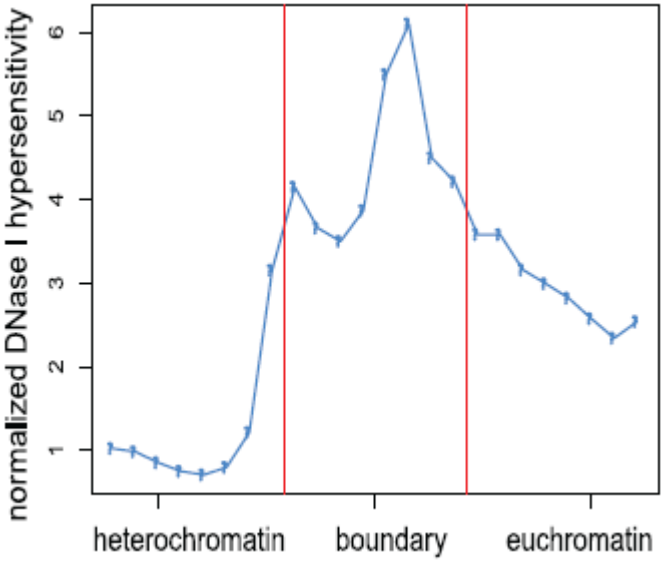
Chromatin State and Boundary Elements

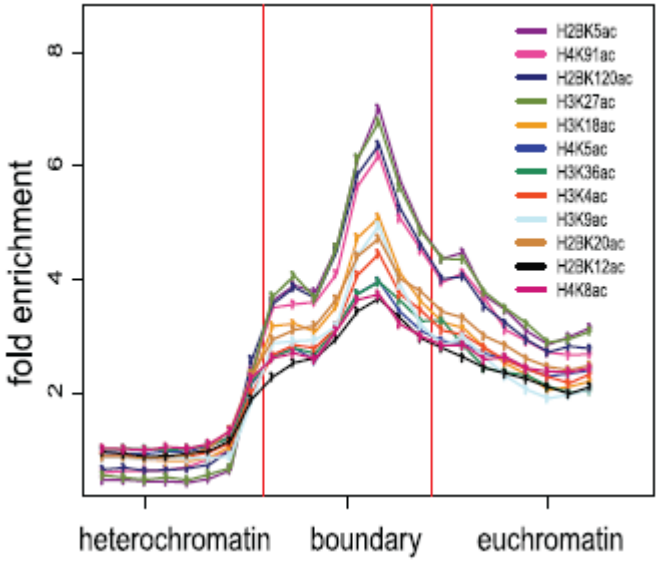
BEAD-1 element:
the only experimentally validated boundary element from human T cells



Chromatin State and Boundary Elements

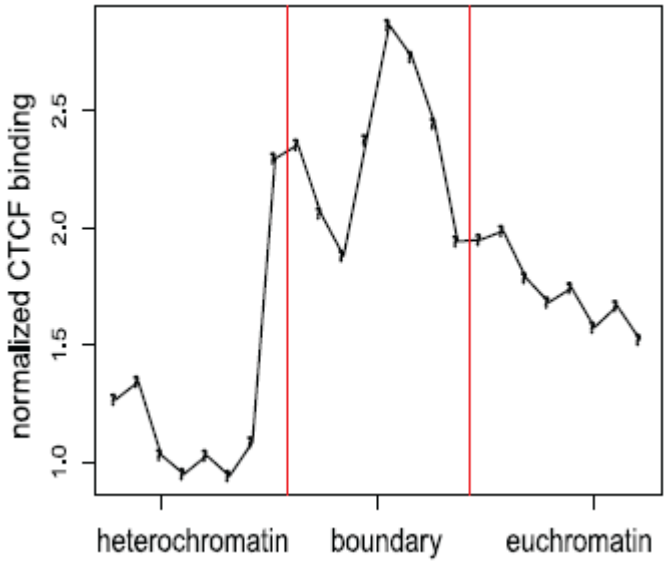
Local Chromatin Signatures of Putative Boundaries





Chromatin State and Boundary Elements

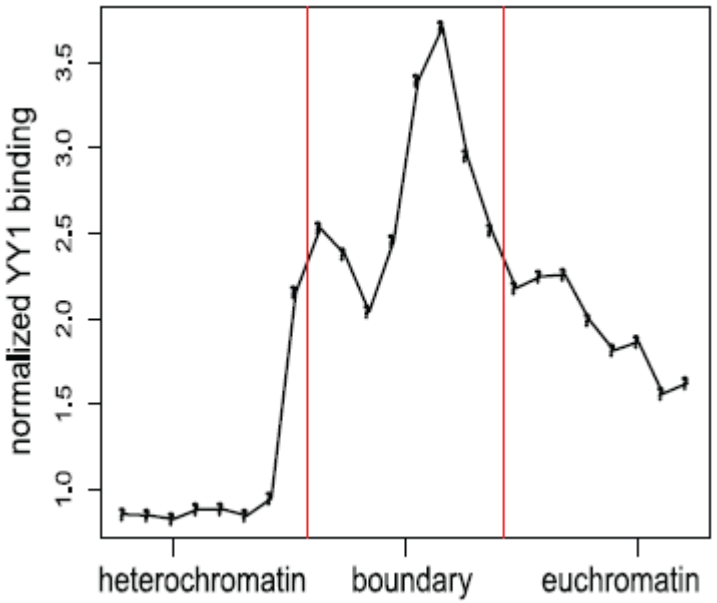
Potential Protein Factors Participating in Boundary Activity



Chromatin State and Boundary Elements

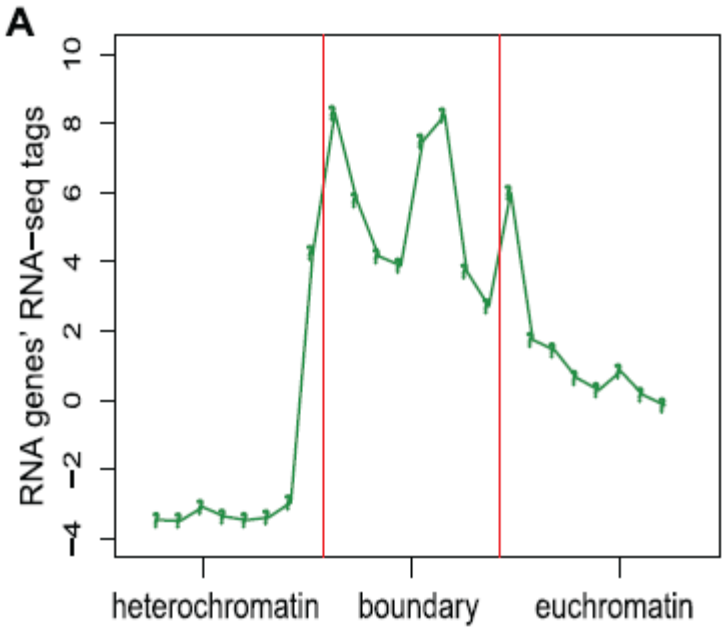
Potential Protein Factors Participating in Boundary Activity

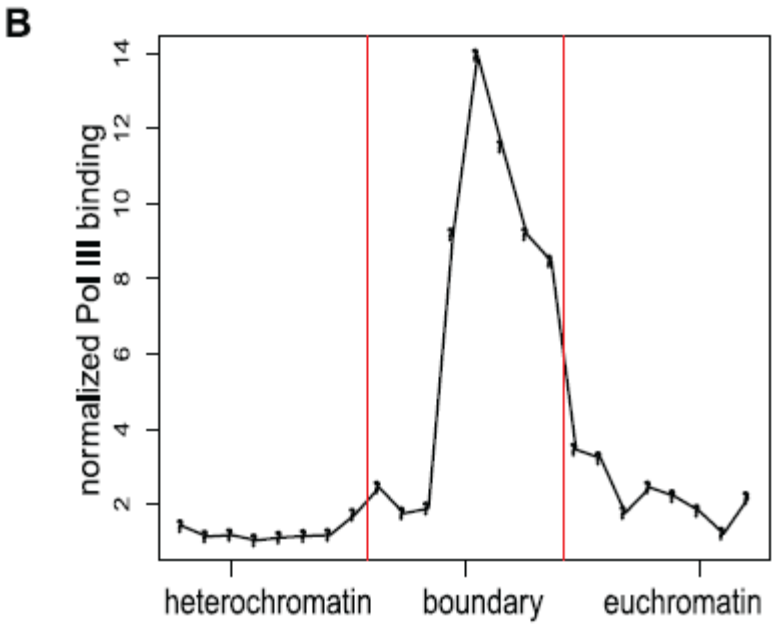
Protein	No.[1]	*P*-value[2]	Annotations[3]
EVI1	382	0.022	Interacts with histone deacetylase, histone methyltransferases and CBP and P/CAF
CEBP	249	2.27E-17	Interacts with CBP and p300 and promotes histone acetylation
YY1	157	1.44E-17	Directs histone deacetylases and histone acetyltransferases to promoter
CREBP1	150	5.87E-24	Essential in H2B and H4 acetylation, can interact with CBP HAT domain
USF	140	2.50E-28	Recruits histone modifications at vertebrate boundary elements



Chromatin State and Boundary Elements

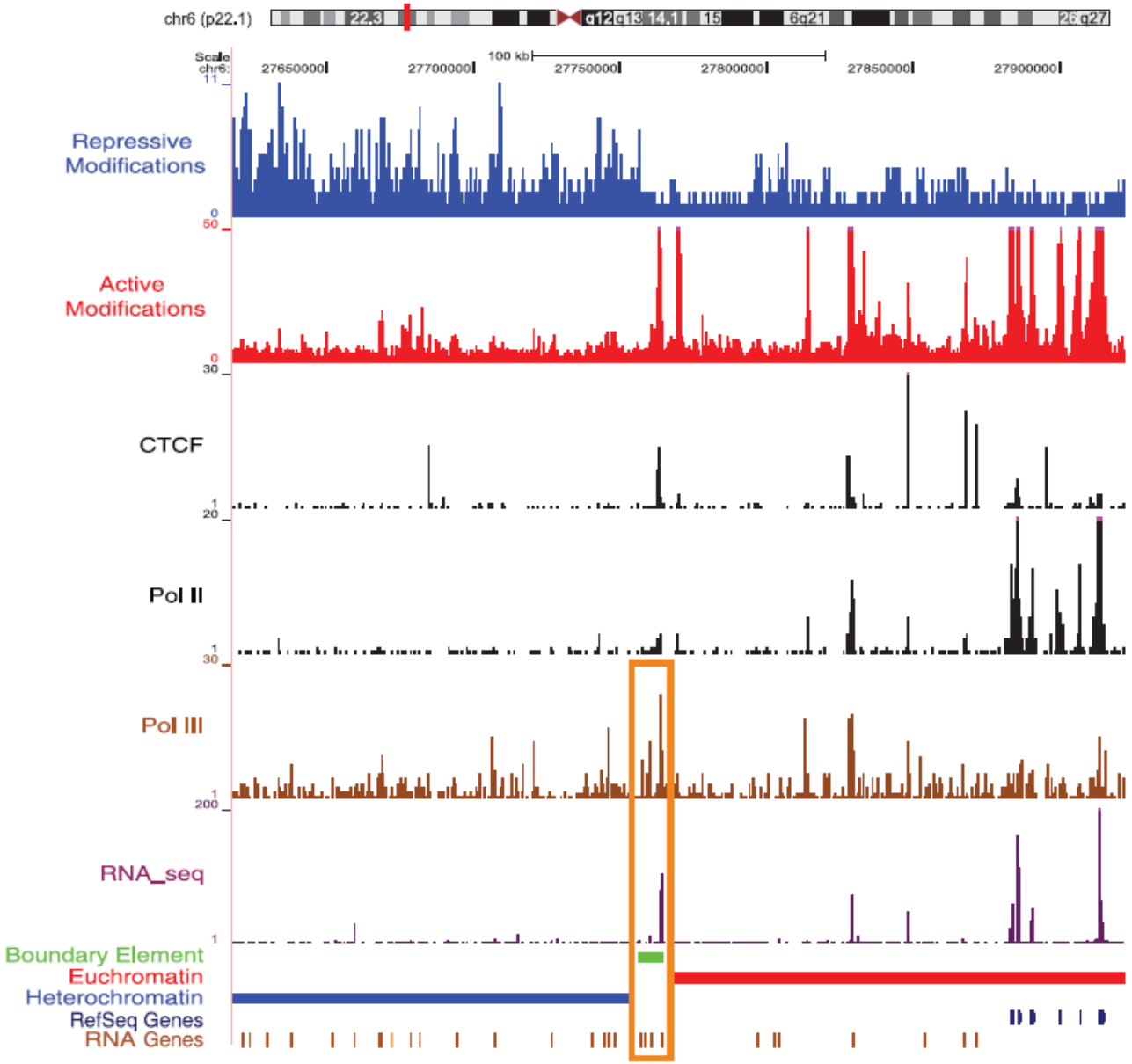
Subset of Putative Boundaries Contain non-coding RNA genes





Chromatin State and Boundary Elements

Subset of Putative Boundaries Contain non-coding RNA genes



Chromatin State and Boundary Elements

Summary:

