Predicting global gene expression from chromatin accessibility in the developing mammalian forebrain

Robin Yeo & Andrew McKay

ABSTRACT New technologies for probing the epigenome provide unparalleled opportunities to ask how epigenetics regulates gene expression. Here we ask whether these new technologies, paired with machine learning approaches, can predict gene expression in the mammalian forebrain. We use publicly available RNA-seq, ChIP-seq, and the newly developed ATAC-seq datasets from mouse embryonic forebrains to curate a collection of features and corresponding gene expression levels. We then apply a series of machine learning techniques to predict gene expression using only epigenetic marks, starting with logistic regression and Support Vector Machines (SVM), performing feature extraction using Random Forest, and finally applying neural networks to the dataset. We find that by bifurcating the data into 'high' and 'low' expression levels, we can classify gene expression levels with 81.8% accuracy. Additionally, using Random Forest classification, we find that the newly developed approach of ATAC-seq is the single most important epigenetic feature in predicting gene expression, and neural networks perform robustly but with surprisingly low accuracy compared to SVMs despite changes to network topology and implementation of L2 regularization.

INDEX TERMS ATAC-seq, ChIP-seq, gene expression, epigenetic regulation, chromatin accessibility, supervised learning

I. INTRODUCTION

The primary means by which a single genome encodes the information for all of the body's cell types with diverse gene expression signatures is by epigenetic regulation. Epigenetic regulation is mediated by various modifications to either DNA (e.g. DNA methylation) or the histone proteins around which the genome is organized (Figure 1). These modifications either reduce expression of the nearby genes by blocking the transcriptional machinery's access to the underlying DNA, or increase gene expression by making a gene locus more accessible. Improving our understanding of epigenetic regulation and its impact upon gene expression is critical to many fields within biology, including cancer, development, and aging.

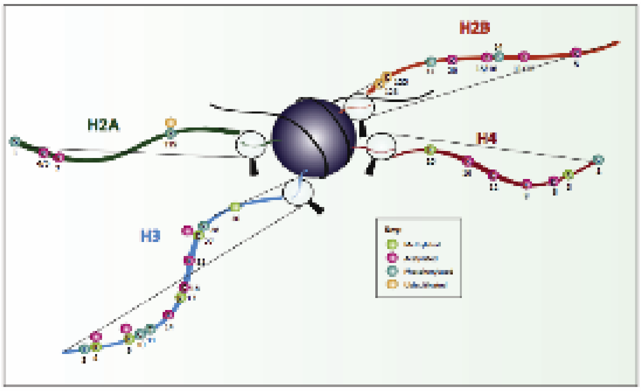

Figure 1: Epigenetic modifications at histone tails

Epigenetic regulation is tightly linked to gene expression, but the causal mechanisms underlying this relationship remain poorly understood. Most histone modifications are comprised of either methylation or acetylation of histone H3, one of the five histone proteins around which DNA is wrapped. Traditionally, epigenetic studies have used a technique called Chromatin Immunoprecipitation Sequencing (ChIP-seq) to sequence DNA regions associated with a given histone modification. While powerful, this provides a relatively myopic view of the epigenome, focusing on one epigenetic marker at a time. Recently, a high-throughput epigenomic assay called Assay for Transposase Accessible Chromatin with Sequencing (ATAC-seq, Figure 2) was developed that allows one to unbiasedly interrogate the "open" or "closed" status of cellular chromatin directly. Together, ATAC-seq and ChIP-seq compliment each other and provide a more comprehensive, richer view of the epigenome and its effects on gene regulation.

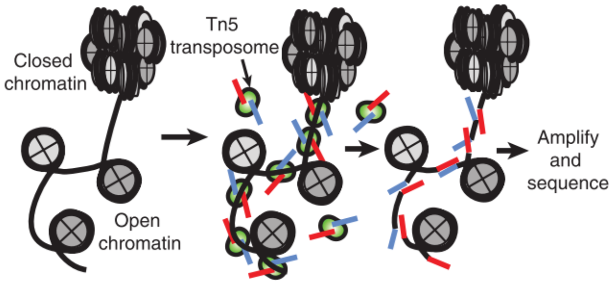

Figure 2: The Assay for Transposase Accessible Chromatin with Sequencing (ATAC-seq) probes open chromatin regions with the Tn5 transposase, resulting in adapter ligation, amplification, and sequencing.

