Evaluating Statistical Correlation Methods on Histone Modifications and Ultraconserved Genomic Regions

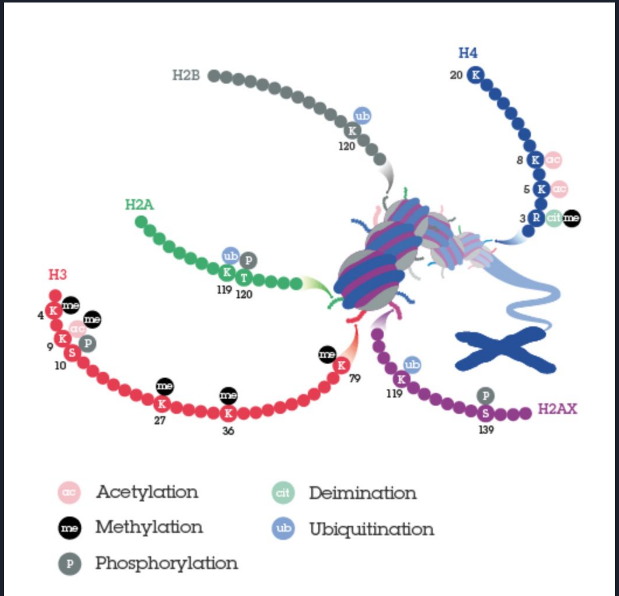
by Alex Porter, Laura Miron, Effie Nehoran, Anjini Karthik, and Armando Banuelos

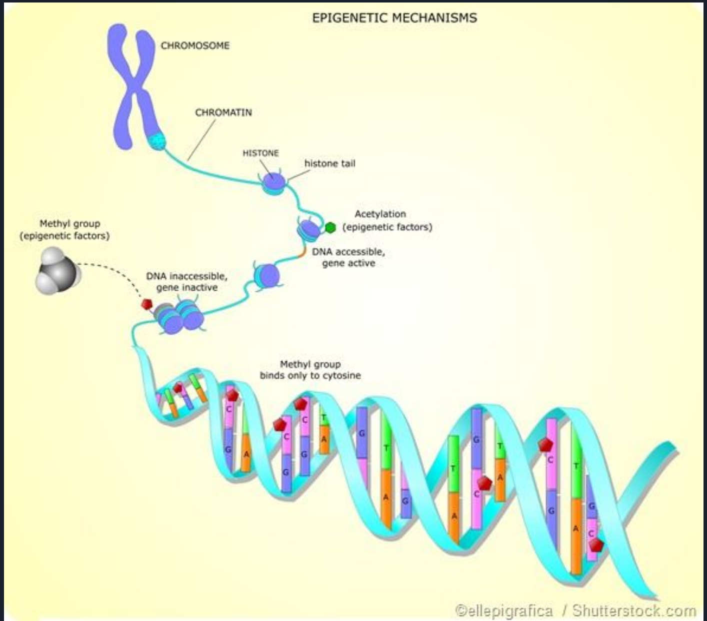
Outline

- **Introduction**
- **Hypothesis from Literature**
- **Methods**
  - Synthetic Data Generation
  - Simple Counting
  - Uniform Random Permutation
  - K-Clustering and **Second Hypothesis**
- **Results and Further Discussion**

What are **histone modifications**?

- Post translational modifications that regulate gene expression
- Can fall into 3 different categories: Enhancers, Promoters, Insulators



