Schreiber *et al. Genome Biology* (2020) 21:82
https://doi.org/10.1186/s13059-020-01978-5

Genome Biology

METHOD Open Access

Completing the ENCODE3 compendium yields accurate imputations across a variety of assays and human biosamples

Jacob Schreiber[1*], Jeffrey Bilmes[1,2] and William Stafford Noble[1,3*]

Abstract

Recent efforts to describe the human epigenome have yielded thousands of epigenomic and transcriptomic datasets. However, due primarily to cost, the total number of such assays that can be performed is limited. Accordingly, we applied an imputation approach, Avocado, to a dataset of 3814 tracks of data derived from the ENCODE compendium, including measurements of chromatin accessibility, histone modification, transcription, and protein binding. Avocado shows significant improvements in imputing protein binding compared to the top models in the ENCODE-DREAM challenge. Additionally, we show that the Avocado model allows for efficient addition of new assays and biosamples to a pre-trained model.

Background

Recently, several scientific consortia have generated large sets of genomic, transcriptomic, and epigenomic data. For example, since its inception in 2003, the NIH ENCODE Consortium [1] has generated over 10,000 human transcriptomic and epigenomic experiments. Similar efforts include Roadmap Epigenomics [2], modENCODE [3], the International Human Epigenome Consortium [4], mouseENCODE [5], PsychENCODE [6], and GTEx [7]. These projects have varied motivations, but all spring from the common belief that the generation of massive and diverse high-throughput sequencing datasets can yield valuable insights into molecular biology and disease.

Unfortunately, the resulting datasets are usually incomplete. In the case of ENCODE, this incompleteness is by design. Faced with a huge range of potential cell lines and primary cell types to study (referred to hereafter using the ENCODE terminology "biosample"), ENCODE investigators made the strategic decision to perform "tiered" analyses. Thus, some "Tier 1" biosamples were analyzed using a large number of different types of sequencing assays, whereas biosamples assigned to lower tiers were analyzed in less depth. This strategy allowed ENCODE to cover many biosamples while also allowing researchers to examine a few biosamples in great detail. In other cases, even for a consortium such as GTEx, which aims to systematically characterize a common set of tissue types across a set of individuals using a fixed set of assays, missing data is unavoidable due to the cost of sequencing and loss of samples during processing. Given the vast space of potential biosamples to study and the fact that new types of assays are always being developed to characterize new phenomena, the sparsity of these compendia is likely to increase over time.

This incompleteness can be problematic. For example, many large-scale analysis methods have trouble handling missing data. Despite the benefit that additional measurements may offer, many analysis methods discard assays that have not systematically been performed in the biosamples of interest. More critically, many biomedical scientists want to exploit these massive, publicly funded consortium datasets but find that the particular biosample type that they study was relegated to a lower tier and hence is only sparsely characterized.

*Correspondence: jmschr@cs.washington.edu; william-noble@uw.edu.
[1]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, USA
[3]Department of Genome Sciences, University of Washington, Seattle, USA
Full list of author information is available at the end of the article

Imputation methods address this problem by filling in the missing data with computationally predicted values. Imputation is feasible in part due to the structured nature of consortium-style datasets, in which data from high-throughput sequencing experiments can be arranged systematically along axes such as "biosample" and "assay." The first epigenomic imputation method to be applied at a large scale, ChromImpute [8], trains a separate machine learning model for each missing experiment, deriving input features from the same row or column in the data matrix, i.e., training from experiments that involve the same biosample but a different assay or the same assay but a different biosample. A second method, PREDICTD [9], takes a more wholistic approach, first organizing the entire dataset into a 3D tensor (assay × biosample × genomic position) and then training an ensemble of machine learning models that each jointly decompose all experiments in the tensor into three matrices, one for each dimension. PREDICTD imputes missing values by linearly combining values from these three matrices. Most recently, a third method, Avocado [10], extends PREDICTD by replacing the linear combination with a non-linear, deep neural network, and by modeling the genomic axis at multiple scales, thereby achieving significantly more accurate imputations without the need to train an ensemble of models.

All three of these existing imputation methods rely upon a common dataset. In creating ChromImpute, Ernst and Kellis utilized what was, at the time, one of the largest collections of uniformly processed epigenomic and transcriptomic data, derived from 1122 experiments from the Roadmap Epigenomics and ENCODE consortia. To allow for direct comparison between methods, both PREDICTD and Avocado relied upon a subset of 1014 of those experiments. Since 2015, however, the amount of available data has increased tremendously. Here, we report the training of Avocado on a dataset derived from the ENCODE compendium that contains 3814 tracks from 400 biosamples and 84 assays (Fig. 1). This ENCODE2018-Core dataset is 3.4 times larger than the original ChromImpute dataset. We demonstrate that this increase in size leads to a concomitant improvement in predictive accuracy.

Furthermore, whereas the ChromImpute dataset included only chromatin accessibility, histone modification, and RNA-seq data, the ENCODE2018-Core dataset also includes ChIP-seq measurements of the binding of transcription factors (TF) and other proteins, such as CTCF and POLR2A (referred to hereafter, for simplicity, as "transcription factors," despite the differences in their biological roles). Accurate prediction of TF binding in a cell type-specific fashion is an extremely challenging and well-studied problem (reviewed in [11]). We demonstrate that by leveraging the large and diverse ENCODE2018-Core dataset, Avocado achieves high accuracy in prediction of TF binding, outperforming several state-of-the-art methods.

Finally, we demonstrate a practically important feature of the Avocado model, namely, that the model can be easily extended to apply to newly or very sparsely characterized biosamples and assays via a simple transfer learning approach. Specifically, we demonstrate how a new biosample or assay can be added to a pre-trained Avocado model by fixing all of the existing model parameters and only training the new assay or biosample factors. We do this using experiments from a second dataset, ENCODE2018-Sparse, that contains 3056 experiments from biosamples that are sparsely characterized and from assays that have been performed in only few biosamples. We find that the model can yield high-quality imputations for transcription factors that are added in this manner, and that these imputations can outperform the ENCODE-DREAM challenge participants even when trained using a

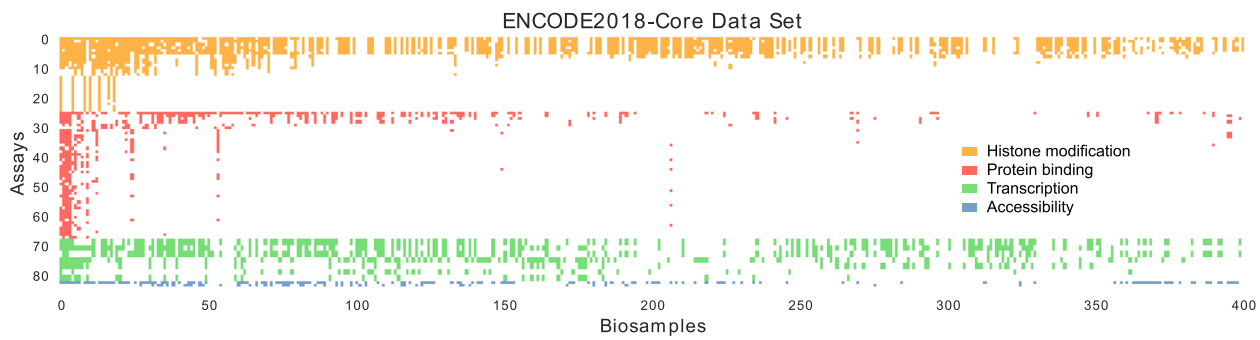

Fig. 1 The ENCODE2018-Core data matrix. In the matrix, columns represent biosamples and rows represent assays. Colors correspond to general types of assays (histone modification ChiP-seq in orange, transcription factor ChIP-seq in red, RNA-seq in green, and chromatin accessibility in blue). Biosamples are sorted by the total number of assays performed in them, and assays are first grouped by their type before being sorted by the number of biosamples that they have been performed in

single track of data. Finally, we find that when biosamples are added using only DNase-seq experiments, the resulting imputations for other assays can still be of high quality.