Although there is growing interest in leveraging accessibility/epigenomic data to predict cell identity and disease states, there exists an unexplored opportunity to leverage machine learning models with these genome-wide datasets to identify key epigenomic features that regulate gene expression. As such, a significant unanswered question is how precisely epigenetic state can predict gene expression, and given this prediction, what epigenetic features are most critical in determining gene expression. By pairing epigenomic datasets (ATAC-seq and multiple ChIP-seq datasets) with the associated gene expression profile of a specific tissue (RNA-seq datasets), we have been exploring how well various machine learning models perform at predicting expression patterns and what this can reveal about the most salient features of the epigenome in regulation gene expression.

II. Data Formatting & Pre-processing

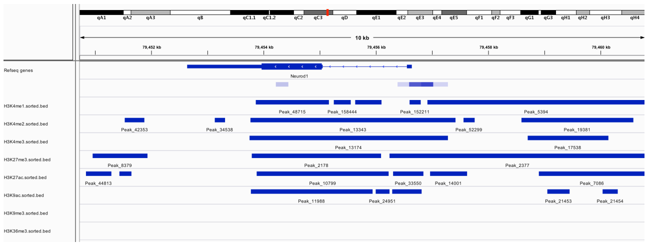

Figure 3: Genome track illustrating a sample distribution of chromatin accessibility peaks and histone modification peaks around the locus of the neurogenic gene Ascl1.

We used publicly available datasets from the ENCODE project (The ENCODE Project Consortium, 2012) which was established to collect and curate "omics" datasets across diverse fields of biology for the use of the scientific community. As graduate students in the genetics/neuroscience laboratory of Dr. Anne Brunet, we decided to perform our analyses using data collected from the developing mouse forebrain (embryonic day 12.5). We collected as many genomics datasets as possible for this stage, resulting in ATAC-seq data for chromatin accessibility (2 biological replicates), 9 ChIP-seq datasets for various histone modifications, and RNA-seq data for gene expression. ATAC-seq and ChIP-seq assays result in millions of DNA sequencing fragments that can be aligned back to the known mouse reference genome to identify which regions are open (in the case of ATAC-seq) or associated with a particular epigenetic histone mark (in the case of ChIP-seq) (Fig 3). By calculating fold-change enrichment of regions in which many reads aligned compared to the background noise level, computational genomic pipelines output "peaks" (specific genomic coordinates) with an associated numerical value indicating signal enrichment (over background).

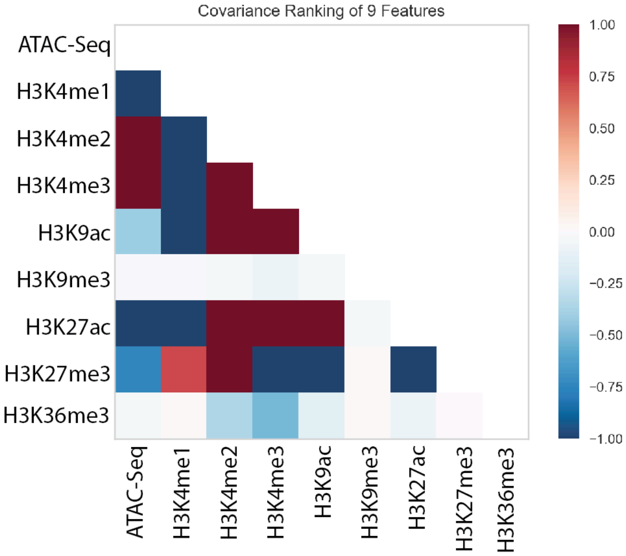

Figure 4: Covariance ranking of features indicated H3K9me3 and H3K36me3 as worth removing for future analysis.

