RESPONSE LETTER

-- Ref1.1.1 – Presentation of in vivo validations --

<ASSIGN>MTG
<PLAN>need to incorporate text into draft
<STATUS>80%

Reviewer Comment	I understand that the authors tested 102 predicted mouse enhancers (plus 31 human orthologs) in transgenic mice, and had another 151 regions from an independent unpublished effort (Moore, in review) available for comparison. This is an unprecedented effort to assess enhancer predictions in vivo, making a systematic and rigorous comparison between the predictions and the experimental outcomes of the in vivo assays highly interesting. However, I find the presentation in the main text and figures not satisfying and partly confusing. For example, what does "61% predicted active rate versus 70% observed active rate" (page 10) mean? I interpret this statement as 61% of the tested regions were predicted to be positive and 70% of the tested regions were found to be positive – there is no indication if the predicted and observed positives actually agree.
Author Response	Thanks the reviewer for pointing this out. We agree that this sentence is a bit confusing and we'll rewrite it. Here we are describing the experimental test result of 62 elements chosen from top, middle and bottom rank of forebrain H3K27ac signal (e.g. how many of them are active in each tier). We made a rough estimation of whether these elements would be active by their overlap with the DHS peaks, but since this estimation is not very relevant, we can remove it to avoid confusion. A rigorous assessment of the our model prediction using these experimental data is presented later in the table and ROC/PR curve of Figure 4. Here we are showing that indeed the highest ranking tier has the highest validation rate, and we provide the detail validation result of each element in the supplementary table.
Excerpt From Revised Manuscript	

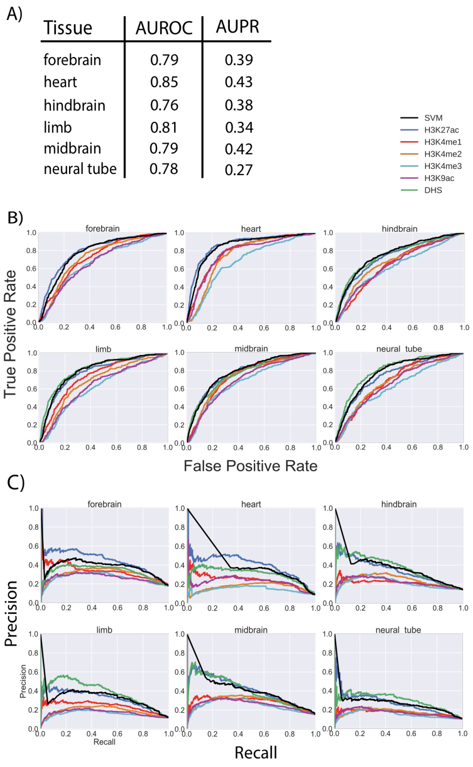
-- Ref1.1.2a – Presentation of in vivo validations --

<ASSIGN> MTG
<PLAN> recalculation and revise figure 4
<STATUS>75%

Reviewer Comment	It is my understanding that the authors have predictions for different mouse tissues and – for each tested candidate – have a readout of activity across the entire embryo, i.e. all tissues. This should allow the rigorous assessment of the prediction accuracy per tissue in comparison to an appropriate random model that accounts for the overall number of active regions per tissue (I assume Fig. 4B and C come close to this, but the corresponding text is confusing – I don't understand what Fig. 4A corresponds to).
Author Response	Thanks to the referee's comment. We have revised the manuscript and Figure 4 to clarify and incorporate the referee's points. In addition, since ENCODE has updated its ChIP-seq processing pipeline and reprocessed many of the ChIP-seq data, we redid our evaluation on mouse with the newly processed data and updated the figures accordingly. Please refer to the excerpt below:
Excerpt From Revised Manuscript	To test the activity of predicted mouse enhancers in vivo, we performed transgenic mouse enhancer assay in e11.5 mice 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. **For each tested candidate, a read out of activity across the entire embryo is collected. The number of transgenic mice that showed the pattern for each tissue is also recorded for reproducibility check (See Methods and Supplement Table S4, S5)**. In addition, we obtained another set of transgenic mouse enhancer assay results from ENCODE Phase III Encyclopedia (Moore et al., in review), 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. … We evaluated the predictability of our matched filter model for each individual histone marks and DHS, as well as the integrated SVM model (Figure 4). **For each tissue, our model ranks all the tested candidate elements with their predicted activity in this tissue using either individual feature or the integrated SVM model. Then the label of each element from experiment read out is used to assess the predictions with ROC and PR curve. One average, the integrated model trained with drosophila STARR-seq data achieves an AUROC of 0.80 and an AUPR of 0.37 for tissue-specific enhancer predictions in mouse (Figure 4A). Unlike AUROC, where the baseline is always 0.50, AUPR is more**

	sensitive to the positive to negative ratio, with a baseline being just the positive rate. Since the positive rate from the experiment varies from 8.8% 17.6% among the tissues, the AUPR has a larger variance compared the AUROC. **Consistent with previous findings from STARR-seq data, when we assess each histone modification signals independently in mice, H3K27ac signal remains best performed histone marks for predicting enhancers. In addition, the DHS signal also performs well as an independent source, as it likely shares some common information with H3K27ac. The integrated model performs similar with the highest prediction feature in each tissue. This is likely due to the fact that the model is trained entirely with drosophila matched filter scores and might not be best optimized in the mammalian systems. We believe that the integrated model would achive better performance when applying our framework directly to mouse tissue STARR-seq dataset when it becomes available.** We also did similar evaluation 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 S16). 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 S16 are probably underestimated.

Figure 4



A)

Tissue	AUROC	AUPR
forebrain	0.79	0.39
heart	0.85	0.43
hindbrain	0.76	0.38
limb	0.81	0.34
midbrain	0.79	0.42
neural tube	0.78	0.27

Figure 4: Performance of matched filters and integrated model for predicting active enhancers in mice. The performance of the *Drosophila*STARR-seq based matched filters and the integrated model for predicting active enhancers identified by transgenic mouse enhancer assays in 6 different tissues of E11.5 mice. A) The AUROC and AUPR for the integrated SVM model in 6 tissues. The weights of the different features in the integrated model is the same as the weights shown in Figure 3 for enhancers. B) The individual ROC curves of each feature and the integrated SVM model for each tissue. C) The individual PR curves of each feature and the integrated SVM model for each tissue.

