|  |
| --- |
| Computational Genomics  HMMicro: predicting miRNA targets from epigenetic data  Wu, Felix\*,1 and Nick Giangreco\*,1  1Department of Systems Biology, Columbia University, New York, NY, USA  \*To whom correspondence should be addressed.  Abstract  **Motivation:** Multiple gene regulatory layers give rise to complex phenotypes, which include redundant patterns to maintain homeostasis and other functions. Micro RNAs (miRNA) play an important regulatory role at the post-transcriptional level by binding longer messenger RNA (mRNA) transcripts and targeting them for degradation. These binding sites are difficult to obtain on a global scale for any given condition or treatment. Computational methods provide an effective alternative for identifying miRNA binding sites across the genome.  **Results:** We developed a Hidden Markov Model (HMM) for predicting miRNA binding sites in HEK293 cells using epigenetic information provided by ENCODE. The HMM predicted miRNA binding sites with high sensitivity and showed a low error rate across the cross-validation scheme. Our results indicate that leveraging known epigenetic information of a cell can potentially aid in predicting novel connections within post-transcriptional regulatory layers.  **Availability:** https://github.com/ngiangre/HMMicro  **Contact:** flw2113@cumc.columbia.edu, npg2108@cumc.columbia.edu |

# Introduction

Knowledge of the processes that give rise to cellular phenotype is a common and critical step in understanding specific biological phenomena such as developmental states, homeostasis, and disease. The phenotype of a cell is a product of the reactions and relationships between many molecular regulatory layers, ranging from transcription factor binding to chromatin confirmation. Each layer provides a regulatory logic, allowing the cell to integrate intracellular cues, environmental signals, and DNA-encoded information to achieve a specific state or activity.

Micro RNAs (miRNAs) – short approximately 22 nt-single-stranded RNA molecules – comprise one such layer and are involved specifically in post-transcriptional regulation. Canonically, miRNAs target messenger RNA (mRNA) transcripts for degradation by primarily binding to their 3'-untranslated regions (3'-UTRs), though evidence exists of binding across the transcript body. These binding sites can be difficult to experimentally identify, although methods do exist (Darnell 2010).

With the advent of next generation sequencing and the popularity of high throughput experimentation, genome-wide assays of many regulatory layers such as histone modifications and transcriptome binding motifs have been made in a wide variety of cellular types and contexts (ENCODE, 2012). We hypothesized that a latent redundancy between the regulatory layers of the eukaryotic cell would allow us to predict miRNA binding sites using this corpus of existing regulatory data.

We present a hidden Markov Model (HMM) for prediction of miRNA binding sites. Focusing on the HEK293 cell line, we trained our model using various epigenetic experiments assayed through the ENCODE consortium, and tested our model using experimentally validated miRNA binding sites for this model system. We show our HMM presents a low prediction error rate and high prediction of known miRNA binding sites. This method can be expanded upon by leveraging the full complement of regulatory annotations associated with a particular cellular phenotype.