We pooled peaks from the two ATAC-seq replicates and thus our initial feature space was composed of the peak location (Chromosome #, Start Position (in bp), End Position (in bp)) and the enrichment signal for chromatin accessibility. In order to enlarge our epigenomic feature space, we annotated accessibility peaks with the presence/absence of various histone marks (included in the resulting feature matrix as signal enrichment over background). This was calculated by assessing when ChIP-seq peaks overlapped the accessibility peak by >50% using the package *BEDTools intersect* (Quinlan and Hall, 2010). In order to investigate these features, we calculated a covariance matrix of the genome-wide distribution of ATAC-seq and ChIP-seq peak enrichments (Fig 4). This allowed us to visualize which epigenomic features were positively or negatively correlated with one another as well as identify that H3K9me3 and H3K36me3 should be removed from our dataset due to low coverage (very few intersecting peaks within accessible regions).

In order to further enlarge our feature space, we used the program *annotatePeaks.pl* from the Homer suite of epigenomic tools (Heinz et al., 2010) to attribute each peak a genomic identity based on its genomic location (i.e. whether it was located in an intron, exon, promoter, …). Finally, to associate each peak with a gene, we fed in their genomic coordinates and returned the ID of the nearest promoter which we then used to attribute peaks to nearly all of the 20,000 genes in the mouse genome. For our feature response vector of gene expression, we used RNA-seq data from the forebrain to get expression levels for each of these genes in order to train a model to predict expression levels based on epigenetic peak information. Since we were primarily interested in assessing how well various machine learning models could classify whether a gene is “on”/ “off”, we categorized gene expression levels as either 0 or 1 using the median FPKM as the threshold to create a dataset with which to train classifiers. We used various unsupervised learning and clustering techniques, such as PCA and RadViz (Fig 5), to visualize the distribution of genome-wide peaks. However, there was no clear delineation of associated gene expression except that the majority of peaks with high H3K4me3 signal were associated with expressed genes (which is consistent with H3K4me3’s known role from the literature in promoting gene expression).

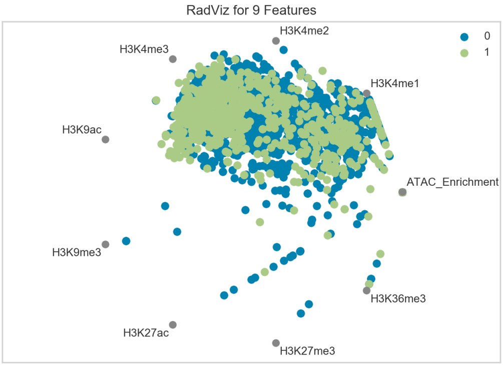

Figure 5: Multivariate data visualization of labelled epigenetic peaks using RadViz

III. Supervised Learning Techniques: Support Vector Machine Classification

We decided to test a number of different supervised learning models in an effort to see which performed best at classifying global gene expression. A starting approach was to classify gene expression using simple logistic regression. For all of the following models, we assed accuracy using 10-fold cross-validation, training the models on 70% of the peaks (39,337 / 56,195 peaks) and calculating accuracy based on the model’s prediction score on the held-out test set containing 30% of the peaks (16,858 peaks). Using a basic logistic regression model, we were able to accurately label 66.48% of the data as either high or low expressing, setting a baseline for our later approaches. Similarly, Nearest Neighbors Classification correctly classified 67.8% of peaks. Based on prediction performances from other genomic machine learning models in the literature, we believed that support vector machine (SVM) classification would be among the best performing models so we decided to apply those to our dataset.

In order to optimize the hyperparameter space, we tested linear, sigmoid and gaussian.kernels (trained with a variety of associated coefficients for each). Ultimately, the Gaussian RBF kernel with the default gamma parameter 1/n performed best with a final optimized accuracy of 74.1% (Table 1), outperforming logistic regression and Nearest Neighbors classification.