As a resource for the community, we have made the AvocadoENCODE imputations publicly available via the ENCODE portal (http://www.encodeproject.org).

Results

Avocado's imputations are accurate and biosample specific

We first aimed to evaluate systematically the accuracy of Avocado's imputed values on the ENCODE2018-Core dataset. One challenge associated with this assessment is that no competing imputation method has yet been applied to this particular dataset, making a direct comparison of methods difficult. Further, the size of the dataset makes training competing methods difficult, with ChromImpute requiring the training of thousands of different models. However, we have shown recently that the average activity of a given assay across many biosamples is a good predictor of that activity in a new biosample [12]. Admittedly, this predictor is scientifically uninteresting, in the sense that it makes the same prediction for every new biosample and so, by construction, cannot capture biosample-specific variation. However, we reasoned that improvement over this baseline indicates that the model must be capturing biosample-specific signal. Furthermore, because the signal from most epigenomic assays is similar across biosamples, the average activity predictor serves as a strong baseline that any cross-cell type predictor must beat. Accordingly, we compare the predictions made by Avocado to the average activity of that assay in the training set that was used for model training.

Overall, we found that Avocado is able to impute signal accurately for a variety of different types of assays. We compared Avocado's imputations to those of the average activity predictor across 37,249,359 genomic loci from chromosomes 12–22 using fivefold cross-validation among epigenomic experiments in the ENCODE2018-Core dataset. Qualitatively, we observed strong visual concordance between observed and imputed values across a variety of assay types (Fig. 2a, Additional file 1: Figure S2). In particular, the imputations capture the shape of peaks in histone modification signal, such as those exhibited in H3K27ac and H3K4me3; the shape of peaks found in assays of transcription factors like ELF1 and CTCF; and the exon-specific activity in gene transcription assays. As our primary quantitative measure, we compute the global mean-squared error (MSE) between the observed and imputed values. This value reduces from 0.0807 to 0.0653 (paired t test p value of 1e−157), a reduction of 19.1%, between the average activity predictor and Avocado (Fig. 2b).

We also compute five complementary quantitative measures. Two measures emphasize the ability of an imputation method to correctly identify peaks in the data. One of these (mse1obs), defined as the MSE in the positions with the top 1% of observed signal, corresponds to a notion of recall. The complementary measure (mse1imp), defined as the MSE in position with the top 1% of imputed signal, corresponds to precision. Three additional measures focus on the MSE in regions of biological activity: the MSE in promoters (mseProm), gene bodies (mseGene), and enhancers (mseEnh). In aggregate, Avocado outperforms the average activity baseline on all six performance measures (p values between 8e−65 for mse1imp and 1e−157 for mseGlobal) (Fig. 2b, c).

When grouped by assay, we find that Avocado outperforms the average activity in 71 of the 84 experiments in our test set according to mseGlobal. Further investigation suggested that these problematic assays were mostly of transcription, indicating a weakness of the Avocado model, or assays that may have been of poor quality (Additional file 2).

The primary benefit of the ENCODE2018-Core dataset, in comparison to previous datasets drawn from the Roadmap Compendium, is the inclusion of many more assays and biosamples. We hypothesized that not only will this dataset allow us to make a more diverse set of imputations, but that these additional measurements will improve performance on assays already included in the Roadmap Compendium. We reasoned this may be the case because, for example, previous imputation approaches have imputed H3K36me3, a transcription associated mark, but have not utilized measurements of transcription to do so. A direct comparison to previous work was not simple due to differences in the processing pipelines and reference genomes, and so we retrained Avocado using the same fivefold cross-validation strategy after having removed all experiments that did not originate from the Roadmap Epigenomics Consortium. Additionally, we removed all RNA-seq and methylation datasets, as they had not been used as input for previous imputation methods. This resulted in 1072 tracks of histone modification and chromatin accessibility.

We found that the inclusion of additional assays and biosamples leads to a clear improvement in performance on the tracks from the Roadmap Compendium. The MSE of Avocado's imputations dropped from 0.115 when trained exclusively on Roadmap datasets to 0.107 when trained on all tracks in the ENCODE2018-Core dataset, an improvement of 7% (p value of 8e−45). When we grouped the error by assay, we observed that tracks appeared to range from a significant improvement to only a small decrease in performance (Additional file 1: Figure S3A). When aggregating these performances across assays, we similarly observe large improvements in the

