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

Jacob Schreiber[1], Deepthi Hedge[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

October 10, 2019

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 modification, protein binding, and chromatin accessibility. While 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 efforts are sparse.

Introduction It has become common practice to quantify various forms of biochemical activity along the human genome by designing and peforming high-throughput assays. Such assays include ChIP-seq, which measures histone modification and protein binding, RNA-seq, which measures transcription, DNase-seq and ATAC-seq, which measure chromatin accessibility, and several others. These assays are performed both by investigators aiming to answer specific research questions, and by large consortia—such as the Roadmap Epigenomics Consortium, the ENCODE Project, and the International Human Epigenomics Consortium—to broadly characterize human epigenomics across a variety of primary cells and tissues. These collections of experiments, often called "compendia", are invaluable for researchers studying the interplay among various forms of biochemical activity or the molecular basis for development and disease.

Despite the value of such compendia, the cost involved in performing experimental characterization means that such compendia are rarely complete. For example, the Roadmap Compendium [1], which was released in 2015, 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, computational strategies have been proposed to complete these compendia by using associations among assays and among biosamples to impute the experiments that have not yet been performed [2–4].

This problem of incompleteness is exacerbated in compendia collected for species other than humans. For example, there are only 1,805 epigenomic experiments mapped to the mouse reference genome mm10 on the ENCODE Project data portal (`https://www.encodeproject.com`) as of October 2nd, 2019.

MSE	Overall	Histone Modifications	Protein Binding	Transcription	Accessibility
Average Activity	0.08811	0.11402	0.09677	0.00257	0.05147
Mouse Only	0.06351	0.08197	**0.06728**	0.00223	0.04326
Mouse + ENCODE2018-Core	0.06319	0.08139	0.06816	0.00207	**0.04262**
Mouse + ENCODE2018-Full	**0.06264**	**0.08060**	0.06804	**0.00189**	0.04264

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.

This is in contrast to the 9,063 experiments mapped to the human reference genome hg38 as of the same date. 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. In particular, the overall characterization of protein binding is much sparser in mice than in human, despite proteins such as transcription factors playing a crucial role in the cell. To illustrate this difference in sparsity, the best characterized human biosample, K562, has 504 protein binding experiments mapped to hg38, while 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, whereas hundreds of human biosamples have been assayed for at least one protein.

Fortunately, many forms of biochemical activity play similar roles in the cell across evolutionarily-related species. For instance, the histone modification H3K4me3 is known to be enriched in active promoters in both humans and mice [**?**], and the transcription factor MYC is associated with cell growth in both species [5]. This similarity suggests that it may be possible for a computational model to transfer knowledge of these types of activity from a species that is well characterized to another that is less well characterized. 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 [6]. In our setting, the epigenomic compendia available for humans could be viewed as a high-resource setting, whereas the sparser compendia available for other species would be the low resource setting.

In this work, we propose an extension to the imputation method Avocado [4] that enables the joint modeling of human and mouse epigenomics. This extension involves merging the human and the mice compendia by taking the union of assays as one axis, the union of biosamples as the second axis, 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 embeddings and the neural network parameters are tied across species (see Methods for details). We anticipated that training these parameters using data from both species would yield better imputation performance in the low-resource species than training using only data from the low-resource species. Furthermore, a key benefit of this approach is that it allows for the imputation of assays that have been performed in the high-resource species but not in the low resource species. This "zero-shot" setting will be particularly useful because it means that the compendia of imputations for new species are not limited by the set of assays that have already been performed in the low resource species. For instance, 64 assays have been performed in both species, but 735 assays have only been performed in humans (excluding assays that involve performing RNA-seq after CRISPR editing or short-hairpin RNA interference).