1. Developed a feature-free method to search boundary elements;

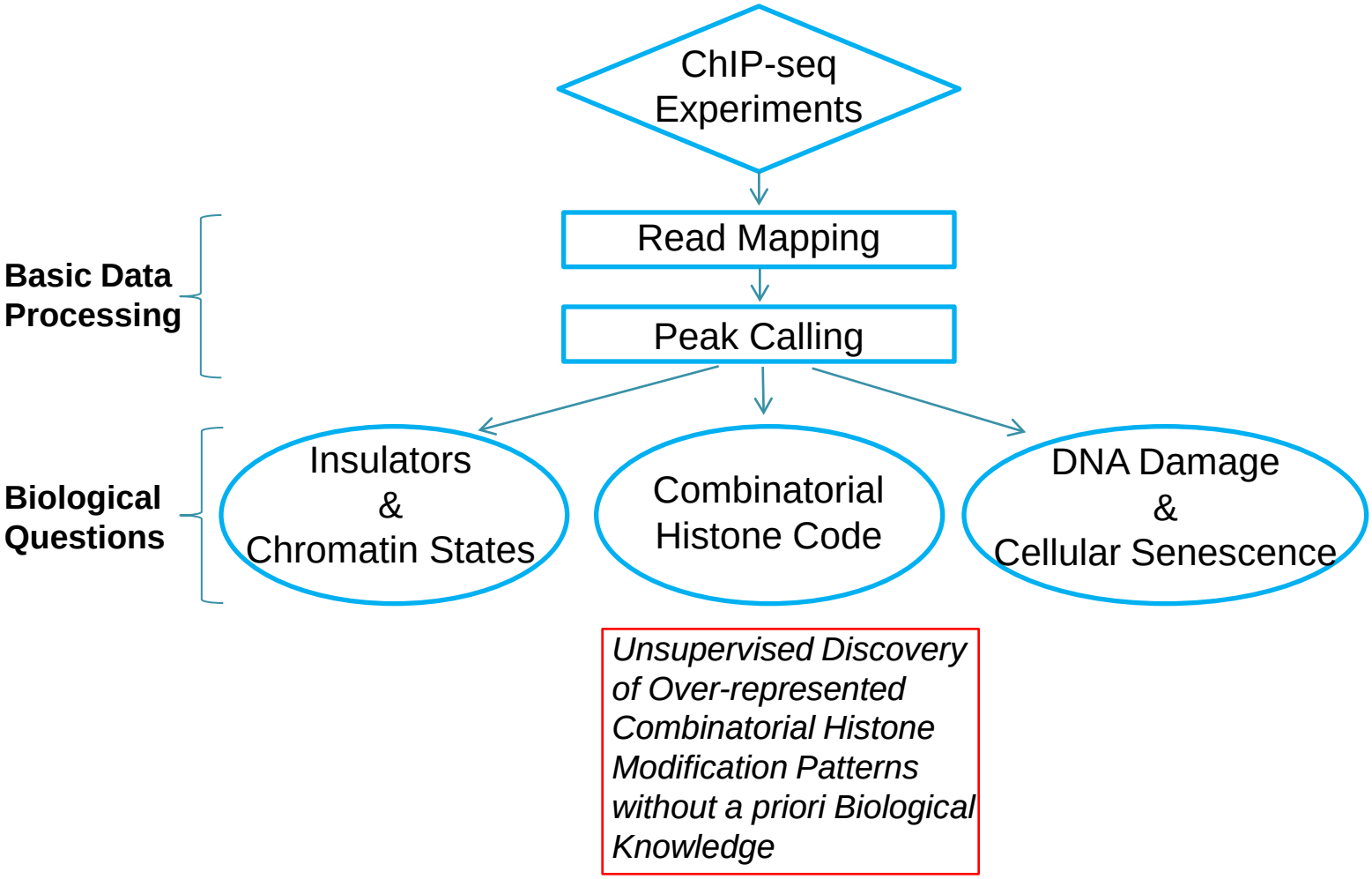
2. The resulted predictions contain both CTCF-related and CTCF-independent boundary elements;

3. BEAD-1 element is found by this method;

4. The candidate boundaries are enriched with DNase hypersensitive sites and a set of active histone modifications;

5. Discovered some enriched protein factor binding motifs;

6. Found a set of non-coding RNA related boundary elements.


ChIP-seq Experiments
Read Mapping
Peak Calling
Basic Data Processing
Biological Questions
Insulators & Chromatin States
Combinatorial Histone Code
DNA Damage & Cellular Senescence
Unsupervised Discovery of Over-represented Combinatorial Histone Modification Patterns without a priori Biological Knowledge

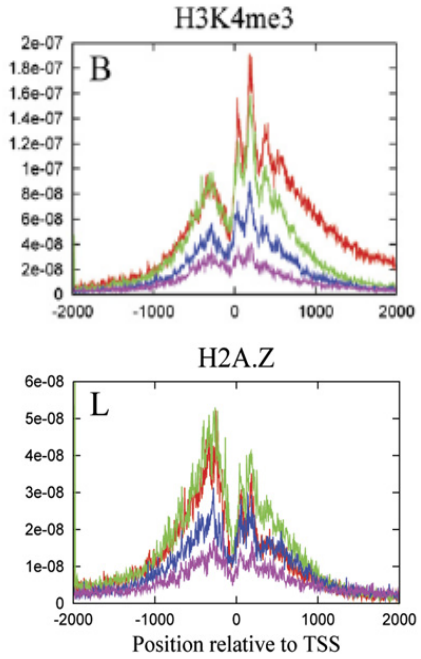
Combinatorial Histone Modification Patterns

Histone Code**:** specific combinations of histone modifications can regulate chromatin structure and gene expressions (histone modification “motif”).

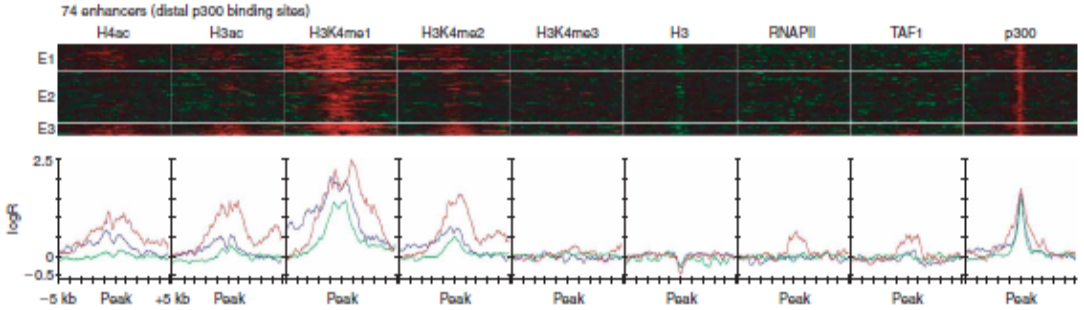
Examples of important histone modification signatures:

promoter: H3K4me3, H2AZ

enhancer: H3K4me1, H3K27ac



Barski et al. 2007 Cell



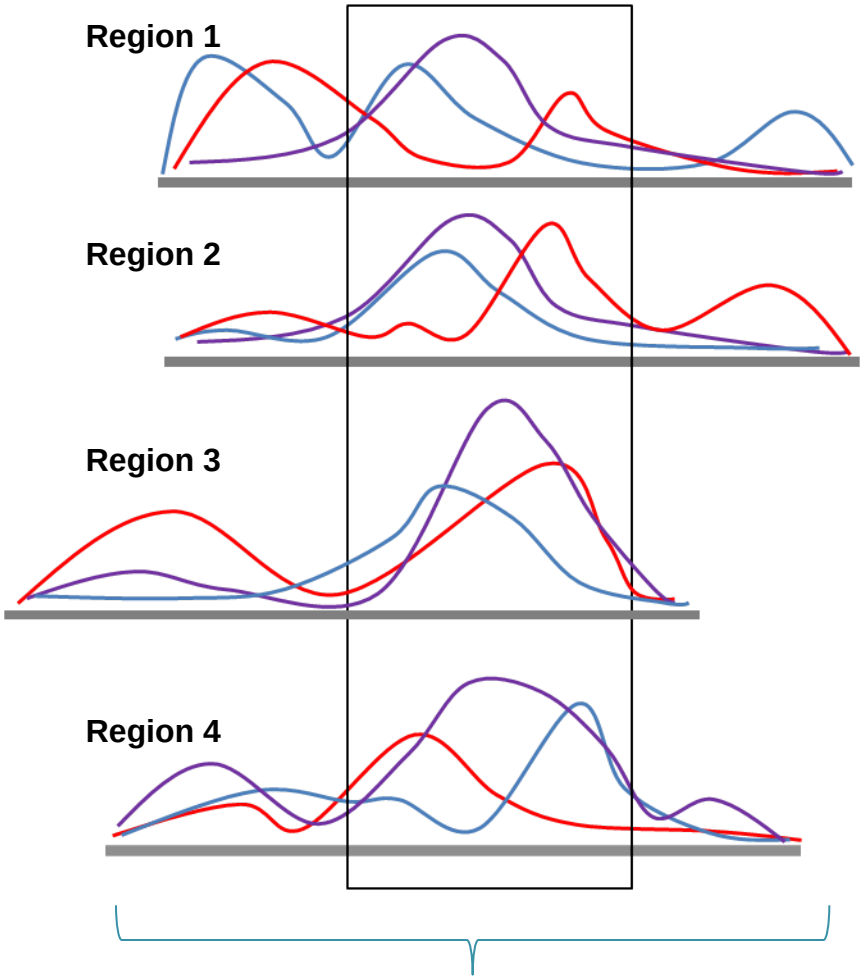
Heintzman et al. 2007
Nature genetics

Question:

Search for novel signatures without restricting on existing biological annotations.

Combinatorial Histone Modification Patterns

basic considerations



Locate sub-regions

Shift the locus

Reverse the strand

Regions with Enriched ChIP-seq Signals

Combinatorial Histone Modification Patterns

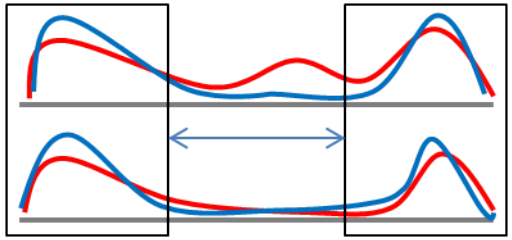
Available Algorithms to Discover Histone Codes:

A. Supervised Methods: Unable to discover novel histone modification patterns

B. Unsupervised Methods:
   a. ChromaSig: simultaneously update the probabilistic model and locations;
   b. CoSBI: coherent and shifted bicluster identification;

Problems with these two unsupervised methods:
1. restrict to patterns with fixed length threshold (e.g. <7kb);
2. only co-located histone modification patterns with single modes;

3. do not consider flexibility of histone modification distributions;



4. greedy search of motif seeds may miss potential patterns (ChromaSig);

Combinatorial Histone Modification Patterns

Available Algorithms to Discover Histone Codes (continue):

Hidden Markov Model based method:

1. Solved the more generalized question: chromatin states;

2. Also provide the spatial relationships between states;

3. Not restricted to specified pattern sizes;

4. Do not need pre-set parameters to initialize the inference.

Ernst & Kellis 2010 Nature Biotechnology
Ernst et al. 2011 Nature

Major differences from HMM based method:

1. Instead of transition probabilities of spatial relationships between patterns, closely spaced patterns are grouped into a single complex pattern.

2. The method and criteria of deciding the number of patterns are different.

3. Patterns consisting with the same marks are classified to be different patterns if their shapes are different.

Combinatorial Histone Modification Patterns

Basic Idea: alignments on histone modification profiles

- ❖ no pre-set region length thresholds needed;
- ❖ no motif seeds needed;
- ❖ gaps: deal with the flexibilities of histone modification distributions;
- ❖ can find patterns with multiple modes;

Combinatorial Histone Modification Patterns

ChAT Algorithm (**Ch**romatin profile **A**lignments with **T**ree clustering)

Step 1: Data Transformation.

Step 2: Pairwise Alignment of combinatorial histone modification profiles:

1. employ dynamic programming;
2. locate similar sub-regions, shift and reverse the regions when necessary;
3. calculate p-values for pairwise similarities;

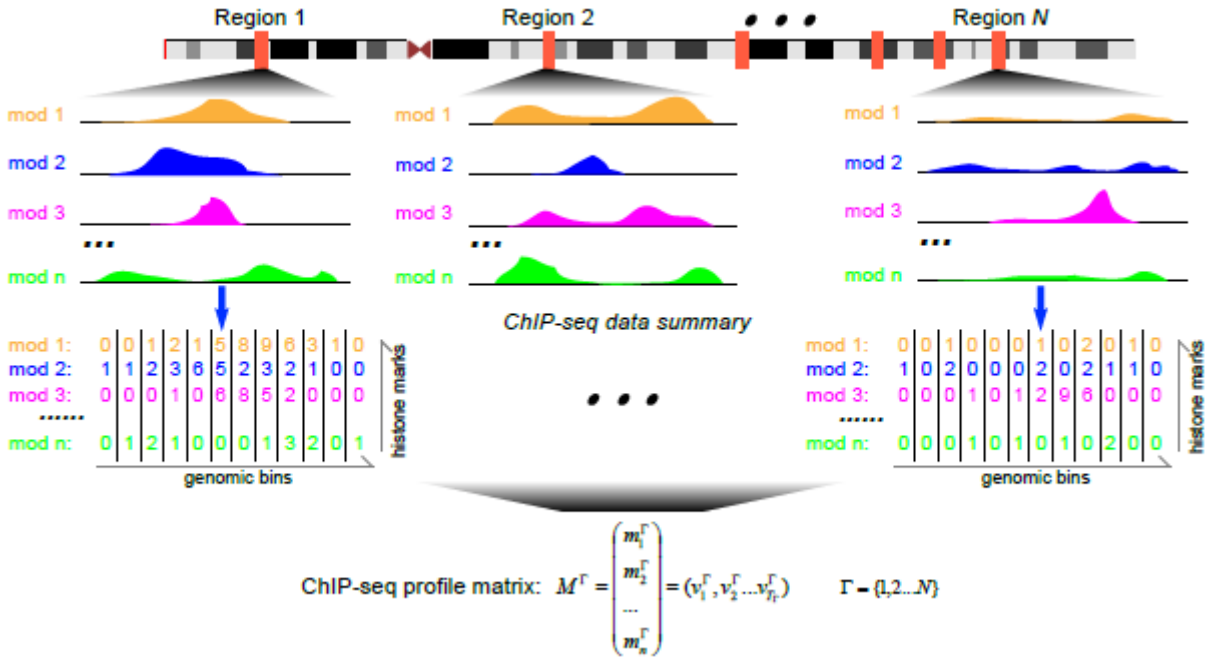
Step 3: Hierarchical Tree of highly similar sub-regions:

1. clustering on the matrix of p-values of pairwise similarity;
2. derive the combinatorial pattern profiles.