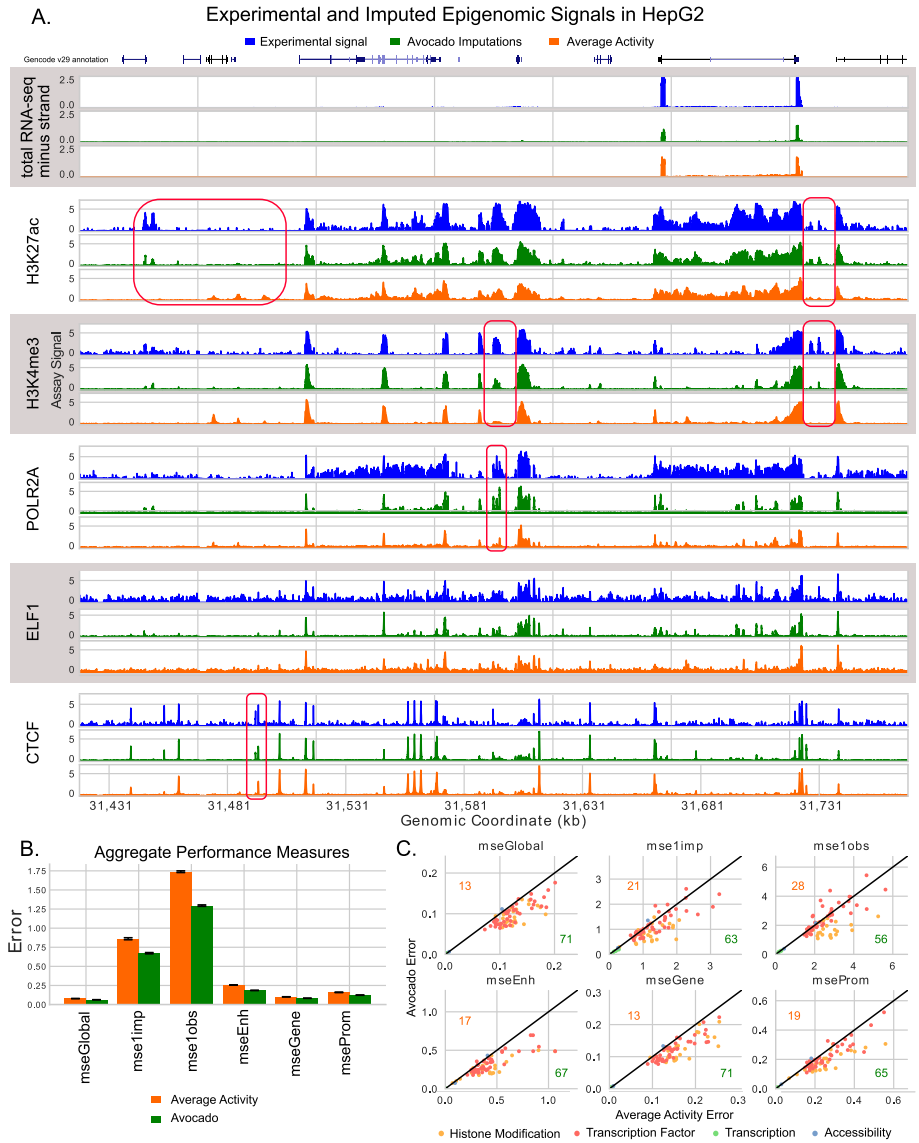

Fig. 2 Avocado imputes epigenomic experiments accurately. **a** Example signal, corresponding imputations, and the average activity of that assay, for six assays performed in HepG2. The figure includes representative tracks for RNA-seq, histone modification, and factor binding. The data covers 350 kbp of chromosome 20. **b** Performance measures evaluated in aggregate over all experiments from all biosamples in chromosomes 12 through 22. Orange bars show the performance of the average activity baseline, and green bars show the performance of Avocado's imputations. **c** Performance measures evaluated for each assay, with Avocado's error (*y*-axis) compared against the error of the average activity (*x*-axis). The number of assays in which Avocado outperforms the average activity is denoted in green for each metric, and the number of assays in which Avocado underperforms the average activity is denoted in orange

performance of most assays and small decreases in a few (Additional file 1: Figure S3B/C). These results indicate that the inclusion of other phenomena does, indeed, aid in the imputation of the original tracks.

Comparison to ENCODE-DREAM participants

Predicting the binding of various transcription factors is particularly important due both to these proteins' critical roles in regulating gene expression and the sparsity with which their binding has been experimentally characterized across different biosamples. For example, of the 43 transcription factors included in the ENCODE2018-Core dataset, only 9 have been performed in more than 10 biosamples. The most performed assay measures CTCF binding and has been performed 136 times, which is almost twice as high as the next most performed assay, measuring POLR2A binding, at 70 assays. In contrast, 13 of the 25 histone modifications in the ENCODE2018-Core

dataset have been measured in more than 10 biosamples, and the top six have all been performed in more than 200 biosamples. The sparsity of protein binding assays is exacerbated in the ENCODE2018-Sparse dataset, where additional 704 assays measuring protein binding have been performed in fewer than five biosamples.

A recent ENCODE-DREAM challenge focused on the prediction of transcription factor binding across biosamples and phrased the prediction task as one of classification where the aim is to predict whether binding is occurring at a given locus (https://www.synapse.org/#!Synapse:syn6131484). The challenge involved training machine learning models to predict signal peaks using nucleotide sequence, sequence properties, and measurements of gene expression and chromatin accessibility. The participants trained their models on a subset of chromosomes and biosamples, and were evaluated based on how well their models generalized both across chromosomes and in new biosamples. We acquired predicted probabilities of binding from the top four teams, Yuanfang Guan [13], dxquang [14], autosome.ru, and J-TEAM [15], for 13 tracks of epigenomic data. Four of the assays, E2F1, HNF4A, FOXA2, and NANOG, were excluded from the ENCODE2018-Core dataset because they had been performed in fewer than five biosamples. Consequently, Avocado could not make predictions for these four assays. Thus, we used only nine tracks for this evaluation.

We compared Avocado's predictions of transcription factor binding to the predictions of the top four models from the ENCODE-DREAM challenge to serve as an independent validation of Avocado's quality. We used both the average precision (AP) and the point on the precision-recall curve where precision and recall are equal (EPR) to evaluate the methods. In order to provide an upper limit for how good Avocado's predictions could be after the conversion process, we included as a baseline the experimental ChIP-seq data that the peaks were called from (called "Same Biosample"). Additionally, we compared against the average activity of that assay in Avocado's training set for that prediction. This baseline serves to show that Avocado is learning to make biosample-specific predictions. Further, when we investigated the training sets for the various experiments, we noted that there were two liver biosamples, male adult (age 32) and female child (age 4), that had similar assays performed in them. To ensure that Avocado was not simply memorizing the signal from one of these biosamples and predicting it for the other liver biosamples, we compare against the signal from the related biosample as well (denoted "Similar Biosample").

We observed that Avocado's predictions outperform all of the challenge participants in all tracks except for CTCF in iPSC and FOXA1 in the liver (Table 1, Additional file 1: Table S1). The most significant improvement comes in predicting REST, a transcriptional factor that represses neuronal genes in biosamples that are not neurons, and the highest overall performance is in predicting CTCF binding. This high performance is due in part to the large number of CTCF binding sites, but is likely also because CTCF binding is similar across most biosamples. Importantly, the REST assay for both liver biosamples was in the same fold, and TAF1 was only performed in one of the liver biosamples, so Avocado's good performance on those tracks is a strong indicator of its performance. Visually, we observe that some of the participants models appeared to overpredict signal values, suggesting that a source of error for these models is their lack of precision, corresponding to rapid drop in precision for predicting REST (Additional file 1: Figure S4). Interestingly, Avocado appears to underperform using the related liver biosample as the predictor for FOXA1, suggesting that perhaps the factors for FOXA1 are poorly trained. However, this result is further evidence that Avocado is not simply memorizing related signal. We also note that in the case of CTCF

Table 1 Comparison of methods on ENCODE-DREAM challenge test set

Biosample	iPSC	PC-3	Liver	Liver	Liver	Liver	Liver	Liver	Liver
Assay	CTCF	CTCF	EGR1	FOXA1	GABPA	JUND	MAX	REST	TAF1
Method									
Yuanfang Guan	0.729	0.600	0.397	0.282	0.353	0.533	0.441	0.319	0.281
dxquang	0.866	0.783	0.274	0.400	0.347	0.260	0.330	0.312	0.264
autosome.ru	0.778	0.486	0.331	0.243	0.342	0.416	0.384	0.264	0.221
J-TEAM	**0.812**	0.747	0.363	**0.462**	0.344	0.415	0.377	0.196	0.272
Avocado	0.723	**0.791**	**0.530**	0.354	**0.396**	**0.660**	**0.574**	**0.477**	**0.384**
Similar biosample	–	–	0.363	0.389	0.226	0.568	0.446	0.408	–
Same biosample	0.741	0.878	0.648	0.716	0.573	0.731	0.622	0.622	0.556
Average activity	0.574	0.735	0.240	0.299	0.253	0.223	0.349	0.124	0.140

The average precision (AP) computed across nine epigenomic experiments in the ENCODE-DREAM challenge test set in chromosome 21. For each track, the score for the best-performing predictive model is in boldface

in iPSCs, the ChIP-seq signal from iPSC appears to underperform two challenge participants, suggesting that the conversion process may limit Avocado's performance.

