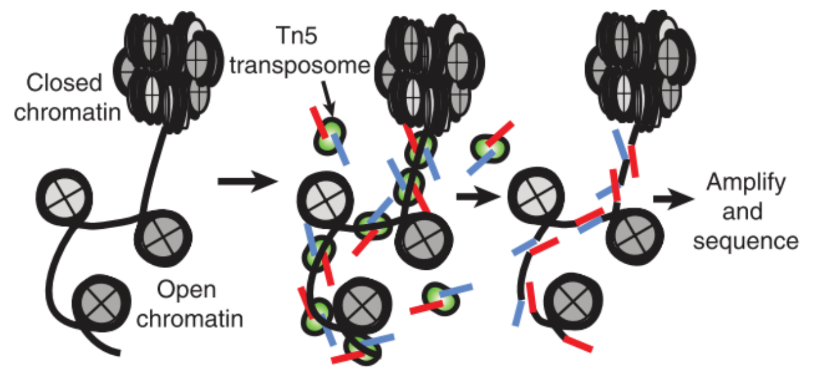
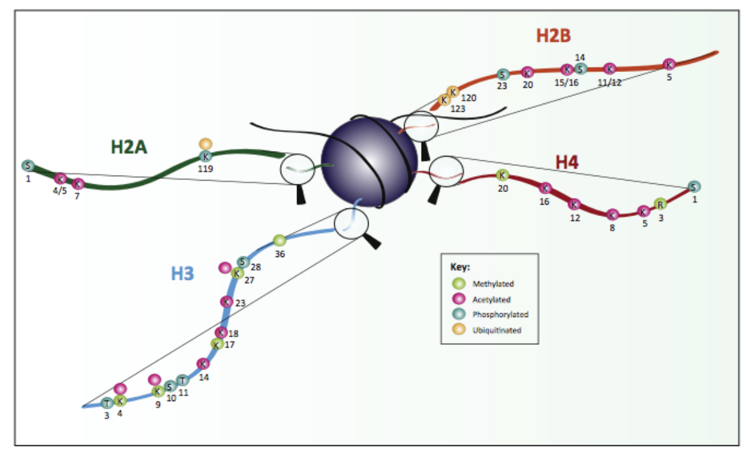
Predicting global gene expression from chromatin accessibility in the developing mammalian forebrain

Robin Yeo and Andrew McKay

Introduction

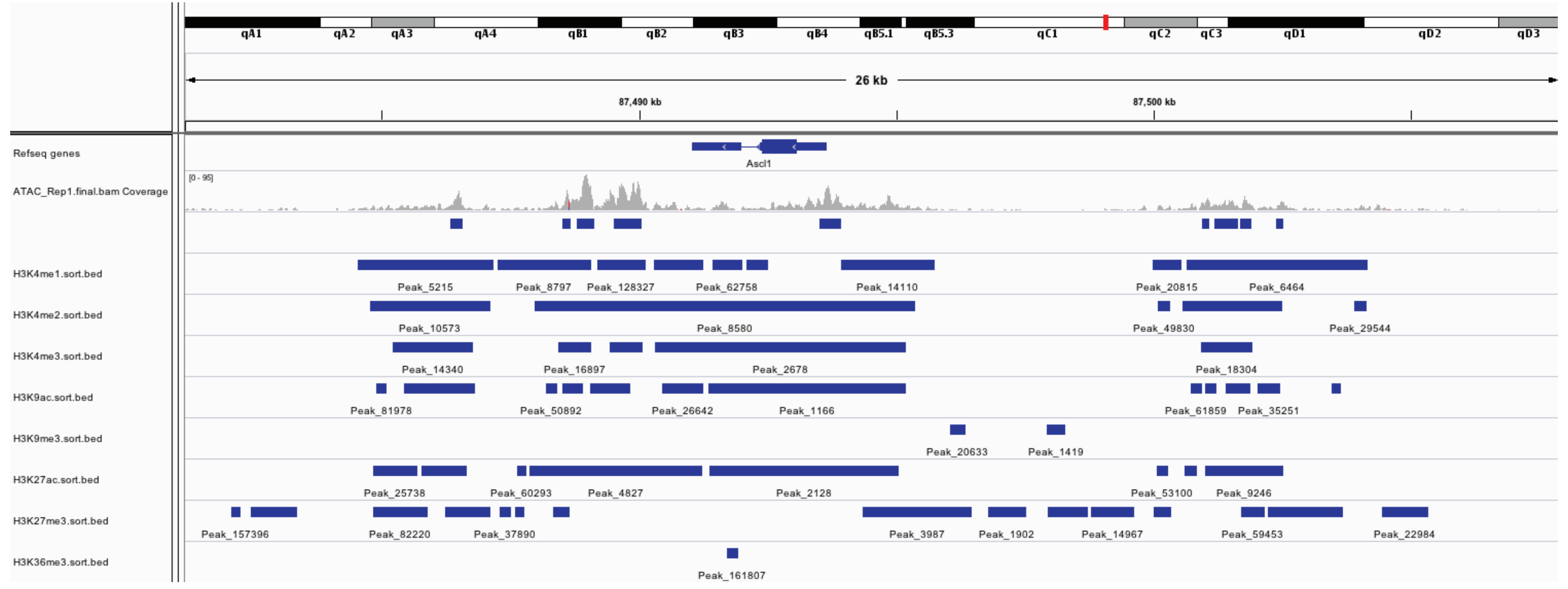
The primary means by which a single genome encodes the information for all of the body’s cell types with diverse gene expression signatures is by epigenetic regulation. Epigenetic regulation is tightly linked to gene expression, but the causal mechanisms underlying this relationship remain poorly understood. 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) was developed that allows one to unbiasedly interrogate the “open” or “closed” status of nuclear 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.