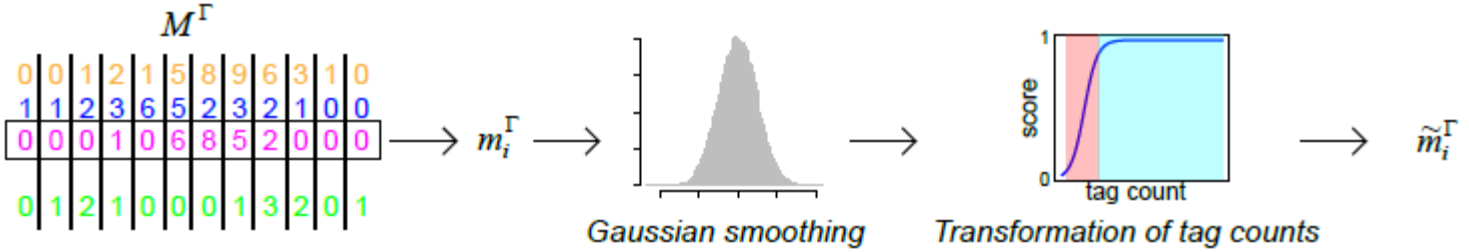
Wang et al. in preparation

Combinatorial Histone Modification Patterns

Data Transformation



1. Gaussian smoothing on the ChIP-seq tag counts;
2. Transformation: suppress the differences in small and large tag count ranges;



3. Closely adjacent regions with enriched ChIP-seq signals are grouped into a single region.

Combinatorial Histone Modification Patterns

Pairwise Alignment of Histone Modification Profiles

1. Dynamic Programming on modification profiles;

2. Cosine similarity between vectors;

$$\tilde{s}_{ij} = \cos(f \cdot \arccos(\frac{v_i^A \bullet v_j^B}{\left|v_i^A\right|\left|v_j^B\right|}))$$

3. Cosine similarity is weighted:
similar vectors with small norms
contribute less than
similar vectors with large norms;

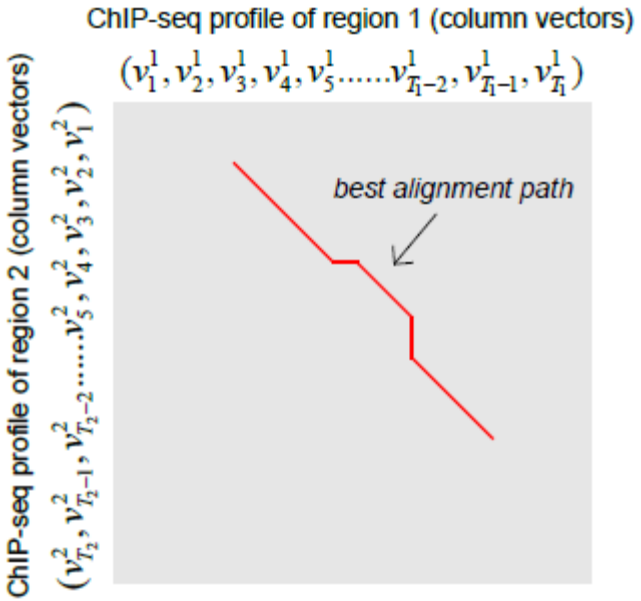
$$s = w \cdot \tilde{s}_{ij}$$

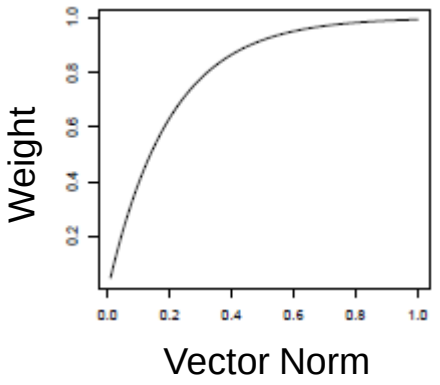
$$w = 1 - e^{-T \cdot m_{ij}}$$

$$m_{ij} = \max\{\left|v_i^A\right|, \left|v_j^B\right|\}$$

4. Gap penalty is proportional to the norm of the vector:
vectors with large norms are penalized
more to align with gaps;

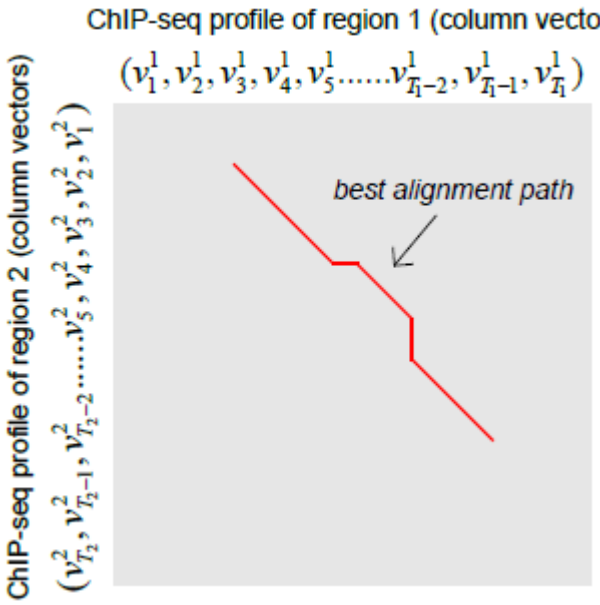
$$g_i^A = k \cdot \left|v_i^A\right|$$

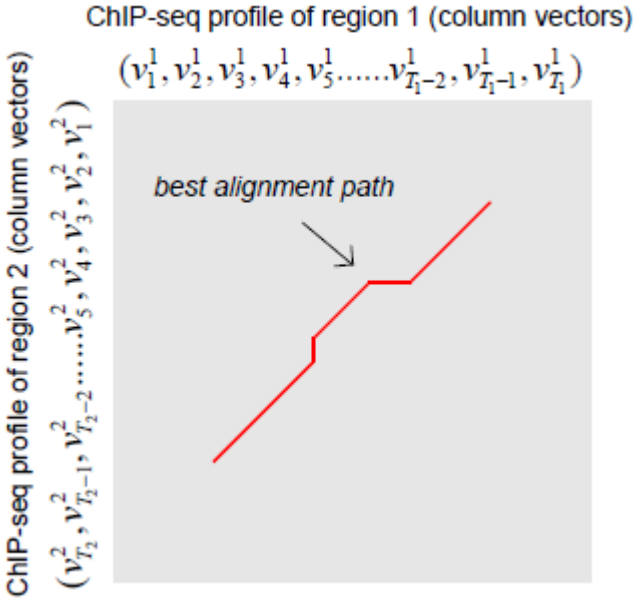




Combinatorial Histone Modification Patterns

5. Each pair of regions are aligned twice: including opposite orientations;





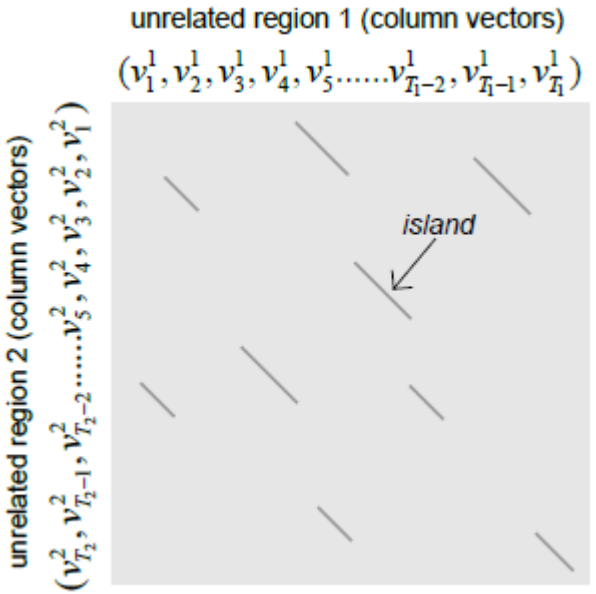
6. p-value is derived based on island method.

$$p \approx 1 - e^{-Kmne^{-\lambda x}}$$

$$\lambda_c = \ln(1 + \frac{1}{S_c})$$

$$K_c = \frac{R_c e^{\lambda_c c}}{A}$$

$$\overline{S_c} = \frac{1}{R_c} \sum_{i \in I_c} [S(i) - c]$$

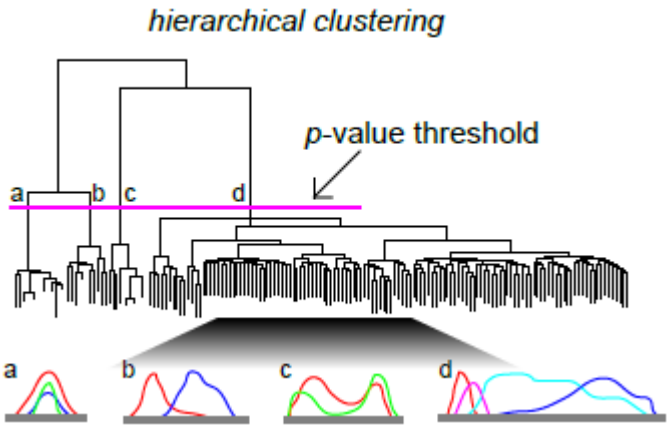


Combinatorial Histone Modification Patterns

Hierarchical tree of highly similar sub-regions

	Region 1	Region 2	. . .	Region *N*
Region 1	p_{11}	p_{12}	.....	p_{1N}
Region 2	p_{21}	p_{22}	.....	p_{2N}
Region 3	p_{31}	p_{32}	.....	p_{3N}
.				
.	⋮	⋮	⋱	⋮
.				
Region *N*	p_{N1}	p_{N2}	.....	p_{NN}

← pairwise p-values



← Use the given p-value threshold to cut the tree

combinatorial histone modification patterns

Combinatorial Histone Modification Patterns

Results:

1. ChAT is applied on the genomic maps of histone modifications in human $CD4^+$ T cells;

2. The resulted combinatorial patterns are classified into 3 size groups: small (<5kb), medium (>5kb, <10kb) and large (>10kb);

3. There are totally 144 small-size patterns;

4. Include patterns with multiple modes;

5. ~50% of those small-size patterns are enriched with DNase hypersensitive sites (fold enrichment > 3);

6. Those small patterns are related to different functional genomic features:
   TSS: transcription start site;
   TTS: transcription termination site;
   p300: transcription co-activator;

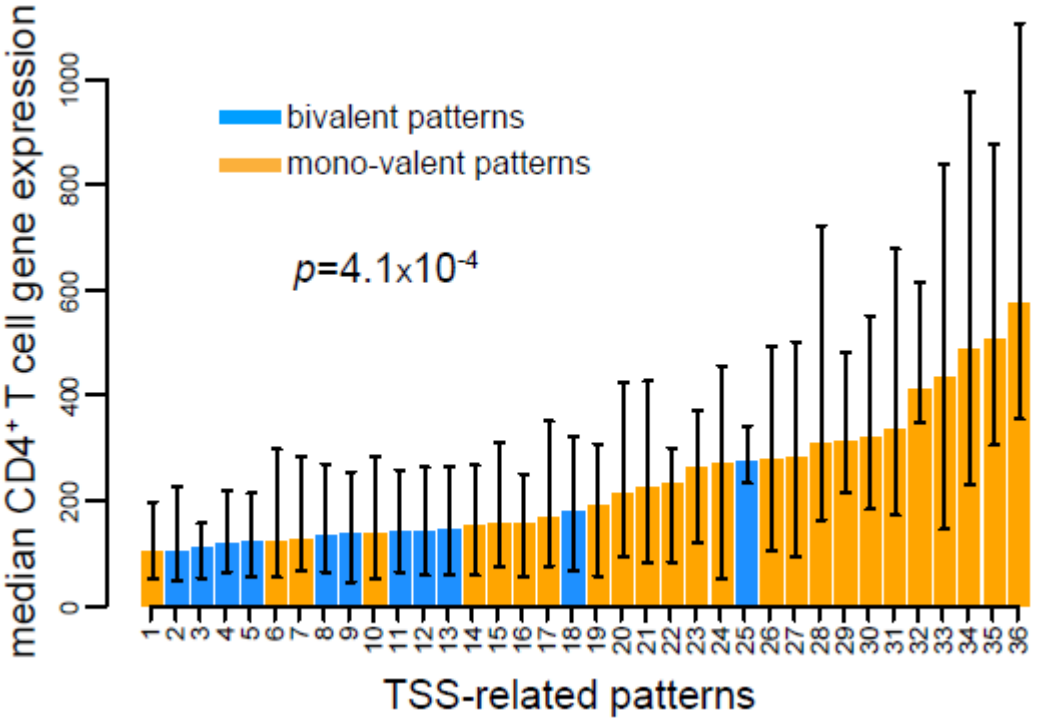
TSS-related patterns

1. There are 36 small-size patterns enriched with TSS (fold enrichment > 3);

2. Characteristic mark: H3K4me3 with various combinations of other marks (e.g. H3K27ac, H3K36ac, H3K4me1);

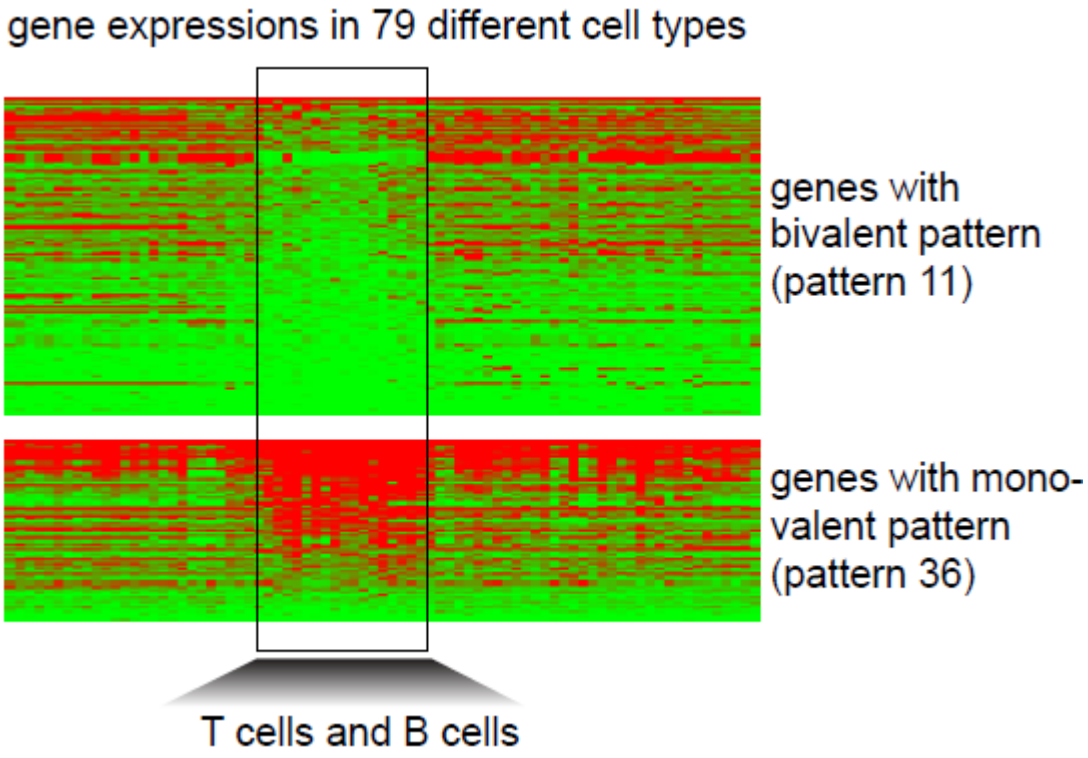
3. Also include bivalent patterns: H3K27me3 & H3K4me3;