We did our best to ensure a fair comparison between Avocado and the challenge participants, but the comparison is necessarily imperfect, for several reasons. Two factors make the comparison easier for Avocado. First, Avocado is exposed to many epigenomic measurements that the challenge participants did not have available, including measurements of the same transcription factor in other cell types. Second, as an imputation approach, Avocado is trained on the same genomic loci that it makes predictions for, whereas the challenge participants had to make predictions for held-out chromosomes. On the other hand, three factors skew the comparison in favor of the challenge participants. First, unlike the challenge participants, Avocado was not directly exposed to any aspect of nucleotide sequence or motif presence. Second, Avocado makes predictions at 25 bp resolution in hg38, whereas the challenge was conducted at 200 bp resolution in hg19. We were able to use liftOver to convert between assemblies, followed by aggregating the signal from 25 bp resolution to 200 bp resolution, but both steps blurred the signal. Third, Avocado is trained to predict signal values directly, whereas the challenge participants are trained on the classification task of identifying whether a position is a peak. Evaluation is done in a classification setting. In particular, Avocado is penalized for accurately predicting high signal values in regions that are not labeled as peaks, exemplifying the discordance between the regression and classification settings. For all these reasons, Avocado would not have been a valid submission to the challenge. Finally, it is perhaps worth emphasizing that whereas the challenge was truly blind, our application of Avocado to the challenge data is only blind "by construction." We emphasize that we did not adjust Avocado's model or hyperparameters based on looking at the challenge results: the comparison presented here is based entirely on a pre-trained Avocado model.

We investigated the effect that these differences may have had on predictive performance. First, we evaluated the performance of Avocado and the challenge participants at predicting the test set challenge tracks on chromosome 17, whose loci were used for training the challenge models. This evaluation resulted in similar trends as in Table 1 (Additional file 3) and suggests that the loci used for evaluation are not a significant factor for Avocado's improved performance over the challenge participants. Next, we removed from Avocado's training set all experiments from biosamples which appeared in the challenge test set, except for those experiments that the challenge participants had—namely, DNase-seq and RNA-seq experiments. This restricted Avocado to only being able to make predictions on the challenge tracks using the same epigenomic information that the participants had. In this setting, we observed poor performance of Avocado on the liver test set tracks, but even better performance on the CTCF tracks in iPSC and PC-3 than the original Avocado model. However, as described in Additional file 3, it was difficult to ensure a fair comparison on the biosamples noted as being from the liver, and these reasons may potentially explain the poor results. Finally, we trained an Avocado model using only DNase-seq and RNA-seq from the biosamples used in the challenge, as well as the transcription factor binding tracks available in the training set. Again, performance on liver biosamples was poor. While performance also degraded on the CTCF tracks, it was still competitive with the top four participants. These results indicate that a source of Avocado's power is leveraging the diverse data in the massive ENCODE compendium.

Extending avocado to more biosamples and assays

Adding new assays

Despite including 3814 epigenomic experiments, the ENCODE2018-Core dataset does not contain all biosamples or assays that are represented in the ENCODE compendium. Specifically, the dataset does not include 667 biosamples where fewer than five assays had been performed, and it does not include 1281 assays that had been performed in fewer than five biosamples. The missing biosamples primarily include time courses, genetic modifications, and treatments of canonical biosamples, such as HepG2 genetically modified using RNAi. However, several primary cell lines and tissues, such as amniotic stem cells, adipocytes, and pulmonary artery, were also not included in the ENCODE2018-Core dataset due to lack of sufficient data. The majority of the missing assays corresponded to transcription measurements after gene knockdowns/knockouts (shRNA and CRISPR assays) or to binding measurements of eGFP fusion proteins. Yet, some transcription factors, such as NANOG, FOXA2, and HNF4A, were excluded as well. We collect these experiments into a separate dataset, called ENCODE2018-Sparse (see the "Methods" section).

We constructed the ENCODE2018-Sparse dataset to attempt to address some of the problems of missingness in ENCODE2018-Core. This sparse version of the data has 99.7% missing entries, in comparson to 88.6% missing in ENCODE2018-Core. Within ENCODE2018-Sparse, we identified four main groups of biosamples: (1) 417 biosamples that only had DNase-seq performed on them, with 58 additional biosamples that had DNase and one or more other assays performed in it; (2) 112 biosamples that had various measurements of transcription performed in them; (3) 7 biosamples that were well characterized by at least 50 sparsely performed assays of transcription factor binding; and (4) biosamples derived from HepG2 and

K562 that were well characterized by various knockouts (Additional file 1: Figure S1).

In general, handling sparsely characterized assays or biosamples in a model like Avocado is challenging. Hence, we designed a three-step process that we hypothesized would allow us to make accurate imputations for additions with few corresponding tracks (Additional file 1: Figure S5). This approach is conceptually similar to our main approach for training Avocado. First, we trained the Avocado model on all 3814 experiments in ENCODE2018-Core. Second, we froze all of the weights in the model, including both the neural network weights and all five of the latent factor matrices. Third, we fit the new biosample or assay factors to the model using only the experimental signal derived from the ENCODE pilot regions. This resulted in a model whose only difference was the inclusion of a set of trained assay or biosample factors that were not present in original model. This training strategy has the benefit of allowing for quick addition of biosamples or assays to the pre-trained model, without requiring retraining of any of the existing model parameters.

In order to test the effectiveness of this approach, we extended Avocado to include assays that were in the ENCODE-DREAM challenge but not in the ENCODE2018-Core dataset. For the four assays that we did not compare against (HNF4A and FOXA2 in the liver, NANOG in iPSC, and E2F1 in K562), all but E2F1 had been performed in a biosample other than the one included in the challenge. Accordingly, we fit these three new assay factors using the procedure above. This fitting was done using HNF4A and FOXA2 from HepG2, and NANOG from h1-hESC. We then used the new assay factors, coupled with the pre-trained network, genome factors, and relevant biosample factors, to impute three remaining tracks in the challenge.

We observed that Avocado's imputed tracks for HNF4A and FOXA2 in the liver were of high quality and outperformed several baselines (Fig. 3). Most notably, both of these tracks outperformed all four challenge participants in their respective settings according to both EPR and AP. Second, both Avocado tracks outperformed simply using the track that they were trained on as the predictor, indicating that the model is leveraging the pre-trained biosample latent factors to predict biosample-specific signal.

However, we also observed that Avocado's imputations for NANOG in iPSCs are of particularly poor quality. Avocado's predictions underperform all four challenge participants. Notably, Avocado also underperforms using the signal from h1-hESC that it was trained on as the predictor. One potential reason for this poor performance is that relevant features of the NANOG binding sites are not encoded in the genomic latent factors. Alternatively, given that Avocado also underperformed the challenge participants at predicting CTCF in iPSC, it may be that the iPSC latent factors are not well trained, leading to poor performance in predictions of any track.

Adding new biosamples

We then tested the ability of the three-step process in Additional file 1: Figure S5 to make accurate predictions for biosamples that the model was not originally trained on. To do so, we began by training biosample factors for 475 biosamples not in the ENCODE2018-Core dataset that had DNase-seq performed in them. We then evaluated Avocado's ability to predict other assays that were performed in these biosamples. A large number of these biosamples had only DNase-seq performed in them, so we also evaluated Avocado's ability to predict DNase-seq as well. We reasoned that because the biosample factors were trained using the ENCODE pilot regions, but the predictions were evaluated in chromosome 20 without retraining the corresponding genomic latent factors, this would be a fair evaluation.

