# Large-scale meta-analysis of chromatin immunoprecipitation-sequenced data to exhibit histone modification relationships

Andrew Nickerson, Larry Kalesinskas, Purvesh Khatri, PhD
Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA 94305

## Introduction

1. **Chromatin immunoprecipitation-sequencing (ChIP-Seq):** method that allows for the identification of histone modification regions using antibody markers.

- **Histones:** control the packing of DNA and influence gene expression.

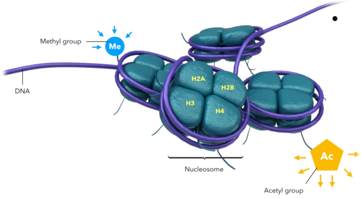


**Figure 1. Histone Modifications.** Histone modifications alter the density/packing of DNA, which can make the DNA either more or less accessible. The four histones are H3, H4, H2A, and H2B.

1. Large public datasets of ChIP-Seq data exist, provided by large consortia (ENCODE, Roadmap).
2. ChIP-Seq analysis has largely focused on single experiments rather than cross-experiment examination.

## Objective

We aim to perform large-scale meta-analysis of ChIP-Seq data to identify and analyze histone modification, cell type, and gene relationships by incorporating publicly available data into a robust database, and applying statistical and network analysis.

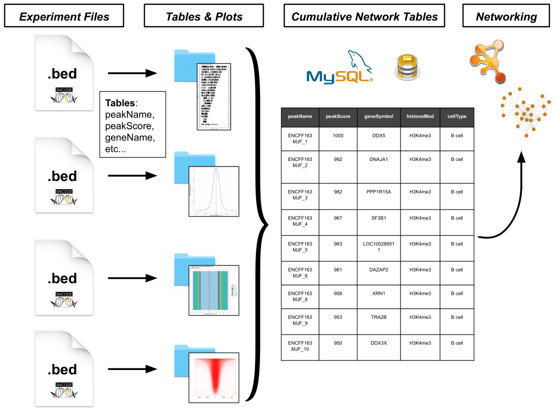
## Methods

- **Systems-level view:** Downloaded and processed 2000 ChIP-Seq experiments (each containing peak data from a single histone modification and cell type pair in humans).
- **Mapping:** Mapped the histone modifications to their nearest genes in hg19.
- **Database:** Consolidated all data into well-designed SQL database called ChIP-Map for quick filtering, querying, and subsetting.
- **Using the database:** Examined multiple B-cell histone modification experiments to find both genes that are highly conserved or unique across histone modifications and to exhibit relationships between histone modifications.

| *Histone Modification* | *Function* |
|---|---|
| H3K4[me1,me2,me3] | transcriptional activation |
| H2AFZ | gene expression influence |
| H3K9ac | transcriptional activation |
| H3K27ac | transcriptional activation |
| H3K27[me1,me2,me3] | transcriptional repression |
| H3K36[me1,me2,me3] | transcriptional activation |
| H4K20me1 | transcriptional silencing |

**Figure 2. Histone Modification Functions.** Shows the function for histone modifications taken from 12 ENCODE experiment files on B-cells.

## Methods (continued)



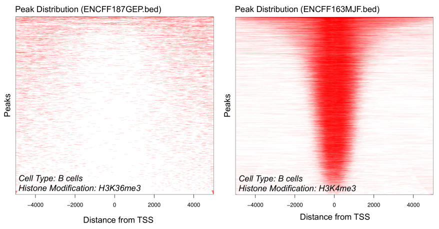
Tables: peakName, peakScore, geneName, etc...

| peakName | peakScore | geneSymbol | histoneMod | cellType |
|---|---|---|---|---|
| ENCFF163 MJF_1 | 1000 | DDX5 | H3K4me3 | B cell |
| ENCFF163 MJF_2 | 992 | DNAJA1 | H3K4me3 | B cell |
| ENCFF163 MJF_3 | 982 | PPP1R15A | H3K4me3 | B cell |
| ENCFF163 MJF_4 | 967 | SF3B1 | H3K4me3 | B cell |
| ENCFF163 MJF_5 | 963 | LOC10028951 1 | H3K4me3 | B cell |
| ENCFF163 MJF_6 | 961 | DAZAP2 | H3K4me3 | B cell |
| ENCFF163 MJF_8 | 958 | XRN1 | H3K4me3 | B cell |
| ENCFF163 MJF_9 | 953 | TRA2B | H3K4me3 | B cell |
| ENCFF163 MJF_10 | 950 | DDX3X | H3K4me3 | B cell |

**Figure 3. Project Pipeline.** Beginning with roughly 2000 ChIP-Seq experiment .bed files from ENCODE, we processed them in R and generated several tables and quality-control figures. From there, we filtered the data and connected it to a database (using MySQL) for querying. The database provided a fundamental base to learn from this data.

## Results

**Figure 4. Quality Control.** Cell Type = B-Cell | Left: Histone Modification = H3K36me3 | Right: Histone Modification = H3K4me3. Quality control figures allowed us to identify early on whether a certain histone modification is likely related to gene transcription.



