



## 36 1 Introduction

37 Since their development, high-throughput chromatin profiling assays such as histone ChIP-  
38 seq, DNase-seq and ATAC-seq have proven crucial for deciphering gene regulatory elements  
39 and characterizing their dynamic activity states across cell types and tissues (together re-  
40 ferred to as “cell types” for the rest of this work). Because each assay makes cell type-specific  
41 measurements, these assays must be performed for each cell type of interest separately. How-  
42 ever, comprehensively profiling a large collection of cell types with assays targeting diverse  
43 attributes of chromatin is prohibitive due to practical constraints on material, cost and per-  
44 sonnel. Hence, even the largest repositories of epigenomic and transcriptomic data are still  
45 incomplete in the sense that they are missing tens of thousands of potential experiments  
46 [1, 2, 3, 4, 5, 6].

47 To address this challenge, predictive models for imputing missing datasets have been  
48 proposed as an inexpensive and straightforward way to obtain complete draft epigenomes [7,  
49 8, 9, 10, 11]. These models leverage the complex correlation structure of signal profiles from  
50 available experiments to impute signal for experiments that have not yet been performed.  
51 Recently, imputation models have been scaled to impute tens of thousands of experiments  
52 [12, 13] spanning dozens of assays in hundreds of human cell types. Although progress has  
53 clearly been made in developing imputation approaches, the field has thus far only explored  
54 a small portion of the space of potential imputation models. Notably, only one of the five  
55 methods surveyed above uses nucleotide sequence as input when making imputations.

56 We organized the ENCODE Imputation Challenge to encourage active development of  
57 imputation models. The challenge consisted of two stages and participants were encouraged  
58 to share ideas and reorganize into new teams between stages. In the first stage, participants  
59 were ranked based on their ability to impute a fixed validation set consisting of experiments  
60 randomly selected from within our data matrix. The second stage also measured imputation  
61 performance on a held-out set, but with two crucial differences from the first stage: first,  
62 the test data was collected during the challenge to ensure a truly prospective evaluation,  
63 and second, the test data was collected almost exclusively for poorly characterized cell types  
64 (only three of the 12 cell types in the test set have more than two training experiments).

65 Our initial expectation was that this challenge would primarily serve as an analysis  
66 of the components of imputation models and, ultimately, identify those that worked well.  
67 However, we found that fairly evaluating the imputations in the second stage was much more  
68 challenging than expected, and so the challenge instead served as an impetus to describe, and  
69 correct, distributional shifts in large collections of genomics data sets. Specifically, we found  
70 that a distributional shift occurs between the more recently collected paired-end data and the  
71 older single-end data available on the ENCODE portal due to small processing differences  
72 that have a big effect. Without correcting for this difference, we found that a baseline method  
73 outperformed all but two of the submissions using the performance measures defined before  
74 the challenge began, and those two submissions only performed marginally better than  
75 the baseline. After correction, more than half of the participants outperformed the same  
76 baseline.

77 We identified three key challenges in fairly evaluating imputation methods. First, dif-  
78 ferences over time in experimental procedure or data processing create distributional shifts  
79 across experiments which must be corrected for ensure a fair evaluation, and this correction  
80 must be more than a simple rescaling of the signal. This concern is particularly important  
81 when dealing with data sources, like the ENCODE Portal, that contain data collected over  
82 long periods of time. Second, while epigenomic imputation is most useful for cell types with

83 few experiments, previous imputation work was evaluated using k-fold or leave-one-out cross-  
84 validation applied to an entire compendium. These evaluation settings over-emphasized the  
85 performance on well-characterized cell types and, unfortunately, good performance on well-  
86 characterized cell types is not always an indicator of performance on poorly characterized  
87 ones. Third, although designing several performance measures is necessary to capture the  
88 many aspects of a high-quality experimental readout, designing these measures without ac-  
89 counting for the first two issues can introduce redundancy in the measures, limiting their  
90 usefulness. We anticipate that giving proper consideration to these three issues in future  
91 works will be crucial for developing imputation methods that perform the best in practice.