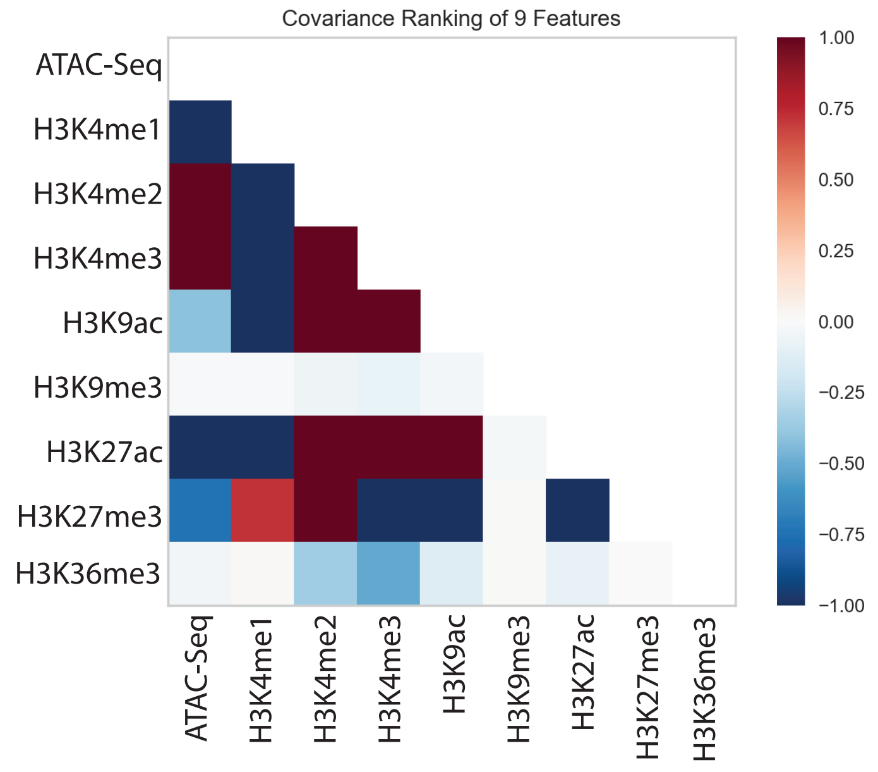


Although there is growing interest in leveraging accessibility/epigenomic data to predict cell identity and disease states, there exists an unexplored opportunity to leverage machine learning models with these genome-wide datasets to identify key epigenomic features that regulate gene expression. With this project, we aim to develop a classification model that predicts gene expression genome-wide using only annotated accessibility peaks and identify which epigenomic features are most predictive of global gene expression using feature inference.