-- Ref1.1.2b – Presentation of in vivo validations --

<ASSIGN>MTG + CY
<PLAN>didn't work on this yet , CY to match IDs in the tbl v the website
<STATUS>25%

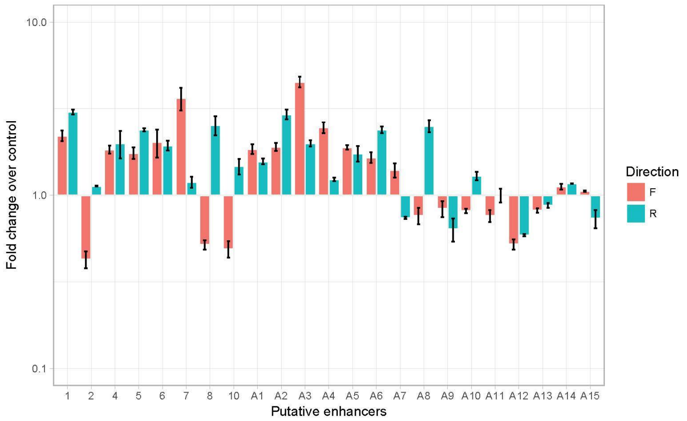
Reviewer Comment	Also, the raw images should be made available either as supplementary information of via a suitable website (e.g. the VISTA database).
Author Resonse	We have made the raw images of these experimental results available through the VISTA enhancer browser.
Excerpt From Revised Supplement	Need to add this to supplement CY: revise supplement table indicate which supplement table and make sure coordinates are searchable on VISTA enhancer browser

-- Ref1.2.1 – Validation in human cell lines: Experimental design--

<ASSIGN> MTG (& Sutton)
<PLAN> Redo experiments and represent the results in figures
<STATUS>70%

Reviewer Comment	I find the presentation of the validation in human cell lines confusing and not sufficiently well controlled. Most importantly, the tests for the individual enhancers don't seem to be replicated, such that one cannot d raw any statistically sound conclusion about the activity of each putative enhancer. Reported are only two numbers (corresponding to the fold change of gene expression of each enhancer in the forward and reverse orientation) in 4 different cell lines (table S7). These numbers often don't agree well and in some cases, the nature of these numbers is unclear. For example, what does "0. 1.06" or "0, 1.73" (note the "." vs. ",") mean - did the forward experiment fail or was the outcome exactly 0? These validations need to be performed in triplicates per cell line and construct such that each region's activity can be rigorously assessed, allowing the subsequent assessment of the predictions for each cell line.
Author Response	We acknowledge the referee's comment. We have revised both the manuscript and the supplement to describe the details of the human cell line validation experiments to make it more clear.

	The original experiment tested each enhancer in all four cell lines in replicates for both forward and reverse orientation. Based on the referee's suggestion, we performed another set of triplicate experiments on these randomly selected putative enhancers in H1-hESC. The triplicate experiment read out is consistent with our previous report. We show the result of each replicate in the supplementary tableXX and a supplementary figure is provided to visualize the data. As the figure shows, the validation experiments are highly reproducible, with the correlation between each pair of replicate being 0.9 and above.
Excerpt From Revised Manuscript and Supplement	Manuscript: We proceeded to validate our STARR-seq based model for predicting human enhancers using a cell-based 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 test putative enhancers after stable transduction of various cell lines, including H1 human embryonic stem cells (hESC) (Figure 5). The predicted enhancers, ranging from 650 to 2500 bp, were PCR amplified from human genomic DNA and inserted immediately upstream of a basal Oct-4 promoter of 142 bp. **Each putative enhancer was tested in all four cell lines in replicates for both forward and reverse orientation. For controls experiments, VSV G-pseudotyped vector supernatants from each were prepared by co-transfection of 293T cells. These were used to transduce the same cell lines, with empty vector and FG12 vector serving as negative and positive controls respectively. Note that the empty vector did have the basal Oct-4 promoter along with the IRES-eGFP cassette. Putative enhancer activity was assessed by flow cytometric readout of eGFP expression 48-72 h post-transduction, normalized to the negative control** Supplement: **The activity of each element is assessed by the read out of FlowJo cytometer. The read-out were normalized to the control and the fold change is represented in table S7, where 0 occurs when the number of positive cells is less than that of the negative control according to FlowJo gating.**

-- Ref1.2.2 – Validation in human cell lines: Experimental design--

<ASSIGN> MTG

<PLAN> Agree and remove

<STATUS> Done

Reviewer Comment	Alternatively, the cell lines for which replicate experiments cannot be performed should be removed to maintain a minimal quality standard for such validation experiments.
Author Response	Thanks and we agree with the referee's suggestion. We have made this clear in the manuscript and removed the two elements for which the experiments cannot be performed from the table.
Excerpt From Revised Manuscript and Supplement	Of these 25 putative enhancers, 23 were successfully PCR-amplified and cloned into the HIV vector in both directions.

-- Ref1.2.3 – Validation in human cell lines: Main text statements--

<ASSIGN> MTG

<PLAN>Remove the text

<STATUS>90%

Reviewer Comment	The same applies to the two statements in the main text (page 11): "a few elements showed significantly higher levels of gene expression in one of the orientations" and "even though some of the elements were preferentially active in one of the cell lines". Both statements are not

	sufficiently supported by data: neither has a systematic comparison been done, nor are the data on which these statements are based replicated. These experiments need to be performed according to minimal quality standards or the statements need to be removed.
Author Response	Here we are describing part of the experiment result rather than making strong statement about the directionality of general enhancer activity. As shown in the figure above, we find that some elements (eg, 7, 8 and A8) have significant different fold change (compared to control) for different directions, and the results are based on three replicates. However, as we are not trying to make strong statement about the directionality of enhancers, we agree to remove this description and present the raw data to the readers. As we clarified under section 1.2.1, the experiments are done in replicates and are normalized under the control.
Excerpt From Revised Manuscript	

-- Ref1.2.4 – Validation in human cell lines: Figure 5 --

<ASSIGN> MTG
<PLAN>break response into 2-3 parts
<STATUS>90%

Reviewer Comment	The presentation is also confusing: for example, figure 5 and the main text state that the Oct4 promoter is used, but also that a "housekeeping promoter is used" (page 11).
Author Response	We have made changes to the description in the manuscript so it is clearer. A minimal basal Oct4 promoter was used in the SIN HIV vector since a primary focus of the work was DNA elements active in hESC.
Excerpt From Revised Manuscript	We proceeded to validate our STARR-seq based model for predicting human enhancers using a cell-based 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 test putative enhancers after stable transduction of various cell lines, including H1 human embryonic stem cells (hESC) (Figure 5). **The predicted enhancers, ranging from 650 to 2500 bp, were PCR amplified from human genomic DNA and inserted immediately upstream of a basal Oct-4 promoter of 142 bp.**

