Zero-shot imputations across species are enabled through joint modeling of human and mouse epigenomics

Jacob Schreiber[1], Deepthi Hegde[2], and William Stafford Noble*[1, 3]

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington
[2]eScience Institute, University of Washington
[3]Department of Genome Sciences, University of Washington

Abstract

Recent large-scale efforts to characterize the human epigenome have produced thousands of genome-wide experiments that quantify various forms of biological activity, such as histone modifications, protein binding, and chromatin accessibility. Although these experiments represent a small fraction of the possible experiments that could be performed, the human epigenome remains the most characterized epigenome of any species. We propose an extension to the imputation approach Avocado that enables the model to leverage the large number of human epigenomic data sets when making imputations in other species. We found that not only does this extension result in improved imputations of mouse epigenomics, but that the extended model is able to make accurate imputations for assays that have been performed in humans but not in mice. This ability to make "zero-shot" imputations greatly increases the utility of such imputation approaches, and enables comprehensive imputations to be made for species even when experimental data are sparse. Further, we found that our extension allows for an epigenomic similarity measure to be defined over pairs of regions across species based on Avocado's learned representations. We show that this score can be used to identify regions with high sequence similarity whose function have diverged.

1 Introduction

A common way for researchers to investigate questions in genomics is by designing and performing high-throughput assays that quantify various forms of biochemical activity along a genome. Such assays include ChIP-seq, which has been used to measure measure histone modification and protein binding, RNA-seq, which measures transcription, and ATAC-seq, which measures chromatin accessibility. The resulting experimental data quantify cell type-specific activity and so can be used to help explain the basis for cellular mechanisms such as differentiation and, when things go wrong, disease. Consequently, individual investigators perform these assays to answer specific research questions, and large consortia—such as the Roadmap Epigenomics Consortium, the ENCODE Project, and the International Human Epigenomics Consortium—perform and collect thousands of them into compendia that broadly characterize human epigenomics across a variety of primary cells and tissues.

Despite the value of this data, the cost of these assays means that such compendia are rarely complete. For example, the Roadmap Compendium [1] contains 1,122 experiments spanning 34

assays and 127 different human cell types and tissues (which we refer to as "biosamples"), making it only 26% filled in. This incompleteness can be an obstacle for computational methods and for investigators studying biosamples that are incompletely assayed. Recently, there have been efforts to build machine learning models that can impute these missing experiments by leveraging learned associations among the experiments that have been performed [2–4].

In practice, this problem of incompleteness is even worse in compendia collected for species other than humans. As of January 24th, 2020, the ENCODE Project portal (`https://www.encodeproject.com`) hosts only only 1,814 epigenomic experiments mapped to the mouse reference genome mm10, in contrast to the 9,111 experiments mapped to the human reference genome hg38. The experiments performed in mice span fewer assays and biosamples than the human experiments, and each mouse biosample is generally less well assayed than a typical human biosample. Perhaps most importantly, the overall characterization of protein binding is far sparser in mouse than in human, despite proteins like transcription factors playing crucial roles in the cell. To illustrate this difference in sparsity, the best characterized human biosample, K562, has 504 protein binding experiments mapped to hg38, whereas the best characterized biosample in mouse, MEL, has only 49 assays mapped to mm10. Further, only 36 mouse biosamples have been assayed for protein binding at all, while hundreds of human biosamples have been assayed for the binding of at least one protein. Because imputation approaches are restricted to the set of assays and biosamples in a given compendium, the lack of protein binding experiments poses a significant challenge that current imputation approaches cannot yet overcome.

Fortunately, many types of biochemical activity play similar roles in the cell across evolutionarily related species. For instance, the histone modification H3K4me3 is enriched in active promoters in both humans and mice [5], and the transcription factor MYC is associated with cell growth in both species [6]. These similarities suggests that it is possible for an imputation model to transfer knowledge of these types of activity across species. The concept of transfer learning has been used in other domains, such as natural language processing, where machine learning models have transferred knowledge from a high-resource setting, such as a language with plentiful annotated training examples, to a low-resource setting, where there are fewer annotated examples [7]. In our setting, the epigenomic compendia available for humans can be viewed as a high-resource setting and the sparser compendia available for other species as the low resource setting.

In this work, we propose an extension to the imputation approach Avocado [4] to enable the joint modeling of human and mouse epigenomics. In its original formulation, Avocado first organizes a compendium of data into a 3D tensor with axes corresponding to the biosamples, assays, and genomic positions along the genome. Then Avocado uses a deep tensor factorization approach to learn latent representations of each of these axes. This process is similar to a standard matrix factorization, except that the dot product operation on the latent factors is replaced with a neural network. Our extension involves merging the human and the mice compendia into a single tensor by taking the union of assays as one axis, many of which have been performed in both species, the union of biosamples as the second axis, which are disjoint across species, and the concatenation of genomic positions as the third axis. This factorization procedure is similar to independently factorizing the human and mouse compendia, except that the assay representations and the neural network parameters are tied across species.