4. Bivalent patterns are associated with lower expression levels.



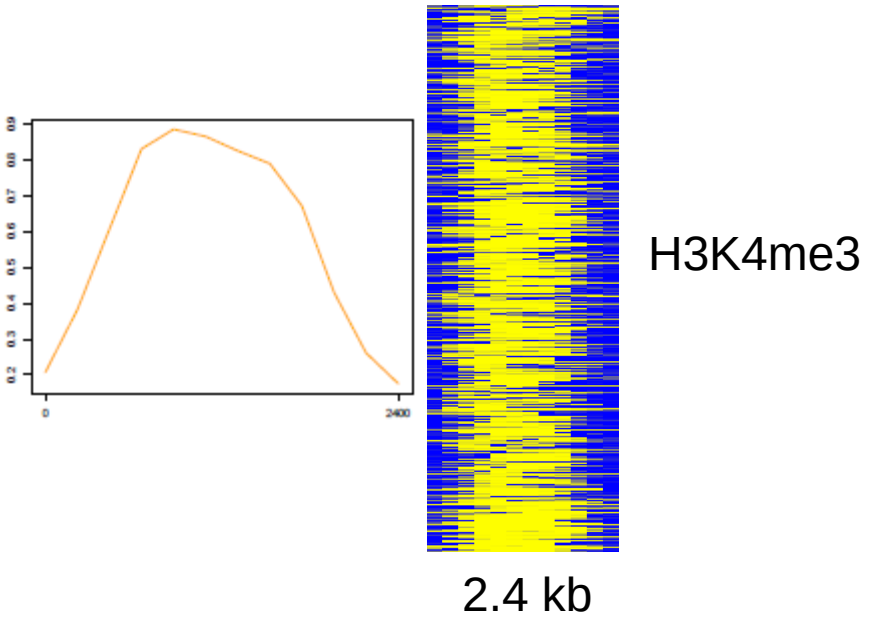
TSS-related patterns

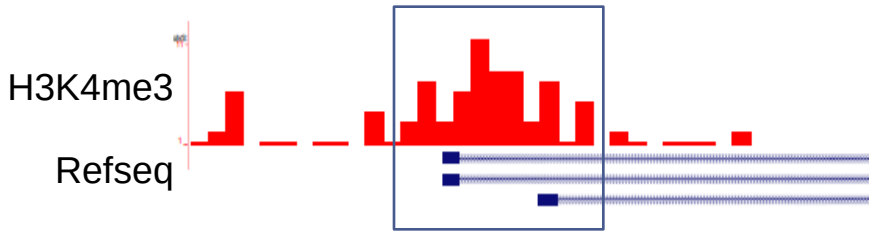
relationship with cell type specific expressions



TSS-related patterns

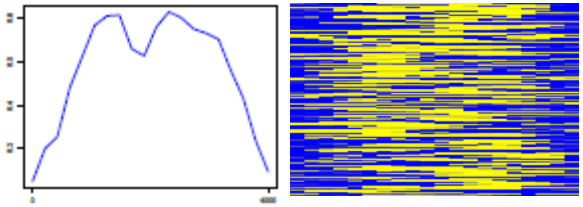
Examples:



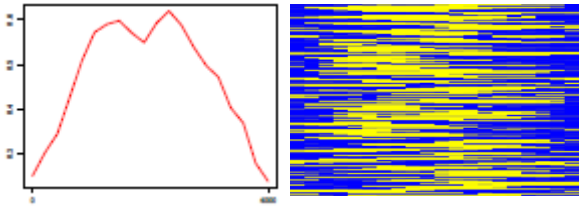


TSS-related patterns

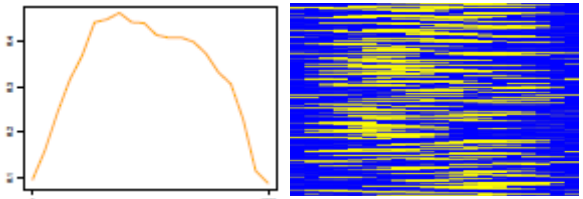
Examples:



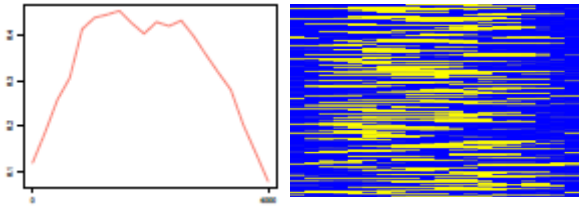
H2AZ



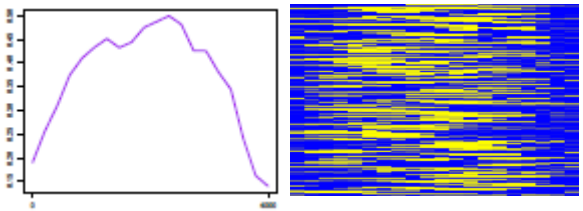
H3K4me3



H3K9me1

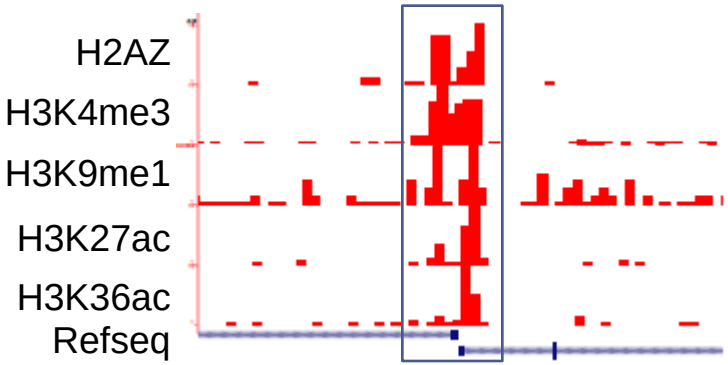


H3K27ac



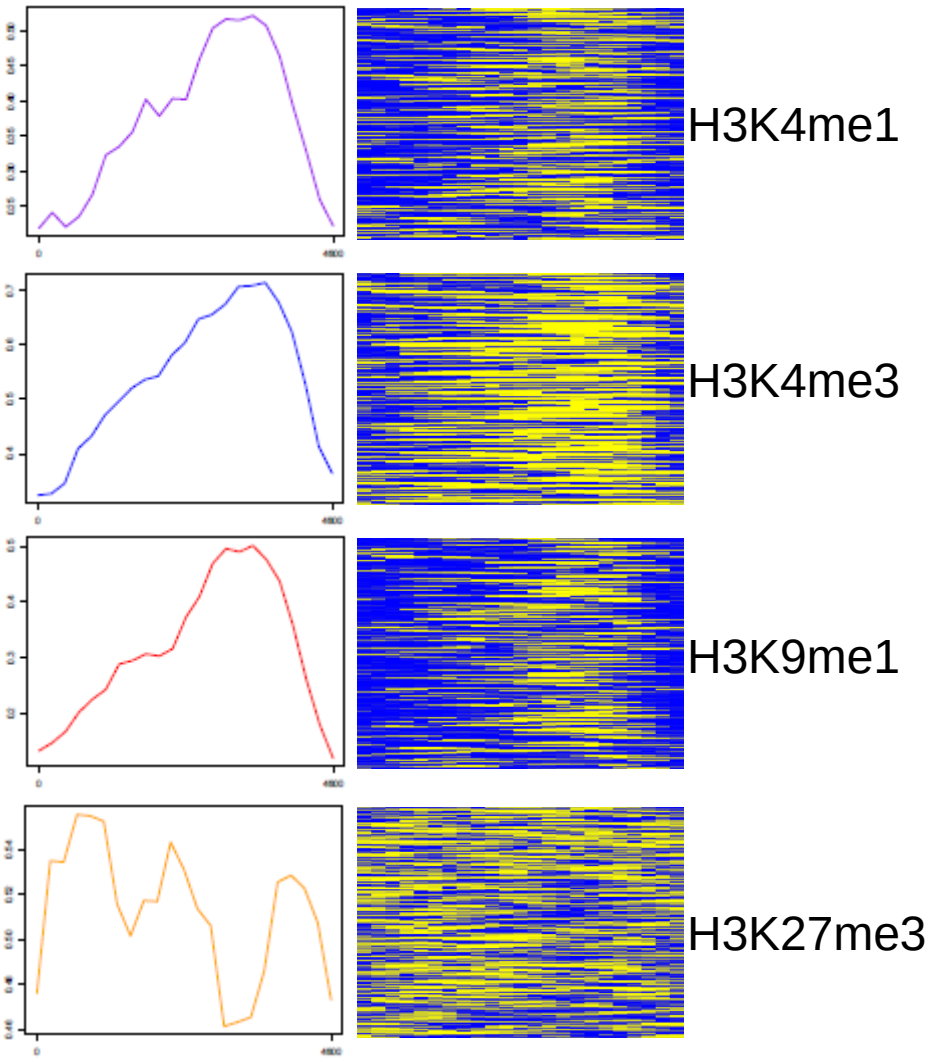
H3K36ac

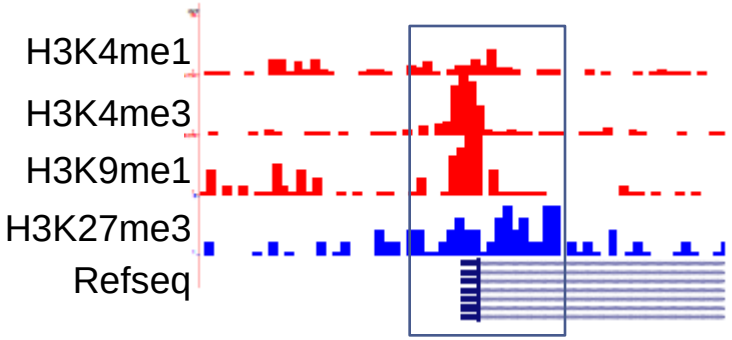
4 kb



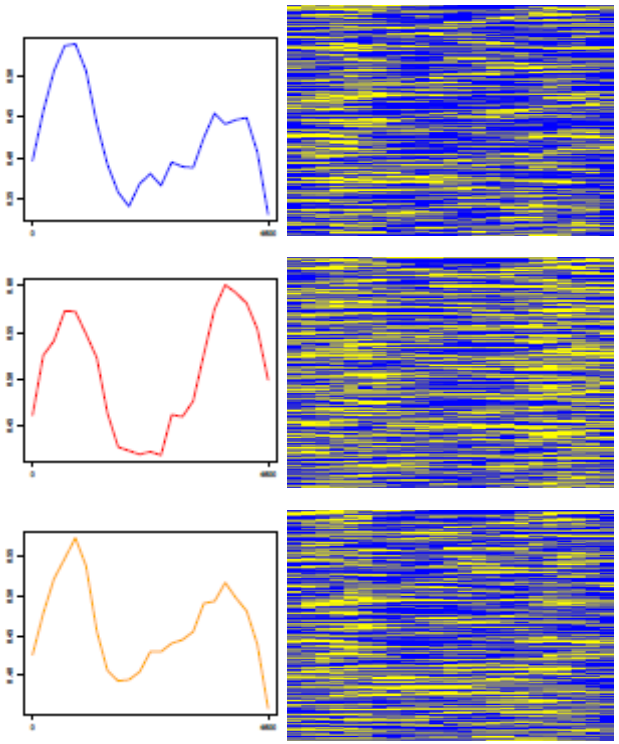
TSS-related patterns

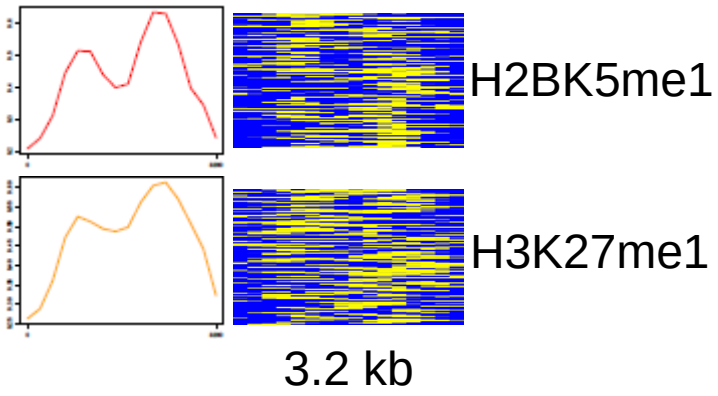
Example of bivalent patterns:

