

Predicting Regulatory Elements from Genome-Wide 3-way Alignments of Human, Mouse, and Rat

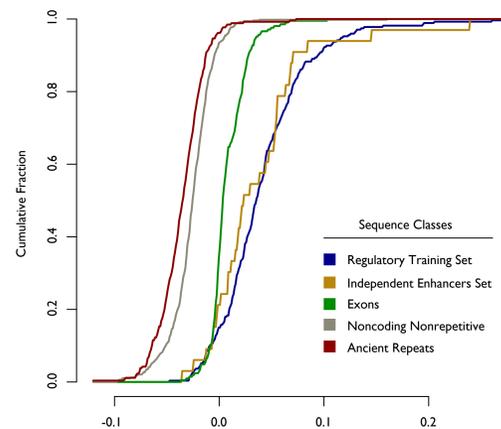
James Taylor, David King, Diana Kolbe, Laura Elnitski, Pallavi Eswara, Jia Li, Webb Miller, Ross Hardison, and Francesca Chiaromonte
Center for Comparative Genomics and Bioinformatics, The Pennsylvania State University

Regulatory Potential (RP) scores are a method to identify cis-regulatory elements using statistical models based on frequencies of short patterns in multiple species alignments.

The RP score for an alignment is calculated using the log ratio of transition probabilities from two fixed order Markov models. The recognition model is trained on a set of experimentally verified human regulatory elements which align in all the species being considered. The background model is trained on a randomly selected set of ancient repeats (approximately the same size as the available regulatory elements).

MODEL SELECTION

In extending the RP score computation from pairwise to multiple species alignments the key challenge was selecting sufficiently simple models for our limited training data. Adding additional species exponentially increases the complexity of input while reducing available training data due to the greater frequency of unsequenced DNA. We addressed this by starting with a state space of all possible alignment columns and collapsing them. First states are clustered based on their frequency in the training data using hierarchical clustering. When the number of states is small enough we switch to a second agglomeration algorithm based on discrimination. For each possible pair of symbols that can be collapsed we train models with that collapse and score the training data with it. To avoid considering order selection at this stage, we train the models with the average transition probabilities over several orders. Whichever collapse maximizes the separation of the regulatory and neutral training data (merit) is selected as the best alphabet at that stage, and the agglomeration continues. Based on the decrease in merit across agglomeration stages we select a set of candidate alphabets. Finally we use leave-one-out cross-validation to select the best of these and the order of the markov models.



DISCRIMINATION

The cumulative distributions for 3-way RP scores on samples of different genomic regions demonstrate the score's discrimination capability. The regulatory training set scores highly, but a disjoint set of independently verified enhancers also scores very well, showing that the score recognizes regulatory elements well. Exons also score highly but are still distinguishable from regulatory elements. Samples of noncoding nonrepetitive DNA and ancestral repeats score the lowest.

PERFORMANCE IMPROVEMENT WITH RAT

We have found that despite the short evolutionary distance between rat and mouse, the RP score from human/mouse/rat alignments performs better than the RP score from human/mouse alignments. Not only is the range of the score higher (as seen in the plots to the right) but leave-one-out cross validation using 2-way and 3-way alignments shows that the 3-way score is able to correctly classify a greater fraction of the training set elements.