We observed good performance of the imputations for these biosamples. Visually, we noticed the same concordance between the imputed and the experimental signal,

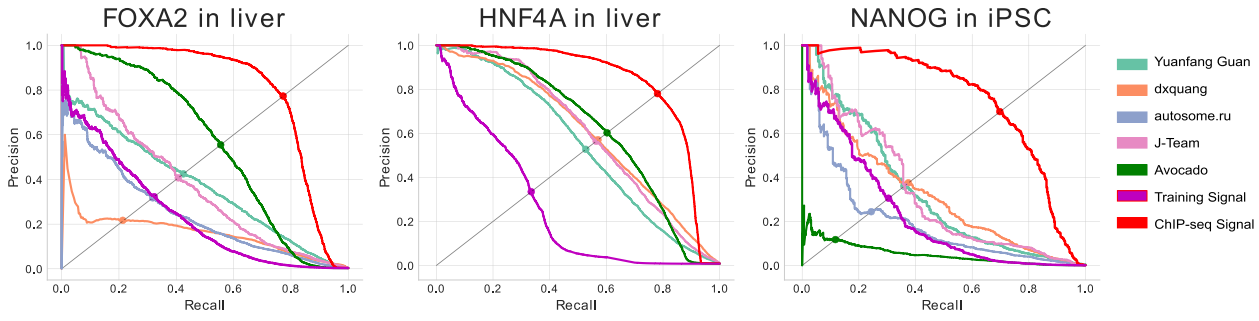

Fig. 3 Avocado's performance when adding new transcription factors to a pre-trained model. Precision-recall curves for three transcription factors that were added to a pre-trained model using a single track of data each from the ENCODE2018-Sparse dataset. Similar to the previous comparisons against the ENCODE-DREAM participants, the evaluation was performed in chromosome 21

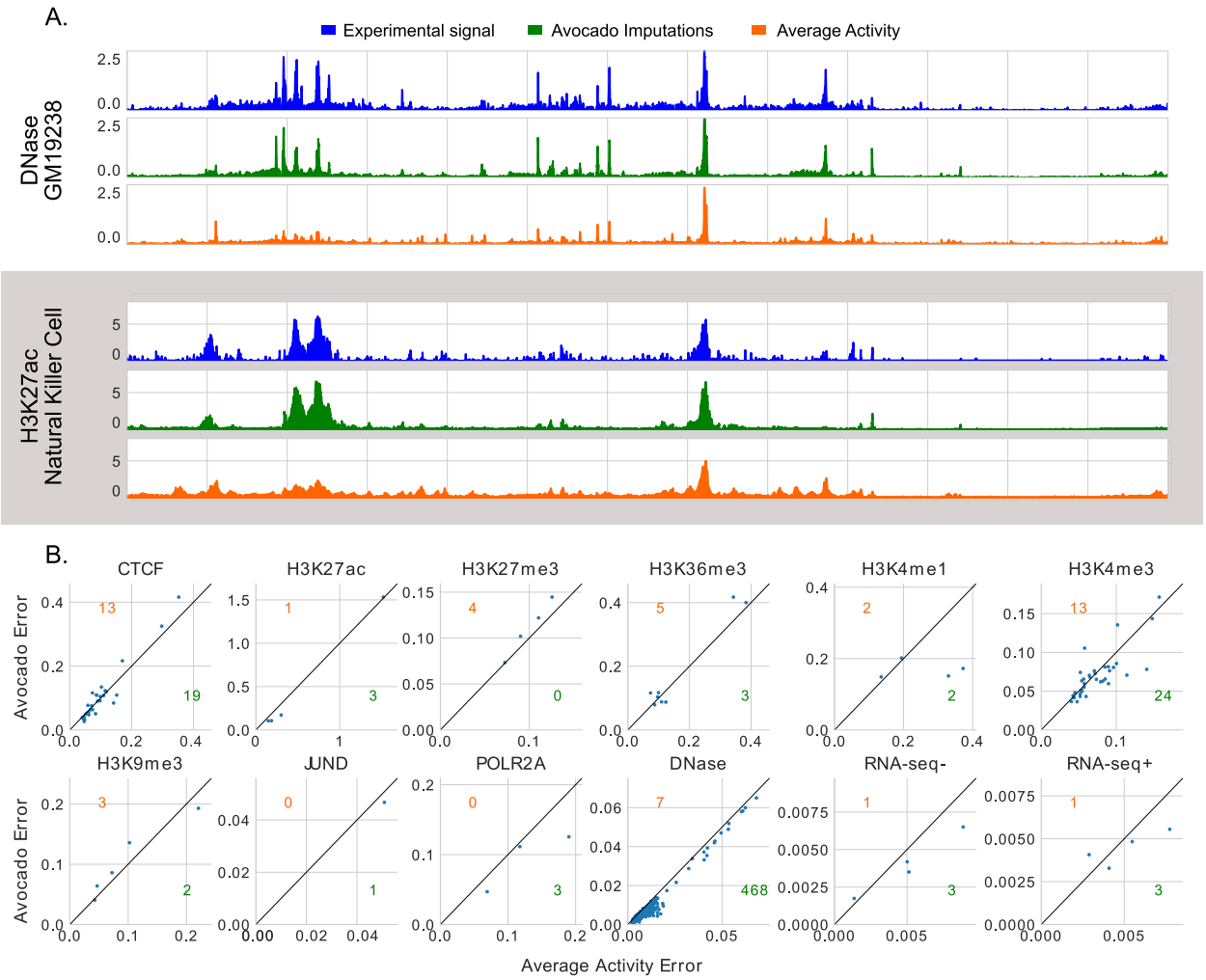

Fig. 4 Imputations and performance when adding biosamples to a pre-trained model. **a** Imputations for two tracks of data in the ENCODE2018-Sparse dataset on chromosome 20 after fitting the biosample factors using only DNase-seq signal from the ENCODE pilot regions. **b** Performance of Avocado at imputing tracks on chromosome 20 after fitting the biosample factors using only DNase-seq signal from the ENCODE pilot regions

and we observed that biosample-specific elements are being captured (Fig. 4a). We then evaluated the performance of Avocado on the mseGlobal metric compared to the average activity baseline for each assay. We observed that Avocado appears to produce high-quality predictions for several assays, including CTCF, H3K27ac, and POLR2A (Fig. 4b). However, for other assays, such as H3K9me3 and H3K36me3, the average activity dominates. It is possible that this phenomenon speaks to the ability of DNase to recover these other approaches. Overall, we observe a decrease in error from 0.027 when using the average activity to 0.024 when using the imputations from Avocado.

While these evaluations have thus far used only DNase-seq to fit new biosamples to the model, it is not necessarily the case that performing a single assay is sufficient to optimally fit new biosamples to a model. Unfortunately, it would be computationally expensive to identify the combination of assays that yielded optimal performance. To investigate whether there was a general trend that biosamples fit with more assays performed better than those fit with fewer assays, we partitioned the 3814 experiments from the fivefold cross-validation on ENCODE2018-Core by the number of assays performed in the biosample of the experiment (Additional file 1: Figure S6). When we plotted the average error of each type of activity, we did not observe a noticable trend between the number of assays used to fit biosample factors and performance at imputing experiments. This evaluation is limited by not considering the composition of experiments used to fit each biosample or by considering experiments that had fewer than four assays performed in the respective biosample. However, these results do not suggest that simply performing more experiments will yield better performance

Table 2 Comparison of approaches for extending Avocado to new cell types and assays

Test set	Retrain from scratch	Fine-tune	Freeze	Fivefold
ENCODE2018-Sparse	0.058	0.056	0.091	–
ENCODE2018-Sparse (w/o shRNA)	0.069	0.068	0.080	–
ENCODE2018-Core	0.049	0.050	0.051	0.050

The MSE of three approaches for adding new biosamples or assay on three test sets. The test sets are half of the experiments in the ENCODE2018-Sparse dataset, that same dataset with shRNA experiments removed, and the fifth fold from the ENCODE2018-Core dataset. The MSE from the fivefold cross-validation is shown for the ENCODE2018-Core test fold as a reference

overall, or that biosamples with many experiments performed in them will necessarily have better imputations than those that are more sparsely assayed.