-- Ref1.2.5 – Validation in human cell lines: Figure 5 --

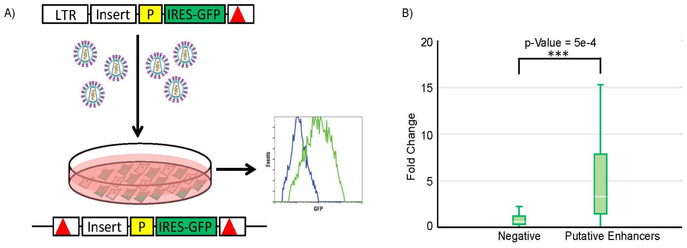
<ASSIGN> MTG

<PLAN>
<STATUS>90%

Reviewer Comment	Figure 5 shows an IRES-GFP construct, which is typically used in combination with a selection marker, yet no such marker is shown and the methods don't indicate selection (which would distort enhancer activity measurements).
Author Response	IRES-eGFP was used downstream of the DNA elements to allow flow cytometric analysis of positive cells after cell transduction. The presence of a selectable marker gene would have needlessly increased the size of the vector, which would be problematic for some of the longer elements. IRES was used so that there would be eGFP translation/readout even if transcription began within the element itself, several kilobases upstream of eGFP start codon.
Excerpt From Revised Manuscript	

-- Ref1.2.5 – Validation in human cell lines: Figure 5 --

<ASSIGN> MTG
<PLAN>
<STATUS>90%

Reviewer Comment	The authors should also comment on the LTRs' promoter function and if this could influence their results.
Author Response	To address concerns regarding the HIV LTR, figure 5 now shows SIN HIV vector structure after genomic integration, with the duplication of ~400 bp deletion of the U3 portion of the LTR. This essentially renders the LTR inactive. However, to take into account possible residual activity (and any activity of the basal Oct4 promoter), all of the transduction data is normalized to that of EV, tested on the same cells.
Excerpt From Revised Manuscript	Figure 5  A) LTR Insert P IRES-GFP Events GFP Insert P IRES-GFP B) p-Value = 5e-4 *** 20 15 10 5 0 Fold Change Negative Putative Enhancers **Figure 5: Enhancer Validation Experiments.** A) Schematic of the enhancer validation experiment flow. At top is the third generation HIV-

	based self-inactivating vector (deletion in 3’ LTR indicated by red triangle), with PCR-amplified test DNA (blue, two-headed arrow indicates fragment cloned in both orientations) inserted at 5’ of a basal (B) Oct4 promoter driving IRES-eGFP (green). Vector supernatant was prepared by plasmid co-transfection of 293T cells. Targeted cells are tranduced and then analyzed by flow cytometry a few days later. Shown below is the expected post-transduction structure of the SIN HIV vector, with a duplication of the 3’ LTR deletion rendering both LTRs non-functional B) Fold change of gene expression of eGFP is compared between negative elements and putative enhancers chosen at random, with p-value measured by Wilcoxon signed-rank test.

-- Ref1.3.1a – Prediction algorithm: Optimization and cross-validation --

<ASSIGN> ANS
<PLAN> split and edited.
<STATUS>Almost done

Reviewer Comment	The brief description of the metaprofile-based predictions on page 6 suggests optimization steps that are not well explained and could break cross-validation if performed incorrectly.
Author Response	Thanks for the referee’s comment. We have clarified and added more details to explain how we did the cross validation in the methods section within the supplemental text. The training data for creating the metaprofile and machine learning models were distinct from the test data during all cross validation tests within the manuscript.
Excerpt From Revised Manuscript (Suppl.)	During the ten fold cross validation with a single histone mark, the profiles are created with 90% of the STARR-seq positives and 10% of the positives are used for testing the accuracy of the model. With the main SVM model within the manuscript, 6 different matched filter profiles are created with 90% of the STARR-seq positives and to train the model while 10% of the positives are used for testing the accuracy of the SVM model.

-- Ref1.3.1b – Prediction algorithm: Templates #1 --

<ASSIGN> ANS
<PLAN> split and edited.
<STATUS>50%

Reviewer Comment	Specifically, the authors state that they “scanned with multiple matched filters with templates that vary in width between the two maxima in the double peak” (page 6).

	How many such templates are used and how many parameters does this add to the model?
Author Response	Thanks to the reviewer for pointing this out. We have added some details to clarify this. We have tried to make this clearer in the text. We have modified the SI to clarify this. and the answer the questions posed by reviewer.
Excerpt From Revised Manuscript (Suppl.)	A single metaprofile or template is used for each epigenetic mark. However, the distance between the two peaks in the peak-trough-peak can vary as shown in the supplementary information. We use a single template with an adjustable parameter set during fitting with matched filter. The width of the region was allowed to vary between 300-1100 basepairs (at steps of 25 basepairs). The width of the template adds a second variable during the fitting of the template to the regions of the genome (in addition to the template itself).

-- Ref1.3.1c – Prediction algorithm: Templates #2 --

<ASSIGN> ANS

<PLAN> split and edited.

<STATUS>50%

Reviewer Comment	Was the template created prior to cross validation or during cross validation?
Author Response	We have modified the methods section in the supplemental text to make this clearer.
Excerpt From Revised Manuscript (Suppl.)	During the ten fold cross validation with a single histone mark, the profiles are created with 90% of the STARR-seq positives and 10% of the positives are used for testing the accuracy of the model. With the main SVM model within the manuscript, 6 different matched filter profiles are created with 90% of the STARR-seq positives and to train the model while 10% of the positives are used for testing the accuracy of the SVM model.

-- Ref1.3.2 – Prediction algorithm: H3K27ac and DHS --

<ASSIGN> MTG

<PLAN>Mostly agree and explain

<STATUS> 80%

Reviewer Comment	I also note that the result that H3K27ac has the highest predictive value and that DHS is partly redundant to H3K27ac is highly confounded by 1. the choosing of templates based on H3K27ac and subsequent application to the other histone modifications (page 12, top paragrah) and 2. the fact that the metaprofile with the two maxima and the dip in-between (plus its width) already captures the DHS signal, which is complementary.
Author	We see the referee's point that there could be different

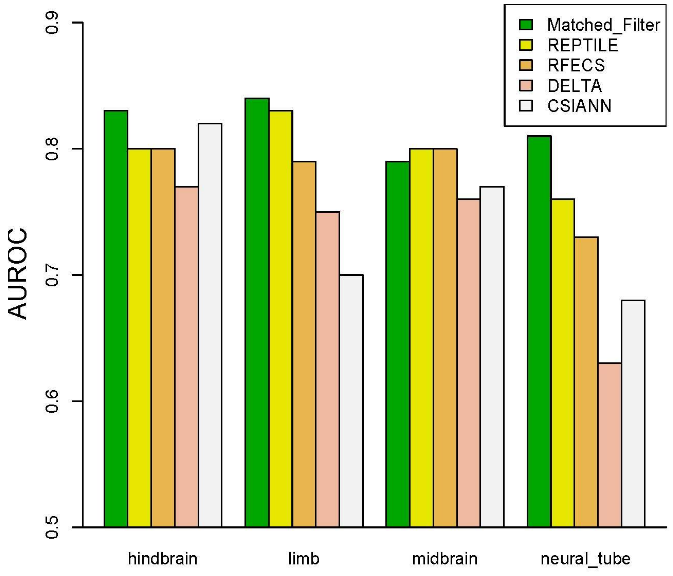
Response	interpretations of the weights of features from trained SVM mode. Specifically: In relation to #1 above, we agree that choosing the template based on H3K27ac could potentially gives H3K27ac more weights. However, H3K27ac has the highest performance even when we compare all histone marks independently. So it's not surprising that the model selects H3K27ac as the highest predictive value. The choose of templates based on H3K27ac is to define a consistent double peak region so the matched filter scores can be calculated on the same region for different histone modifications. However, the double peaks of these histone modifications usually align very well, so the templates based on H3K27ac should not introduce large bias to the weights of different features in the SVM model. As for #2 the redundancy between DHS and H3K27ac, we agree that the dip in between the two maxima is usually where the DHS peak would occur, which provides good explanation for the redundancy. We have added this discussion in the manuscript as shown below.
Excerpt From Revised Manuscript	According to the model, the acetylations (H3K27ac and H3K9ac) are the most important feature for predicting active regulatory regions. The DHS matched filter performed well as an individual feature (AUPR in Figure 2) to predict enhancers, but had a lower weight among the six features likely due to the fact that the information in DHS is redundant with the information contained within the histone mark, **eg. the DHS peaks usually occur at the trough region between two maxima in the histone signal. Despite the redundancy, combination of the DHS and histone signals is more predictive of regulatory activity as the complementary signals are strengthened compared to the uncorrelated noise in each signal.**