SVMs	**Classification Score**
Linear	0.667
Sigmoid	0.540
Gaussian RBF (gamma = 1/n)	0.741

Table 1: Prediction accuracies for various SVM kernels

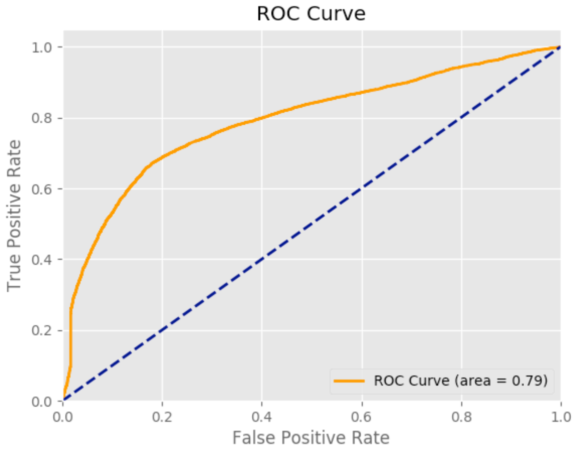

Figure 6: ROC curve for best performing SVM with Gaussian RBF kernel

IV. Random Forests & Feature Importance

Besides SVMs, the best performing published machine learning models for genomic prediction from the literature were often random forests. Aside from robust prediction scores in the genomics field, random forest classifiers benefit from allowing one to easily extract the relative importance of the input features for prediction. This is especially important as one of our primary goals in designing a machine learning model for gene expression prediction was to gain novel insights into how epigenetic features are associated with nearby gene expression. We first decided to optimize the hyperparameter space by varying the number of trees included in the forest and number of features considered when deciding to branch (log2 or sqrt(2)) (Fig 7). Hyperparametric optimization was performed using 10-fold cross-validation holding out 30% of the data as a test set. Surprisingly, optimizing the hyperparameter space only slightly improved the model's performance, resulting in a final test accuracy of 81.8% with an area under the ROC curve of 0.89 for the best performing model (1075 trees with log2 features considered per branching).

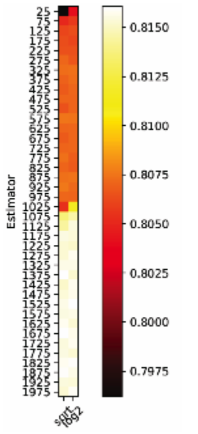

Figure 7: Optimization of Random Forest approach

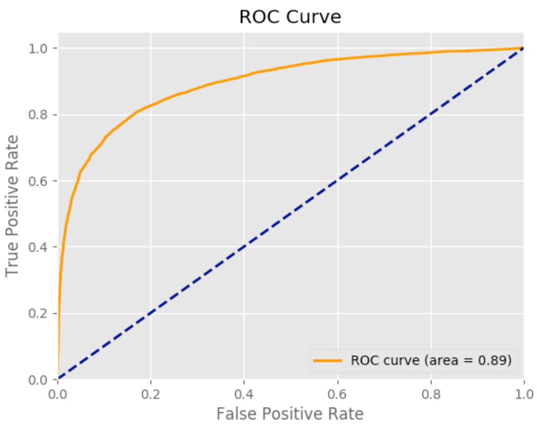

Figure 8: ROC curve for Random Forest approach

One of the principle benefits of random forests as a supervised learning classifier is the unambiguous output of feature importance. Extracting the top features underlying the optimized random forest classifier revealed that chromatin accessibility and H3K4me3 were the single best predictors of gene expression which is encouraging given their known role in regulating active promoters. Interestingly H3K27ac, a known mark of active cis-regulatory enhancers, was also among the top features indicating that accessibility of distal genomic regulatory units is helpful in predicting gene expression. Surprisingly, feature annotation (whether a peak is located in an intron, exon, intergenic region or promoter) was among the least important features for this model. We initially believed that including this information in the model would be helpful for prediction given the known contributions of different genomic regions to nearby gene regulation.