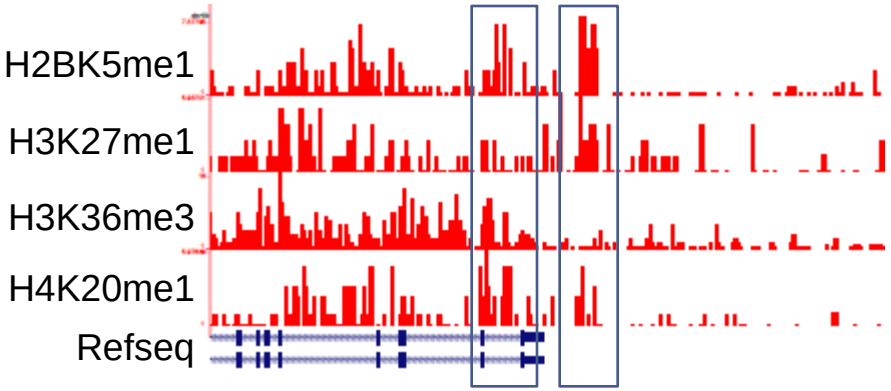




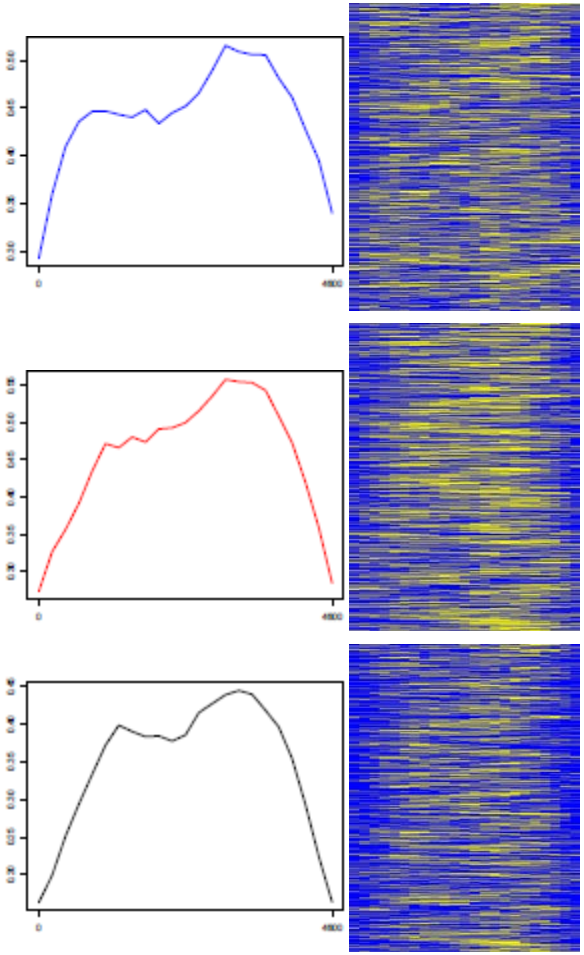
TTS-related patterns



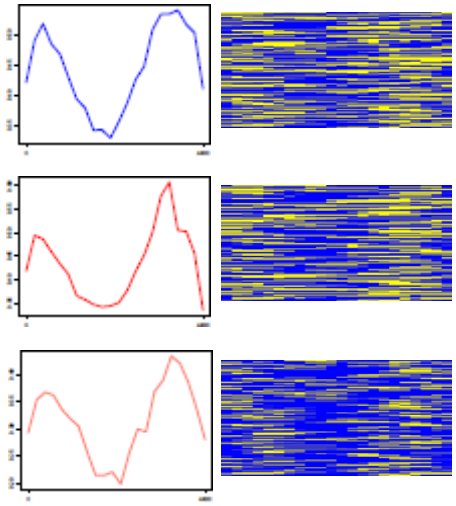




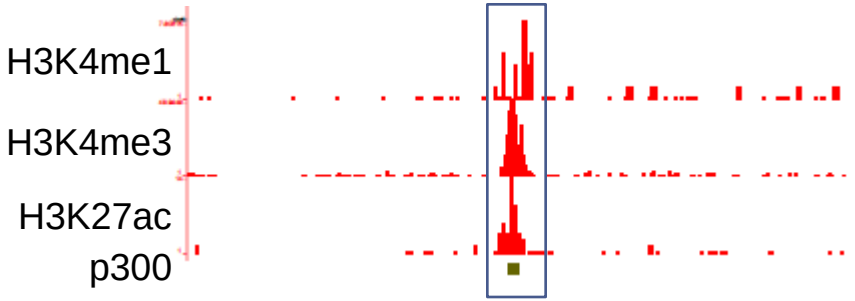
p300-related patterns



4.6 kb

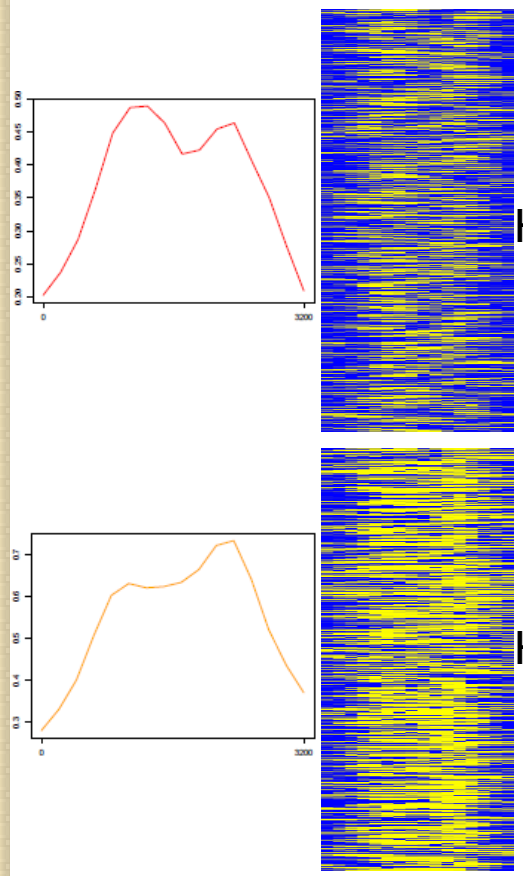


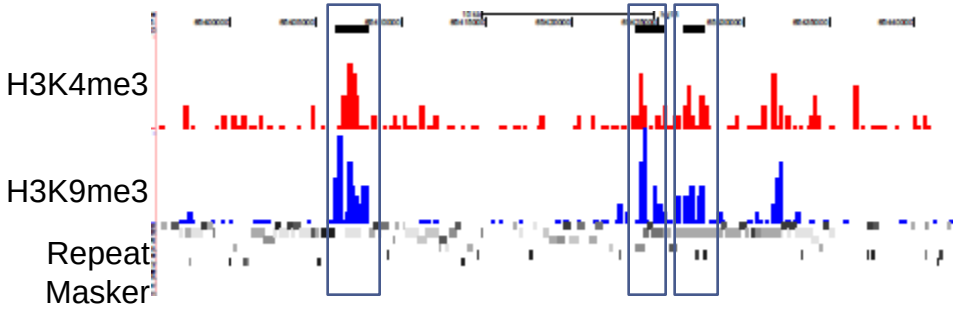
4.2 kb



H3K9me3-H3K4me3 bivalent pattern

1. Basically are not DNase hypersensitive;
2. 77.0% overlap with L1 retrotransposons;
3. 25.7% overlap with Alu retrotransposons;





H3K9me3-H3K36me3 bivalent pattern

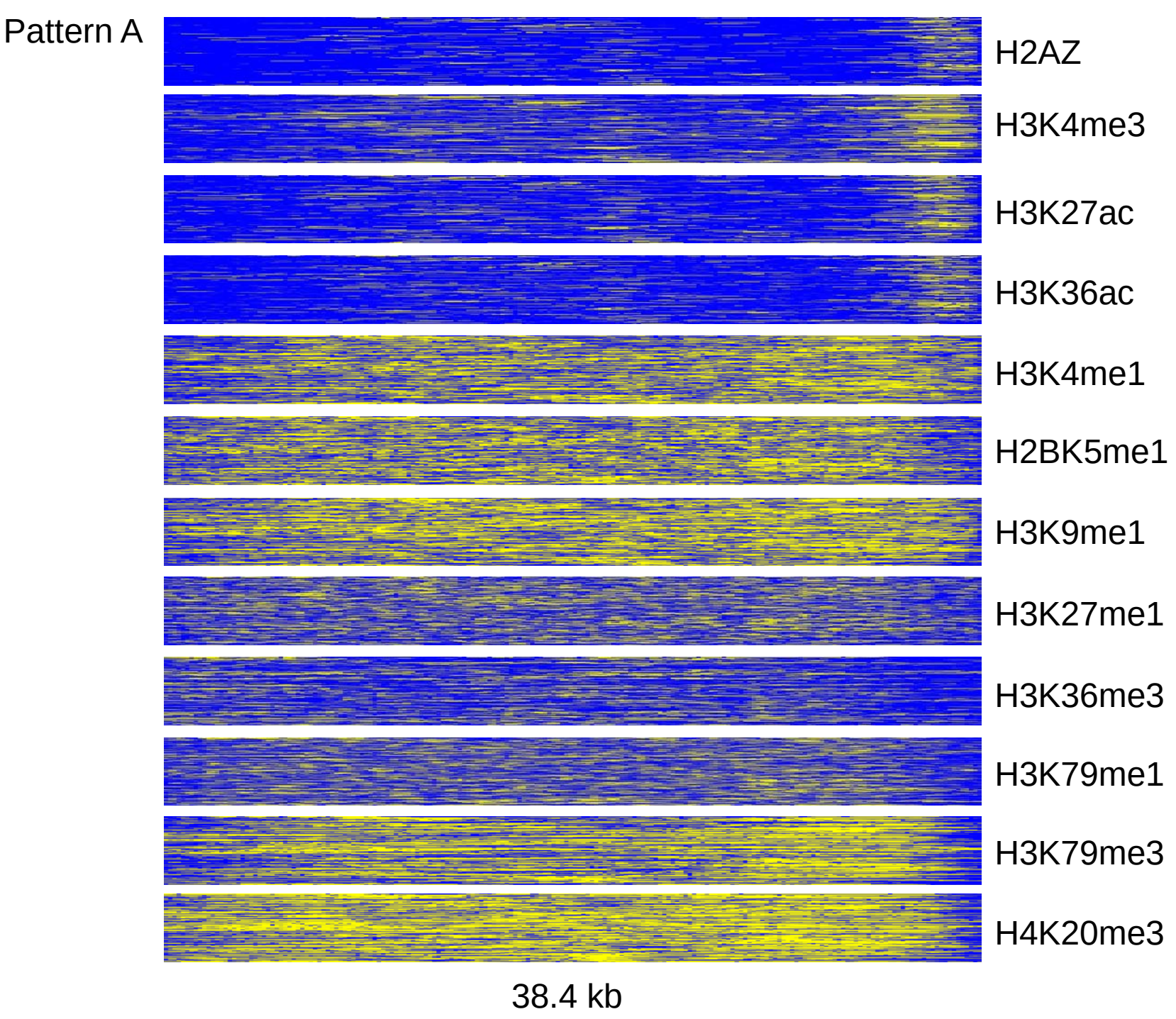
1. Basically are not DNase hypersensitive;
2. 68.4%% overlap with L1 retrotransposons;
3. 38.9% overlap with Alu retrotransposons;

Large-size patterns

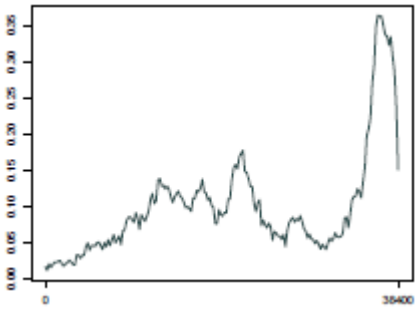
1. Generally two kinds of large patterns:

    a. Contiguous repressive marks;
    b. Gene-body related patterns;

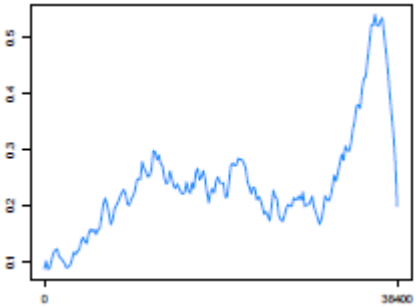
2. Gene-body related patterns resemble the “K4-K36” domain with variations of other marks.


Pattern A
H2AZ
H3K4me3
H3K27ac
H3K36ac
H3K4me1
H2BK5me1
H3K9me1
H3K27me1
H3K36me3
H3K79me1
H3K79me3
H4K20me3
38.4 kb