-- Ref1.4 – Comparison with previous methods --

<ASSIGN> MTG
<PLAN> recalculation/Exclude midbrain
<STATUS> 50%

Reviewer Comment	The authors compare their approach to chromHMM and SegWay, which are both not built for enhancer prediction but rather to segment the genome into different types of regions. A more relevant comparison to a supervised machine learning approach (Capra, ref 64) is presented only superficially in the methods section and without any

	(supplementary) figure.
Author Response	With the referee’s suggestion, we did more comparison with other published methods, and we have included the results in our manuscript as shown below. In our original submitted manuscript, we compared our method with ChromHMM and SegWay because the ChromHMM and SegWay enhancer annotations of the Roadmap Epigenetics samples has been used in many publications to define enhancer regions. We want to compare with them to show that our framework provides a better set of enhancers readily available for related studies.
Excerpt From Revised Manuscript	In addition to the comparison with unsupervised segmentation based methods, we also compared with other published enhancer prediction tools, including CSIANN, a neural network based approach; DELTA, an ensemble model integrating different histone modifications; RFECS, a random forest model based on histone modifications, and REPTILE, a more recent published method that integrates histone modifications and whole genome bisulfite sequencing data. We used their published results and compared their methods with our model on the same experimental data reported in their paper(\cite()). The comparison was done in a tissue specific manner for all four mouse tissues with all required ChIP-seq and DNase experiment data available. For 3 out of 4 tissues in the comparison, our prediction shows higher AUROC than the other four published methods. In midbrain, the AUROC for our prediction is slightly lower than REPTILE and RFECS, possibly due to the data quality of the DNase experiment performed in midbrain. The DNase experiment of mouse E11.5 stage midbrain is marked as low spot score in ENCODE. We found that while 75% to 81% of the genome regions has DNase signals in the other three tissues, only 52% of the genome regions show DNase signal in the experiment in midbrain. It is also worth noticing that our model is trained using the drosophila STARR-seq data whereas the other methods were trained directly with mouse data. We believe that our method would have better performance if mouse STARR-seq data could be applied for training in our framework.

-- Ref1.5 – Critique to main text and referencing --

<ASSIGN>ANS
<PLAN>Rewrite the response. Will do when we move to making changes in main text.
<STATUS>25%

Reviewer Comment	The main text needs to be substantially revised to improve clarity and avoid repetitiveness. While some parts explain fundamental basics in great detail, such as the difference between ROC and PR statistics (pages 5-6), other more important details are missing. For example, it only becomes obvious in the methods but not in the main text (page 5) that only STARR-seq enhancers with a H3K27ac and DHS peaks are considered (page 3 in the supplement).

Author Response	We thank the reviewer for pointing this inconsistency and have added critical details to the main text of the manuscript.
Excerpt From Revised Manuscript	As STARR-seq quantifies enhancer activity in an episomal fashion, all STARR-seq peaks may not be active in the native chromatin environment. Stark and coworkers showed that the STARR-seq peaks that occur in enriched DNase hypersentivity or H3K27ac modifications tend to be near active genes while other STARR-seq peaks tend to be associated with enrichment of repressive marks such as H3K27me3. Hence, we took the overlap of the STARR-seq enhancers with H3K27ac and/or DHS peaks to get a high confident set of enhancers that are active in vivo.

Comment [1]: +anurag.sethi@gmail.com fix _Assigned to Anurag Sethi_

-- Ref1.6 – Negative control regions --

<ASSIGN> MTG
<PLAN>Reword
<STATUS>85%

Reviewer Comment	The restriction of the STARR-seq enhancers to those that intersect with H3K27ac and DHS peaks (supplement page 3, see also my last point) and the selection of negatives as "randomly chosen regions in the genome with H3K27ac signal that had the same width distribution of the distance between double peaks near STARR-seq peaks (supplement pages 3-4) makes me wonder how H3K27ac can be the most predictive feature: if the negatives controls are chosen to match the positives in H3K27ac signals (which is a very powerful control), the predictive value of H3K27ac should be minimal or even zero. In this respect, the results are strange and the authors need to investigate the reasons for this outcome.
Author Response	Thanks the referee for the comment. For negative regions we match the width distribution which is essentially selecting regions that has similar lengths to the enhancers. These regions does not have the same H3K27ac signals in terms of the signal strength and pattern, but mostly have some background H3K27ac signals that the model would learn to distinguish from. We didn't choose non-STARR-seq peaks with no H3K27ac signal as they wouldn't provide enough information for training. Based on the comment, we have made it more clear how we select the negatives in this section of supplement as reproduced below.

Excerpt From Revised Manuscript	The negatives are randomly chosen non-STARR-seq-peak regions in the genome that had the same lengths distribution as the enhancers from the STARR-seq. We require most of the regions contain some H3K27ac signals, since negatives with no H3K27ac signal at all wouldn't provide enough information for training.

-- Ref1.7.1 – Minor comments: Title and Abstract --

<ASSIGN>
<PLAN>didn't work on this yet
<STATUS>To discuss later

Reviewer Comment	The message that the authors' approach is trained on Drosophila enhancers und functions successfully across different species does not come across very clearly in the title and abstract, which could be improved.
Author Response	To discuss Current: A framework for supervised enhancer prediction with epigenetic patternrecognition and targeted validation across organisms
Excerpt From Revised Manuscript	

-- Ref1.7.2 – Minor comments: Reference --

<ASSIGN> MTG
<PLAN>Fix the reference
<STATUS>Done

Reviewer Comment	The referencing of manuscripts is broken and needs to be fixed: several references seem to not be correctly formatted (e.g. "cite 31, 50" on page 5, "linear SVM [54]" on page 7 points to the wrong paper, "(see Supplement)" on page 12 is an unclear reference).
Author Response	We thank the referee for pointing out the formatting issue and we've fixed the citations accordingly.
Excerpt From Revised Manuscript	The STARR-seq studies on *Drosophila* cell-lines provide the most comprehensive MPRA datasets as the whole genome was tested for regulatory activity within these assays and these assays were performed with multiple core promoters [31, 49]. We built an integrated model with combined matched filter scores of the most informative epigenetics marks (H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac, and DHS) associated with active regulatory regions using a linear SVM [59].