Joint optimization improves imputations in mice We began our evaluation by comparing the performance of models trained using different amounts of epigenomic data. 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 a joint optimization procedure (see Methods for details) on both

Average Activity	ENCODE2018-Core	ENCODE2018-Full	Other Proteins	MSE
✓				0.09677
	✓		✓	0.09252
		✓	✓	0.08570
	✓			0.11867
		✓		0.10446

Table 2: **Zero-shot imputation performance with different training data sets.** The mean squared error (MSE) computed across all protein binding experiments as a result of various forms of zero-shot imputation.

mouse and human epigenomic data, where the human data came from the ENCODE2018-Core data set, which is a collection of 3,814 human epigenomic experiments. The third model is similar to the second model except that it used the ENCODE2018-Full data set, which is a superset of the ENCODE2018-Core data set that includes 6,870 experiments. We consider both data sets to assess the effect that including more human experiments has on performance. Lastly, as a baseline, we calculated the average activity at each genomic position of each assay. Improvement relative to this baseline generally indicates biosample-specific activity. These models, and the average activity baseline, were each evaluated by three-fold cross-validation on the 1,145 mouse epigenomic experiments.

We observed that leveraging human epigenomic data sets led to a small but consistent improvement in performance in comparison to using only mouse epigenomic data (Table 1). Overall, using more human epigenomic data, in the form of ENCODE2018-Full, gave a larger gain (Wilcoxon signed-rank test p-value = 2.8e-15) than using less human epigenomic data, in the form of ENCODE2018-Core (p-value = 3.9e-08). Proportionally, the largest improvement is observed in the transcription-based experiments, which may be because the human epigenomics data set contains a large number of transcription-measuring experiments and includes assays not seen in the mouse data set such as CAGE and RAMPAGE. Interestingly, we observe a small decrease in performance at predicting protein binding in mice when using human epigenomics data. Potentially, this decrease in performance could arise from proteins that bind in different contexts on 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 0.163 MSE to 0.104 MSE. This result supports our hypothesis that joint modeling can yield improved performance when an assay is sparsely characterized in mice but well characterized in humans.

Joint optimization enables zero-shot imputations Encouraged by the overall improvement in performance from models that used human epigenomic data, we next investigated the ability of our approach to make zero-shot imputations 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 were divided into three folds, and cross-validation was performed to evaluate performance. 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 Avocado was capable of making biosample-specific imputations even in the zero-shot setting (Table 2). Models trained using the ENCODE2018-Core data set yielded imputations 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 imputations manage to learn not only the general locations where protein binding occurs, which is measured by the average activity, but

biosample-specific signal as well. 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 of other proteins included in training.

Discussion Taken together, these results suggest that our proposed joint optimization approach is capable of leveraging epigenomic measurements from well characterized species to improve imputations in more poorly characterized species. Because our approach is conceptually simple, we anticipate that it would be straightforward to apply to a variety of species, not just mice as we do here.

We did not observe an increase in performance for all assays when we included human epigenomic data during training. Potentially, because the model uses tied assay representations, this decreased performance could indicate forms of activity that differ across species. Quantifying the difference in performance one observes when modeling data from both species versus a single species may result in a data-driven way to identify those phenomena in the cell that do differ between the species.

Some proteins are from the same family, or bind to each other, and so exhibit similar binding profiles along the genome. We intentionally did not take protein families into account when constructing folds for our zero-shot imputation setting. Our reasoning was that, because it is not the case that entire families of proteins are assayed together, it is not a realistic evaluation setting to hold out these entire families. Rather, it is generally the case that the binding of some proteins in a family have been assayed, and one remains interested in imputing the binding behavior of the other family members.

We anticipate that these imputations will be of great use to investigators, and that this approach will increase the utility of currently available human epigenomics experiments and existing imputation methods.

Methods

Data sets In total, we downloaded and processed 8,015 epigenomic experiments from the ENCODE project (`https://www.encodeproject.org`). These experiments were partitioned into three data sets. The first was the ENCODE2018-Full data set, which comprised of the 6,870 epigenomic experiments that measured activity in humans. The second was the ENCODE2018-Core data set, which comprised 3,814 of those experiments included in ENCODE2018-Full that included only biosamples that had at least five assays performed in them, and assays that had been performed in at least five biosamples. The third was the 1,145 epigenomic experiments that measured activity in mice. All experiments were processed in the same manner as previous work involving Avocado.

Avocado 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. In our experiments, this meant that chromosome 19 from the mouse genome 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.

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] Jason Ernst and Manolis 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. *bioRxiv*, 2018. https://www.biorxiv.org/content/early/2018/07/08/364976.

[5] 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.

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