92 Accordingly, this work focuses on characterizing the effect that these issues had on eval-  
93 uating imputation methods, with the goal of providing guidance on how to fairly evaluate  
94 such methods in the future. When collecting a test set, one should ensure that processing  
95 steps have been uniformly applied to raw data and that the data have been collected using  
96 similar procedures. When differences in processing arise that cannot be undone, we propose  
97 handling distributional shifts by using a quantile normalization approach that separately  
98 normalizes signal in peaks and signal in background. We also propose a set of new perfor-  
99 mance measures that focus on orthogonal aspects of imputation performance. Finally, we  
100 note that performance that does not generalize from well characterized cell types to poorly  
101 characterized ones does not have a simple fix like the other issues do. Rather, this disparity  
102 can only be evaluated by explicitly including both well- and poorly-characterized cell types  
103 in the evaluation. At a higher level, one should ensure that at least one setting used to  
104 evaluate their approach matches how they expect the method to have the most impact in  
105 practice, namely, on poorly characterized cell types.

## 106 2 Methods

### 107 2.1 The ENCODE Imputation Challenge

108 We acquired candidate imputation models by hosting the ENCODE Imputation Challenge  
109 (<https://www.synapse.org/encodeimpute>), a public challenge for imputing epigenomic pro-  
110 files, which began on February 20th, 2019, and concluded on August 14th, 2019. The  
111 challenge evaluated how well predictive models could impute held-out epigenomics experi-  
112 ments using other functional genomics experiments and nucleotide sequence as input (see  
113 challenge site for more details). Overall, we acquired 267 data sets from the ENCODE  
114 Portal to use as the training set, 45 data sets from the ENCODE Portal to use as a valida-  
115 tion set, and performed 51 new experiments to use as a test set for prospective evaluation  
116 (Figure 1, Additional File 1).

117 The challenge was divided into two stages. In the first stage, participants were provided  
118 with the training and validation data sets as well as a real-time public leaderboard of perfor-  
119 mance on the held-out validation set. Team BrokenNodes and Hongyang\_Li\_and\_Yuanfang\_Guan  
120 tied for first place at the conclusion of the first stage (Additional File 2 Supplementary Fig-  
121 ure S1, Additional File 3). In the second stage, the teams were allowed to re-organize, and  
122 participants were encouraged to refine their models using lessons learned from the first stage.  
123 The winners of the second stage, and of the entire challenge, were the top three teams based  
124 on performance on the held-out prospective test set, which the teams did not have access  
125 to.

126 The challenge was well attended with 196 people signing up on Synapse. Eight teams  
127 submitted results for the first round. After teams merged before the second round, 23



128 imputation models were submitted. Of these models, only one did not submit the full set of  
129 required imputations. Although our method for calculating team rankings as a part of the  
130 challenge accounted for missing imputations, our subsequent analyses excluded this model.

## 131 2.2 Performance Measures

132 Prior to the start of the challenge, we specified nine different performance measures to be  
133 used in the challenge. These performance measures included (1) the genome-wide mean-  
134 squared-error (MSE), (2) the genome-wide Pearson correlation, (3) the genome-wide Spear-  
135 man correlation, (4) the MSE calculated in promoter regions defined as  $\pm 2\text{kb}$  from the  
136 start of GENCODEv38 annotated genes [14], (5) the MSE calculated in gene bodies from  
137 GENCODEv38 annotated genes, (6) the MSE calculated in enhancer regions as defined by  
138 FANTOM5 annotated permissive enhancers [15], (7) the MSE weighted at each position by  
139 the variance of the experimental signal for that assay across the training set, (8) the MSE  
140 at the top 1% of genomic positions ranked by experimental signal, and (9) the MSE at the  
141 top 1% of genomic positions ranked by predicted signal. We note that 8 and 9 make a  
142 calculation similar to recall and precision, respectively.

143 We used a multi-stage process, originally developed for the ENCODE Transcription  
144 Factor Binding Challenge [16], to aggregate these performance measures into a single score  
145 to determine the challenge winners. First, ten equally-sized bootstraps were drawn from the  
146 pool of all genomic positions, and each of the nine performance measures was calculated for  
147 each team on each of the bootstraps for each experiment. For each bootstrap-experiment  
148 pair, the scores were converted to rankings across teams for each performance measure,  
149 and these rankings were then averaged across performance measures. This resulted in a  
150 score for each team in each bootstrap-experiment pair. This score was then converted back  
151 into a ranking over teams for each bootstrap-experiment pair. Next, these rankings were  
152 aggregated across experiments by calculating  $\frac{1}{|E|} \sum_{e \in E} \min(0.5, r_e)$  where  $E$  is the set of  
153 all experiments,  $e$  is an individual experiment, and  $r_e$  is a team's ranking on experiment  
154  $e$  divided by the number of teams. Finally, a rank was calculated across teams for each  
155 bootstrap, and the 90th percentile score, i.e. the second-best bootstrap rank, was used to  
156 determine the winners. This procedure is implemented at [https://github.com/ENCODE-](https://github.com/ENCODE-DCC/imputation_challenge)  
157 [DCC/imputation\\_challenge](https://github.com/ENCODE-DCC/imputation_challenge).