Evaluating alternate training methods

Our strategy for incorporating new biosamples and assays into a pre-trained Avocado model involves first freezing almost all of the parameters of the model. In practice, large consortia and other providers of imputations are likely to be interested in this approach because it would allow for continuous incorporation of new biosamples and assays without affecting the imputations that have already been released. However, it is unlikely that keeping these parameters frozen during training would yield performance as high as updating them using the new data, because the new experiments may point to interesting loci, novel forms of activity, or important cell type-specific signatures that are not captured in the frozen parameters. To test this hypothesis, we compared the performance of our strategy for incorporating new biosamples and assays to two alternate approaches: retraining Avocado from scratch and fine-tuning a pre-trained Avocado model (see the "Incorporating new experiments" section).

We first simulated the setting where one has trained a model on a set of "original" experiments and would now like to extend the model to include assays and biosamples contained in a set of "additional" experiments. We used four of the five folds used in the "Avocado's imputations are accurate and biosample specific" section from the ENCODE2018-Core dataset as the original experiments, and half of the experiments from ENCODE2018-Sparse, after filtering, as the set of additional experiments. This filtering step consisted of removing all experiments where the assay or biosample had only been performed once. We created two separate test sets: the first was the second half of the experiments in the ENCODE2018-Sparse dataset, and the second was the fifth fold from the ENCODE2018-Core dataset to use as validation that the models were still performing well on the original data. The experiments from ENCODE2018-Sparse were split such that, for each assay, the biosamples were evenly partitioned into the training and test sets.

Overall, we found that our strategy of freezing parameters underperformed both retraining Avocado and fine-tuning a pre-trained model (Table 2). In particular, we observed a large difference in performance between the freezing strategy and the other two strategies on the ENCODE2018-Sparse test set. However, upon inspecting the errors more closely, we observed that the majority of errors on the second half of the Sparse datasets come from short-hairpin RNA-seq (shRNA) experiments (Additional file 1: Figure S7). These experiments involve knocking out a target gene using RNA interference and are not present at all in the ENCODE2018-Core dataset. Thus, it makes sense that a model that had been trained on ENCODE2018-Core and then had most of its parameters frozen would perform poorly at imputing shRNA experiments, because neither the genome factors nor the neural network was trained using this type of activity. When we remove these experiments from the ENCODE2018-Sparse test set, we find that the gap in performance between the different methods diminishes significantly. Further, we observe that the error of the frozen model decreases, whereas the errors of the other models increases, confirming that models that were able to train on this type of activity could capture it well. We then validated the resulting models by checking their performance on the fifth fold of the ENCODE2018-Core dataset that had been held out. We observed that the models all performed similarly both to each other and to the split in the original fivefold cross-validation when the fold used here as the test set was held out.

Discussion

To our knowledge, we report here the largest imputation of epigenomic data that has been performed to date. We applied the Avocado deep tensor factorization model to 3814 epigenomic experiments in the ENCODE2018-Core dataset. The resulting imputations cover a diverse set of biological activity and cellular contexts and are publicly available at http://www.encodeproject.org. Due to the cost of experimentation and the increasing sparsity of epigenomic compendia, we anticipate that imputations of this scale will serve as a valuable community resource for characterizing the human epigenome.

We used multiple independent lines of reasoning to confirm that Avocado's imputations are both accurate and biosample specific. First, we compared each imputed data

track to the average activity of that assay and found that for almost all assays, Avocado's imputations were more accurate. A current weakness in Avocado's imputations is imputing transcription, likely due to the sparse, exon-level activity of these assays along the genome. Second, we compared imputations of transcription factor binding tracks to the predictions made by the top four models in the recent ENCODE-DREAM challenge. In almost all cases, the Avocado imputations were significantly more accurate than the imputations produced by the challenge participants. Notably, Avocado is not exposed to nucleotide sequence at all during the training process, and so its ability to correctly impute transcription factor binding is based entirely on local epigenomic context, rather than binding motifs.

Ongoing characterization efforts regularly identify new biosamples of interest and develop assays to measure previously uncharacterized phenomena. These efforts aid in understanding the complexities of the human genome but pose a problem for imputation efforts that must be trained in a batch fashion. Given that it took almost a day to fit the Avocado genomic latent factors for even the smallest chromosome, retraining the model for each inclusion is not feasible. We demonstrated that by leveraging parameters that had been pre-trained on the ENCODE2018-Core dataset, new assays and biosamples could be quickly added to the existing Avocado model. In contrast, extending imputations to cover a single new assay using ChromImpute would require training a new model for each of the 400 biosamples in the ENCODE2018-Core dataset, or each of the >1000 biosamples in the combined ENCODE2018-Core/ENCODE2018-Sparse dataset. Our observations suggest that not only is the Avocado approach computationally efficient, with three new assays taking only a few minutes to add to the model, but that the resulting imputations are highly accurate.

One potential reason that this pre-training strategy works well is that the genomic latent factors efficiently encode information about regions of biological activity. For example, rather than memorizing the specific assays that exhibit activity at each locus, the latent factors may be organizing general features of the biochemical activity at that locus. We have previously demonstrated the utility of Avocado's latent genomic representation for several predictive tasks [10]. Investigating the utility and meaning of the latent factors from this improved Avocado model is an ongoing work.

Notably, however, the encoding of relevant information in the latent factors may lead to a potential weakness in Avocado's ability to generalize to novel biosamples or assays. Specifically, if the signal in a novel biosample or assay is not predictable from the tracks that were used to train the initial genomic latent factors, then it is unlikely that Avocado will make good imputations for the new data. For example, if a transcription factor is dissimilar to any factors in the training set, then the genomic latent factors may not have captured features relevant to the novel factor. This may explain why Avocado fails to generalize well to NANOG.

A strength of large consortia, such as ENCODE, is that they are able to collect massive amounts of experimental data. This amount of data is only possible because many labs collect it over the course of several years. Inevitably, this results in some data that is of poor quality. While quality control measures can usually identify data that is of very poor quality, they are not perfect, and the decision of what to do with such data can be challenging. Unfortunately, data of poor quality poses a dual challenge for any large-scale imputation approach. When an imputation approach is trained on low-quality data, then the resulting imputations may be distorted by the noise. Furthermore, when the approach is evaluated against data that is of poor quality, imputations that are of good quality may be incorrectly scored poorly. Thus, when dealing with large and historic data sources, it is important to ensure the quality of the data being used.