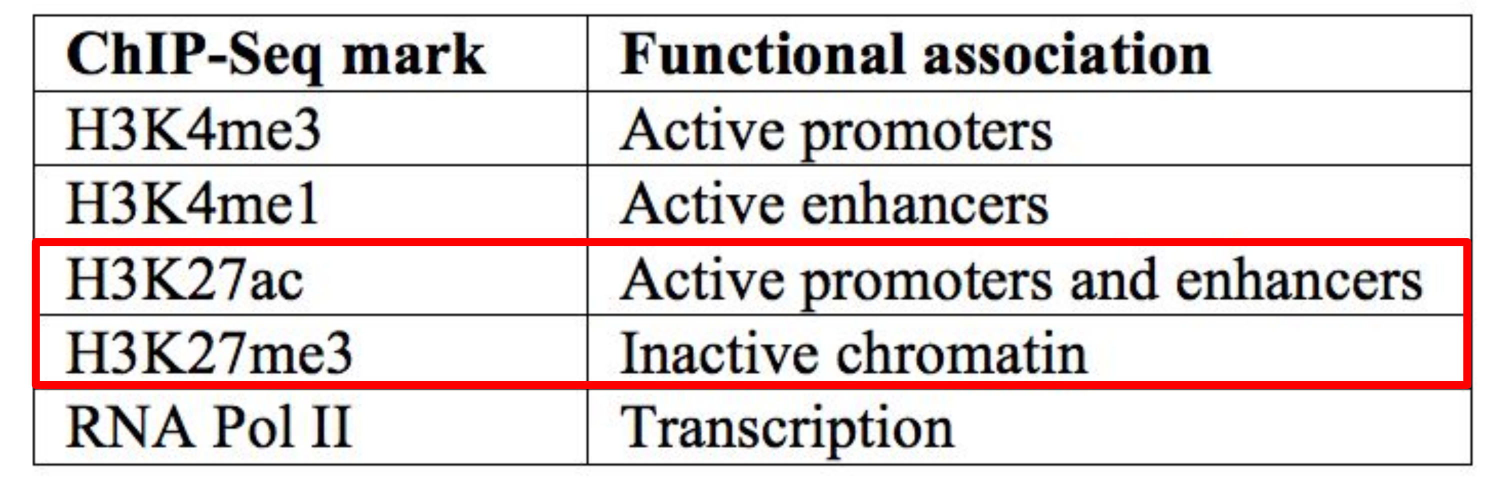


©ellepigrafica / Shutterstock.com

“Histone Modifications: A Guide”; http://www.abcam.com/epigenetics/histone-modifications-a-guide

Why do **Histone Modifications** matter?

- Transcriptional Activation/Inactivation
- Chromosome Packaging
- DNA Damage/Repair
- DNA Replication
- Gene Silencing



ChIP-Seq mark	**Functional association**
H3K4me3	Active promoters
H3K4me1	Active enhancers
H3K27ac	Active promoters and enhancers
H3K27me3	Inactive chromatin
RNA Pol II	Transcription

Why do we care about studying the **evolutionary conservation** of histone modifications?

“The majority of genetic variants associated with complex traits lie in non-coding regions of the genome, with many lying some distance away from the nearest protein-coding locus[1]. This observation implies that many variants affecting the risk of common, complex diseases are likely to exert their effect by altering the regulation of genes rather than by directly affecting gene and protein function.” (Graham R. S. Ritchie)

“That the landscape of genetic risk for psychiatric disorders will be largely non-coding aligns with overwhelming evidence that changes in gene-regulatory regions are a major contributor to risk for complex genetic traits: The majority (~93%) of disease-associated markers emerging from GWAS findings lie within the non-coding genome” (C.L. Bar)

Barr, C. L., and V. L. Misener. “Decoding the Non-Coding Genome: Elucidating Genetic Risk Outside the Coding Genome.” *Genes, brain, and behavior* 15.1 (2016): 187–204. *PMC*. Web. 13 Mar. 2018.

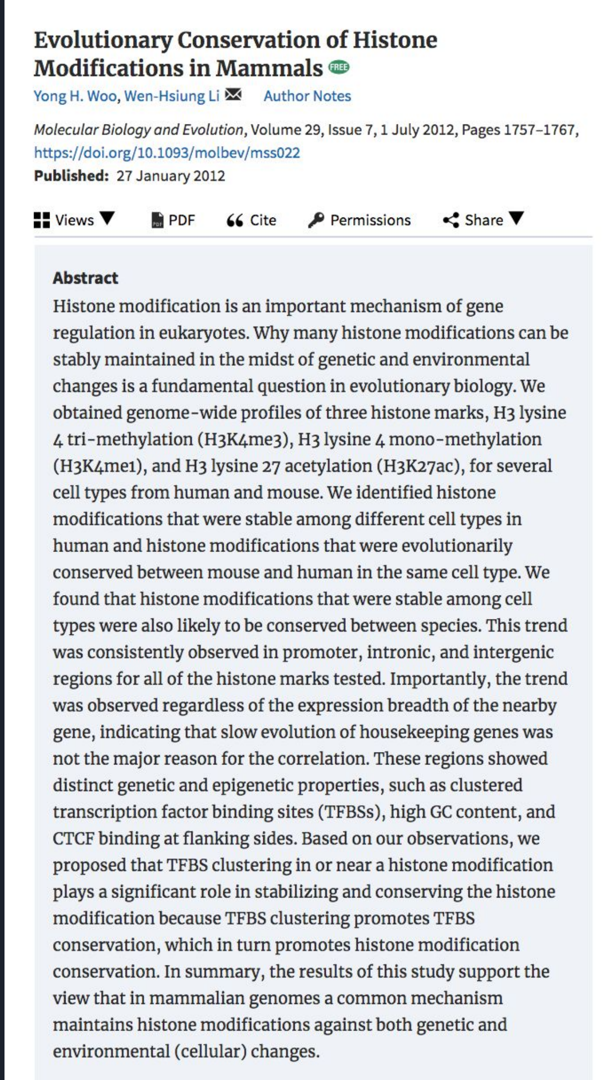
Ritchie, Graham R. S. et al. “Functional Annotation of Non-Coding Sequence Variants.” *Nature methods* 11.3 (2014): 294–296. *PMC*. Web. 13 Mar. 2018.

