**Application of the model to the mammalian genomes: The epigenetic underpinnings of active regulatory regions are highly conserved in evolution**

Owing to the high conservation of the transcription machinery across higher eukaryotes [citation to ENCODE series papers], our model could be applied to predict enhancers within mammalian species including mice and human. These predictions were subsequently tested with targeted validation experiments.

We started with genome-wide predictions of regulatory regions in mouse. To account for tissue specificity of enhancers, the predictions are based on tissue-specific epigenomics signals. We applied our model to six different tissues, including forebrain, midbrain, hindbrain, limb, heart and neural tube, all at mouse e11.5 stage. These tissues are among the most assayed tissues in mouse ENCODE, so a rich amount of epigenetic signals for each tissue are made available by the mouse ENCODE consortium. In addition, a few of the e11.5 enhancers corresponding to each of these tissues have been tested and available from the VISTA database, for which we can compare with. Using our model, we predicted XXX-XXX number of enhancers in individual tissues, promoter VS enhancer?, ranging from 800 – 1500bp.

Similarly, we did genome wide prediction of regulatory regions in ENCODE top tier human cell lines, including H1-hESC, GM12878, K562, HepG2 and MCF-7. For each cell line, we utilized the 6-parameter integrated model to predict active enhancers and promoters based on the epigenetic datasets measured by the ENCODE consortium[53]. In H1-hESC, for example, we predicted 43463 active regulatory regions, of which 22828 (52.5%) are within 2kb of the TSS and are labeled as promoters. A large proportion of the predicted enhancers are found in the introns (30.41%) and intergenic regions (13.93%) (Figure *S14).* The predicted promoters and enhancers are significantly closer to active genes than might be expected randomly (Figure S15). By comparing the matched filter predicted enhancers and promoters with chromatin states predicted by chromHMM [30] and SegWay [27], we observe that a majority of the predicted enhancers and promoters are also predicted to be enhancers and promoters by chromHMM and SegWay respectively (Figures *S16 to S19).*

**Validation in Mouse tissues**

To test the activity of predicted mouse enhancers in vivo, we performed several rounds of enhancer validation experiments. In E11.5 mice, we performed transgenic mouse enhancer assay for 133 regions in heart and forebrain, including 102 regions selected based on the H3K27ac signals rank of corresponding mouse tissues, and 31 regions selected by an Ensemble approach from human homolog sequences (See Methods and Supplement table S3, S4). In addition, we obtained another set of transgenic mouse enhancer assay results from ENCODE Phase III Encyclopedia [Reference to the main encyclopedia paper], which assessed 151 regions in mouse e11.5 hindbrain, midbrain and limb. The combined results from these two large sets of validations, as well as any previously tested tissue-specific e11.5 enhancers from VISTA database, allow us to comprehensively evaluate our enhancer predictions in all six e11.5 mouse tissues.

Among the first 102 tested regions, 62 are selected based on forebrain H3K27ac signal rank, with 20, 22, 20 regions being in the top, middle and bottom rank respectively. Another 40 regions are selected by heart H3K27ac signal rank with half of them coming from the top rank and the other half coming from the middle rank. The bottom rank regions from heart are not included due to technical issues. Consistently, the observed active rate of assessed regions decrease from top tier to bottom tier. The validation result suggested a great prediction accuracy of our model: 61% predicted active rate versus 70% observed active rate for top tier, 45% predicted active rate versus 32% observed active rate for middle tier, and 34% predicted active rate versus 35% observed active rate for bottom tier in forebrain, etc) (See supplement table S5). For the other 31 human homolog sequences, 12.9% and 9.7% of the assessed regions are active in heart and forebrain respectively. The lower active rate is likely due to the fact that these human sequences are less well behaved in mouse tissues compared to their original native environment.

In total, we pulled together 1253 tissue specific positive regions and 8631 tissue specific negative regions from the assays. Despite the significantly unbalanced positive to negative ratios from VISTA mouse enhancer database, the 6-parameter integrative SVM models learned using balanced STARR-seq data from S2 cell-line were highly accurate at predicting active enhancers and promoters in mouse (Figure 4 and Figure S12).

In addition, the individual matched filter prediction also displayed similar accuracy for predicting enhancers and promoters in mouse tissues as in the original S2 cell-line. We also assessed the accuracy of our model using the regulatory elements identified by the transduction-based FIREWACh assay in mouse embryonic stem cells (mESC) [36]. With the same metaprofiles, the predictions are based on epigenetic signals of mESC available from ENCODE website. Again, we observe similar results for individual histone marks and combined SVM model (Figure S12). As the *in vivo* and FIREWACh assays utilized a single core promoter to validate regulatory regions, the performance of the different models in Figures 4 and S12 are probably underestimated.

**Validation in human cell lines**

We then proceeded to validate our model on predicted human enhancers using *in vitro* transduction assay. A third generation, self-inactivating HIV-1 based vector system in which the eGFP reporter was driven by the DNA element of interest was used to validate putative enhancers after stable transduction of various cell lines, including H1 hESC (Figure 5). The predicted enhancers, ranging from 650 to 2500 bp, were PCR amplified from human genomic DNA and inserted just upstream of a basal Oct-4 promoter of 142 bp (a housekeeping promoter is used so that the activity of the putative enhancers should be similar across different cell lines). VSV G-pseudotyped vector supernatants from each were prepared by co-transfection of 293T cells, and were used to transduce the various cell lines, with empty vector and FG12 vector serving as negative and positive controls, respectively. Putative enhancer activity was assessed by flow cytometric readout of eGFP expression 48-72 h post-transduction, normalized to the negative control.

A total of 25 predicted intergenic enhancers were randomly selected for validation (Supplementary Table S5). These predictions were chosen randomly to ensure that these truly represented the whole spectrum of predicted enhancers and not just the top tier of predicted enhancers. Of these 25 putative enhancers, 23 were successfully amplified and cloned into the HIV vector.  To measure the distribution of gene expression in the absence of enhancer, we also amplified and cloned 25 non-repetitive elements with similar length distribution that were predicted to be inactive using the same HIV vector.  All positive and negative DNA elements were transduced and tested for activity in both forward and reverse strand orientations since enhancers are thought to function in an orientation-independent manner. Functional testing was performed in HOS, TZMBL, and A549 cell lines in addition to H1-hESCs.

Insertion of twelve of the 23 putative enhancers into the HIV vector resulted in a significant increase in eGFP expression (P-value < 0.05 over the distribution of gene expression for negative elements) in the H1-hESCs (Supplementary Table S5). While most of the positive enhancers displayed a significant increase in gene expression irrespective of their orientation, a few elements showed significantly higher levels of gene expression in one of the orientations (Supplementary Table S6).  In contrast, the negatives displayed much lower levels of gene expression typically (Figure 5 and Supplementary Figure S20). In addition, most of these elements increased gene expression of GFP in the four different cell lines even though some of the elements were preferentially active in one of the cell lines. Overall, 16 of the 23 tested predictions displayed a statistically significant increase in gene expression of the reporter gene in at least one of the cell lines (Supplementary Table S5 and Supplementary Figure S20). Given the promoter specificity of enhancers in such assays, we would anticipate that some of the elements that could not be validated in this particular vector would function as enhancers in a more natural biological context.