Conclusion

In this work, we describe the training of an imputation approach that can predict a variety of epigenomic phenomena, including histone modification, protein binding, transcription, and chromatin accessibility, across hundreds of human biosamples. The resulting model is capable of imputing 33,600 genome-wide epigenomic experiments, representing the largest imputation effort performed to date both in terms of the number of tracks imputed and in terms of biological phenomena that are jointly modeled. We found that these tracks were of high quality, with a 19.5% decrease in overall error when compared to the strong average activity baseline. Empirically, the imputations of transcription factor binding significantly outperformed the top participants in a recent ENCODE-DREAM transcription factor binding challenge, further indicating their quality.

We anticipate that this work will be impactful in several ways. The simplest application of these imputations is to enable analyses or prediction in biosamples where the required epigenomic experiments have not yet been performed. Another approach is to look for inconsistencies between the imputed and primary data for experiments that have been performed, with the anticipation that these regions may prove biologically interesting. Further, one could use imputed tracks where there is no corresponding experimental data to determine what experiments should be performed next, priotizing imputed tracks that appear to encode interesting phenomena.

The imputation approach offered by Avocado has great potential to be extended both to precision medicine and to single-cell datasets. In the precision medicine setting, a biosample is sparsely assayed in a variety of individuals, and the goal is to correctly impute the inter-individual variation, particularly in regions associated with disease. We anticipate that Avocado could be either applied directly in this setting, with biosamples including the annotation of the individual they came from, or extended to accommodate a 4D data tensor, where the fourth dimension corresponds to distinct individuals. In the single-cell setting, the biosample axis would be replaced with a cell axis where each entry would correspond to a single cell. This approach could potentially be used as a computational co-assay, leveraging a shared genomic axis to impute multiple types of experiments in each individual cell.

Methods

Avocado

Avocado topology

Avocado is a multi-scale deep tensor factorization model. The tensor factorization component comprised five matrices of latent factors that encode the biosample, assay, and three resolutons of genomic factors at 25 bp, 250 bp, and 5 kbp resolution. Having multiple resolutions of genomic factors means that adjacent positions along the genome may share the same 250 bp and 5 kbp resolution factors. We used the same model architecture as in the original Avocado model [10], with 32 factors per biosample, 256 factors per assay, 25 factors per 25-bp genomic position, 40 factors per 250-bp genomic position, and 45 factors per 5-kbp genomic position. The neural network model has two hidden dense layers that each have 2048 neurons, before the regression output, for a total of three weight matrices to be learned jointly with the matrices of latent factors. The network uses ReLU activation functions, $\mathrm{ReLU}(x) = \max(0, x)$, on the hidden layers, but no activation function on the prediction.

Avocado training

Avocado is trained in a similar fashion to our previous work [10]. This procedure involves two steps, because the genome is large and the full set of genomic latent factors cannot fit in memory. The first is to jointly train all parameters of the model on the ENCODE pilot regions, which comprise roughly 1% of the genome. After training is complete, the neural network weights, the assay factors, and the biosamples are all frozen. The second step is to train only the three matrices of latent factors that make up the genomic factors on each chromosome individually. In this manner, we can train comparable latent factors across each chromosome without the need to keep then all in memory at the same time.

Avocado was trained in a standard fashion for neural network optimization. All initial model parameters and optimizer hyperparameters were set to the defaults in Keras. In this work, Avocado was trained using the Adam optimizer [16] for 8000 epochs with a batch size of 40,000. This is longer than our original work, where the model was trained for 800 epochs initially and 200 epochs on the subsequent transfer learning step. Empirical results suggest that this longer training process is required to reach convergence, potentially because of the large diversity of signals in the ENCODE2018-Core dataset. When adding in additional biosample or assay factors, due to the small number of trainable parameters, the model was trained for only 10 epochs with a batch size of 512. Due to the large dataset size, one epoch is defined as one pass over the genomic axis, randomly selecting experiments at each position, rather than one full pass over every experiment.

The model was implemented using Keras (https://keras.io) with the Theano backend [17], and experiments were run using GTX 1080 and GTX 2080 GPUs. For further background on neural network models, we recommend the comprehensive review by Schmidhuber [18].

Data and evaluation

ENCODE dataset

We downloaded 6870 genome-wide tracks of epigenomic data from the ENCODE project (https://www.encodeproject.org). These experiments were all processed using the ENCODE processing pipeline and mapped to human genome assembly hg38, except for the ATAC-seq tracks, which were processed using an approach that would later be added to the ENCODE processing pipeline. The values are signal p value for ChIP-seq data and ATAC-seq, read-depth normalized signal for DNase-seq, and plus/minus strand signal for RNA-seq. When multiple replicates were present, we preferentially chose the pooled replicate; otherwise, we chose the second replicate. The experimental signal tracks were then further processed before being used for model training and evaluation. First, the signal was downsampled to 25 bp resolution by taking the average signal in each 25 bp bin. Second, an inverse hyperbolic sin transformation was applied to the data. This transformation has been used previously to reduce the effect of outliers in epigenomic signal [9, 19].

We divided these experiments into two datasets, the ENCODE2018-Core dataset and the ENCODE2018-Sparse dataset. The ENCODE2018-Core dataset contains 3814 experiments from all 84 assays that have been performed in at least five biosamples, and all 400 biosamples that have been characterized by at least five assays. Hence, $\sim$88.6% of the data in the ENCODE2018-Core data matrix is missing. The ENCODE2018-Sparse dataset contains 3056 experiments, including 1281 assays that have been performed in fewer than five biosamples and

667 biosamples that have been characterized by fewer than five assays, yielding a matrix that is ∼ 99.7% missing.

We adopted a similar strategy to Durham et al. for partitioning these experiments into folds for cross-validation. Specifically, we partitioned entire genome-wide experiments into five folds such that a model would be trained on all genomic loci and then evaluated on its ability to predict entirely held-out experiments, because this is the most realistic evaluation setting. However, randomly assigning tracks to each fold may inadvertently leave some folds without seeing some assays or some biosamples, meaning that the model would not learn anything for those embeddings and thus perform poorly on imputation. Unfortunately, even after only keeping experiments from assays and biosamples where five experiments had been performed, it is not always possible to partition a set of experiments into folds such that each assay and biosample are seen. Thus, Durham et al. adopted a simple optimization approach that randomly assigned experiments to partitions and evaluated each partition by the total number of biosamples and assays covered by each partition. We empirically found that this approach underperformed a simple greedy approach that uses a counter to sequentially assign folds to random experiments within biosamples, one biosample at a time, preserving the location in the cycle from one biosample to the next.

ENCODE-DREAM challenge datasets

For our comparisons with the ENCODE-DREAM challenge participants, we acquired from the challenge organizers both genome-wide model predictions from the top four participants and the binary labels (https://www.synapse.org/#!Synapse:syn6112317). The predictions and labels were defined at 200 bp resolution, with a stride of 50 bp, meaning that each 50 bp bin was included in four adjacent bins. The labels corresponded to conservative thresholded irreproducible discovery rate (IDR) peaks called from multiple replicates of ChIP-seq signal.