We demonstrate that this joint optimization procedure allows Avocado to make higher quality and more comprehensive imputations for mice than one could by modeling mouse epigenomics alone. First, we show that incorporating human data into the standard imputation task improves

model performance, particularly for experiments that measure transcription. Then we show that this procedure allows for the imputation of assays that have been performed in humans but not in mice. In machine learning terminology, this is an example of a "zero-shot" problem, where a model is asked to make predictions despite not having any labeled training data for a particular class. Specifically, only 64 assays have been performed in both mice and humans, but we aim to make zero-shot predictions for a further 735 assays have been performed only in humans (excluding assays that involve performing RNA-seq after CRISPR editing or short-hairpin RNA interference).

Lastly, we show that jointly modeling human and mouse epigenomic data can encode epigenomic state into a latent representation in a manner that is shared across species. Conveniently, these representations can be learned regardless of the number of experiments in each species and without the need to either explicitly match biosamples between species or be restricted to orthologous regions. We exploit these representations to define a similarity score between pairs of regions that is based on their epigenomic similarity. To demonstrate that this similarity score differs from simple sequence similarity, we highlight an epigenomically dissimilar pair of regions with high sequence similarity that corresponds to an exon in the gene EEF1D that is alternatively spliced and potentially an alternate promoter in humans but not in mice.

2 Methods

Data sets In total, we downloaded and processed 7,986 epigenomic experiments hosted on the the ENCODE project portal (`https://www.encodeproject.org`). These experiments were partitioned into three data sets: (1) the ENCODE2018-Full data set comprising the 6,870 epigenomic experiments that measure activity in humans, (2) a subset of 3,814 of those experiments, called the ENCODE2018-Core data set, that only include experiments from biosamples or assays that had at least five experiments performed in them, and (3) the 1,116 epigenomic experiments that measure activity in mice.

The data processing pipeline is similar to previous work involving Avocado [4, 8]. All experiments were processed using the ENCODE Processing Pipeline and mapped to either human genome assembly hg38 or mouse genome assembly mm10. The signal values were $-\log 10$ p-values for the ChIP-seq and ATAC-seq data, read-depth normalized signal for DNase-seq, and normalized read coverage for the RNA-seq experiments. When multiple replicates were present for an experiment, we preferentially chose the pooled replicate; otherwise, we chose the second replicate if two had been performed, and the first (and only) replicate otherwise. The data were then further processed before model training. First, the signal was downsampled to 25 bp resolution by taking the average signal in each non-overlapping 25 bp window. Second, an inverse hyperbolic sine transformation was applied to the downsampled data. This transformation has been used previously to reduce the effect of ourliers in epigenomic signal [3, 9] and is similar to a log transformation except that it is defined at 0 and is almost linear in the range between 0 and 1. The transformed tracks are used both for training and evaluating the models.

Calculation of average activity "Average activity" here refers to the average signal value across a set of training tracks at each position in the genome. Specifically, in the context of three-fold cross-validation, for each fold the average activity is calculated for each assay using the experiments in the other two folds (the training set). This baseline is generally much stronger than the

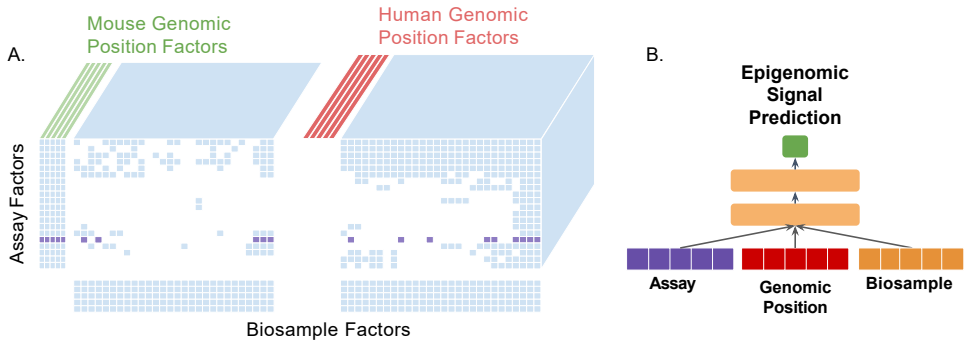

Figure 1: **The cross-species Avocado model.** (A) A schematic of the mouse and human compendia, each represented as a 3D tensor of data, aligned on the assay axis. The Avocado model learns latent representations of each dimension of these tensors but maintains a single shared assay representation across both compendia. (B) The neural network component of the model takes in factor values for a single biosample, assay, and genomic position at a time and predicts the signal for that assay in that biosample at that position. Both the biosample and the genomic position factors come either from human or mouse compendia.