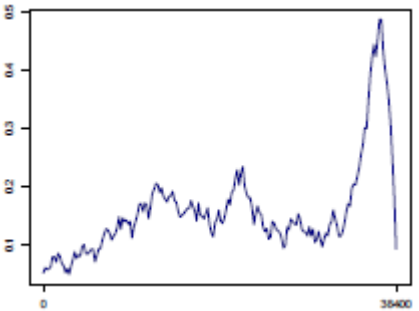
Pattern A



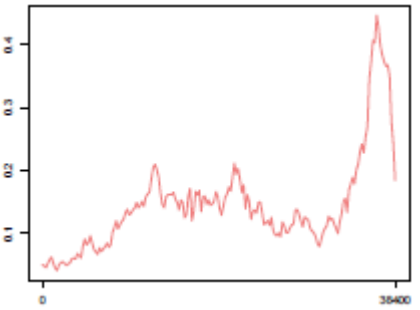
H2AZ



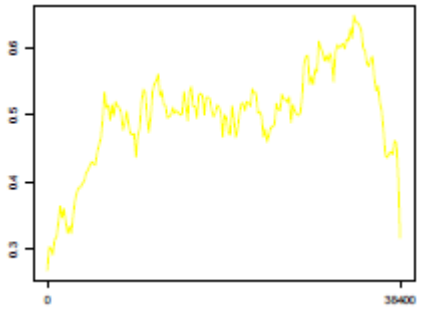
H3K4me3



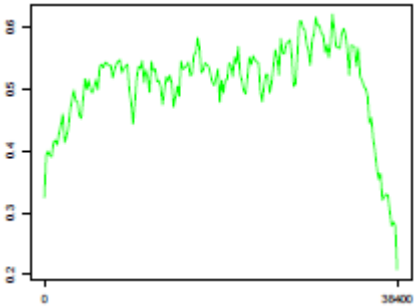
H3K27ac



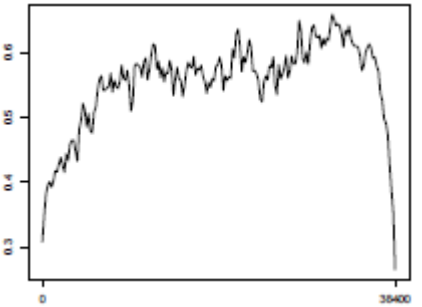
H3K36ac



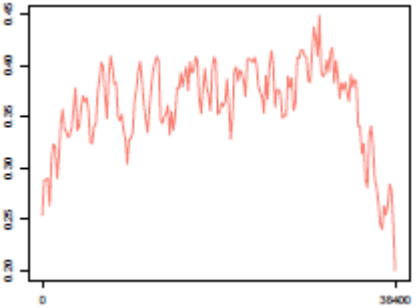
H3K4me1



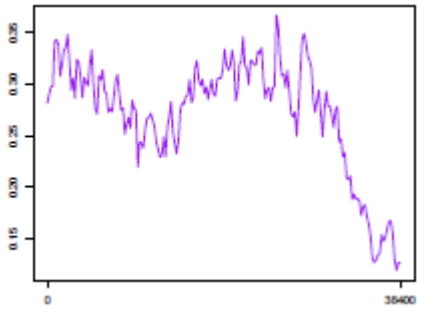
H2BK5me1



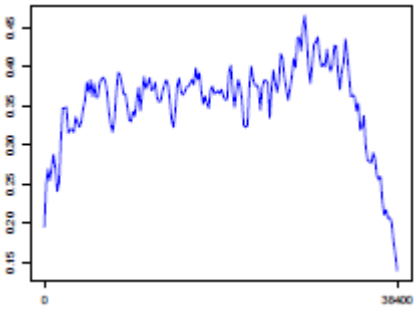
H3K9me1



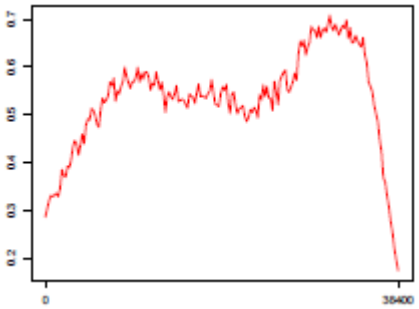
H3K27me1



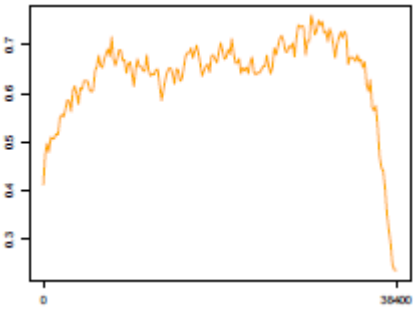
H3K36me3



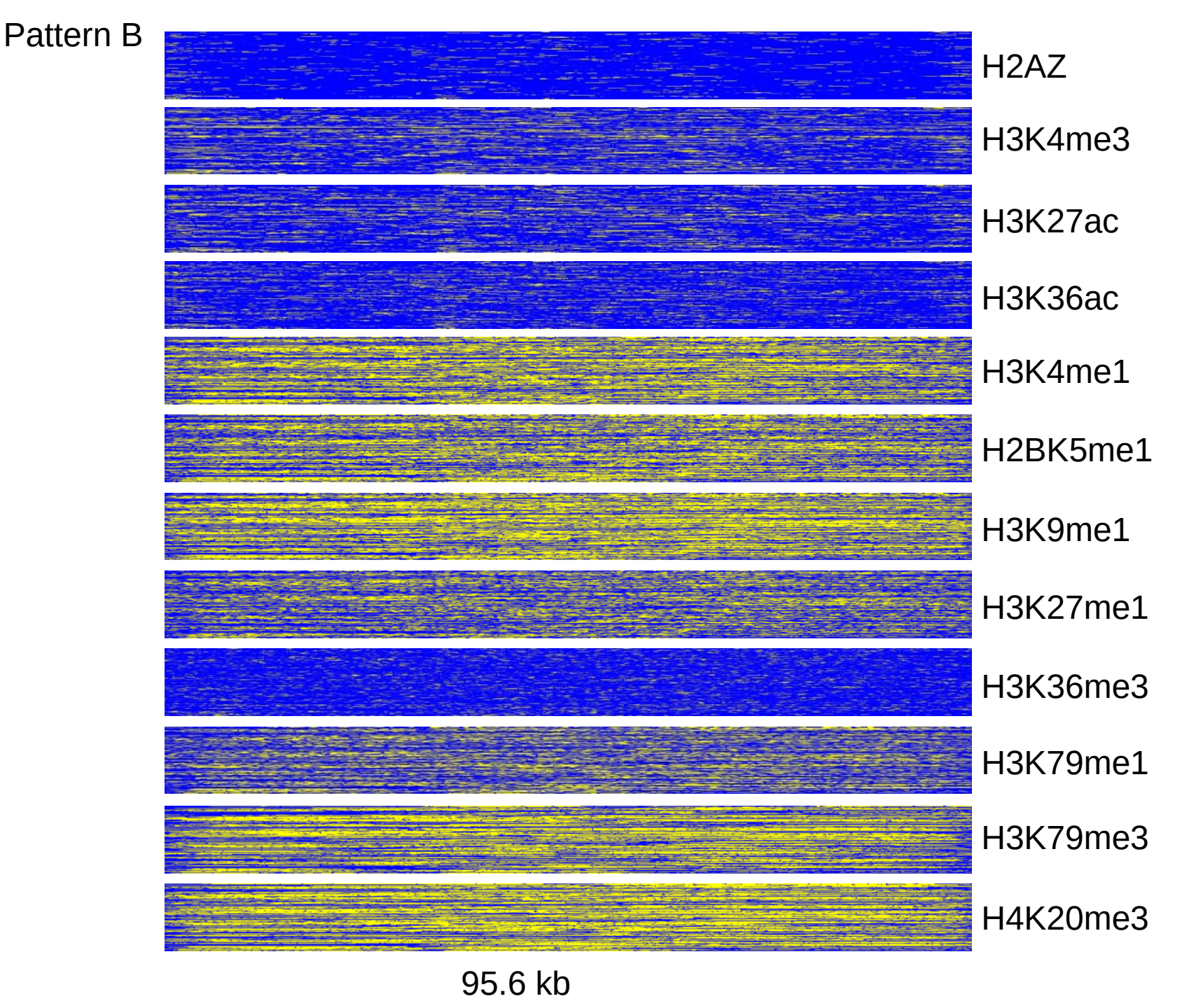
H3K79me1



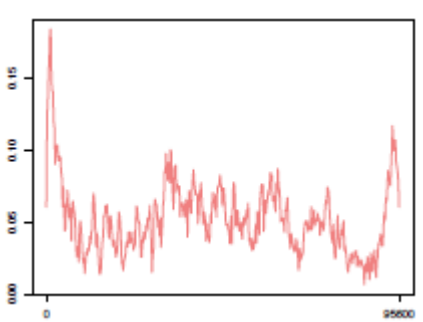
H3K79me3

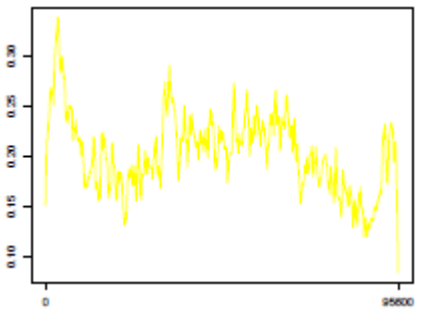


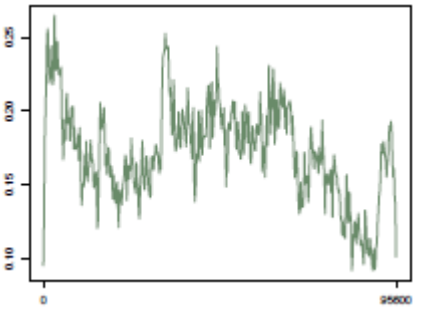
H4K20me3

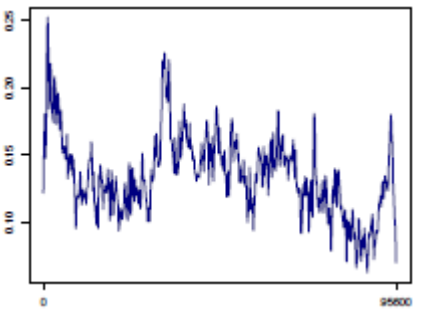

Pattern B
H2AZ
H3K4me3
H3K27ac
H3K36ac
H3K4me1
H2BK5me1
H3K9me1
H3K27me1
H3K36me3
H3K79me1
H3K79me3
H4K20me3
95.6 kb