## 158 2.3 Baseline Methods

159 The methods submitted by the participants were compared to two baseline methods. The  
160 first baseline was the average activity, which is a straw-man imputation approach that  
161 simply predicts the average training set signal at each position in the genome across all cell  
162 types for a given assay type [17]. Consequently, this approach cannot make cell type-specific  
163 predictions. However, it represents the simple rule that regions of the genome that always  
164 exhibit peaks in signal and that regions of the genome that never exhibit peaks will continue  
165 to do so in other cell types. The second baseline was the Avocado model, using the same  
166 model architecture and training procedure described by Schreiber et al [12]. Importantly,  
167 this model was not tuned for this data set—it was applied as-is using the default settings  
168 and hyperparameters.

169 Although we had initially expected that ChromImpute [7] would serve as a baseline in  
170 this challenge, for logistical reasons ChromImpute was not applied to the challenge data  
171 until well after the challenge concluded. Because the participants did not have access to

172 these predictions, as they did with the other two baselines, we did not include ChromIm-  
173 pute in the original rankings or analysis. However, we have included a ranking of methods  
174 that includes ChromImpute in a re-analysis of the challenge participants using six of the  
175 measures used to evaluate the original ChromImpute method [7], as a reference (Additional  
176 File 2 Supplementary Figure S2/S3). These measures emphasize the relative distribution  
177 of signals, and included Pearson correlation, three measures quantifying percentage overlap  
178 between positions exhibiting high signal, and AUC measures for predicting peaks in ob-  
179 served signal from imputed signal values and vice versa. In order to obtain team ranks on  
180 these measures, we first ranked each team’s prediction for each test track on each measure  
181 separately. We then averaged ranks across metrics and re-assigned integer ranks in each  
182 track for each team. Each team’s final rank was then computed from the average of their  
183 predictions’ track ranks for the 51 test tracks.

## 184 2.4 Quantile Normalization

185 We developed a three-step quantile normalization method for normalizing signal across  
186 genomics experiments. Because signal distributions differ significantly across assays, we  
187 applied this normalization separately for each assay. Importantly, the normalization is also  
188 done separately for signal in peak and background regions (as defined by MACSv2 peak calls  
189 for the experiment [18]) to account for peaks spanning differing proportions of the genome  
190 across cell types. In the first step, quantiles are derived separately from each training set  
191 experiment. That is, if there are  $N$  training set experiments,  $M_p$  peak quantile bins, and  
192  $M_b$  background quantile bins, one would extract  $Q_p \in \mathcal{R}^{N, M_p}$  and  $Q_b \in \mathcal{R}^{N, M_b}$ . Quantiles  
193 are extracted by ranking all signal values for an experiment (in peaks or outside of peaks,  
194 respectively), binning those ranks into either  $M_p$  or  $M_b$  equally sized bins, and assigning  
195 to each bin the average signal value from positions within the bin. In the second step, an  
196 average is taken across experiments for each quantile bin to construct reference quantiles  
197  $R_p \in \mathcal{R}^{M_p}$  and  $R_b \in \mathcal{R}^{M_b}$ . Finally,  $R_p$  and  $R_b$  are applied to the test set tracks, with  
198  $R_p$  being applied only within signal peaks and  $R_b$  being applied only within background  
199 regions. Because peak regions are more complex and span a larger dynamic range than the  
200 background, we set  $M_p$  to be 1000 and  $M_b$  to be 50. Given that this procedure is designed  
201 to combat distributional shift, we note that it should be applied to test set experiments  
202 before evaluation.

