Team: Augur (Kelley Paskov, Greg McInnes, Christine Tataru, and Nate Stockham)

Review of:

Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, *26*(7), 990–999. http://doi.org/10.1101/gr.200535.115

Overview:
The goal of the Basset paper is to learn cell-type specific DNA accessibility from sequence using a deep convolutional neural net. This is an important problem because variants in noncoding regions are more likely to affect phenotype if they are DNA accessible. By understanding and predicting DNA accessibility, Basset gives researchers a way to better annotate non-coding regions in the genome, which in turn can be used to improve the power of GWAS studies. Basset improves upon previous methods by producing nucleotide-level accessibility measurements across the genome, giving researchers more fine-grained and comprehensive information than publicly available regulatory annotations.

Data:
The paper used DNase-seq data from 164 cell types collected from ENCODE and the Roadmap Epigenetics Consortium. Their dataset consisted of approximately 2 million DNase I hypersensitive sites (DHS) each known to be either accessible or inaccessible. A median of 8.2% of the sites were accessible across cell types. For each site, a 600bp region centered at the site was extracted from the reference genome and used as input to the model.

Model Architecture:
Input data is one-hot encoded before being fed into the neural net. The neural net architecture used in the paper consists of three consecutive convolutional layers, each with a convolution, a ReLU, and a pooling layer. Next, output from the final convolutional layer is fed into two consecutive fully-connected hidden layers. Finally, the sigmoid transformation is performed to produced 164 outputs each corresponding to the predicted probability of accessibility of the site in the 164 cell types. Stochastic gradient descent was used to learn model parameters using the binary cross entropy loss function as the objective. Training was stopped when accuracy no longer increased on a small validation set over 12 consecutive passes.

Performance and Feature Analysis:
Basset outperforms gkm-SVM, a gapped SVM model, when comparing AUC across each cell type. Furthermore, Basset seems to produce biologically relevant features. By nullifying individual filters, the paper was able to recover known protein binding motifs. Also, by adding known protein binding motifs to the center of input sequences, the model predicted the expected increase in accessibility. The paper also performed in silico saturation mutagenesis to understand the effect of any individual mutation on DNA accessibility, across the genome. They were able to further validate their model by noting that mutations that disrupt known motifs also decrease predicted accessibility. Finally, the paper defined a score they call SNP Accessibility Difference (SAD) based on Basset's saturation mutagenesis data and compared this score with an orthogonal statistical method called PICS to try to predict in the context of a GWAS, which SNP from a set of SNPs in LD is most likely to be causal. The predictions by both Basset and PICS seemed to largely agree.

Computation:
Since Basset must be trained on a particular cell type, the paper tried a warm-start method in order to decrease training time. They pretrained their model on 149 of their cell-types and then

attempted to retrain the full 164 cell-type model using these parameters as a warm start. Using this method, they were able to achieve similar accuracies to their full model, but the second round of specialized training required only a single training pass through the new data. This method could be extremely useful to researchers hoping to train Basset on their cell-type of interest.

Extensions:
There are several interesting directions for future work. More work could be done to validate and better understand the features being used by Basset. In particular, it would be interesting to use DeepLIFT to analyze the motifs being learned by each convolutional layer. Experimenting with alternative architectures could also achieve improved performance. In particular, adding bypass layers could give Basset the flexibility to learn motifs of differing complexity. Finally, the saturation mutagenesis experiments could be extended to analyze the effect of pairs of SNPs on the accessibility prediction. Perhaps these higher-order interactions between SNPs could give us additional insights when analyzing GWAS.