Comparison to ENCODE-DREAM predictions

Avocado's predictions had to be processed in several ways to make them comparable with the data format for the challenge. First, because Avocado's predictions are in hg38 and the challenge was performed in hg19, the UCSC liftOver command (https://genome.ucsc.edu/cgi-bin/hgLiftOver) was used to convert the coordinates across reference genomes. Unfortunately, many of the 25-bp bins in hg38 mapped to the middle of bins in hg19, blurring the signal. Further, ∼ 27% of positions on chromosome 21 of hg38 could not be mapped to positions in hg19, so those positions were discarded from the analysis. Lastly, because the challenge was performed at 200 bp resolution, the average prediction in the 200-bp region was used as Avocado's predictions for that bin. We then filtered out all regions that were marked as "ambiguous"' by the challenge organizers. These regions included both the flanks of true peaks as well as regions that were considered peaks in some, but not all, replicates.

The evaluation of each model was performed using both the average precision, which roughly corresponds to the area under a precision-recall curve, and the point along the precision-recall curve of equal precision and recall (EPR). The EPR corresponds to setting the decision threshold so that the number of positive predictions made by the model is equal to the number of positive labels in the dataset. This is also called the "break-even point." A strength of EPR, in comparison to taking the recall at a fixed precision, is that it accounts for the true sparsity in the label set. For example, if it is known beforehand that an experimental track generally has between 100 and 200 peaks across the entire genome, then a reasonable user may use the top 150 predictions from a model. However, if an experimental track had between 10,000 and 20,000 peaks, then a user may use the top 15,000 predicted peaks.

Calculation of average activity

In several of our experiments, we compared model performance against the average activity of an assay. In all instances involving the ENCODE2018-Core dataset, "average activity" refers to the average signal value at each locus across all biosamples in the training set for that particular experiment. Because the predictions across the entire ENCODE2018-Core dataset are made using five-fold cross-validation, the training set differs for tracks from different folds. This approach ensures that the track being predicted is not included in the calculation of average activity which would make the baseline unfair. In instances involving the ENCODE2018-Sparse dataset, "average activity" refers to the average activity across all tracks of that assay that were present in the entire ENCODE2018-Core dataset.

Incorporating new experiments

We evaluated the performance of three approaches for handling the incorporation of additional biosamples or assays into a model: retraining the model from scratch, fine-tuning the parameters of a pre-trained model, and freezing most parameters of a pre-trained model and training the remaining subset. These approaches were evaluated using four of the five folds from the ENCODE2018-Core dataset as the set of "original" experiments and half of the experiments in the ENCODE2018-Sparse dataset as the "additional" experiments. When training Avocado from scratch, the model was trained on both the original and additional experiments for a total of 8000 epochs, just like our normal training approach. When fine-tuning a pre-trained model, we first created a pre-trained model by training Avocado for 6000 epochs

on just the original experiments and then training on both the original and the additional experiments for additional 2000 epochs. This ensured that differences in performance between the retrained model and the fine-tuned model did not arise simply due to a different number of epochs of training. Lastly, we trained Avocado for 8000 epochs on just the original experiments, froze the neural network and genomic position parameters, and proceeded with training the assay and biosample factors using only the additional experiments for 100 epochs.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-020-01978-5.

Additional file 1: Supplemental figures and tables.

Additional file 2: Underperforming imputations. Follow-up analysis of those tracks whose imputations underperform the average activity baseline.

Additional file 3: Further analyses of ENCODE challenge results. Follow-up analyses showing the performance of Avocado on the ENCODE TF Binding challenge when trained using different subsets of experiments to investigate the source of its strong performance.

Additional file 4: The review history.

Peer review information
Anahita Bishop was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Acknowledgements
We would like to thank Giancarlo Bonora, Timothy Durham, Ritambhara Singh, and Gürkan Yardımcı for many productive discussions, as well as Anshul Kundaje and Akshay Balsubramani for providing data from the ENCODE-DREAM challenge.

Review history
The review history is available as Additional file 4.

Authors' contributions
WN and JS designed the experiments. JS acquired the data and performed the experiments. WN and JS wrote the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported in part by an NSF IGERT grant (DGE-1258485) and by NIH awards U24 HG009446 and U01 HG009395.

Availability of data and materials
The ENCODE2018-Core and ENCODE2018-Sparse datasets, as well as the resulting model, can be found at https://noble.gs.washington.edu/proj/avocado. The authors place no restrictions on the download and use of our generated datasets or model. The imputated genome-wide tracks from our model can be found on the ENCODE portal (https://www.encodeproject.org) under the accession ENCSR481OSA.
The code used to train and use Avocado can be found at https://www.github.com/jmschrei/avocado [20] under an Apache v2 license. Release v0.1.0 corresponds to the submission of this manuscript and can be found at https://github.com/jmschrei/avocado/releases/tag/v0.1.0 or on Zenodo at https://zenodo.org/record/3549064 [21]. The code is written in Python and is platform independent.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, USA. [2]Department of Electrical Engineering, University of Washington, Seattle, USA. [3]Department of Genome Sciences, University of Washington, Seattle, USA.

Received: 28 May 2019 Accepted: 26 February 2020
Published online: 30 March 2020

References

1. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) project. Science. 2004;306(5696):636–40.
2. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317–30.
3. The modENCODE Consortium. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science. 2010;330:1775–87.
4. Bujold D, Morais DA, Gauthier C, Cote C, Caron M, Kwan T, Chen KC, Laperle J, Markovits AN, Pastinen T, Caron B, Veilleux A, Jacques PE, Bourque G. The international human epigenome consortium data portal. Cell Syst. 2016;3:496–9.
5. Yue F, The Mouse ENCODE Consortium. A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014;515:355–64.
6. Akbarian S, et al. The PsychENCODE project. Nat Neurosci. 2015;18:1707–12.
7. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. 2017;550:204–13.
8. Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat Biotechnol. 2015;33(4):364–76.
9. Durham TJ, Libbrecht MW, Howbert JJ, Bilmes JA, Noble WS. PREDICTD: PaRallel Epigenomics Data Imputation with Cloud-based Tensor Decomposition. Nat Commun. 2018;9:1402.
10. Schreiber JM, Durham TJ, Bilmes J, Noble WS. 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.
11. Lai X, Stigliani A, Vachon G, Carles C, Smaczniak C, Zubieta C, Kaufmann K, Parcy F. Building transcription factor binding site models to understand gene regulation in plants. Mol Plant. 2019;12:743–763.
12. Schreiber JM, Singh R, Bilmes J, Noble WS. A pitfall for machine learning methods aiming to predict across cell types. bioRxiv. 2019. https://www.biorxiv.org/content/10.1101/512434v1.
13. Li H, Quang D, Guan Y. Anchor: trans-cell type prediction of transcription factor binding sites. Genome Res. 2019;29(2):281–92.
14. Quang D, Xie X. FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. Methods. 2019;166:40–47.
15. Keilwagen J, Posch S, Grau J. Accurate prediction of cell type-specific transcription factor binding. Genome Biol. 2019;20(9):. https://doi.org/10.1186/s13059-018-1614-y.
16. Kingma D, Ba J. Adam: a method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations; 2015. https://iclr.cc/archive/www/doku.php%3Fid=iclr2015:accepted-main.html.
17. Theano Development Team. Theano: a Python framework for fast computation of mathematical expressions. arXiv e-prints, abs/1605.02688. 2016.
18. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 61:2015.
19. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods. 2012;9(5):473–6.
20. Schreiber JM. Avocado. GitHub. https://github.com/jmschrei/avocado.
21. Schreiber JM, Durham TJ, Bilmes J, Noble WS. Avocado source code. Zenodo. 2019. https://doi.org/10.5281/zenodo.3549064.

Publisher's Note