## 203 2.5 Data Processing

204 We processed the DNase and ATAC-seq experiments using a uniform pipeline [19]. First,  
205 FASTQ files containing read sequences and quality scores for the training and validation  
206 sets experiments were downloaded from the ENCODE Portal, and FASTQs for the test  
207 set experiments were acquired from our own experiments. For ATAC-seq experiments (but  
208 not DNase-seq), we first trimmed adapters and then mapped reads to the hg38 reference  
209 human genome using the Bowtie2 [20] aligner. After mapping, reads were filtered to re-  
210 move unmapped reads and mates, non-primary alignments, reads failing platform/vendor  
211 quality checks, and PCR/optimal duplicates (-F 1804). Reads mapping reliably to more  
212 than one location (MAPQ < 30), i.e., multi-mapping reads, were removed. Duplicate reads  
213 were then marked with Picard MarkDuplicates [21] and removed. For single-end DNase  
214 data sets, a single read was chosen from a set of duplicate reads, whereas for paired-end  
215 data sets, read-pairs were chosen if any one of the two reads in the pair was unique. Al-

216 though this is the standard approach for de-duplicating single-end and paired-end data, this  
217 step had unintended consequences for the challenge, which we describe in Section 3.2. For  
218 ATAC-seq data, 5' ends of filtered reads on the + and - strand were shifted by +4 and  
219 -5 bp respectively to account for the Tn5 shift. Reads from biological and technical repli-  
220 cates were merged. We normalized the sequencing depth across data sets by subsampling  
221 them to a maximum of 50 million reads (after excluding reads mapping to mitochondria).  
222 Although there are several ways to represent the signal from sequencing experiments, e.g.,  
223 read-counts and fold-change, we chose to use the statistical significance of the fold-change to  
224 be consistent with previous imputation literature [11, 12, 7, 10]. We used the MACSv2 peak  
225 caller to compute the fold-enrichment and statistical significance. MACSv2 was applied to  
226 smoothed counts (150 bp smoothing window) of read-starts (5' ends of reads) at each posi-  
227 tion in the genome relative to the expected number of reads from a local Poisson-simulated  
228 background distribution. We filtered out all peaks that overlapped with the ENCODE Ex-  
229 clusion list consisting of abnormal high signal regions [22]. We provided the genome-wide  
230 signal tracks containing the statistical significance of enrichment (i.e., the  $-\log_{10}$  p-values)  
231 at each basepair in the genome. The processing pipeline is open-source and available at  
232 <https://github.com/ENCODE-DCC/atac-seq-pipeline>.

233 Next, we processed the histone ChIP-seq experiments using the ENCODE processing  
234 pipeline [23]. For each experiment we downloaded FASTQ files from the ENCODE Portal  
235 for at least two replicate experiments and a control experiment. All reads were mapped to  
236 the hg38 reference human genome using the BWA aligner [24]. After mapping, the process  
237 was similar to the ATAC-seq/DNase-seq pipeline. Reads were filtered to remove unmapped  
238 reads and mates, non-primary alignments, reads failing platform/vendor quality checks, and  
239 PCR/optical duplicates (-F 1804). Multi-mapping reads ( $\text{MAPQ} < 30$ ) were also removed.  
240 Duplicates were identified using Picard MarkDuplicates and subsequently removed, with the  
241 same single-end vs. paired-end differences as mentioned for DNase data sets. Reads from  
242 the biological and technical replicates were then merged. We normalized the sequencing  
243 depth across data sets by subsampling each to a maximum of 50 million reads. We used  
244 the MACSv2 peak caller to calculate fold-enrichment and statistical significance of counts  
245 of extended ChIP-seq reads (reads were extended in the 5' to 3' direction based on the  
246 predominant fragment length), relative to the number of extended reads from the control  
247 experiment, and filtered out peaks that overlapped with the ENCODE Blacklist [22]. The  
248 statistical significance of the enrichment was computed using a local Poisson null distribution  
249 whose mean parameter is estimated from the control experiment. For the purposes of this  
250 challenge, we provided the genome-wide signal tracks containing the statistical significance of  
251 enrichment (i.e., the  $-\log_{10}$  p-values) at each basepair in the genome. The processing pipeline  
252 is open-source and available at <https://github.com/ENCODE-DCC/chip-seq-pipeline2>.