Why do we care about studying the **evolutionary conservation** of histone modifications?

If we understand how certain diseases are retained evolutionarily via histone modification, we are one step closer to uncovering a cure.

Hypothesis from Literature:

From Woo & Li (2012), we expect a high level of correlation between histone modifications (specifically H3K27ac) conserved evolutionarily and conserved across cell type.



Evolutionary Conservation of Histone Modifications in Mammals

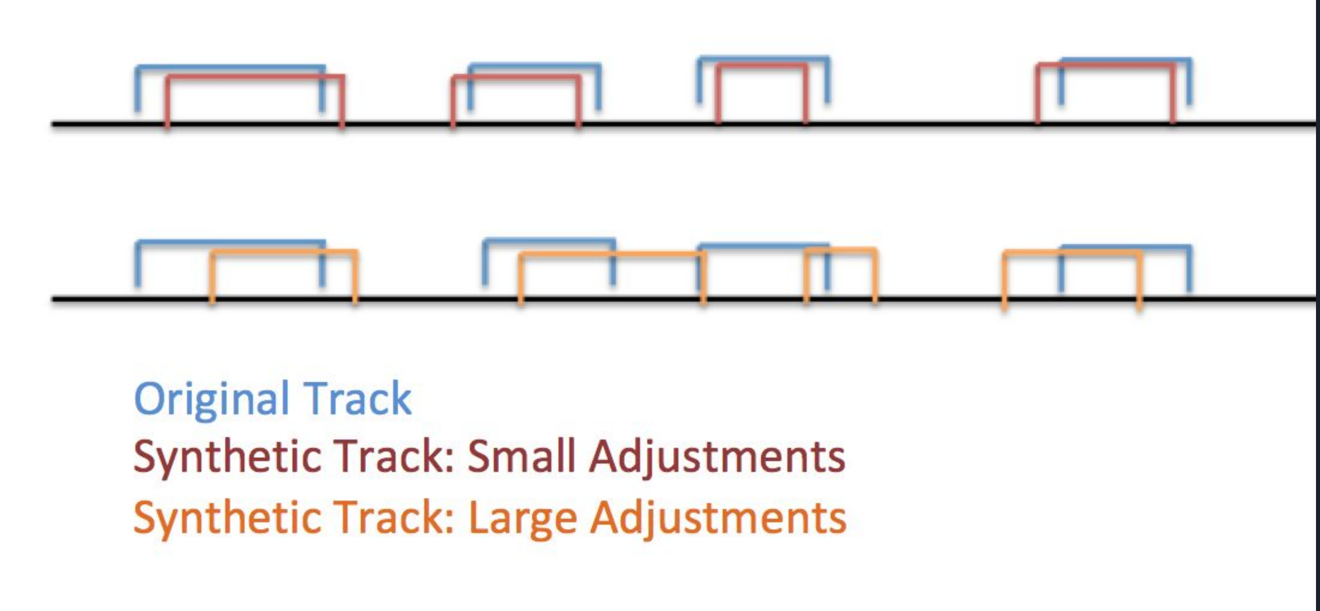
Yong H. Woo, Wen-Hsiung Li   Author Notes

Molecular Biology and Evolution, Volume 29, Issue 7, 1 July 2012, Pages 1757–1767, https://doi.org/10.1093/molbev/mss022
Published: 27 January 2012

Abstract

Histone modification is an important mechanism of gene regulation in eukaryotes. Why many histone modifications can be stably maintained in the midst of genetic and environmental changes is a fundamental question in evolutionary biology. We obtained genome-wide profiles of three histone marks, H3 lysine 4 tri-methylation (H3K4me3), H3 lysine 4 mono-methylation (H3K4me1), and H3 lysine 27 acetylation (H3K27ac), for several cell types from human and mouse. We identified histone modifications that were stable among different cell types in human and histone modifications that were evolutionarily conserved between mouse and human in the same cell type. We found that histone modifications that were stable among cell types were also likely to be conserved between species. This trend was consistently observed in promoter, intronic, and intergenic regions for all of the histone marks tested. Importantly, the trend was observed regardless of the expression breadth of the nearby gene, indicating that slow evolution of housekeeping genes was not the major reason for the correlation. These regions showed distinct genetic and epigenetic properties, such as clustered transcription factor binding sites (TFBSs), high GC content, and CTCF binding at flanking sides. Based on our observations, we proposed that TFBS clustering in or near a histone modification plays a significant role in stabilizing and conserving the histone modification because TFBS clustering promotes TFBS conservation, which in turn promotes histone modification conservation. In summary, the results of this study support the view that in mammalian genomes a common mechanism maintains histone modifications against both genetic and environmental (cellular) changes.