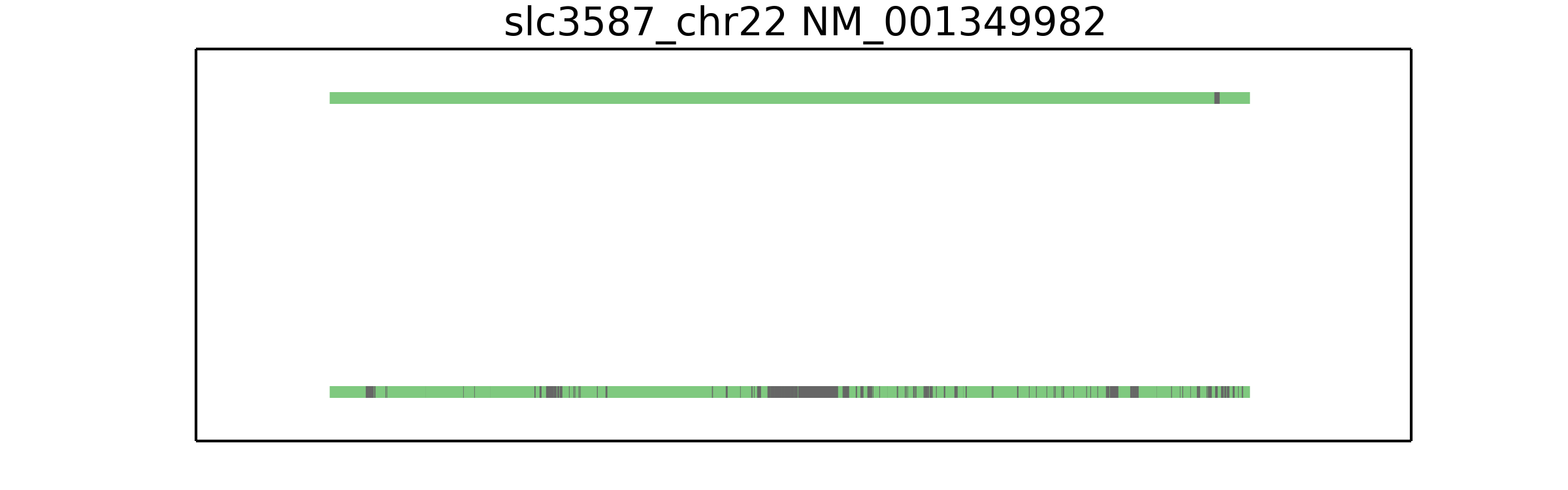
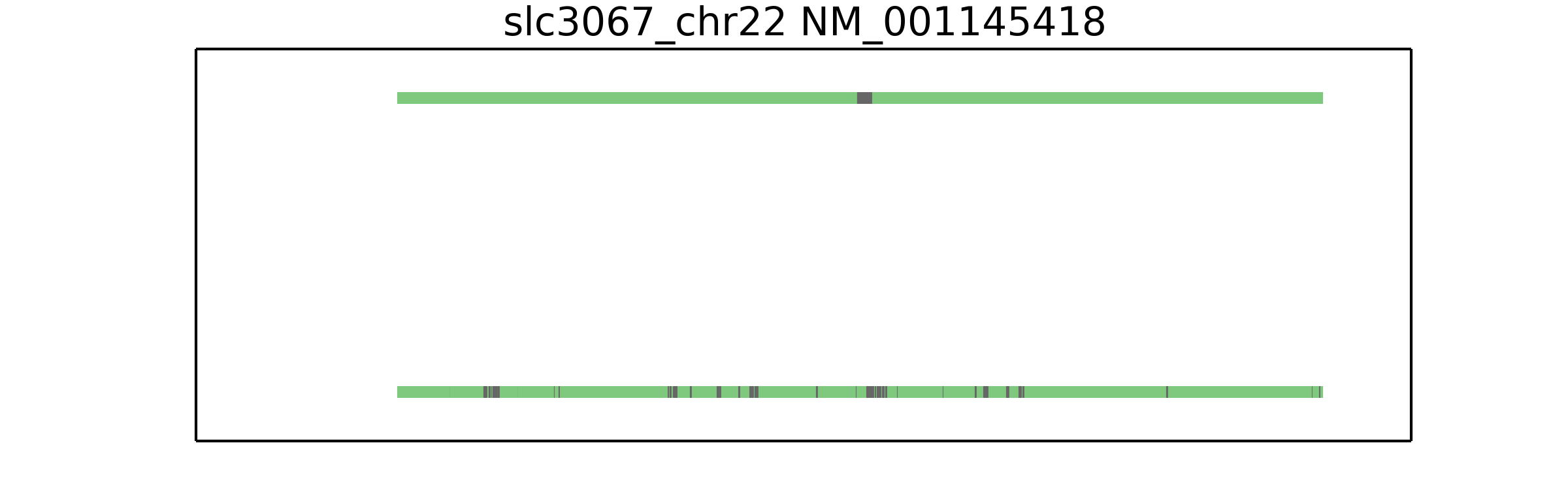
# Methods

We obtained HEK293 epigenetic signals at nucleotide resolution on chromosome 22 from ENCODE (ENCODE, 2012). We retrieved and processed the raw data via remote server using the R package *DeepBlueR* (Albrecht, 2016). We obtained normalized read signal (wiggle file signal) for 8 experiments (7 ChIP-Seq assays against KAP1, POL2RA, TCFL2, ZNF263, CTFC, ELK4, and H3K4me3) at nucleotide resolution across chromosome 22. We also developed in-house R scripts for querying and performing additional processing.

We utilized a set of experimentally identified miRNA binding sites from argonaute (AGO) protein PAR-CLIP data by Hafner et al. (Hafner, 2010). Since miRNAs direct AGO binding, assaying its binding distribution across transcripts using PAR-CLIP is a standard way of assaying for novel miRNA binding sites. Since ENCODE epigenetic signal data uses hg19 reference genome coordinates while the PAR-CLIP data was aligned to the hg18 reference genome, we lifted over the set of miRNA binding sites to the hg19 reference using the UCSC online liftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). All hg19 and refseq annotations were downloaded from the UCSC Genome Browser and the UCSC Table Browser (Karolchik, 2004; Kent, 2002).