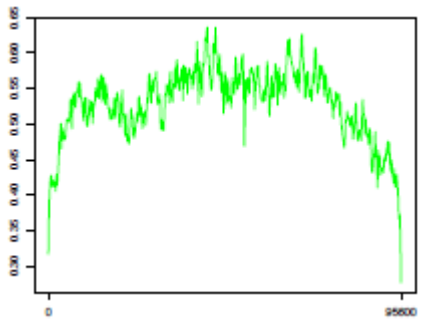
Pattern B

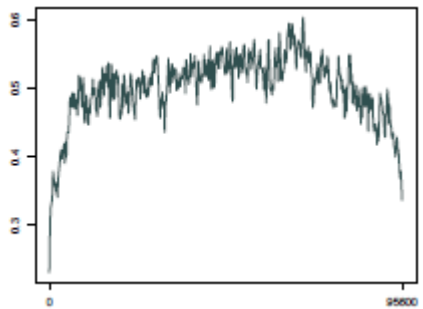

H2AZ

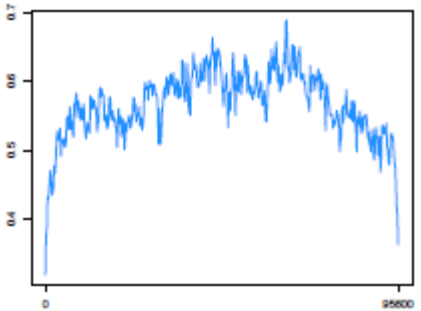

H3K4me3

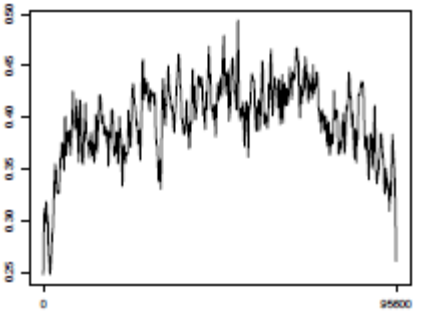

H3K27ac

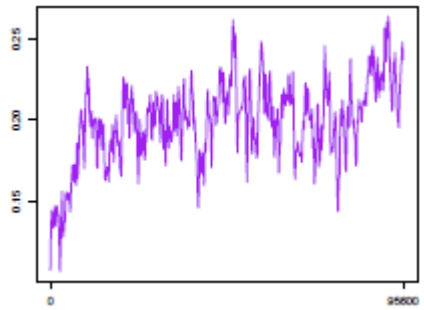

H3K36ac

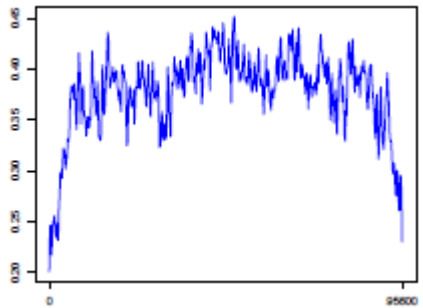

H3K4me1

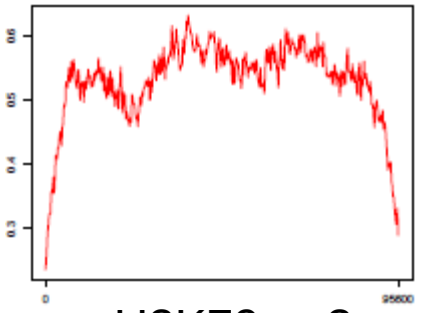

H2BK5me1

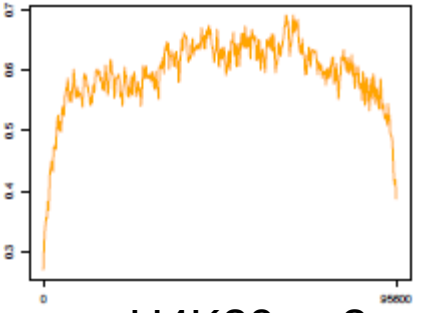

H3K9me1


H3K27me1


H3K36me3

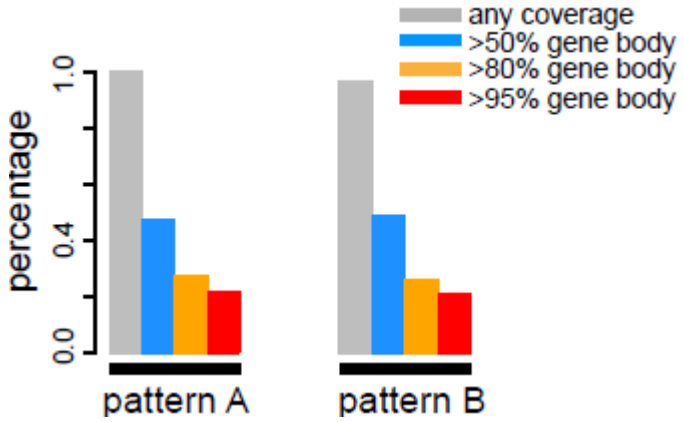

H3K79me1


H3K79me3


H4K20me3

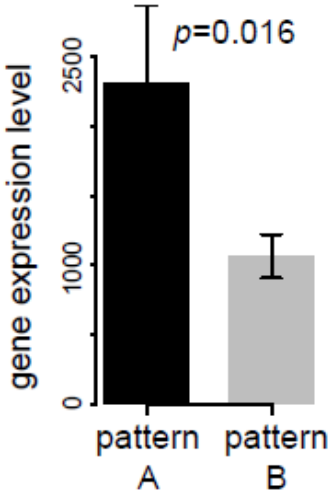
Large-size patterns

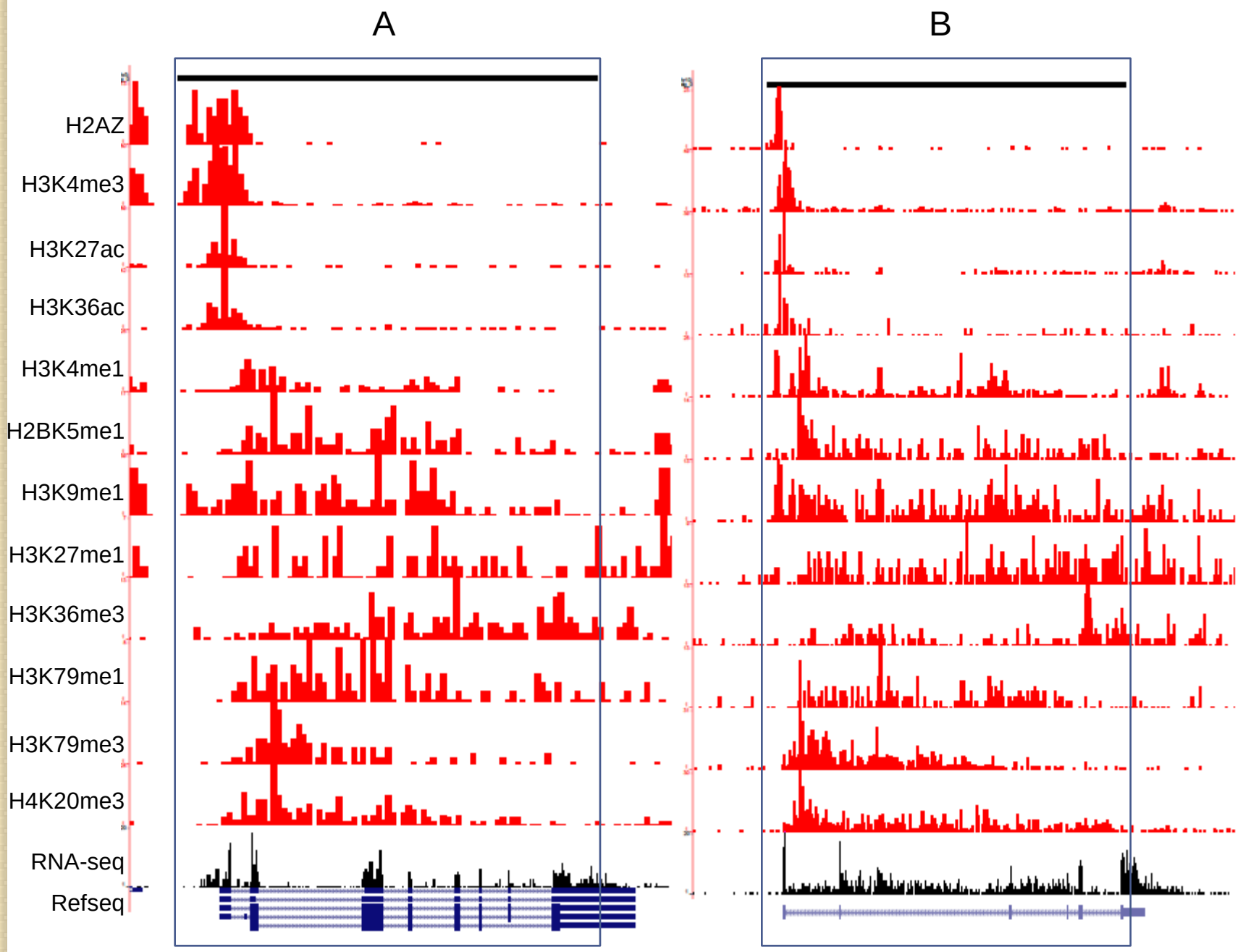
1. Both of the two pattern examples can mark gene bodies;

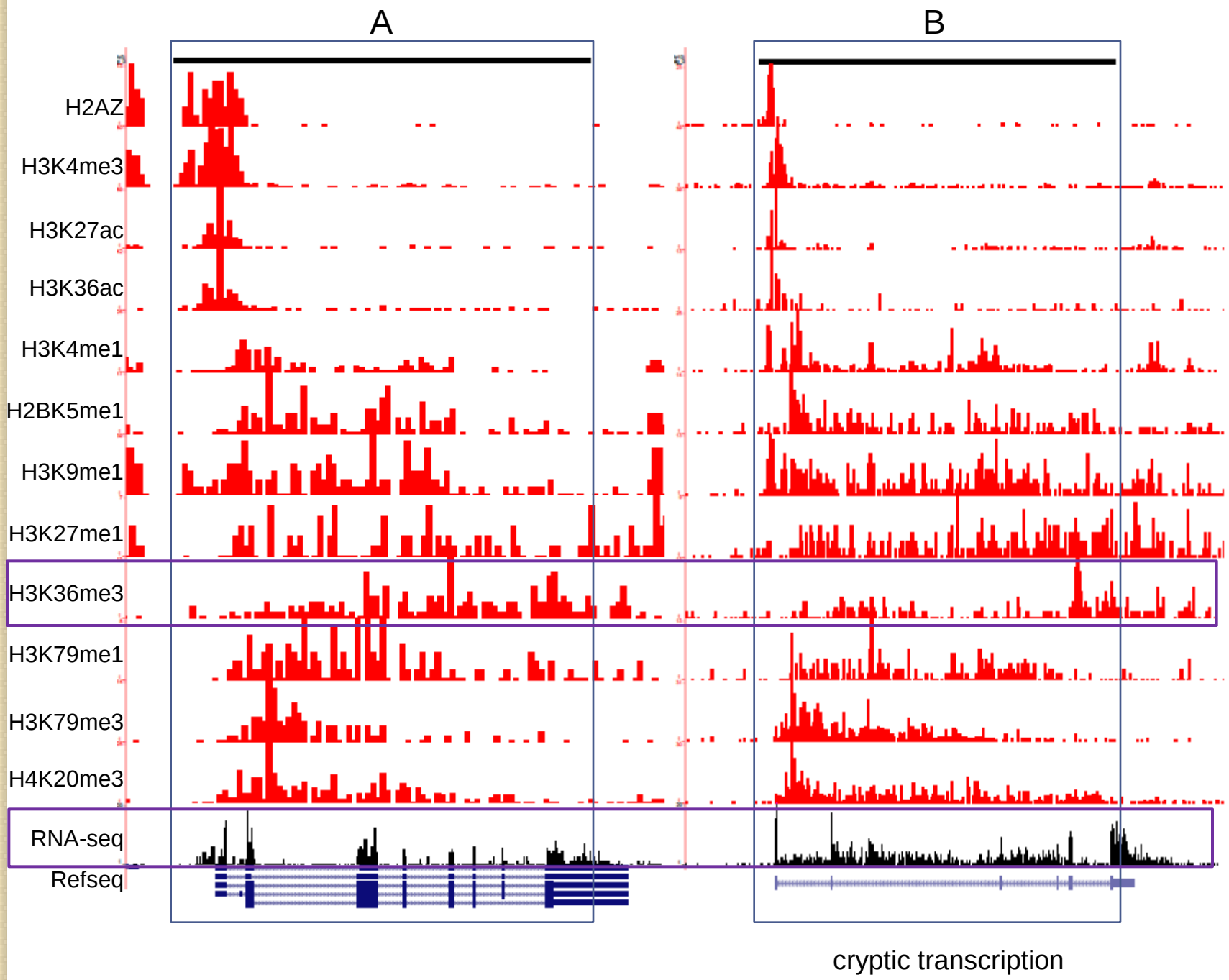


2. Difference between Pattern A and Pattern B: lower levels of H3K36me3 in Pattern B;

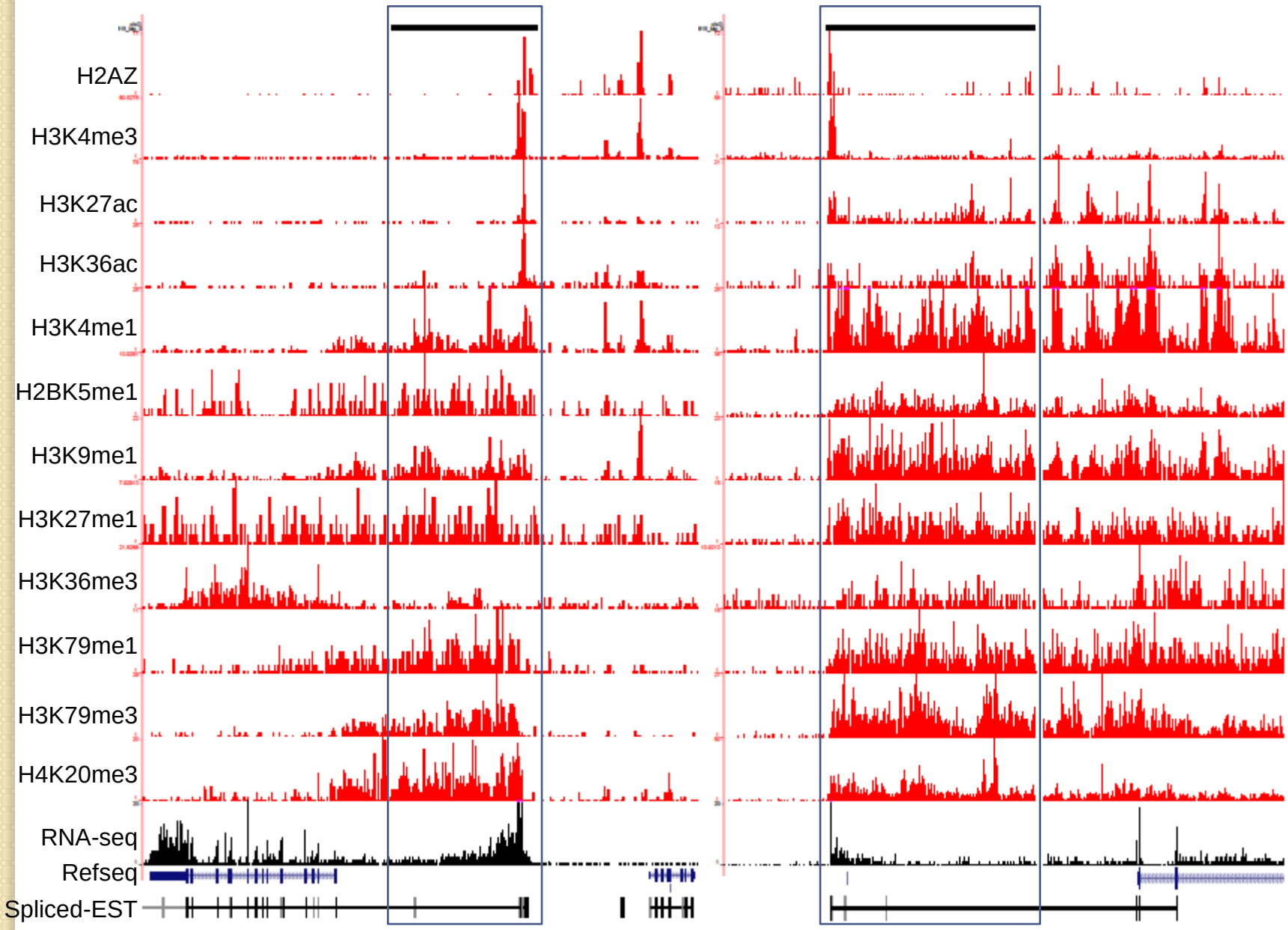
3. Different gene expression levels associated with these two patterns.




A
B
H2AZ
H3K4me3
H3K27ac
H3K36ac
H3K4me1
H2BK5me1
H3K9me1
H3K27me1
H3K36me3
H3K79me1
H3K79me3
H4K20me3
RNA-seq
Refseq


A
B
H2AZ
H3K4me3
H3K27ac
H3K36ac
H3K4me1
H2BK5me1
H3K9me1
H3K27me1
H3K36me3
H3K79me1
H3K79me3
H4K20me3
RNA-seq
Refseq
cryptic transcription

These large patterns have the potential to find un-annotated genes or long non-coding RNAs.



Combinatorial Histone Modification Patterns

Summary:

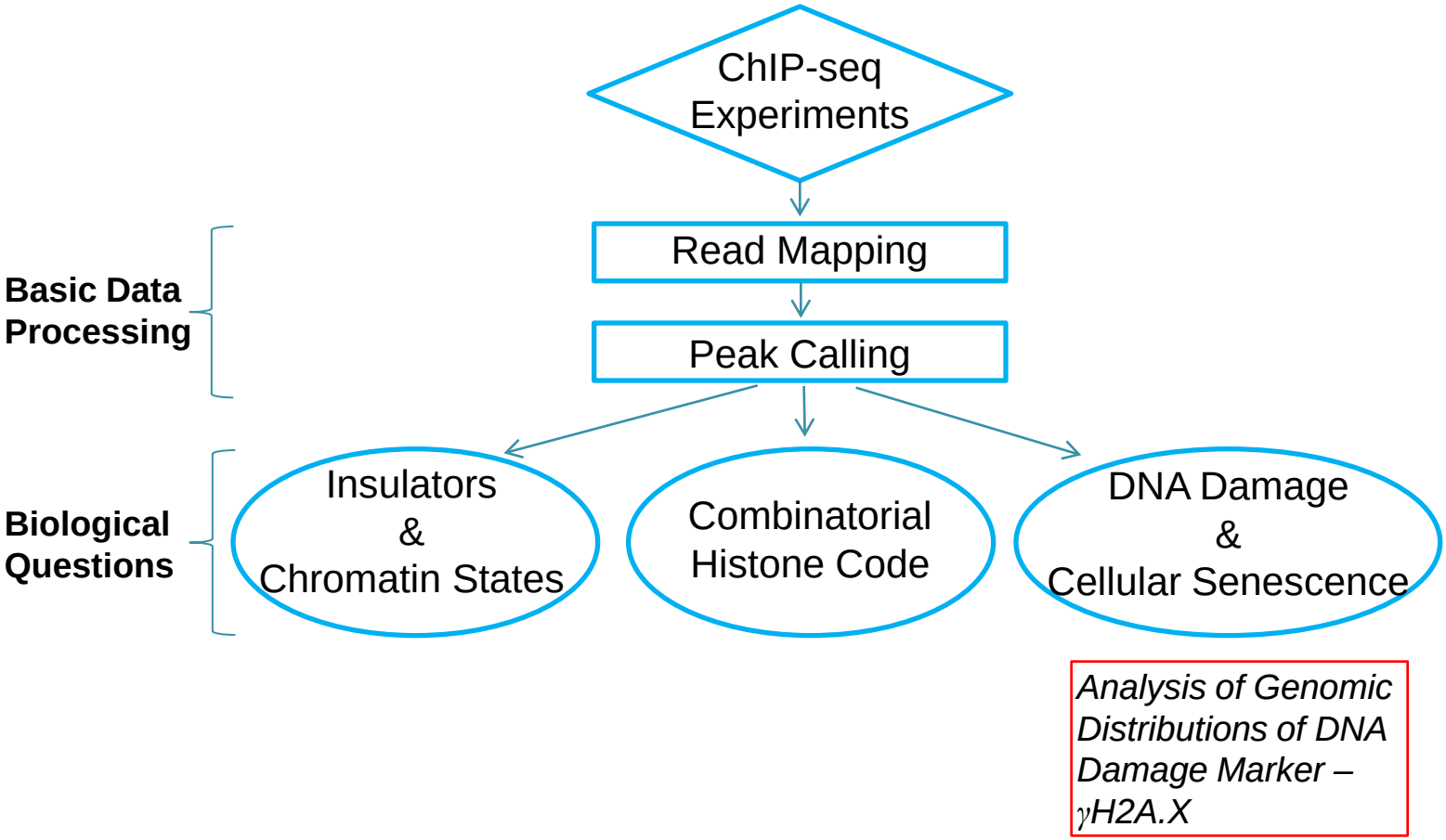
1. Developed a profile-alignment based algorithm to search combinatorial histone modification patterns;