Methods: Synthetic Data


Original Track
Synthetic Track: Small Adjustments
Synthetic Track: Large Adjustments

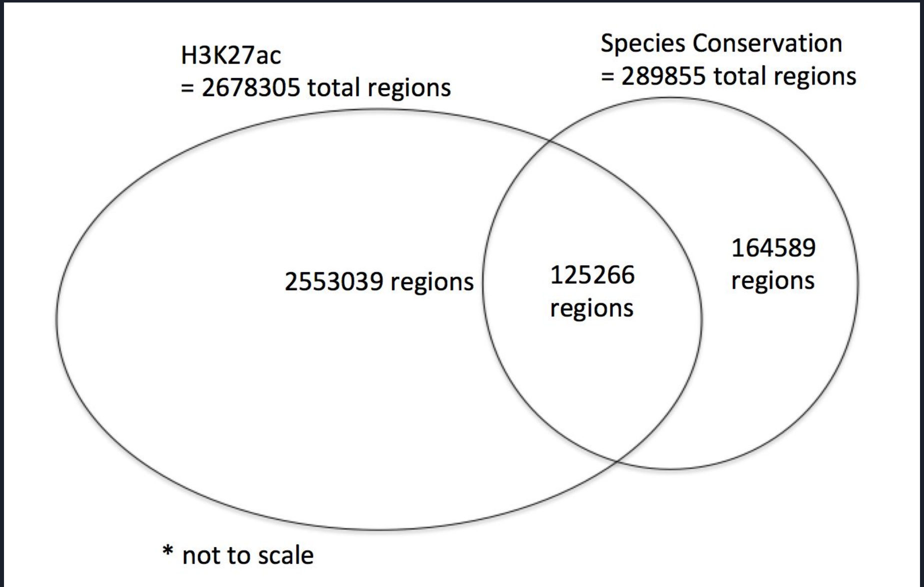
Methods: Data Selection

1. Species conservation track:
   a. Human (hg19)
   b. Mouse
   c. Canine
2. H3K27ac conserved in human cell types: intersection of histone modification tracks on
   a. Muscle cell
   b. Umbilical cell
   c. Stem cell
   d. Lung cell
   e. Skin cell

Methods: Simple Counting (Synthetic Data)

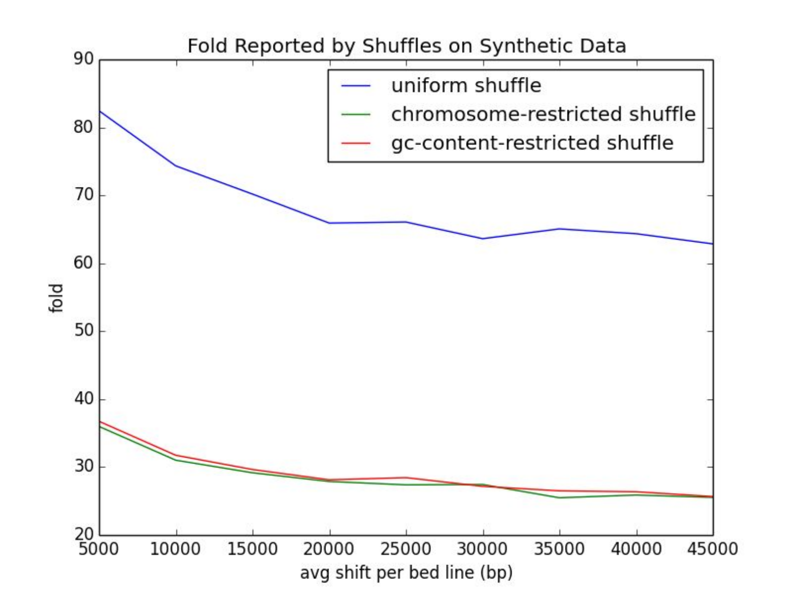
Average Shift Per Bed Line (bp)	Ratio (intersection / sum of non-intersecting regions)
2,500	.61306
5,000	.55895
50,000	.42937
500,000	.33535
5,000,000	.32407
50,000,000	.29719