simple average signal value across all loci, which is a more traditional baseline. Formally, the average activity AA for an assay a from the set of all training set experiments of that assay E at position i in the genome is calculated as $AA_{a,i} = \frac{1}{|E|} \sum_{e \in E} e_i$

Avocado model We kept the general topology of the Avocado model the same as previous work, but in some experiments we made two modifications to enable the joint modeling of two species. The first modification is that we treated the genomic axis as the concateation of positions from both the mouse and human genome (Figure 1). In our experiments, this meant that either mouse chromosome 11 or 19 was concatenated to the ENCODE Pilot Regions from the human genome. The second modification is that the total set of experiments modeled was the union of the mouse experiments and the human experiments. This meant that the biosample axis contained the union of mouse and human biosamples and that the assay axis contained the union of assays performed in mice and humans.

These modifications required changes in the training strategy. In its original formulation, Avocado would sequentially sample positions along the genomic axis and randomly select an experiment at each position to train on. However, this strategy would not work with two disjoint sets of experiments that were performed on disjoint sets of loci. Thus, for each genomic position, an experiment was selected at each locus from the set of experiments performed on that locus, i.e., mouse experiments were selected for positions on the mouse chromosome and human experiments were selected for positions in the ENCODE Pilot Regions. This procedure is similar to simply performing a separate factorization for each species except that the assay embeddings and the neural network parameters are tied across species. We also observed empirically that permuting the order of the genomic positions that were sampled, rather than passing over them sequentially, led to better convergence of the model.

Model evaluation We evaluate the presented models on only mouse chr19 and chr11 for computation efficiency, where chr19 was the smallest chromosome and chr11 was randomly selected. These evaluations were primarily done using the mean-squared-error (MSE) criterion. For our initial cross-validation experiment we randomly split experiments into three folds. For the protein

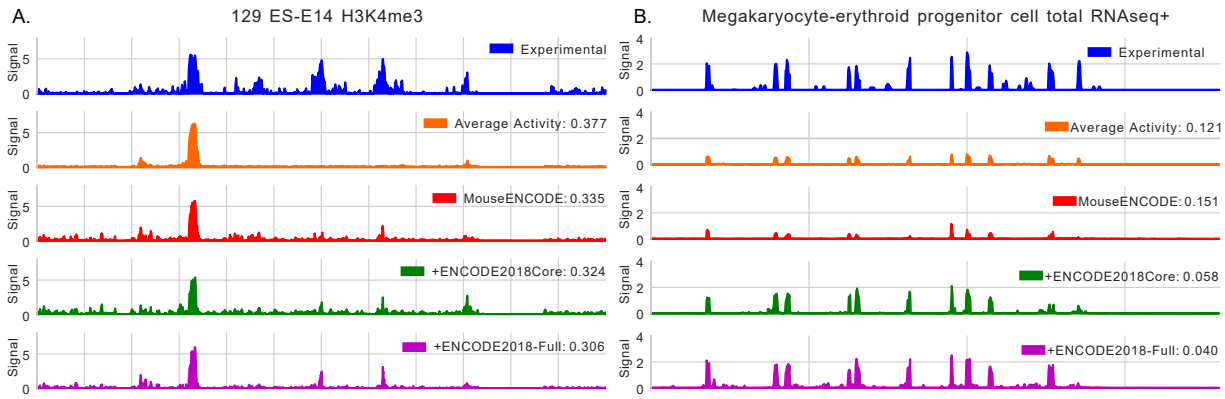

Figure 2: **Examples of real and imputed signal.** (A) An example of experimental signal for H3K4me3 in ES-E14 cells, the average activity baseline (orange), and the imputed signal from three different Avocado-based imputation models. The Avocado-based models used only mouse epigenomic data (red), mouse and the human data in the ENCODE2018-Core data set (green), or mouse and the human data in the ENCODE2018-Full data set (magenta). The MSE of each approach for the visualized region is shown in the legend. (B) The same as A except for predicting total RNA-seq in megakaryocyte-erythroid progenitor cells.

binding sections, proteins were included or excluded during cross-validation.

Identification of mm10–hg38 aligned regions We extracted pairs of regions with high sequence similarity from the mm10ToHg38 chain file (`http://hgdownload.soe.ucsc.edu/goldenPath/mm10/liftOver/`). This file contains a series of "chains," which are the boundaries of adjacent local alignments. Each chain begins with a line that indicates the starting position for each species of the local alignments. Subsequently lines indicate the locations of each of the aligned regions within the chain. Documentation can be found at `https://genome.ucsc.edu/goldenPath/help/chain.html`. We took all aligned regions with high sequence similarity that were longer than 500 bp and mapped to the positive strand on both genomes. The requirement for the positive strand is not necessary for our analysis but made extracting the aligned regions easier. This process resulted in 3,188 pairs of regions.