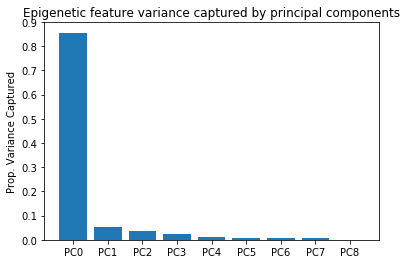
The HMM was developed using the *Pomegranate* python package [<http://github.com/jmschrei/pomegranate>]. We developed the HMM in three steps: pre-processing, training and testing. In the processing step, we computed the principal components of the epigenetic matrix. Also, using experimentally validated miRNA binding sites (Hafner, 2010), we modeled the distribution of start sites across chromosome 22 to identify an appropriate prior distribution of the binding state. In the training step, we used the Baum-Welch algorithm to learn emission parameters (epigenetic characteristics) and refine the prior transition (binding and non-binding miRNA states) probabilities from the training dimension-reduced epigenetic matrix. In the testing step, we applied a 5-fold cross validation to test the HMM. The code for these analyses are available at GitHub (https://github.com/ngiangre/HMMicro).

**Fig. 2. True vs. predicted miRNA binding sites along transcript body.** miRNA binding (purple) and non-binding sites (green). miRNA binding site names from Hafner et al. are given as well as transcript RefSeq id. True binding site location along the body (x axis) is given in the upper display of each plot and predicted on the bottom



# Results

## Multiple epigenetic layers contain redundant information



**Fig. 1. Principal components analysis shows epigenetic variation captured by few di-mensions..** Reducing dimensions of the HEK293 epigenetic features showed majority of variance captured in the first few components.

We investigated the amount of variation present in normalized epigenetic signal from 7 different ChIP-Seq experiments across chromosome 22 in HEK293 cells. This gave an indication of the amount of redundancy across the epigenetic layers. In agreement with our hypothesis of redundant patterns, we observed the first principal component captured about 85% of the variation across the epigenetic layers (Fig. 1).

### 3.2 *De novo* miRNA binding site discovery using a Hidden Markov Model

We developed a HMM that would take into account multiple epigenetic features of our HEK293 cell model system to guide predicting of miRNA binding patterns. Briefly, we defined transition probabilities between binding and non-binding states along chromosome 22, as well as train emission probabilities based on the Baum-Welch algorithm. All parameters and models used are defined and outlined in the code on the github repository.

To test our HMM, we conducted 5-fold cross-validation. We found a low median error (calculated as the number of incorrectly predicted states normalized by the total number of states) in our validation strategy (Table 1).

Clearly, improvements can be made to the HMM model as it currently stands. Future efforts would include feature data from RNA-binding protein experiments as well as expand the analysis to other chromosomes or ideally genome-wide. Moreover, other supervised learning models might be more suited for this problem and would be worth exploring.

**Table 1.**5-fold cross validation results

|  |  |
| --- | --- |
| *k* | Median error |
| 1 | 11.649 |
| 2 | 0.642 |
| 3 | 0.646 |
| 4 | 10.784 |
| 5 | 0.643 |

Acknowledgements

We would like to thank Itsik Pe’er and the Spring 2017 Computational Genomics class at Columbia University for feedback and support.

Funding

This work has been supported by the training grant from the Integrated Program in Biomedical, Cellular, and Molecular Sciences at Columbia University Medical Center.

*Conflict of Interest*: none declared.

References

Albrecht,F. et al. (2016) DeepBlue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets. *Nucleic Acids Research*, 44(Web Server issue): W581-W586.

Darnell R. (2010) HITS-CLIP: panoramic views of protei-RNA regulation in living cells. *Wiley Interdisciplinary Reviews: RNA*. 1(2): 266-286.

ENCODE consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414): 0028-08.

The Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature*, 409(6822): 860-921.

Hafner,M. et al. (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1): 129-41.

Karolchik,D. et al. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, 32(Database issue): D493-6.

Kent,W.J. et al. (2002) The human genome browser at UCSC. *Genome Research*, 12(6): 996-1006.

Schreiber J. (2016) pomegranate. <http://github.com/jmschrei/pomegranate>.

Yates,L.A. et al. (2013). The long and short of microRNA. *Cell*, 153(3): 516–519.