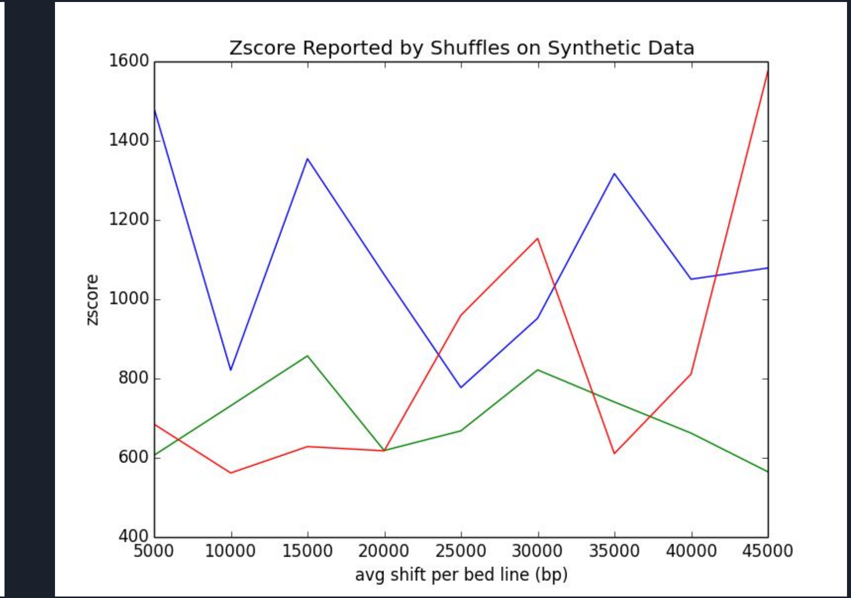
Methods: Simple Counting



- 5% of H3K27ac modifications and 43% of species conservations fall into overlaps of the two
- This implies most histone modifications do not overlap species conservation but a significant number of species-conserved regions include H3K27ac modifications

Methods: Permutation and OverlapSelect on Synthetic Data





Methods: Permutation and OverlapSelect

	Fold	Zscore
Uniform random shuffle	1.710	121.3
Chromosome-restricted shuffle	1.713	124.7
Gc-content-restricted shuffle	1.678	142.3

Why restrict the shuffle to not be uniform?

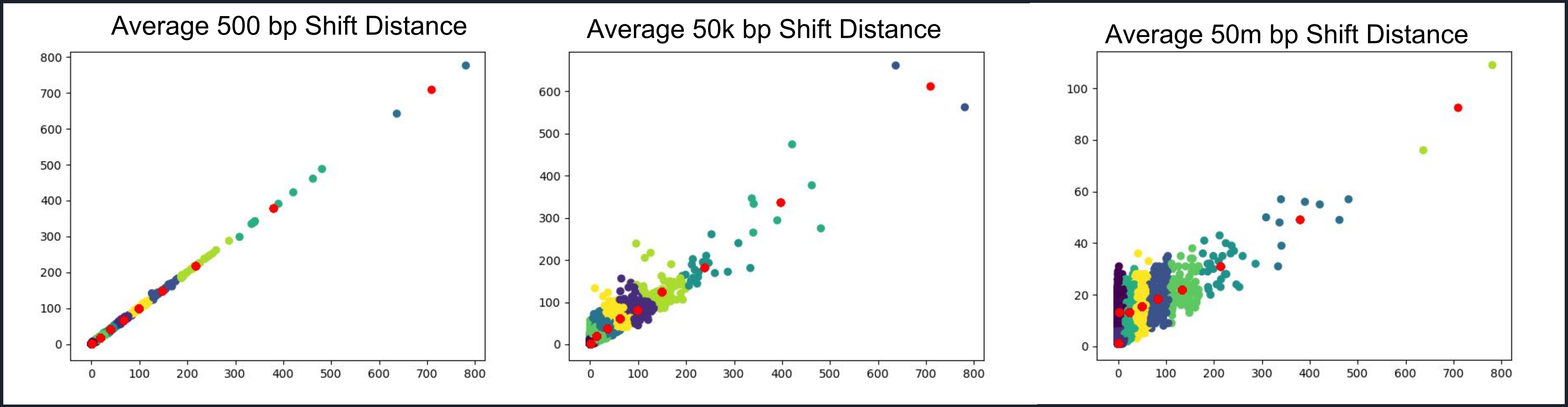
- What if ultraconserved elements and H3k27ac were not evenly distributed across chromosomes, but had no causal relationship?