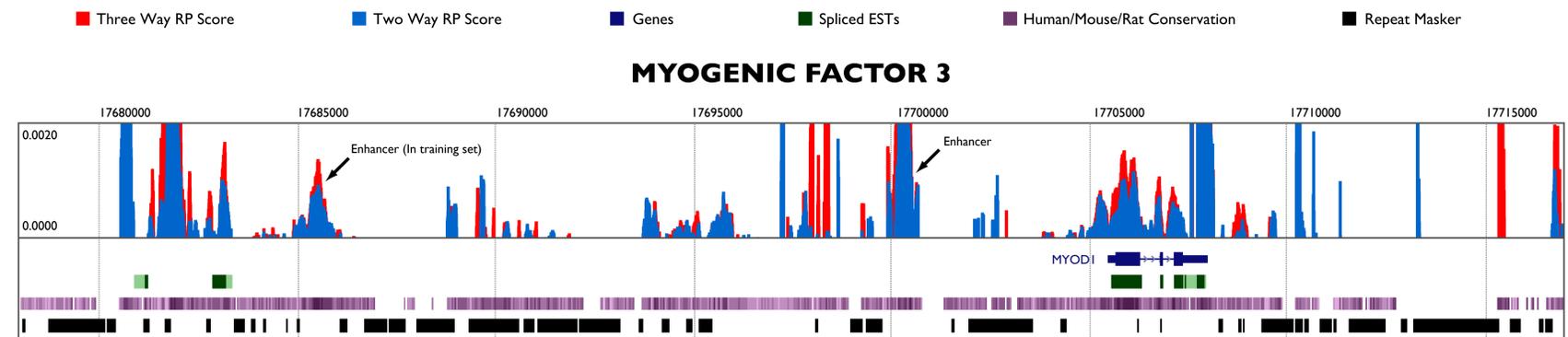
CROSS VALIDATION COMPARISON

| | 2-way score | 3-way score |
|----------------------|-------------|-------------|
| Regulatory Correct | 139 | 184 |
| Regulatory Incorrect | 134 | 89 |
| Neutral Correct | 222 | 201 |
| Neutral Incorrect | 38 | 59 |
| Success Rate | 68% | 72% |

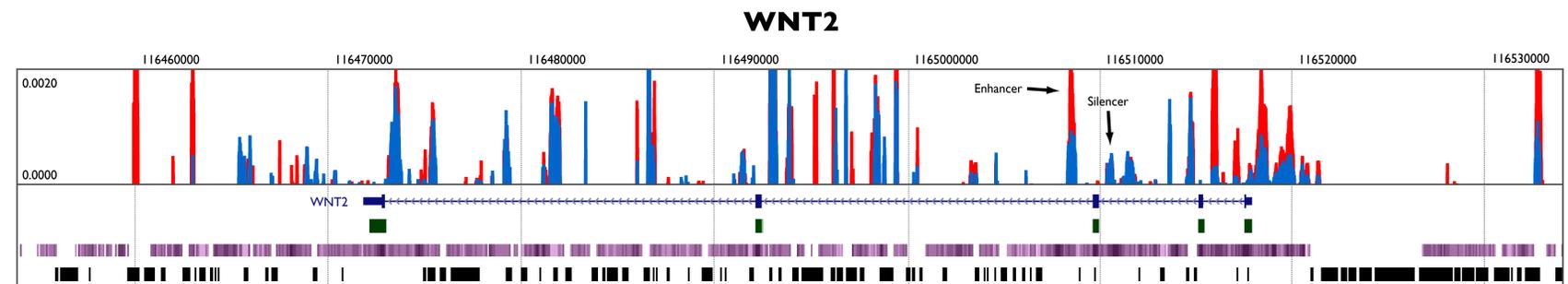
CALIBRATION

Performance studies were conducted on alignments encompassing the beta globin complex, a 70kb region containing experimentally well-studied regulatory regions. Ten sites were chosen as controls, comprising promoters, upstream sequences, and distal regulatory elements known collectively as the locus control region that are important for the regulation of beta globins. Optimum thresholds for RP scores were determined by a minimum number of false predictions and a maximum number of true predictions across the scoring range. At the optimum threshold, a minimum error rate was obtained that is related to the average failure rate of a prediction. For the 3-way RP score this is ~26%. This compares favorably to predictions based on conservation between human and mouse, which show a minimum error rate of ~53%.