3 Results

3.1 Joint optimization improves imputations in mice

Our primary hypothesis is that an imputation approach that jointly models epigenomic state in humans and mice will perform better in the less characterized species than an approach that considers each species independently. Accordingly, our first evaluation involves imputing epigenomic experiments in mice. We used three-fold cross-validation to measure model performance, where the three folds came from partitioning the 1,116 experiments in MouseENCODE2019. While some models were trained using both human and mouse epigenomic experiments, the models were only evaluated on their ability to impute mouse epigenomic experiments.

We evaluated three different Avocado models that were trained using different numbers of epigenomic tracks. The first model was trained using only epigenomic experiments from mice and so represented model performance on the standard imputation task. The second model was trained using the joint optimization procedure (see Methods for details) on the same amount of mouse data

MSE	Overall	Histone Modifications	Protein Binding	Transcription	Accessibility
Average Activity	0.10030	0.13021	0.10957	0.00296	0.05512
Mouse Only	0.07125	0.09199	**0.07628**	0.00260	0.04618
Mouse + ENCODE2018-Core	0.06992	0.09004	0.07654	0.00217	**0.04542**
Mouse + ENCODE2018-Full	**0.06986**	**0.08984**	0.07683	**0.00209**	0.04617

Table 1: **Imputation performance with and without including human epigenomic data.** The mean squared error (MSE) computed both overall across all experiments and for each of the four main forms of biological activity in our data set. For each measure, the score for the best-performing model is in boldface.

as the first model but also the smaller of the two human data sets. The third model is similar to the second model except that it was trained using the larger of the two human data sets. We train models using two human data sets to assess the effect that including additional human experiments has on performance. Importantly, when performing cross-validation, models that include human epigenomic data are given access to the entirety of the human data sets for each fold.

As a baseline for the imputation approaches described above, we calculated the average activity of each assay (see Methods for details). The average activity for an assay is the average signal at each position exhibited by the training set experiments that are of that assay. The average activity baseline represents a simple rule that regions of consistently high or low signal in the training set will exhibit similar behavior in the test set [10]. Accordingly, improvement over the average activity baseline generally indicates the prediction of cell-type specific activity.

A visual inspection of the imputations made during cross-validation (Figure 2) showed that, consistent with prior work on epigenomic signal imputation, a source of error was locations that are peaks in some, but not all, biosamples [4]. These locations can be identified as those where the average activity is lower than experimental signal, i.e. those where the signal in this biosample is above average. When we visualized the imputations of H3K4me3 in HS-E14 cells (Figure 2A) and the imputation of total RNA-seq in progenitor cells (Figure 2B), we observed that the model trained on mouse data alone generally makes imputations that are similar, but not identical, to the average activity. In contrast, the models trained using human data appeared to make more accurate predictions at these biosample-specific regions, suggesting that our joint training procedure reduces the risk of fitting too closely to the average activity in the less well characterized species.

We then comprehensively calculated the overall performance of each of the models during cross-validation using MSE. The two models that leveraged human epigenomic data sets produced a small but consistent improvement in performance in comparison to the model that used only mouse epigenomic data (Table 1, Wilcoxon signed-rank test p-values of 5.9e-16 and 7.4e-17 respectively). Overall, using more human epigenomic data, in the form of ENCODE2018-Full, gave a larger decrease in MSE relative to the model trained using only mouse epigenomic data than using only the ENCODE2018-Core data set. Proportionally, the largest improvement is observed in the experiments that measure transcription, likely because the human epigenomics data sets contain a large number of transcription-measuring experiments. A similar trend was observed when building models that predict epigenomic state across species using nucleotide sequence alone [11]. Interestingly, we observe a small decrease in performance at predicting protein binding when using human epigenomics data. Potentially, this decrease in performance could arise from proteins whose binding affinities differ in the human and mouse genomes. We found that the greatest improvement came when imputing the histone modification H3K79me2, which had been assayed 6 times in mice and 37 times in humans, with a decrease from 0.128 MSE when only using mouse