Correcting for hidden correlations is a huge challenge in statistical analysis of biological relationships

- With data that has not been previously studied, we are not even aware of what correlations we need to correct for

Methods: K-Means Clustering

- Data points = 10k bp intervals of synthetic data chromosomes
- Non-red coloring = cluster assignment
- Red circles = cluster centers

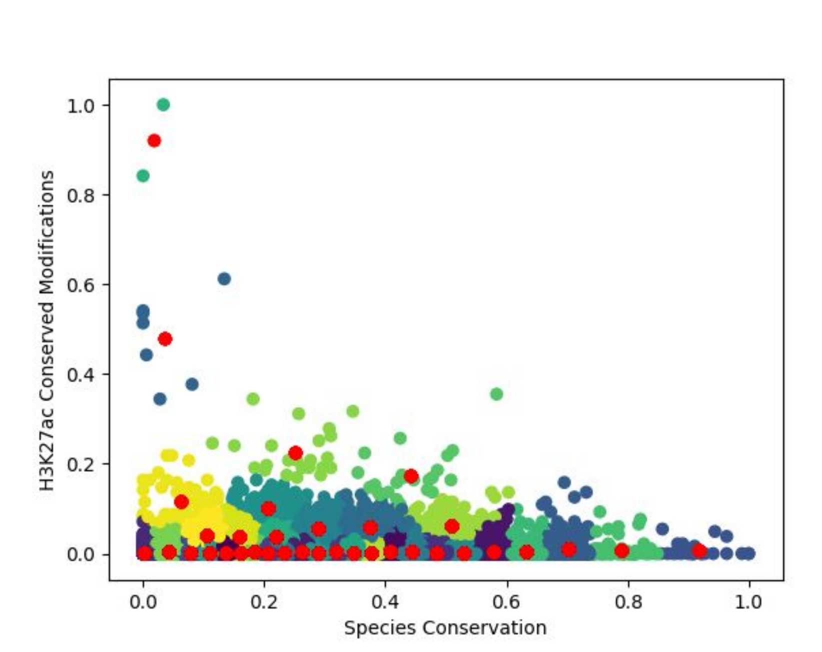

Average 500 bp Shift Distance

Average 50k bp Shift Distance

Average 50m bp Shift Distance

Methods: K-Means Clustering

- Data points = 10k bp intervals of hg19 genome
- Non-red coloring = cluster assignment
- Red circles = cluster centers