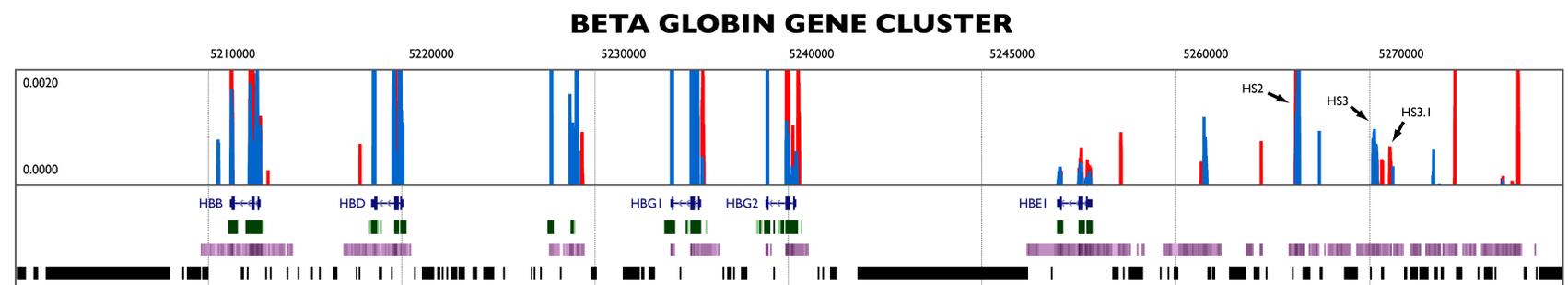
The following plots show 2-way and 3-way RP scores on several regions in the human genome. Some known cis-regulatory elements have been annotated. Elements in the training set have also been marked. RefSeq Genes, Repeat Masker repeats, and human/mouse/rat conservation annotations are also included for reference.



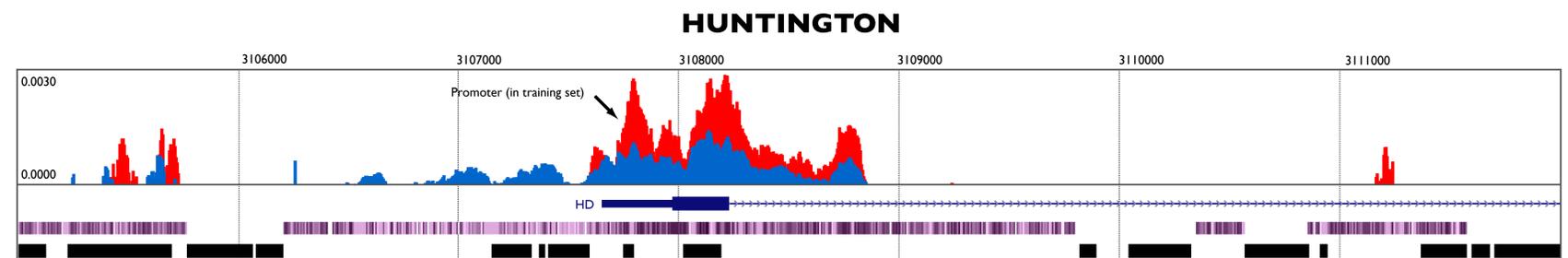
MYOD1 encodes a helix-loop-helix transcription factor and is critical to skeletal muscle lineage determination and damaged tissue repair. It has an enhancer ~20kb upstream which is contained in our training set, but also has another enhancer ~5kb upstream which has a high RP score.



WNT2 is expressed in the thalamus and is in a family of genes in oncogenesis and development of several organs and systems, and some evidence suggests a link to autism. There are two verified regulatory elements near the third exon, both with significant RP scores.



This gene cluster contains the genes that encode beta globin. Upstream of the genes (in the locus control region) are several hypersensitive sites which the RP score recognizes.



HD is a widely expressed gene important to normal development. A mutation in this gene is the cause of Huntington's Disease. This plot shows the promoter region of HD. Other high scoring sites are also shown which may be worth investigating further.