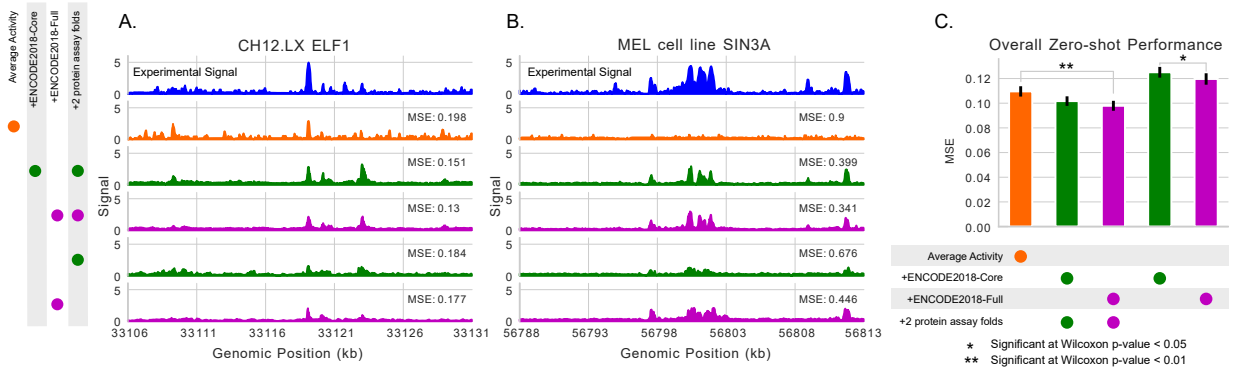

Figure 3: **Examples and evaluation of zero-shot imputations.** (A) The experimental signal, average activity, and imputed signal of four models for binding of the protein ELF1 in CH12.LX. The MSE for each approach compared to the experimental signal in the displayed window is also shown. The legend to the left shows the set of experiments used to train each model, with the top two using human experimental data as well as two of the three folds of proteins in mice, whereas the bottom two only use human experimental data. (B) The same as (A) but measuring the binding of the protein SIN3A in the MEL cell line. (3) The performance of each approach overall on 140 tracks of protein binding data on chr19 and chr11. Two statistically significant relationships are shown.

epigenomic data to 0.108 MSE when also using the ENCODE2018-Full data set. Conversely, we found that the greatest decrease in performance came from imputing the binding of ZC3H11A, which has been assayed only 3 times in mice and 2 times in human, with an increase from 0.041 to 0.049 MSE. These results support our hypothesis that joint modeling can yield improved performance when an assay is sparsely characterized in mice but well characterized in humans.

3.2 Joint optimization enables zero-shot imputations

Encouraged that our joint training procedure led to an improvement in overall performance, we hypothesized that the same procedure could allow a model to make predictions in one species for assays that have only been performed in the other species. We refer to this as the "zero-shot" setting because the model has no training data in mice for the assays that it is making imputations of. This setting is particularly relevant because there are 735 assays, most of which measure protein binding, that have been performed in humans that have not yet been performed in mice.

We investigated the zero-shot performance of models trained in two related settings, both of which involved imputing protein binding. The first setting involved holding out the experiments from some, but not all, protein binding assays performed in mice, while keeping the experiments that were performed in humans. In this setting, protein binding assays, rather than experiments, were divided into three folds for cross-validation. The second setting, which was more challenging, involved holding out all protein binding experiments in mice. While the first setting is the more realistic one, because some protein binding experiments have already been performed in mice, the second setting allows us to investigate the extent to which the tied assay and neural network parameters can be utilized. We focused our evaluations on imputing protein binding experiments due both to the importance that protein binding plays in regulating gene expression and because there are many proteins whose binding has been characterized in humans but not in mice.

We found that these models were capable of outperforming the average activity baseline even in the zero-shot setting. Notably, models in both settings were capable of imputing signal peaks both at regions with high average activity (Figure 3A) and low average activity, indicating biosample-specific peaks (Figure 3B). Models trained using the ENCODE2018-Core data set yielded imputa-

tions that beat the average activity baseline from the previous cross-validation, and models trained using the ENCODE2018-Full data set yielded even better performance. These results indicate that the model managed to learn not only the general locations where protein binding occurs, which is measured by the average activity, but also biosample-specific signal. This result is particularly striking because the model has not been exposed to nucleotide sequence, motif presence, or examples of that particular protein binding in other mouse cell types. Thus, the model is able to discern where particular proteins bind both from their epigenomic context and from the binding profiles of other proteins included in training.

3.3 Unified genomic embeddings identify epigenomic divergence

A key property of Avocado is that the learned latent representations encode a compression of epigenomic information that is useful for more than just imputing epigenomic signal. In initial work where we trained Avocado using the Roadmap compendium we showed a clustering of biosamples by anatomy type in the learned biosample representation and distinct clusters of promoters and enhancers in the learned genome embedding [4]. We hypothesized that training Avocado using epigenomic data from multiple species would result in representations of epigenomic context that were comparable regardless of the number or composition of epigenomic experiments available for each species. Learning comparable representations in such a manner would greatly simplify the epigenome alignment problem [12, 13].

We tested this hypothesis by extended the strategy proposed in our initial work for training Avocado to the setting of modeling multiple species. As before, we began by fitting a model to the ENCODE2018-Full data set using only the ENCODE Pilot Regions. Then we froze the biosample and assay embeddings as well as the neural network parameters and fit only the genome embeddings for each human chromosome. At this point, we add in the additional step of taking the frozen assay embedding and neural network parameters and training the biosample and genome embeddings for each mouse chromosome individually.