## 253 3 Results

### 254 3.1 The ENCODE Imputation Challenge

255 Participants submitted 23 models to the second stage of the challenge (see Section 2.1).  
256 Each group was allowed to submit up to three models to encourage inclusion of unorthodox  
257 solutions with at least one submission. As a result, the models encompassed a diverse range  
258 of strategies (see Table 1). The models differed primarily along three axes. The first axis  
259 was the signal preprocessing, with almost every method further preprocessing the data from  
260 the given  $-\log_{10}$  signal p-values. The second axis was the data sources used to construct

Name	Model	Norm	Inputs			
			Sequence	Functional	Average	Avocado
Aug2019Impute						
BrokenNodes/v2	KNN	arcsinh		✓		
BrokenNodes/v3	KNN	arcsinh		✓		✓
CostaLab v2						
CUImpute1/CUWA/ICU	ensemble	arcsinh		✓	✓	✓
Guacamole/Lavawizard	DTF	arcsinh		✓	✓	
HLYG/v1/v2	GBT	quantile	✓	✓	✓	
imp/imp1	DTF+AE	Cauchy		✓		
KKT-ENCODE	CNN	arcsinh	✓			
LiPingChun	DTF	arcsinh		✓	✓	
NittanyLions	KNN			✓		
NittanyLions2	KNN	quantile		✓		
SongLab	CNN	log1p		✓		
SongLab2	HMM			✓		
SongLab3	CNN	log1p		✓	✓	✓
UIOWA	CNN	quantile	✓	✓		

Table 1: **Methodologies of imputation methods.** The table lists the modeling strategies and input features used by each of the models, as reported by the teams. The models include k-nearest neighbors (KNN), deep tensor factorization (DTF), autoencoders (AE), convolutional neural networks (CNN), hidden Markov models (HMM), and gradient-boosted decision trees (GBT). The authors of Aug2019Impute and CostaLab v2 did not describe their methods.

261 input features. Most methods followed previously published methods by only using assay  
 262 measurements as inputs (denoted “functional” in Table 1). However, five of the methods  
 263 used nucleotide sequence as input, eight methods used the average activity baseline, and  
 264 three used Avocado’s imputations. The third axis was the manner in which the underlying  
 265 tensor structure of the data was modeled. Some methods explicitly modeled the data as  
 266 a tensor (e.g., imp and Lavawizard), whereas other methods only implicitly modeled the  
 267 structure through rule-based approaches or similarity methods (e.g., the HLYG and KNN-  
 268 based approaches).

269 An initial inspection of the imputations revealed that most methods captured the general  
 270 shape of the signal well. Examples drawn from H3K27ac in brain microvascular endothe-  
 271 lial cells and DNase-seq in DND-41 cells (Figure 2A/B, Additional File 2 Supplementary  
 272 Figure S4) suggest two sources of error: the misprediction of a small number of peaks re-  
 273 lative to the total number of true peaks, and the misprediction of the precise signal value  
 274 within correctly predicted peaks. Focusing on the misprediction of peaks, we noted that  
 275 some methods made similar mistakes as the average activity baseline, whereas others made  
 276 similar mistakes as the Avocado baseline (gray highlights in Figure 2A/B). Unsurprisingly,  
 277 methods that used Avocado’s imputations as input had the highest genome-wide correlation  
 278 with Avocado’s predictions (it is worth noting that CUImpute1 only used Avocado’s impu-  
 279 tations for some, but not all, assays). In contrast, methods that explicitly used the average  
 280 activity did not always exhibit higher correlation with it than other methods (Supplemen-  
 281 tary Figure S5). This finding suggests that, because the average activity can be directly  
 282 derived from the training set, many types of models are able to implicitly learn it even when  
 283 not explicitly trained on it.

284 Next, we comprehensively evaluated the methods using a battery of performance mea-  
 285 sures that were specified at the beginning of the challenge (see Section 2.2, Additional File 4).  
 286 We found that performance on these measures depended heavily on the imputed assay (Fig-  
 287 ure 2C/D). For instance, most models exhibited four orders of magnitude higher MSE on