-- Ref1.7.3 – Minor comments: BG3 cells --

<ASSIGN> MTG
<PLAN>Fix in the manuscript
<STATUS>Done

Reviewer Comment	On page 7, it seems that the authors conclude from a good performance in BG3 cells that the SVM model 'is applicable across species'. Please note that BG3 cells are also Drosophila cells.
Author Response	Thanks for pointing this out. Indeed, the validation experiments described later in the paper shows that the model is applicable across species, but the BG3 cell line validation here is to show that our model is applicable across different cell lines.
Excerpt From Revised Manuscript	The model is highly accurate at predicting active enhancers and promoters in the S2-cell line (Figure S6), indicating our framework of combining epigenetic features with a linear SVM model to predict enhancers is applicable **across different cell lines.**

-- Ref1.7.4 – Minor comments: Term correction --

<ASSIGN> MTG
<PLAN>Fix in the manuscript
<STATUS>Done

Reviewer Comment	“impute chromatin status” (page 12) should be “segment the genome based on chromatin features” or similar.
Author Response	We have rephrased the sentence as shown in the excerpt below.
Excerpt From Revised Manuscript	We first did the comparison with ChromHMM[63], a well known method to segment the genome based on chromatin features

-- Ref1.7.5 – Enhancer-specific factors --

<ASSIGN> ANS
<PLAN>
<STATUS> 80%

Reviewer Comment	The differential distribution of factor binding between enhancers and promoters (page 12 and figure 6) shows many signals for promoters but only very few (and relatively weak ones) for enhancers. Are there no enhancer-specific factors?
Author Response	Thank you for the question. There are some TFs that preferentially bind to enhancers as compared to promoters and

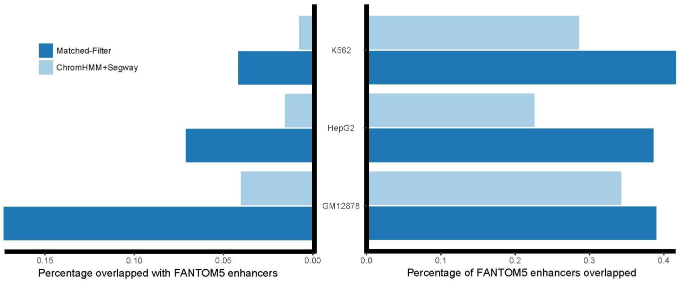
	we have expanded the text in the results to include a discussion of enhancer-specific TFs.
Excerpt From Revised Manuscript	As expected, TATA-binding proteins bind to most of the predicted active promoters according to our model. In comparison, among the TFs with experimentally measured ChIP-seq experiment, there is no single TF that binds to a majority of predicted enhancers. This indicates that unlike promoters, there is no single set of TFs that bind to a majority of active enhancers. Instead, the TFs that bind to active enhancers tend to bind to smaller subsets of enhancers. This could explain why, unlike promoters, it has been hard to find a single sequence signature associated with active enhancers in a tissue. However, a few of the TFs (for example, POUF1 and BCL11A) do bind preferentially to enhancers as compared to promoters according to our model.

-- Ref2.1a – Comparison with FANTOM5 and ENCODE --

<ASSIGN> MTG 80%
<PLAN>Compare
<STATUS>80%

Reviewer Comment	Page 3: "In addition to the small numbers, the validated enhancers were typically selected based on conserved noncoding regions [17] with particular patterns of chromatin [18], transcription-factor binding, [19] or noncoding transcription [20]." Since the FANTOM5 Atlas is the most comprehensive collection of transcribed enhancers across different primary cells and tissues, I would like to see a comparison of the model predictions in human to the enhancer dataset of the FANTOM5 Atlas dataset taking into account cell-type/tissue specificity. In a similar fashion, what is the overlap with the integrative ENCODE annotation proposed by Hoffman et al. NAR 2013.
Author Response	Thanks to the referee for this point. The FANTOM5 Atlas contains a good set of transcribed enhancers, although there is only a relatively small number of transcribed enhancers detected in each cell. Based on the referee's suggestion, we've checked our predictions against the FANTOM5 enhancer set and compared our overlap with the annotation provided by Hoffman et al, NAR 2013. We included the result in the supplement as reproduced below:
Excerpt From Revised Manuscript (in suppl.)?	For predictions in human we compared with the integrative annotation of ChromHMM and Segway using CAGE-defined enhancers from FANTOM5 Atlas. We checked the overlap

<table>
<tr>
<td></td>
<td>
between our predictions with the FANTOM5 enhancers and compared that of the integrative annotation provided by Hoffman et al, NAR 2013 in a cell-type specific manner. The FANTOM5 Atlas has included three human cell lines from ENCODE project with enhancer predictions from both methods: GM12878, K562 and HepG2. We found that the percentage of overlap for our predicted enhancers is more than three times higher than that of the combined ChromHMM and Segway enhancers in each of these cell lines. Despite the fact that our framework predicted a smaller number of enhancers, the exact number of overlap is still higher for our predictions. Around 40% of the CAGE-defined enhancers overlap with our predicted enhancers, while 23% to 34% overlap with the enhancers predicted by integrative ENCODE annotation method.



The cell-type specific percentages of overlap between FANTOM5 enhancers and two sets of predicted enhancers are shown in the bar plots. The left panel bar plot shows the fraction of overlap over the total number of enhancers predicted in each method. The right panel shows the fraction of overlap over the total number of FANTOM5 enhancers.
</td>
</tr>
</table>

-- Ref2.1b – Saturation analysis--

<ASSIGN>MTG

<PLAN>Refer to 2.3

<STATUS>done

Reviewer Comment	Assuming that the size of training datasets is the only limiting factor for achieving high discrimination performance, what is the minimum number of samples that guarantees good performance in the deployed method?
Author Response	We performed detailed saturation analysis under comment 2.3.