To assess whether the learned representations were comparable across species, we first inspected promoter regions in both the mouse and human genomes. We took 1,000 random protein coding genes from the GENCODE v33 annotation file and an equal number of random protein coding genes from the GENCODE M23 annotation file for mm10 and extracted the corresponding $\pm$1kb windows surrounding the respective transcription start sites. As a reference for distance in the projected space, for each promoter we also extracted a "decoy" region that was on the same chromosome and the same size but at a randomly selected position. When we projected all 4,000 of these regions using principal component analysis (PCA), we observed that while the promoters were separated from decoy regions, neither the promoters nor the decoy regions aligned across speices (Figure 4A). We suspected that this may be because the mouse chromosomes had two free axes being fit simultaneously (genomic position and biosample) and so a linear shift occured across both axes. To factor this effect out, we ran PCA separately for the examples from each species, and observed a much better degree of alignment between both the promoter and the decoy regions across species (Figure 4B). A projection of the biosample embeddings shows that those from the human and mouse models do not overlap, indicating that a shift had occured (Figure 4C). Overall, these results suggested that, despite this shift, our approach led to similar structure being encoded in the learned genome representations.

Having shown that similar structure is being encoded across species in the genomic represen-

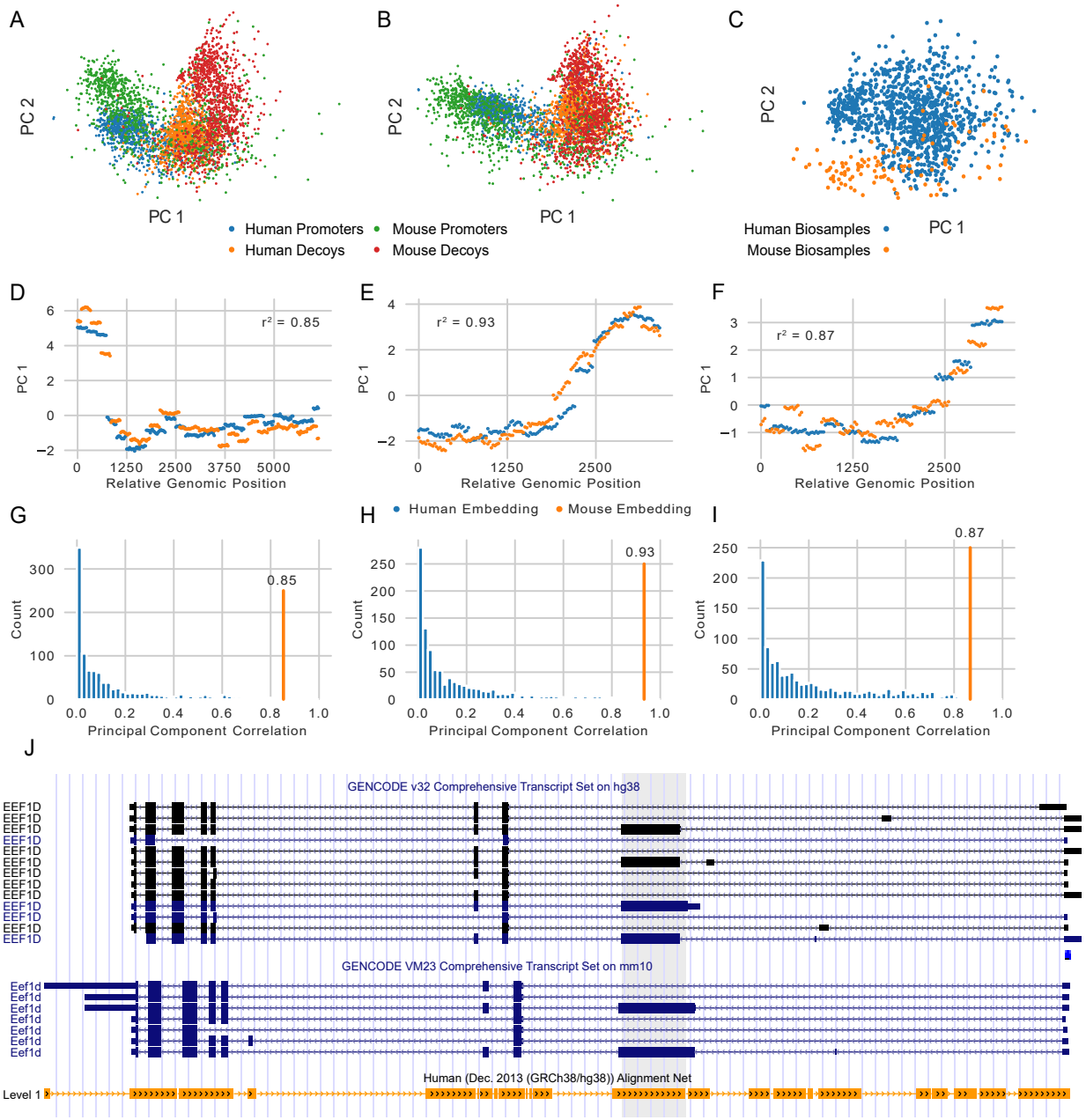

Figure 4: **Genome embeddings across species.** (A) The first and second principal components of a PCA projection of the Avocado embeddings (excluding 5 kbp factors) for human promoters and decoy regions are shown superimposed on those components for mouse promoters and decoy regions. (B) The same as (A) except as a single PCA projection of both human and mouse regions instead of separate PCA projections per species. (C) A projection of the biosample embeddings learned from the original human model and from the mouse model. (D-F) The first principal component of the Avocado embeddings (excluding 5 kbp factors) for three pairs of regions whose sequences align across mice and humans. (G-I) The correlation of the first principal components for the three regions shown in (B-D) is displayed as the orange line and the histogram shows the correlation between 1000 randomly selected mouse regions of equal size with the aligned human region. (J) Gene annotations for EEF1D on hg38 (top) and mm10 (bottom). The exon of interest is highlighted in a grey box. Below the annotations, a chain that shows aligned regions of the genes is shown. The visualization is a modified version of images from the UCSC Genome Browser.

tations, we used the representations to develop a measure of epigenomic similarity between a pair of regions. Our procedure for calculating similarity involves first defining a pair of regions, one in human and one in mice, that are of equal size. Then, similar to our approach for aligning promoter and decoy regions, we extract the first principal component separately for each of the two regions. The principal component can serve as a coarse-grained summary of epigenomic activity at each genomic position in the region. Finally, we calculate the squared Pearson correlation between the two principal components—squared because the sign can flipped for principal components—and use this correlation as our measure of similarity.