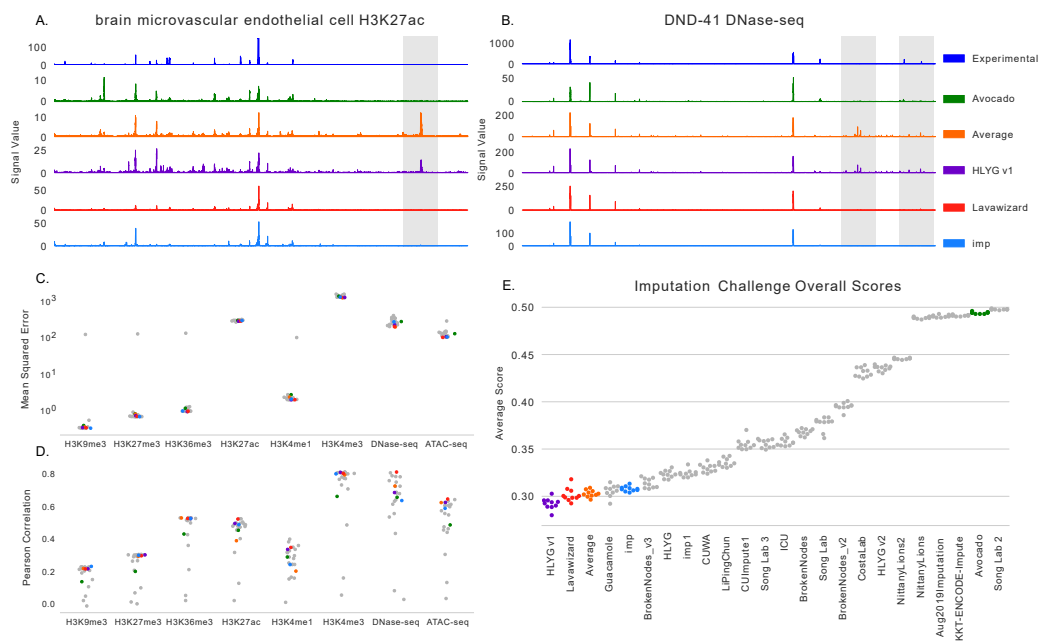


Figure 2: **Results from the ENCODE Imputation Challenge.** (A) The H3K27ac signal for brain microvascular endothelial cells that is observed (in blue), from baseline methods, and from the winning three teams in the challenge. (B) The same as (A) except for DNase-seq signal in DND-41 cells. (C) The average MSE for each method across test set tracks and bootstraps but partitioned by assay type. (D) The same as (C) except for Pearson correlation. (E) The overall score, calculated as described in Section 2.2, across all test set tracks and performance measures shown for each bootstrap for each team. The baseline methods and winners are colored.

288 H3K4me3 than on H3K9me3. However, several assays that exhibited the highest MSE also  
289 exhibited the highest Pearson correlation, indicating that the scale of MSE across assays  
290 is likely more related to the dynamic range of the assay rather than the accuracy of the  
291 imputations. Unsurprisingly, a projection of all imputed and experimental tracks clustered  
292 predominately by assay type (Silhouette Score = 0.4601), as opposed to by cell type (SS  
293 = -0.4028) or imputation method (SS = -0.3133, Supplementary Figure S6). Accordingly,  
294 we used a rank-based transform to account for differences in dynamic range when calculat-  
295 ing global performance measures across experiments (see Section 2.2) to ensure that assays  
296 with large dynamic ranges did not dominate the evaluation. After calculating the global  
297 performance of each method, we found that there was a gradient of methods that performed  
298 increasingly well, and a set of methods that performed relatively poorly (Figure 2E). The  
299 best performing methods, and hence the winners of the challenge, were Hongyang Li and  
300 Yuangfang Guan v1 (abbreviated as “HLYGv1”) in first place, Lavawizard and Guacamole  
301 (two similar methods from the same team) tied for second place, and imp in third place.

302 Given the diverse modeling strategies of the winning teams, our primary take-away from  
303 these results is that there does not appear to be a single key insight that led to good overall  
304 performance on the measures used in the challenge. HLYGv1 used nucleotide sequence  
305 as input, but so did KKT-ENCODE and UIOWA Michaelson; all three models submitted  
306 by Hongyang Li and Yuanfang Guan used gradient boosted trees (GBTs), yet their models  
307 exhibited both good and poor performance. However, these results do suggest certain models  
308 to be wary of: convolutional neural networks and k-nearest neighbor models underperformed  
309 deep tensor factorization (DTF) and GBT models. This is likely because the similarities  
310 used by KNN models are a less sophisticated version of the representations learned by tensor  
311 factorization approaches, and that the specific structure presented in the data is not well  
312 modeled by simple applications of convolutions.

313 However, when we compared model performance to the baseline methods, we made two  
314 important observations. First, almost every team outperformed the Avocado baseline, as  
315 one might expect because the participants had access to the Avocado model and predictions  
316 during the development process, and because the default settings were used for Avocado  
317 despite them being tuned for significantly larger amounts of training data. Second, the  
318 average activity baseline performed extremely well, coming in third in our ranking and  
319 first place in five of the nine performance measures used (Additional File 4). Both of  
320 these observations are a reversal from the first round in the challenge, where Avocado  
321 outperformed all the participants but almost all the participants outperformed the average  
322 activity baseline (Supplementary Figure S1). This reversal in performance between the  
323 two baselines is partially because the evaluation setting changed from overrepresenting well  
324 characterized cell types to focusing on poorly characterized ones and, as we will see later,  
325 partially due to the performance measures used for the challenge.