2. Discovered both small-size and large-size patterns;

3. Many small-size patterns are related with functional genomic features;

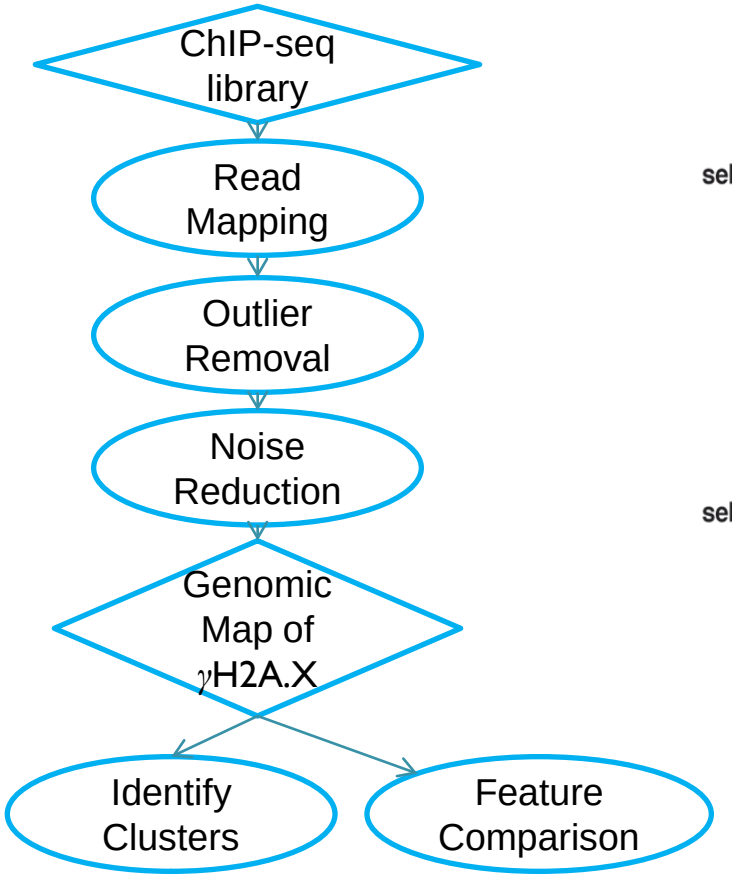
4. Some bivalent patterns;

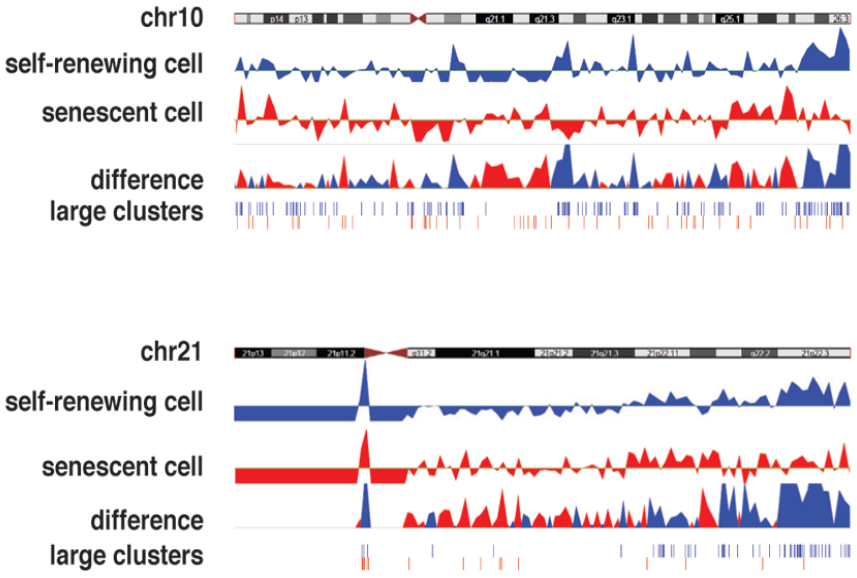
5. Some large-size patterns could be used to annotate genes or non-coding RNAs.


ChIP-seq Experiments
Read Mapping
Peak Calling
Basic Data Processing
Biological Questions
Insulators & Chromatin States
Combinatorial Histone Code
DNA Damage & Cellular Senescence
Analysis of Genomic Distributions of DNA Damage Marker – γH2A.X

Analysis of DNA damage marker: γH2A.X

1. DNA damage is highly related to cell senescence;
2. DNA damage sites: γH2A.X ;
3. Analyzed the genomic ChIP-seq data of γH2A.X in self-renewing and senescent human cells;

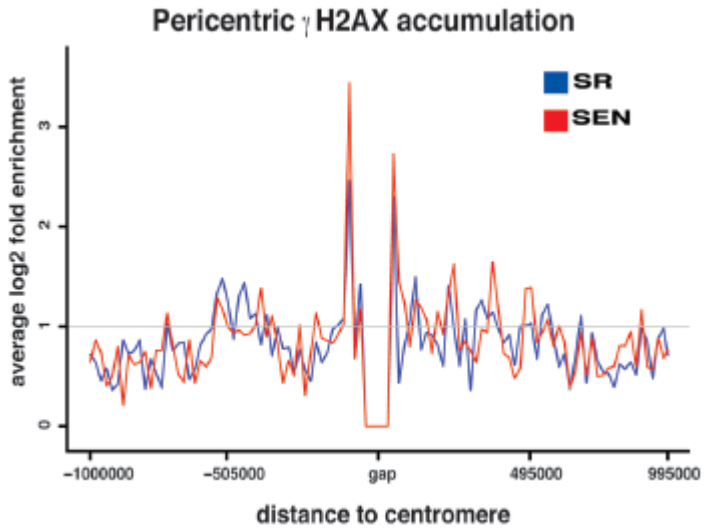




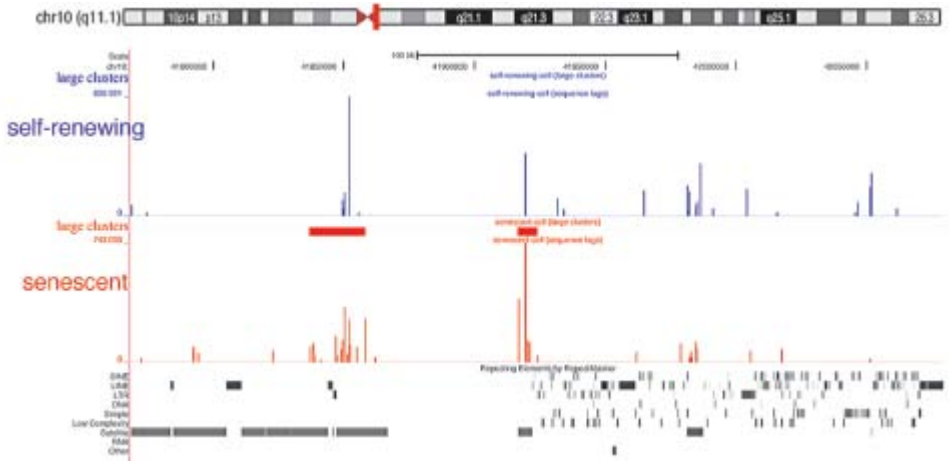
Wang et al. 2011 Cell Cycle
Wang et al. 2012 in preparation

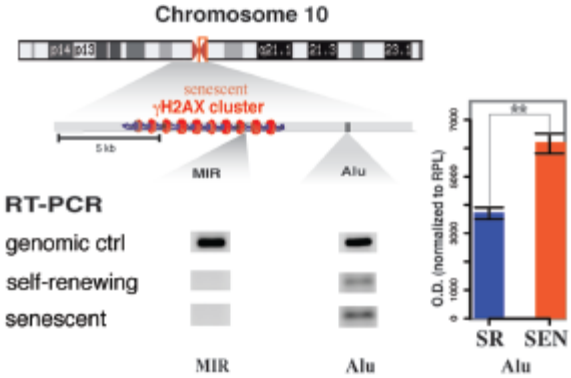
Analysis of DNA damage marker: γH2A.X

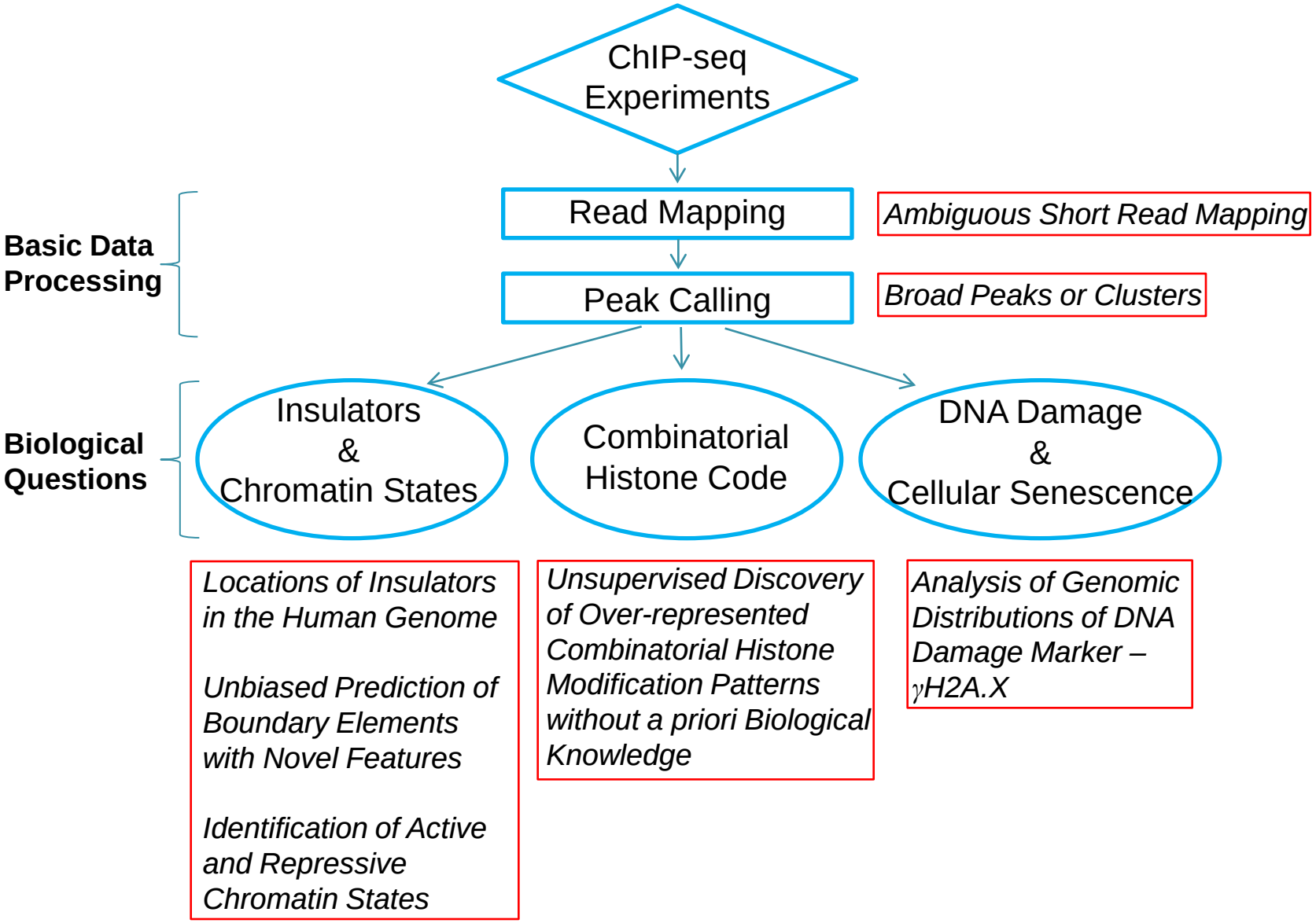
Compare γH2A.X distribution with genomic features: peri-centromeric regions.



Identify clusters of γH2A.X






ChIP-seq Experiments
Read Mapping
Ambiguous Short Read Mapping
Peak Calling
Broad Peaks or Clusters
Basic Data Processing
Biological Questions
Insulators & Chromatin States
Combinatorial Histone Code
DNA Damage & Cellular Senescence
Locations of Insulators in the Human Genome
Unbiased Prediction of Boundary Elements with Novel Features
Identification of Active and Repressive Chromatin States
Unsupervised Discovery of Over-represented Combinatorial Histone Modification Patterns without a priori Biological Knowledge
Analysis of Genomic Distributions of DNA Damage Marker – γH2A.X

On-going Projects:

1. Develop method to detect genomic sites with strong nucleosome phasing;

2. Comparative analysis of transcriptomes (RNA-seq) for self-renewing and senescent cells;

3. Look for histone modification signatures of HIV integration sites.

Future directions of interest

Analyze the dynamics of chromatin states along cell differentiation and infer the causal factors;

Develop computational methods to analyze Hi-C datasets and integrate 3D chromatin structure information with histone modification maps;

Integrate epigenetic information into gene regulatory network analysis.

Acknowledgement:

Advisor:
Prof. I. King Jordan;

Collaborators:
Prof. Victoria V. Lunyak;
Prof. Lluis Montoliu;
Cristina Vicente-Garcia;
Elbert Lee;

Members of Jordan Lab:
Andrew Conley;
Daudi Jjingo;
Aswathy Sebastian;
Kevin Lee;
Neha Varghese;
Robert Petit;
Angela Pena;