This similarity measure is complementary to a sequence-based similarity measure, e.g. alignment score, because the genome representations were not learned using nucleotide sequence. Thus, a natural comparison is to compare the epigenomic and sequence-based similarity scores for pairs of regions. Because it is likely that regions that have dissimilar sequence will also have dissimilar epigenomic state, we focus our search on regions with high sequence similarity. We extracted 3,188 aligned regions from the mm10ToHg38 chain file (see Methods for details) and began by visualizing the first principal component for three of the longest aligned regions (Figures 4B-D). We note that the principal components appear to overlap well for all three of the displayed regions and that the squared Pearson correlation values are high. However, to make sure that these values are significant, we perform a permutation test by calculating the epigenomic similarity score between the same human region and 1000 randomly selected and equally sized regions on the same mouse chromosome. For each of the three regions, we see that the epigenomic similarity score on these random regions is generally much smaller (Figures 4E-G), indicating that the similarity scores at these aligned regions are statistically significant (p-value < 0.01).

Next, we consider the more interesting setting where sequence-based similarity and epigenomic similarity differ. This form of analysis may be particularly useful for identifying the large number of orthologous regions that have functionally diverged [14]. To automatically find these regions, we calculate the epigenomic similarity score between all 3,188 pairs from the same set. We identified a representative pair of regions with low score (0.063) that corresponded to a region on chain 9 that mapped chr8:143,589,453-143,589,975 on hg38 to chr15:75,903,147-75,903,669 on mm10. This pair of regions corresponds to the largest exon on the EEF1D gene (Eef1d on mice), which encodes a protein that is a subunit of elongation factor 1 involved in transferring aminoacyl tRNAs to the ribosome [15]. Interestingly, while this exon is alternatively spliced in both humans and mice, it only serves as the initial promoter for a splice variant in humans, suggesting a difference in regulation that is not conserved between species.

4 Discussion

In this work, we introduce and analyze an optimization approach for training imputation models like Avocado by jointly modeling human and mouse epigenomics. Our experiments show that the impact of this extension can be significant. Not only does leveraging the large number of human epigenomic experiments result in improved imputations of mouse epigenomics, but the same procedure allows for zero-shot imputations to be made for assays that have been performed in human but not in mouse.

A straightforward application of these imputations will be to help prioritize the order of future epigenomic experiments in mice. We have previously described an approach for experiment

prioritization based on minimizing redundancy within imputed versions of the experiments [16]. Applying that approach to the imputations generated here would likely be even more valuable because they include proteins that have never been assayed in mice. Thus, the ordering would reflect not only what further biosamples to perform assays in, but what assays should even be performed.