Jaccard similarity on located gene presence between histone modifications

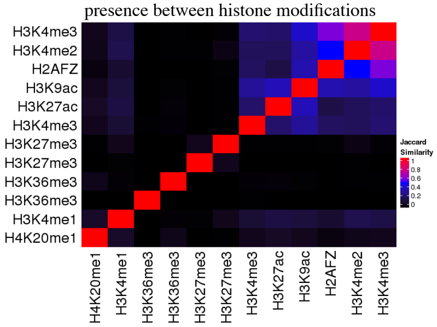


**Figure 5. Jaccard Similarity.** Cell Type = B-Cell; peak score threshold = 250; distance to TSS threshold = ±1000; taken from 12 ENCODE experiment files on B-Cells. Jaccard similarity reveals that certain histone modifications share enrichment for the same set of genes pairwise.

## Results (continued)

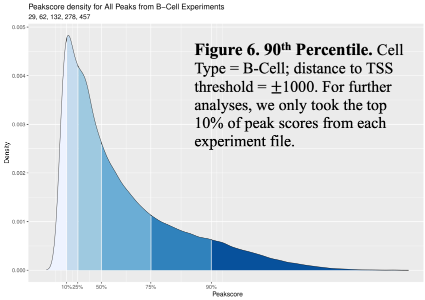


**Figure 6. 90th Percentile.** Cell Type = B-Cell; distance to TSS threshold = ±1000. For further analyses, we only took the top 10% of peak scores from each experiment file.

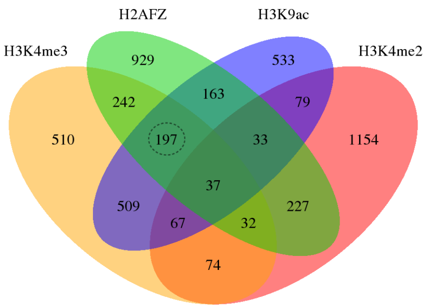


**Figure 7. Overlapping Histone Modifications.** Cell Type = B-Cell; peak score threshold = 90th percentile per experiment; distance to TSS threshold = ±1000. Four histone modifications show significant overlap in the genes they were mapped near.

Gene set enrichment analysis on intersecting genes

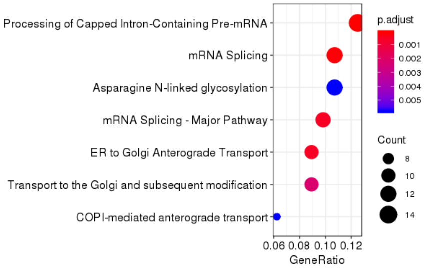


**Figure 8. Gene Set Enrichment.** Cell Type = B-Cell; peak score threshold = 90th percentile per experiment; distance to TSS threshold = ±1000; intersection between H3K4me3, H2AFZ, and H3K9ac (197 genes). Gene set enrichment analysis reveals the hypothetical function of a set of genes. The gene set enrichment analysis for histones H3K4me3, H2AFZ, and H3K9ac shows significant enrichment for mRNA splicing-related genes.

## Results (continued)

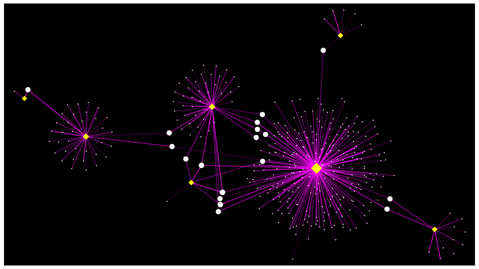


**Figure 9. Network Analysis.** Cell Type = B-Cell; peak score threshold = 800; distance to TSS threshold = ±1000; taken from 12 (7 shown) ENCODE experiment files on B-Cells. Network analysis reveals genes that are conserved across histone modifications.

## Conclusion

1. We processed and mapped the ChIP-Seq data to generate quality-control figures to check validity of our data.
2. We were able to successfully design a robust database (ChIP-Map) to cumulate the ChIP-Seq peak data.
3. By producing a Jaccard similarity heatmap, a Venn diagram, and network from several B-Cell experiments (with various histone modifications), we were able to draw similarities between histone modifications and conserved genes.
4. Gene set enrichment analysis allowed us to see the function of overlapping genes across several histone modifications (for example, mRNA splicing).

## Future Direction

1. Further network analysis to identify important histone/gene/cell-type relationships.
2. Development of a publicly-accessible web app to allow others to easily explore our data.

## Selected References

1. Histone Modification Associated with Initiation of DNA Replication. (2017, April 14). National Cancer Institute - Center for Cancer Research.
2. Northrup, D. L., & Zhao, K. (2011, June 24).
3. Yu G, Wang L, He Q (2015). Bioinformatics, 31(14), 2382-2383. doi: 10.1093/bioinformatics/btv145.
4. Zhao, Y., and Garcia, B.A. (2015). Cold Spring Harbor Perspectives in Biology 7, a025064.

## Acknowledgements

This work was supported by the Stanford Institutes of Medicine Summer Research Program (SIMR), its division for Biomedical Informatics Research, and members of the Khatri Lab.