Data processing



We collected publically available epigenomic datasets (ATAC-seq and ChIP-seq) for the developing mouse forebrain from the ENCODE Project Consortium in order to develop a model to predict global gene expression patterns (RNA-seq) from the chromatin landscape. The 2 biological ATAC-seq replicates were pooled and the resulting peaks were annotated based on their relative genomic position (e.g. exonic, intronic, intergenic, …) and their overlap with a set of 9 histone modifications. Every annotated accessibility peak was then associated with a gene (based on nearest transcription start site) whose expression level was marked “on” or “off” (thresholded by the median gene read count) in order to build a classification model. Based on low coverage, we eliminated 2/9 of the ChIP-seq datasets from our final analyses as they did not yield any improvement in prediction accuracy.



Supervised Learning Techniques: SVM Classifiers

We decided to test a number of different supervised learning models in an effort to see which performed best at global gene expression. An starting approach was to classify gene expression using logistic regression. Using a cross-validation test set of 30% we were able to accurately label 66.48% of the data as either high or low expressing, setting a baseline for our later approaches. We tested a number of other supervised learning techniques (such as Nearest Neighbors Classification before moving on to suppoprt vector machines (SVM) classification

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i$$

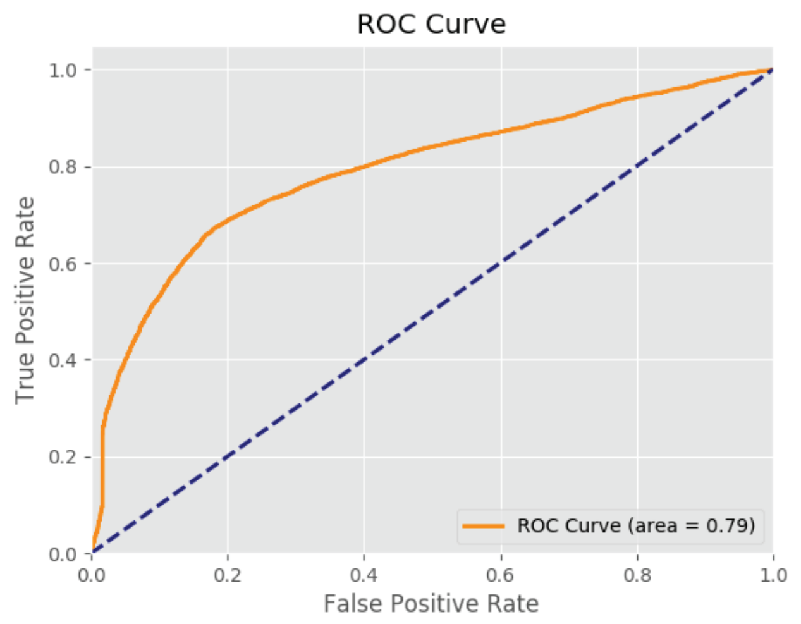
$$\text{subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i,$$

$$\zeta_i \geq 0, i = 1, ..., n$$

- linear: $\langle x, x' \rangle$.
- polynomial: $(\gamma\langle x, x' \rangle + r)^d$. d
- rbf: $\exp(-\gamma\|x - x'\|^2)$. γ is s
- sigmoid $(\tanh(\gamma\langle x, x' \rangle + r))$,

Given the robust nature of SVM classification, we decided to test multiple models, varying the kernel choice using 10-fold cross-validation holding out 30% of the data as a test set. The Gaussian RBF kernel with the default gamma parameter 1/n performed best with a final optimized accuracy of 74.1%.