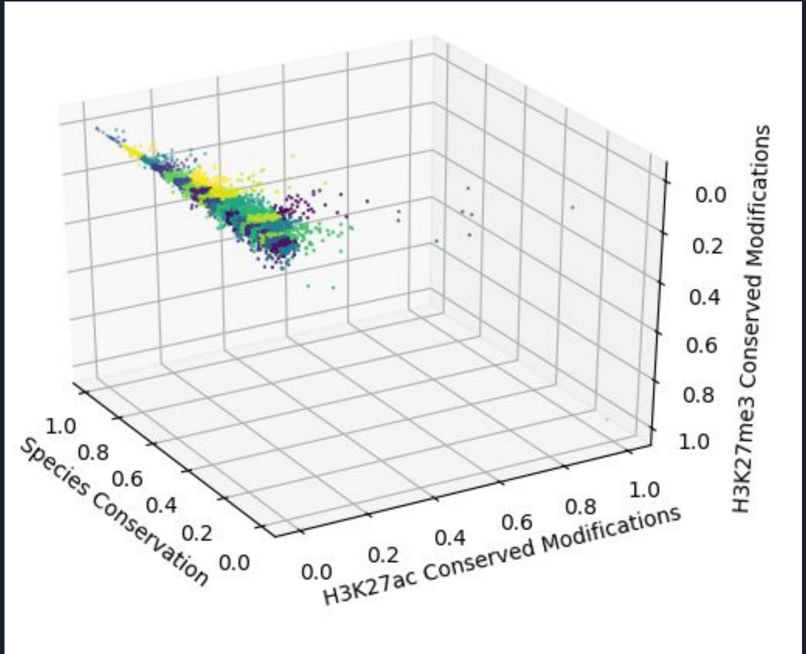


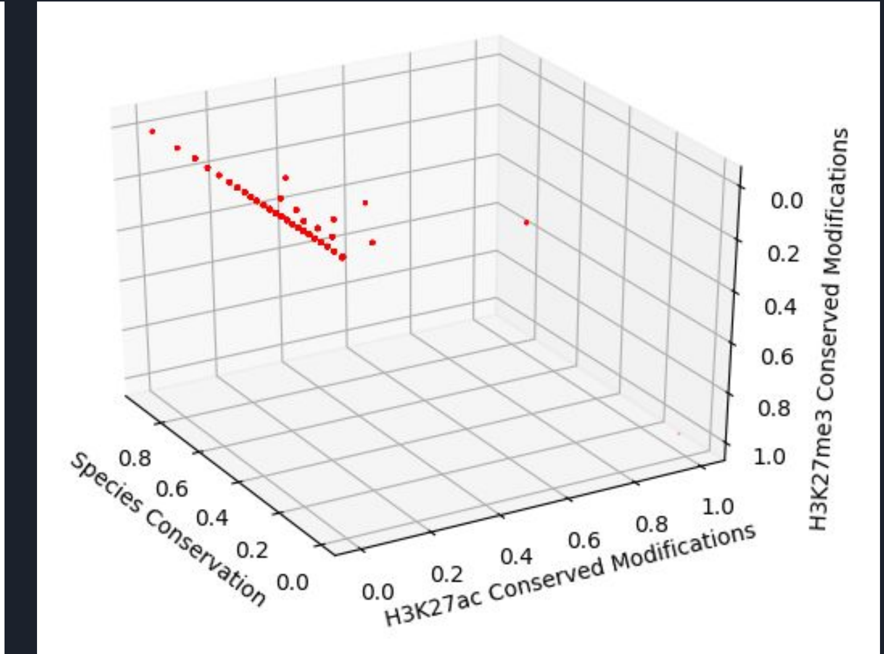
Our Hypothesis: Correlation with H3K27me3

- We hypothesized a correlation between the species conserved regions, H3K27ac, and a third input track, H3K27me3
- We expect other histone modifications in same places to have non-zero correlation (positive or negative)
- H3K27me3 appears in the same places as H3K27ac and antagonizes it
    - Methylation recondenses chromatin, but acetylation decondenses it to make it accessible for transcription
- We tested Hk27me3 across the same cell types compared to H3K27ac and species conservation and hypothesized a negative correlation