Excerpt From Revised Manuscript	

-- Ref2.2 – Method justification --

<ASSIGN> ANS
<PLAN>
<STATUS> 60%

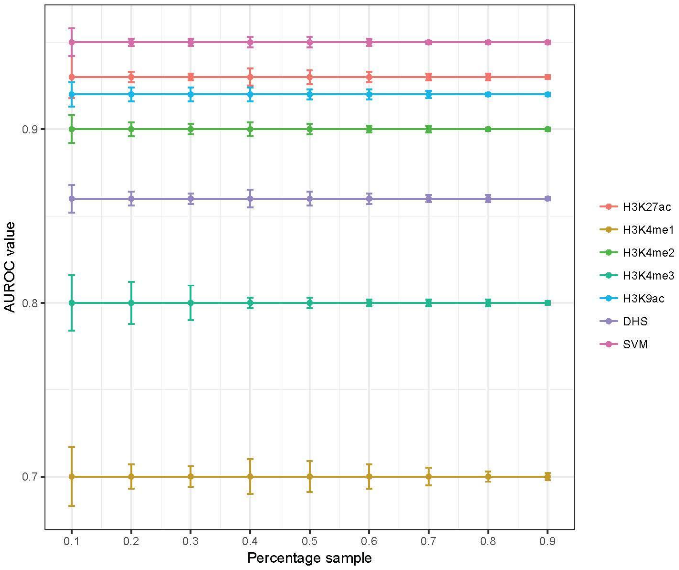
Reviewer Comment	Page 3: "For example, two widely used methods for predicting enhancers were based on the fact that these elements are expected to contain a cluster of transcription factor binding sites [24] and their activity is often correlated with an enrichment of particular post-translational modifications on histone proteins [27, 30]." In a similar fashion one can argue that the authors use STARR-seq peaks that overlap with DHS or H3K27ac peaks to identify active regulatory regions in the genome. See comment below. This requires much better justification.
Author Response	We acknowledge that we are utilizing information from epigenetic marks to define our positives. Due to the biases present within different massively parallel regulatory assays, it is difficult to define the positives for training utilizing information from a single regulatory assay. We have defined the training positives by overlapping STARR-seq peaks with epigenetic marks as these were shown by Alexander Stark to be more accurate at identifying active enhancers and promoters in the genome than using all STARR-seq peaks as explained in the manuscript now. In addition, unlike previous methods that just looked for enrichment of histone marks or DNase hypersensitivity as a predictor for active enhancers, we look for the occurrence of a template in the presence of noise for predicting enhancers.
Excerpt From Revised Manuscript	While STARR-seq identifies regions that could be potential enhancers or promoters, it does not guarantee that the region will be active or repressed in that cell-type as the activity of the region is tested in a plasmid. In machine learning models, the training data should be as well annotated as possible. As our attempt is to use the cleanest set of experimentally verified enhancers that could be active in a cell-type specific fashion, we used the experimentally active STARR-seq peaks that overlapped with DHS or H3K27ac peaks as our training data as these are more correlated with active regions in the genome as per the STARR-seq study. While we do utilize information from epigenetic marks to define our positives for training, we differ from previously published enhancer prediction methods as all our positives have

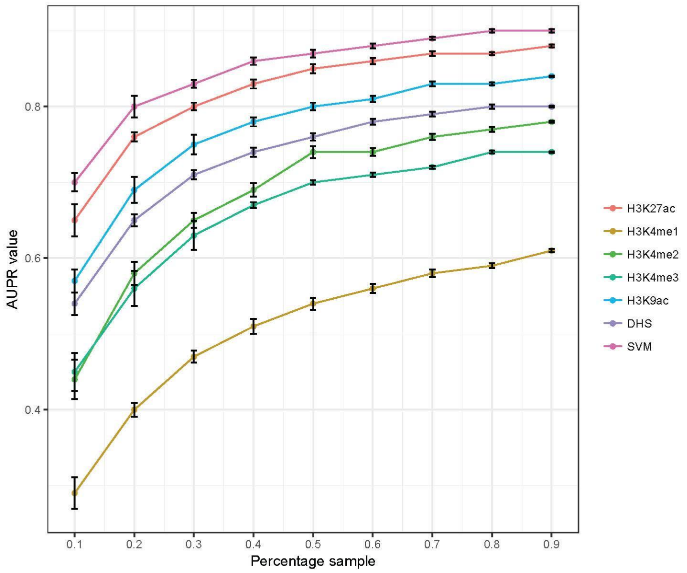
	already displayed potential enhancer activity in an experimental assay.

-- Ref2.3 – Training and test data --

<ASSIGN> MTG
<PLAN>Saturation analysis
<STATUS>80%

Reviewer Comment	Page 3: "However, the optimal method to combine information from multiple epigenetic marks to make cell-type specific regulatory predictions remains unknown. For the first time, using data from several MPRAs, we have the ability to properly train our models based on a large number of experimentally validated enhancers and test the performance of different models for enhancer prediction using cross validation" By no means this is an optimal method. This may only be considered optimized but under very specific constraints. Most of the existing methods for the prediction of regulatory regions based on epigenetic markers such as RFECS, ChromaGenSVM, DEEP, CSI-ANN, Chromia, DELTA and others including the proposed method apply heuristic techniques to identify solutions that are close to the best possible answer. So, they are optimized. The sub-optimality of the achieved solutions using epigenetic markers is not due to the training procedure of the methods, but mainly due to the variability of the epigenetic profiles across different cells or developmental stages. However, the problem-solving technique (e.g., heuristic or analytic) is not related by any means to the proper training of the method, meaning that a method is properly trained as long the training data are completely independent from the testing. **Following, the previous points, the authors need to provide more evidence about the effect of the number of training samples on the performance maximization and make clear in their manuscript that the testing data are completely independent from the training.**
Author Response	Thanks for the comment. In our original text, we didn't mean to claim that our method is the optimal method. Here, our goal is to build a framework with small number of inputs requirement to ensure that we had a widely applicable method that could be used across species. Our advantage was to use large scale STARR-seq experimental data to train the model, which was not used in previous methods. As suggested by the referee, we did a saturation analysis where we down-sampled the training data to demonstrate the effect of the training sample size on model performance. We included the result of this analysis in the supplement as reproduced below. For each cross-validation performed in this paper, the test

	dataset is completely separated from the training dataset. We have made that clear in the main manuscript and supplement as well. In addition, the many independent sources of validation performed in this paper shows that the model has good ability to generalize and has wild applications.
Excerpt From Revised Manuscript	To evaluate the impact of the training sample size on model performance, we did a saturation analysis where we down sampled the training data to different levels of fractions and evaluated the model performance on the remaining data. For each fraction level, we did a 10-fold cross-validation (see methods) and then took the average of the ten output result. The result shows that the average AUPR increases with increasing size of training data, and it starts to saturate for our SVM model with 80%-90% of the experimental data for training. In contrast to that, the average AUROC remain comparable with varying training size, but the performance variances decrease with increasing training data size. [[In methods section: The metaprofile and SVM models are trained on x% of samples and tested on the rest of the data, so the testing data is completely independent from the training.]] 

	 Figure SXX: Evaluating model performance with varying training data size. Model performance measured by A) area under the ROC curve (AUROC) B) area under the PR curve (AUPR) with different fractions of training data used. Error bar indicates the standard deviation from 10-fold cross-validation.