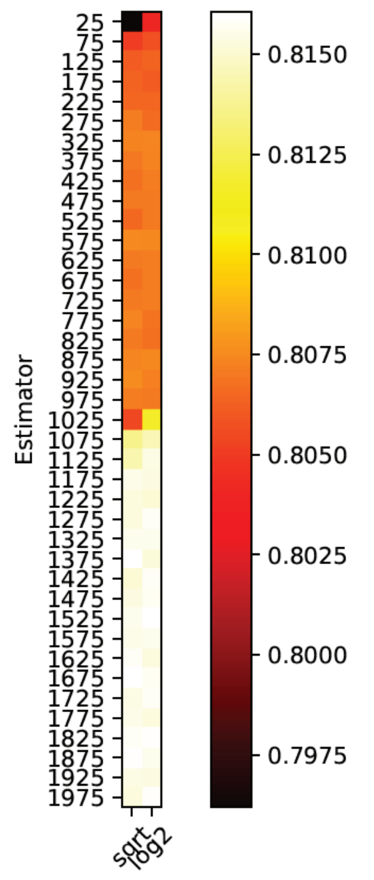
SVMs	Classification Score
Linear	0.667
Sigmoid	0.540
Gaussian RBF (gamma = 1/n)	0.741

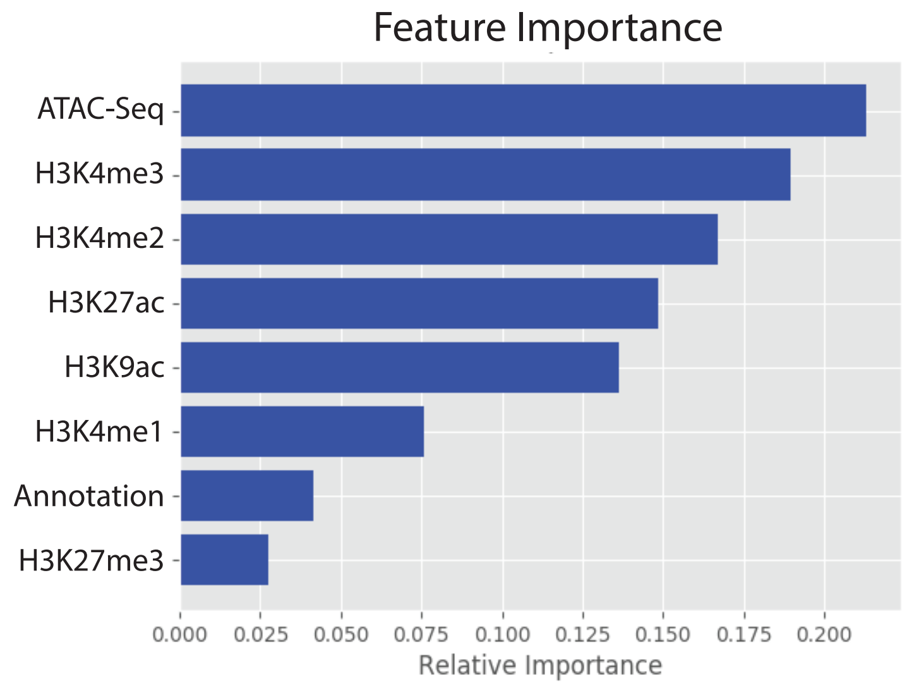


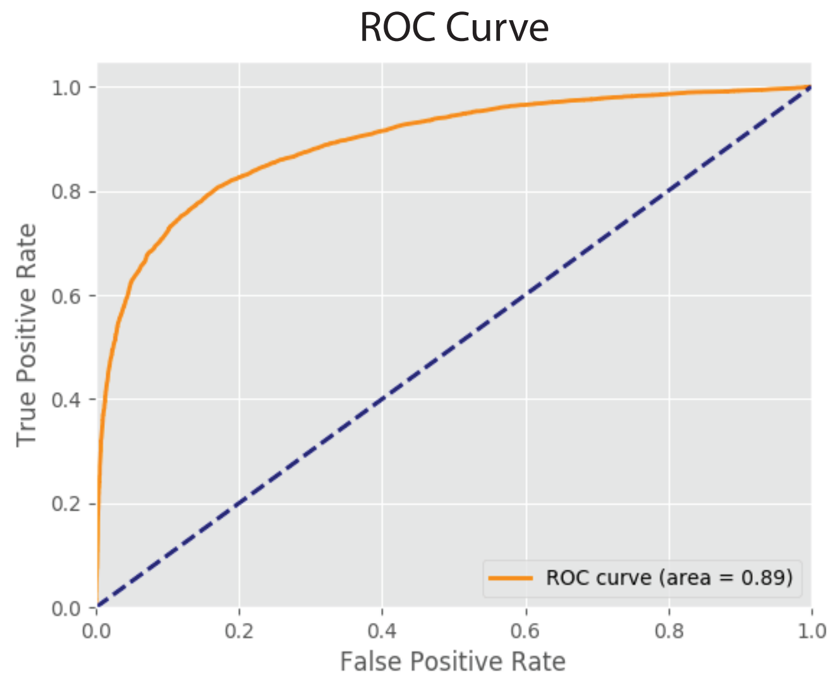
Random Forests & Feature Importance

In order to train random forest classifiers to predict gene expression, we first decided to optimize the hyperparameter space of number of trees included in the forest and number of features considered when deciding to branch (log2 or sqrt(2)). Hyperparametric optimization was performed using 10-fold cross-validation holding out 30% of the data as a test set. The best performing model resulted in an overall test accuracy of 81.8% with an area under the ROC curve of 0.89.

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.