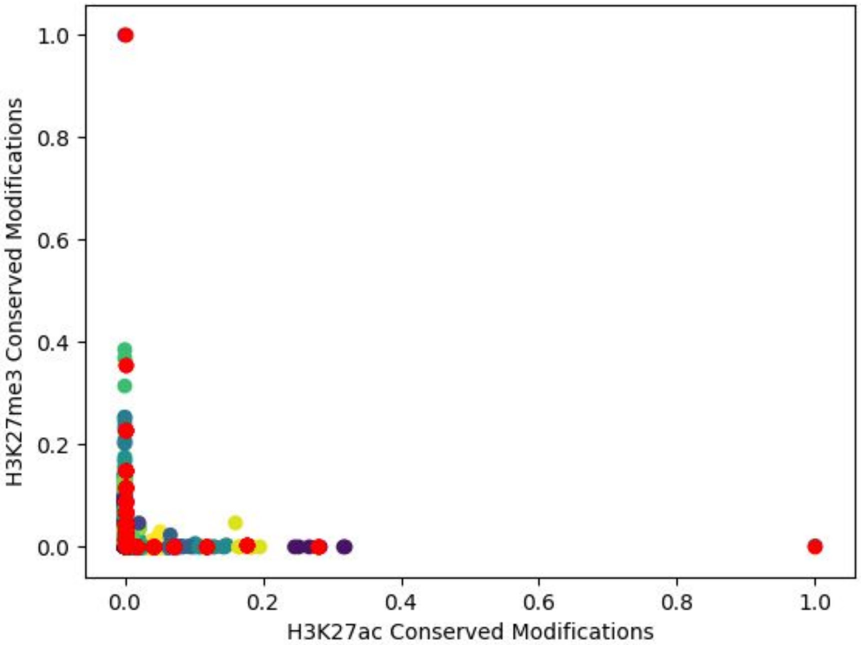
K-Means Clustering on Triplets of Data

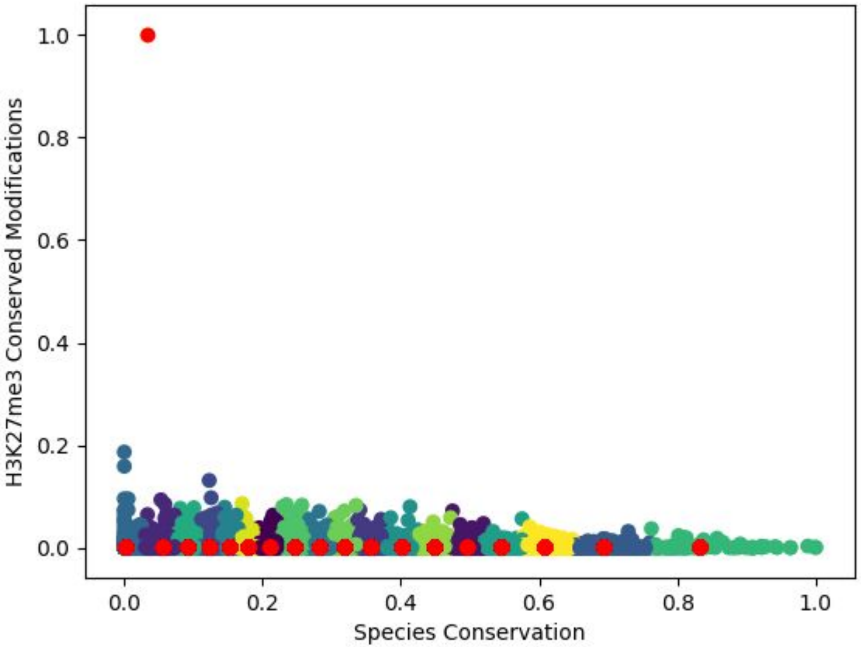
- Clusters exist with both species conservation and H3K27ac conserved modifications
- Significantly fewer genomic regions with H3K27me3 conserved modifications and species conservation
- Negative correlation between H3K27ac and H3K27me3 conserved modifications





K-Means Clustering on Triplets of Data





Results and Further Discussion

- Correlation between species conservation and inter-cell conserved H3K27ac modifications
  - Validation of Woo & Li (2012): conservation of histone modifications is due to its proximity to the transcription factor binding factor
  - Intuitively, the conservation of histone modifications across humans, canines, and mice holds since histone modifications are essential for gene expression and for the survival of certain traits due to natural selection
- Intersection with H3k27me3:
  - Negative correlation
  - Expected because the two are located in close proximity on H3 and antagonize each other in function

Questions?

Supplementary Slides

Synthetic Data Creation

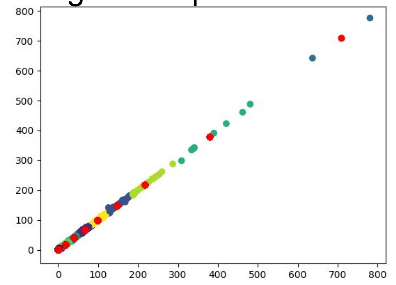
- Goal: create a modification of a track file of genomic intervals (bp start and end pairs) which is correlated to the original but not exactly the same
- Set maximum interval shift value s_max
    - Higher s_max values will create lower correlation between input track and generated track
- For each interval in the input track file:
    - Sample shift value s from uniform(-s_max,s_max)
    - Sample scaling factor r from uniform(-0.1,0.1)
    - Shift original interval by s bp’s and resize it to (1+r)*original size
    - Write shifted and scaled interval to new .bed file
- We chose hg19 coding exons for chr2 as input track but this method can be used starting with arbitrary track of genomic intervals

Methods: K-Means Clustering

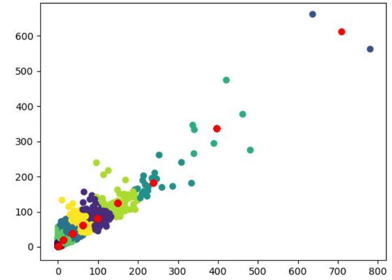
- Data points = 10kb bp intervals of synthetic data chromosomes
- Non-red coloring = cluster assignment
- Red circles = cluster centers
- Number of clusters chosen based approximately on convergence of clustering score

Synthetic Noise (all for size change in (0,0.01))	Clustering score (at k=10)
500 average shift distance	0.0372084011649
50k average shift distance	0.0466791505287
50m average shift distance	0.0530719930251

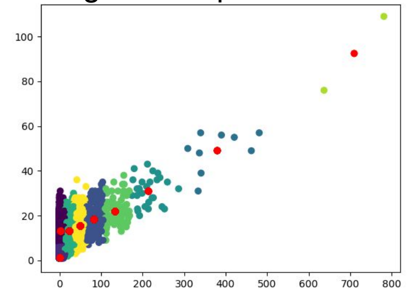
Average 500 bp Shift Distance



Average 50k bp Shift Distance



Average 50m bp Shift Distance


100
80
60
40
20
0
0
100
200
300
400
500
600
700
800