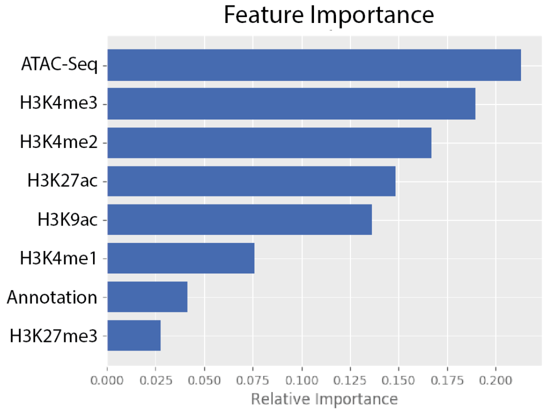

Figure 9: Feature importance identified by Random Forest classification

IV. Neural Networks

A promising area of machine learning is neural networks, and especially deep neural networks. We applied various neural network architectures to our dataset and assessed accuracy. We found that over a spectrum of hidden layer sizes and activation functions, networks with two hidden layers performed better than networks with three hidden layers for our binary classification problem, with three layered networks topping out at an accuracy of 59% and two layered networks correctly classifying at an accuracy of 68%. We found this to be surprising and suspected overfitting. Plotting accuracy over training epochs (Fig 10), we found that accuracy indeed increased well beyond our highest tested accuracy, and implemented L2 regularization. This brought the accuracy of larger networks (>2 hidden layers) in line with our two hidden layer topography, but was not able to increase the accuracy beyond 68% when tested on our 30% test set.

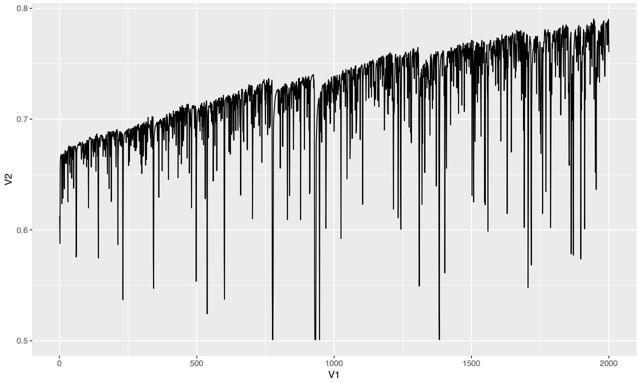

Figure 10: Example training session of neural network plotting accuracy versus training epoches.

Focusing on the two layered networks architecture, we explored the parameter space and found that the classification accuracy was highly robust to changes in the number of nodes and the activation function used, with the most accurate network containing 148 nodes in the first layer with a sigmoid activation function and 184 nodes in the second layer with a tanh activation function. However, we did find correlations between layer nodes, activation functions, and accuracy. The number of nodes in hidden layers 1 and 2 were significantly negatively correlated with accuracy (-0.0010352 and -0.0011397, respectively), while sigmoid and tanh activation functions were consistently positively correlated with higher accuracy for both layers.