-- Ref2.4.1 – Exclude Marks (Figure 2) --

<ASSIGN> ANS
<PLAN>Redo the calculation and response
<STATUS>

Reviewer Comment	Figure 2 requires more information: The authors assessed the performance of the deployed matched filter algorithm by predicting active STARR-seq peaks, and they concluded that H3K27ac is the most informative predictor. However, H3K27ac together with DHS has been used for the selection of the active STARR-seq peaks. **Thus, the authors should exclude those two markers and repeat the analysis without them.**
Author Response	Thanks the reviewer for the comment. We have created a new model that utilized all STARR-seq peaks to create metaprofiles for different epigenetic marks and redone the ROC and PR calculations by training with all STARR-seq peaks without taking

	any additional information from epigenetic assays and show H3K27ac is the most informative predictor. We have added a figure to evaluate the performance of this model in the Supplemental Information.
Excerpt From Revised Manuscript	SI figure will be added.

--- Ref2.4.2 – Cross-validation Figure 2 --

<ASSIGN> MTG

<PLAN>Refer to 2.3 saturation analysis

<STATUS>80%

Reviewer Comment	Another more technical comment is about usage of 10-fold cross validation. If the number of training and testing sample is large enough 10-fold cross validation is not necessary. 5-fold cross validation is sufficient or even 2-fold cross validation assuming big numbers of training and testing data (e.g., more than few thousands).
Author Response	We thank the referee for the comment. We agree to the referee that the 5-fold or even 2-fold cross validation might be sufficient. This can be viewed from the saturation analysis under the above section 2.3. We added this point in the supplement shown below.
Excerpt From supplement	The result shows that the average AUPR increases with increasing size of training data, and it starts to saturate for our SVM model with 80%-90% of the experimental data for training. In contrast to that, the average AUROC remain comparable with varying training size, but the performance variances decrease with increasing training data size. Therefore, instead of doing 10-fold cross validation, a 5-fold cross validation might be sufficient with this size of data, as a 5-fold cross validation uses 80% of the data for training and the remaining 20% of the data for testing. Even a 2-fold cross validation could work as the AUPR is close to saturation with 50% of the data for training.

--- Ref2.4.3 – Minor comment Figure 2 --

<ASSIGN> MTG

<PLAN>Use high resolution PDF

<STATUS>80%

Reviewer Comment	Finally, there is a minor comment about the quality of Figure 2 and some other figures. In my pdf many of them appear a bit blurry.
Author Response	We used the original PDF of figure 2 but we apologize it looks a bit blurry upon upload. We'll make sure it is upload in full

	resolution and is in the clear form.
Excerpt From Revised Manuscript	

-- Ref2.5 – Feature selection --

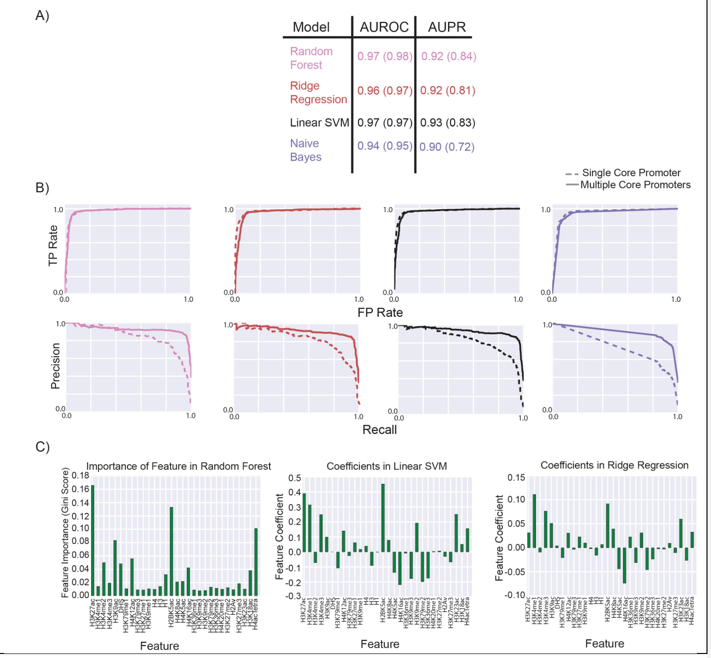
<ASSIGN> ANS

<PLAN>

<STATUS>50%

Reviewer Comment	I need more justification about the selection of six predictors for the development of the integrated model. I agree that the selected epigenetic marker datasets are widely available for many cell-lines from publicly available resources. Without doubt, this way increase the utilization of the method in new cases. **My question is why six and not another combination out of the 30?** Continuing the previous comment about optimality of the heuristically identified solutions, is there any guarantee that the integration of the selected six predictors is optimized? For example, one can apply an exhaustive search algorithm and find the best combination. One also can argue that since the performance differentiation with Random Forests is small, the latter classifier is more effective since it integrates an "out-of-bag" feature selection technique. For example, this is the biggest advantage of RFECS method that pooled together multiple epigenetic markers and identifies the most informative. **Authors have to elaborate more on the available dimensionality reduction techniques to select the best combination of predictors.** To keep it as simple as possible, combining filtering techniques such as mRMR or Gini index with the linear SVM is quite powerful and provides interpretable results.
Author Response	Thanks to the referee for the question. The 30 histone marks we tested are from drosophila experiments, and most of them of them does not have available data even in top tier tissues and cell lines for mouse and human. We have created SVM as well as random forest models with all 30 epigenetic marks and added the performance of these models to supplement. Using these models you can identify the 6 epigenetic marks that provide the most information for enhancer prediction. In our model, we chose these 6 histone marks because we wanted to test the applicability of the model trained with fly data for predicting active enhancers and promoters in mouse and human tissues. We didn't seek to pursue an optimal combination of all histone marks. While optimality of marks could potentially be used to identify other histone marks that provide complementary information about activity of enhancers and

	promoters, it could potentially reduce the applicability of the model to mouse and human tissues and cell-lines. We select the features to which are both widely available and have good individual performance. Based upon GINI index for the random forest model (Supporting Information), H3K27ac and H3K9ac are two of the epigenetic marks whose matched filters provide the best performance among the thirty marks for identifying active enhancers and promoters. In addition, H3K4me1 and H3K4me3 marks provide the ability to distinguish between promoters and enhancers (Figure 3). In addition, DHS and H3K4me2 are also widely used within the literature to identify enhancers and promoters. The set of histone marks selected in our model is in agreement with REFCS, where H3K4me1, H3K4me2, H3K4me3 are identified as the most predictive histone marks, with H3K9ac following as the commonly available highly predictive histone mark. They also adopted H3K27ac as it is the most commonly available histone mark with prior knowledge of being predictive for enhancers, although H3K27ac is not among the top important histone marks in their importance analysis. Also, we allow our model to be flexible so even one of the histone mark is missing the model still works.