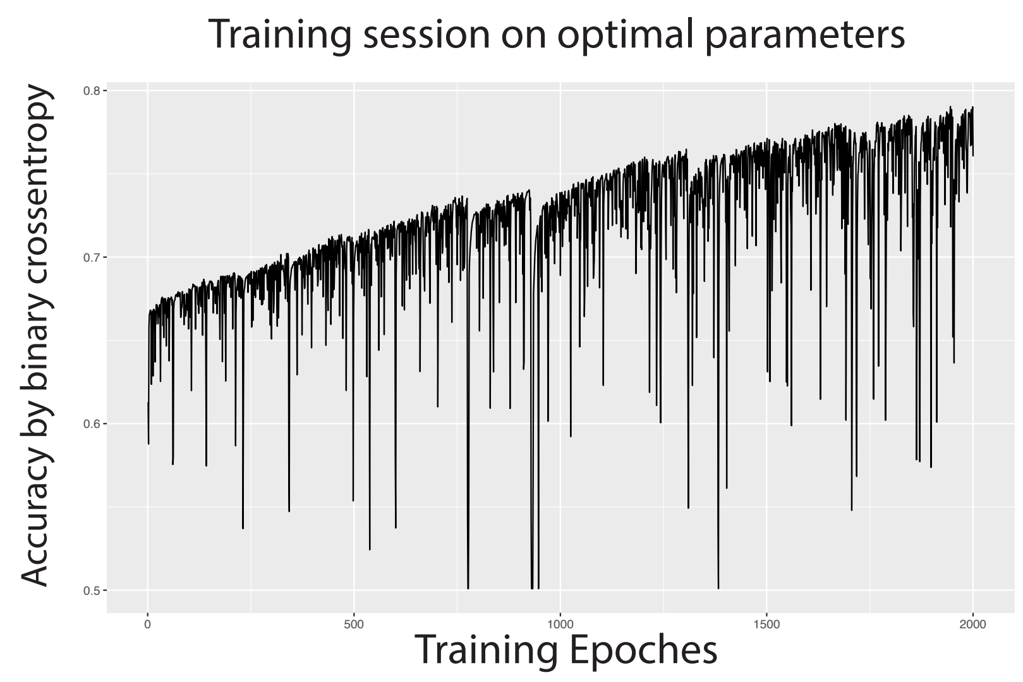


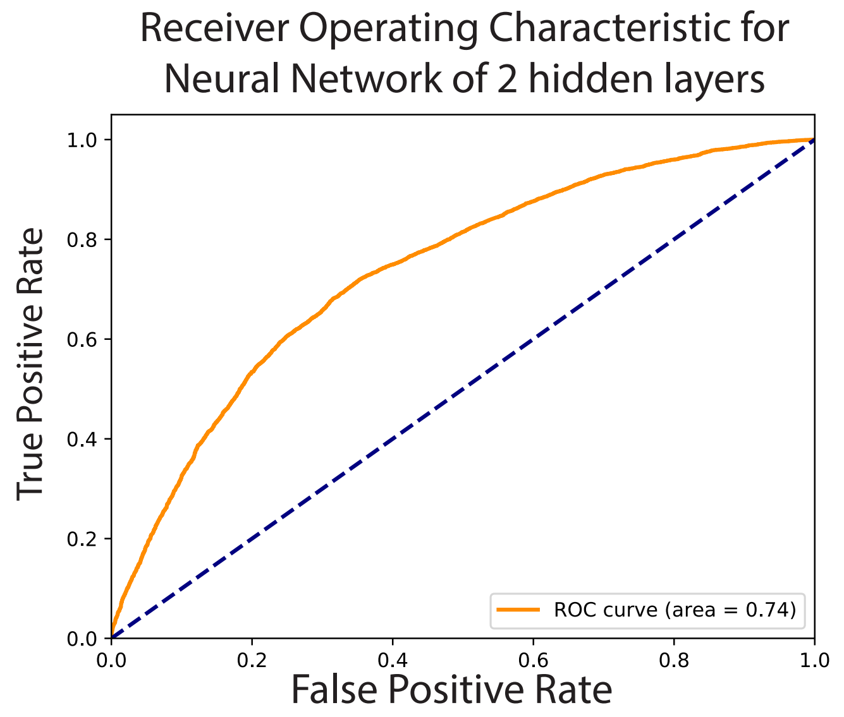




Neural Networks

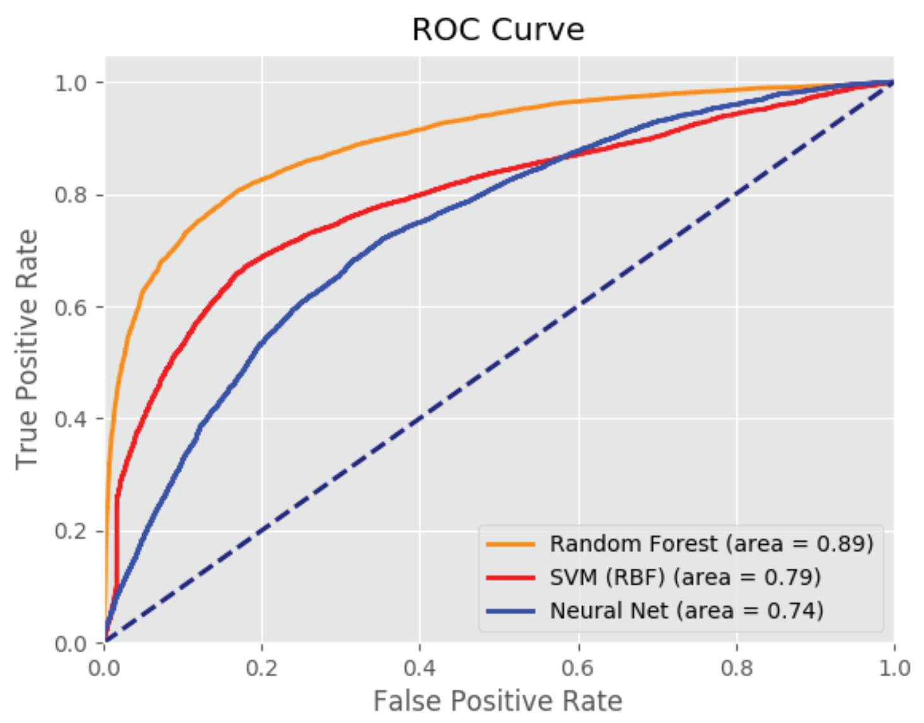
A promising area of machine learning is neural networks, and especially deep neural networks. We applied various neural network architectures to our dataset and assessed accuracy. We found that over a spectrum of hidden layer sizes and activation functions, networks with two hidden layers performed better 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% using a 30% hold out for our test set. 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.





Conclusions

Here we have demonstrated that epigenetic information alone is sufficient, using machine learning approaches, to correctly classify gene expression levels at the binary level with an accuracy as high as 81.8% using a Random Forest approach. Furthermore, in interrogating which features are most important for this accuracy, we found that ATAC-seq is highly informative, while gene annotations were surprisingly uninformative. In future work we would like to extend these findings to other datasets, with the hopes that our approach is highly generalizable, and look for exceptions to the rule in datasets that may contain unusual deviations in gene expression versus epigenetic information with biological significance.



References

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

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

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

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