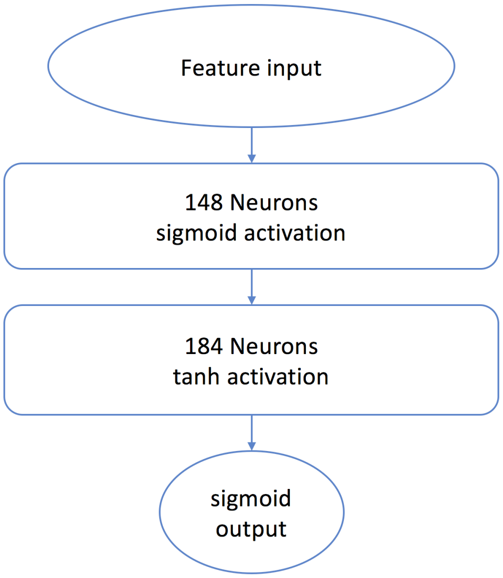

Figure 11: Optimal neural network structure as determined empirically

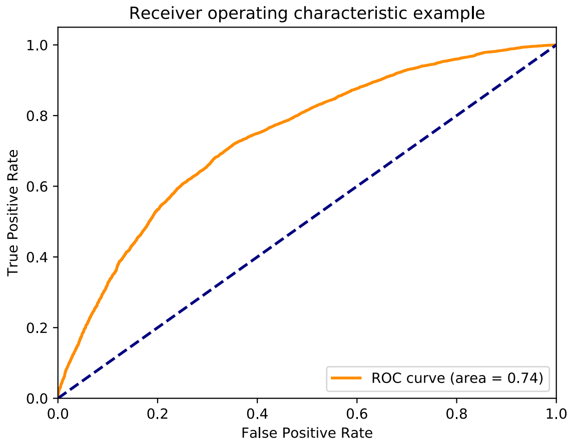

Figure 12: ROC curve for Neural Network implementation

VI. Conclusion & Future Directions

Here we have demonstrated that epigenetic information alone is sufficient, using machine learning approaches, to correctly classify whether a gene is expressed or not with an accuracy as high as 81.8% using a Random Forest approach (Fig 13). Furthermore, in interrogating which features contribute most to prediction, we found that ATAC-seq and H3K4me3 were highly informative, while gene annotations were surprisingly uninformative. Surprisingly, neural network implementations failed to improve our accuracy beyond that of the SVM with Gaussian RBF kernel or the Random Forest (Fig 13). This was despite widespread changes to the network topography and implementation of L2 regularization.

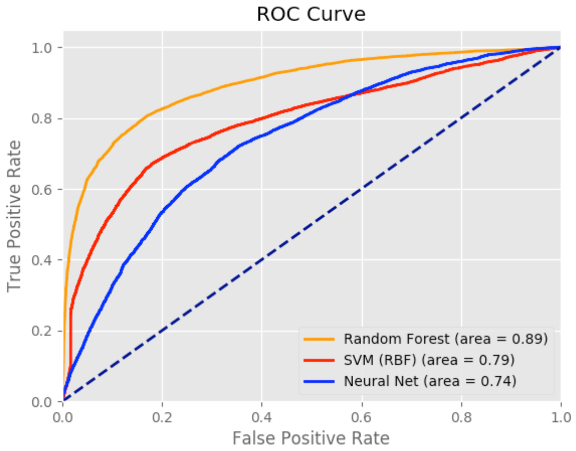

Figure 13: Overlaid ROC curves for the tested approaches

In future work we would like to extend these findings to other datasets, with the hopes that our approach is highly generalizable. We would like to go beyond the binary classification as well, exploring multi-class thresholding of the dataset, as well as regression. We predict that while neural networks may not perform better than Random Forest and SVM on this particular dataset, it may be the dominant approach when considering regression. Finally, we would like to apply our approaches to other publicly available datasets and look for exceptions to the rule that may contain unusual deviations in gene expression versus epigenetic information with biological significance.

References:

Aaron R. Quinlan, Ira M. Hall; BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics*, Volume 26, Issue 6, 15 March 2010, Pages 841–842, https://doi.org/10.1093/bioinformatics/btq033

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213–8. http://doi.org/10.1038/nmeth.2688

Heinz S, Benner C, Spann N, Bertolino E et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell 2010 May 28;38(4):576-589. PMID: 20513432

Pedregosa *et al.*, Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830, 2011.

Rendeiro, A. F., Schmidl, C., Strefford, J. C., Walewska, R., Davis, Z., Farlik, M., Bock, C. (2016). Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. Nature Communications, 7, 11938. JOUR. Retrieved from http://dx.doi.org/10.1038/ncomms11938

The ENCODE Project Consortium. (2012). An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature*, *489*(7414), 57–74. http://doi.org/10.1038/nature11247