Excerpt From Revised Manuscript (Figure in Supplement)	Figure S6 

-- Ref2.6 – Definition of promoters and enhancers --

<ASSIGN> ANS and MTG

<PLAN>More calculations ### Leave out this section, to be finished

<STATUS>

Reviewer Comment	Separation of active STARR-seq peaks to promoters and enhancer based on the distance from known TSSs is the adopted practice, however it is too "quick and dirty". The truth is that, it is very difficult to discriminate sharply enhancers from promoters based on the distance from TSSs since promoters have frequently function of enhancers and vice versa, and both of them share similar transcriptional architecture and have similar properties (ref. PMID: 26073855). **From a technical point of view and based on the existing results, I would like to see the performance of the deployed method by varying the distance from TSS for selecting enhancers and promoters for testing.** In the extreme case the binary classification problem is transformed to one-class classification problem that the method should handle. **An alternative way is to repeat the analysis, using appropriate CAGE-defined promoter and enhancer datasets that coincide with STARR-**

	seq peaks. There are also data from studies such as “Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay” or “High-throughput functional testing of ENCODE segmentation predictions” that could be used as baseline for benchmarking the performance of the method in a more orthogonal way.
Author Response	The referee is making a reasonable point. We have varied the promoter definition from a distance of 500-2500 bp upstream and downstream of transcription start sites and evaluated the sensitivity of our results to the cutoff. While accuracy of enhancer predictions reduce as the distance cutoff is increased, the importance of different histone marks for the enhancer model remains similar as the distance is increased. We have included a supplemental figure to display these results.
Excerpt From Revised Manuscript	Figure to be added.

-- Ref2.7 – Comparison analysis for human cell lines --

<ASSIGN> MTG and CY
<PLAN> To compare with other methods on FANTOM; ANS to find hela predictions
<STATUS> 25%

Reviewer Comment	Page 9: “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 (all available through our website)”. Following my previous comment, I would like to see the comparison analysis with CAGE-defined enhancers and promoters for some cell-specific cases, comparison with the integrative ENCODE annotation proposed by Hoffman for all top-tier cell-lines as well as comparison with other studies (see previous papers) that validated the regulatory activity of different segments in K562, HepG2 or H1-hESC cell-lines.
Author Response	Thanks for the suggestion. As the referee suggested in section 2.1a, we did a comparison with the integrative ENCODE annotation using the CAGE-defined enhancers in a cell-type specific manner. We find that our predictions has higher percentage of overlap with the transcribed enhancers from FANTOM5 Atlas. We also did a comparison with CAGE defined promoters too. We show that again our prediction has higher percentage of overlap with CAGE promoters and we included the result in the supplement as reproduced below.

Excerpt From Revised Manuscript	We also compared the overlap of our predicted promoters and the CAGE defined promoters, with the overlap between the integrative annotation and the CAGE defined promoters. We found that 70% of our predicted GM12878 promoters overlap with CAGE defined GM12878 promoters, whereas only 37% of the integrative annotations overlap. In K562 65% of our prediction overlaps versus 51% of the integrative annotation, and in HepG2 it is 63% versus 33%. Again, the enhancers predicted using our framework has higher percentage of overlap with FANTOM5 Atlas promoters.

-- Ref2.8 – Comparison with previous methods --

<ASSIGN> CY and MTG

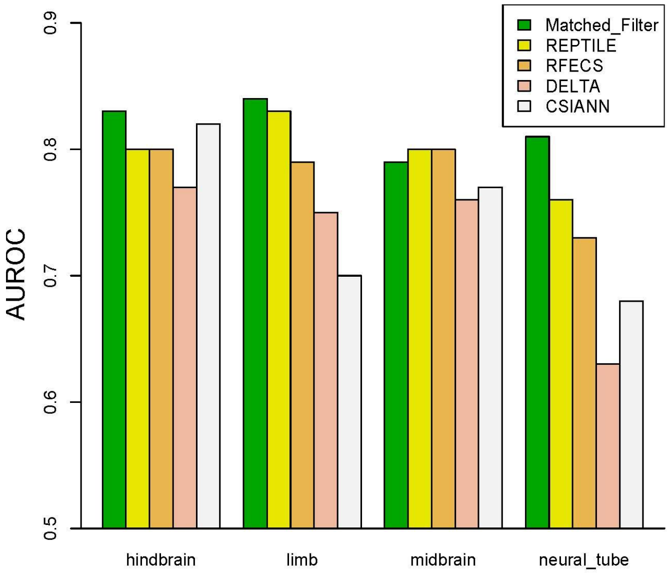
<PLAN> TF-binding comparison - to be finished

<STATUS>

Reviewer Comment	The comparison analysis is limited to ChromHMM and Segway. However, there are more methods available such as RFECS, DEEP, CSI-ANN that provide predictions for top tier ENCODE cell-lines. I would like to see a comparison analysis similar to the one presented in Figure 5 of the RFECS paper. Are the predictions of the competitor methods supported by same TF-binding sites? This might reveal that STARR-seq peaks that overlap with specific TFs such as p300 or CBP provide a better training dataset. Related to the comparison with ChromHMM and Segway. Both ChromHMM and Segway are based on probabilistic graphical models (HMM and Bayes). They should include a method of different type for example using SVM or Random Forest that is more close to what they have been developed.
Author Response	We compared with ChromHMM and SegWay as their enhancer annotation has been used in many publications as a way to define enhancer regions. Based on the referee's suggestion, we also did more comparison with other published methods, and we have included the results in our manuscript as shown below.
Excerpt From Revised Manuscript	In addition to the comparison with unsupervised segmentation based methods, we also compared with other published enhancer prediction tools, including CSIANN, a neural network based approach; DELTA, an ensemble model integrating different histone modifications; RFECS, a random forest model based on histone modifications, and REPTILE, a more recent published

Comment [2]: +chengfei.yan@yale.edu help!
Assigned to chengfei.yan

method that integrates histone modifications and whole genome bisulfite sequencing data. We used their published results and compared their methods with our model on the same experimental data reported in their paper(\cite()). The comparison was done in a tissue specific manner for all four mouse tissues with all required ChIP-seq and DNase experiment data available. For 3 out of 4 tissues in the comparison, our prediction shows higher AUROC than the other four published methods. In midbrain, the AUROC for our prediction is slightly lower than REPTILE and RFECS, possibly due to the data quality of the DNase experiment performed in midbrain. The DNase experiment of mouse E11.5 stage midbrain is marked as low spot score in ENCODE. We found that while 75% to 81% of the genome regions has DNase signals in the other three tissues, only 52% of the genome regions show DNase signal in the experiment in midbrain. It is also worth noticing that our model is trained using the drosophila STARR-seq data whereas the other methods were trained directly with mouse data. We believe that our method would have better performance if mouse STARR-seq data could be applied for training in our framework.


AUROC
hindbrain
limb
midbrain
neural_tube
Matched_Filter
REPTILE
RFECS
DELTA
CSIANN