Our approach assumes that the assays measure phenomena that are conserved across species. While this is generally true for evolutionarily related species, such as humans and mice, it is not always true. A source of error in the imputations may come from cases where these forms of activity differed between mice and humans. Potentially, instead of using the same assay representations across species, one could use a soft-tying approach where the model learns species-specific assay representations that are regularized to be similar to each other. On the other hand, careful analysis of errors made by the current approach may yield a data-driven way of identifying what forms of activity differ across species and where they differ along the genome. Regardless, it will be important to keep in mind that this procedure will likely need to be modified to work well on evolutionarily distant pairs of species.

Some proteins that are co-factors or from the same family exhibit similar binding profiles along the genome. We intentionally did not take these relationships into account when constructing folds for our zero-shot imputation setting. Our reasoning was that it is not the case that these sets of proteins (co-binders or families) are assayed together and so it is not a realistic assumpton that experimental data will or will not be available for all members. Rather, it is generally the case that the binding of some proteins in these sets have been assayed and a researcher is interested in imputing the binding behavior of the other family members. Indeed, for the general task of imputing the binding of proteins, it would be expected that an algorithm would make use of close relationships. One would be disappointed to, for instance, give a model an experiment profiling MYC binding and see poor performance at predicting MAX binding.

An element of this work that deserves further inspection is the process for learning representations when jointly modeling mouse and human epigenomics. Our preliminary work resulted in an epigenomic similarity score that can measure the degree to which a pair of regions are epigenomically similar. Because the epigenomic similarity score is not directly based on nucleotide sequence, it serves as a complementary measure to sequence similarity. Indeed, we found that some pairs of regions with aligned sequence have a very low epigenomic similarity score, suggesting a divergence in function in these regions. A comprehensive analysis of these regions will likely help researchers better identify and understand regions whose function have diverged. We anticipate that designing optimization strategies that result in improved comparability of the representations across species will be an important piece of future work.

Imputations, models, and latent factors produced by this project will be made freely available at `https://github.com/jmschrei/avocado`.

References

[1] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, and Michael J Ziller. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

[2] J. Ernst and M. Kellis. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature Biotechnology*, 33(4):364–376, 2015.

[3] T. J. Durham, M. W. Libbrecht, J. J. Howbert, J. A. Bilmes, and W. S. Noble. PREDICTD: PaRallel Epigenomics Data Imputation with Cloud-based Tensor Decomposition. *Nature Communications*, 9, 2018.

[4] J. M. Schreiber, T. J. Durham, J. Bilmes, and W. S. Noble. Multi-scale deep tensor factorization learns a latent representation of the human epigenome. *Genome Biology*, 2020. In press.

[5] C. W. Hanna, H. Demond, and G. Kelsey. Epigenetic regulation in development: is the mouse a good model for the human? *Human Reproduction Update*, 24:556–576, 2018.

[6] J. P. Morton and O. J. Sansom. Myc-y mice: From tumour initiation to therapeutic targeting of endogenous myc. *Molecular Oncology*, 7(2):248–258, 2013.

[7] J. Hana, A. Feldman, and C. Brew. A resource-light approach to Russian morpholoy: Tagging Russian using Czech resources. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 222–229, 2004.

[8] J. M. Schreiber, J. Bilmes, and W. S. Noble. Completing the ENCODE3 compendium yields accurate imputations across a variety of assays and human biosamples. *Genome Biology*, 2020. In press.

[9] M. M. Hoffman, O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes, and W. S. Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, 9(5):473–476, 2012.

[10] J. M. Schreiber, R. Singh, J. Bilmes, and W. S. Noble. A pitfall for machine learning methods aiming to predict across cell types. *bioRxiv*, 2019. https://www.biorxiv.org/content/10.1101/512434v1.

[11] D. R. Kelley. Cross-species regulatory sequence activity prediction. *bioRxiv*, 2019.

[12] J. Lu, X. Cao, and S. Zhong. EpiAlignment: alignment with both dna sequence and epigenomic data. *Nucleic Acids Research*, 47:W11–W19, 2019.

[13] R. F. Lowdon, H. S. Jang, and T. Wang. Evolution of epigenetic regulation in vertebrate genomes. *Trends Genet*, 32:269–283, 2016.

[14] W. H. Gharib and M. Robinson-Rechavi. When orthologs diverge between human and mouse. *Brief Bioinform*, 12:436–441, 2011.

[15] T. Kaitsuka and M. Matsushita. Regulation of transcription factor eef1d gene function by alternate splicing. *Int J Mol Sci*, 16:3970–3979, 2015.

[16] J. M. Schreiber, J. Bilmes, and W. S. Noble. Prioritizing transcriptomic and epigenomic experiments by using an optimization strategy that leverages imputed data. *bioRxiv*, 2019. https://www.biorxiv.org/content/10.1101/708107v1.