### 326 3.2 Accounting for distributional shift

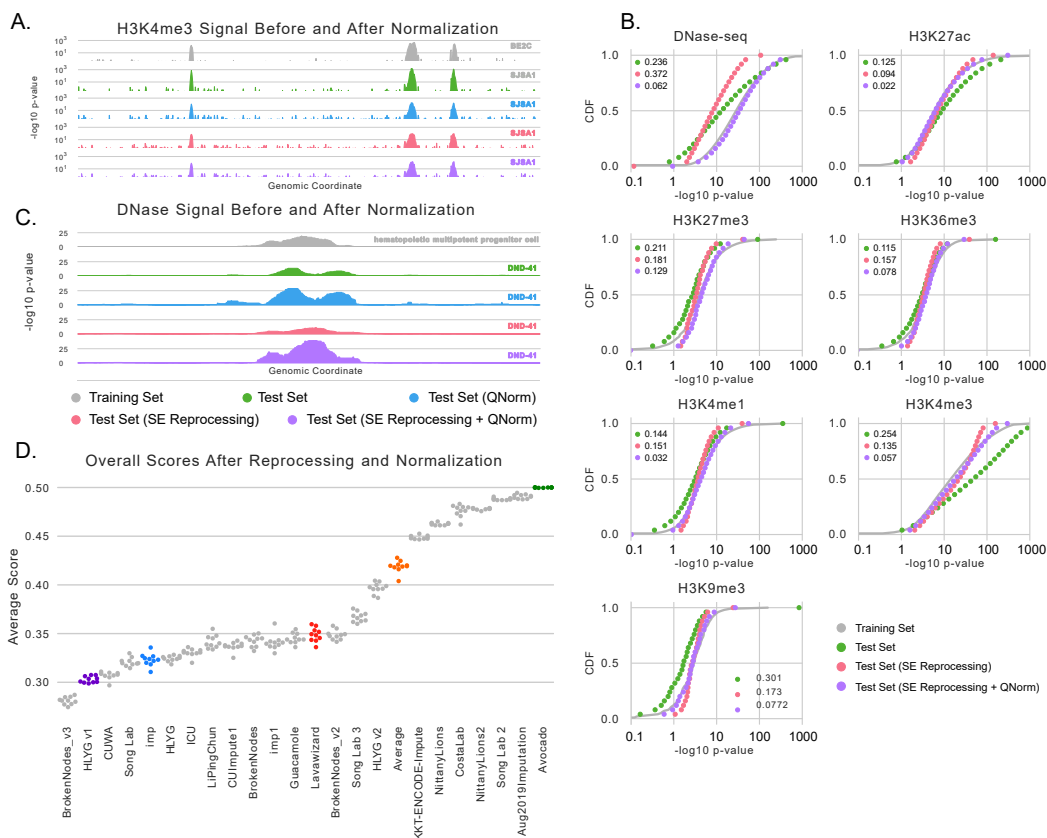
327 A visual inspection of the test set experiments revealed significant distributional differences  
328 in peak signal values between the training and test sets for some assays (Figure 3A). Most  
329 obviously, the signal values within H3K4me3 peaks from test set experiments were gen-  
330 erally much higher than the signal values within peaks from training and validation set  
331 experiments (Figure 3B). Although one would expect a locus to exhibit different signal in  
332 different cell types because of real biology, one would also expect that the distribution of  
333 signal values within peaks across entire experiments would be similar for experiments of

334 the same assay. Because distributional shifts have major ramifications for the scale-based  
335 performance measures used in the challenge, we next investigated the source of these distri-  
336 butional differences.

337 After considering several potential covariates that could explain this distribution shift,  
338 including multiple measures of experimental quality (Additional File 2, Supplementary Fig-  
339 ure S7), we found that the primary driver was a subtle difference in how the test set ex-  
340 periments were processed. By design, the test set experiments were performed during the  
341 challenge to ensure a truly prospective evaluation. However, experimental methods have  
342 changed in the many years since the training data were collected. Most notably, collecting  
343 paired-end data is now the standard approach for ENCODE data sets because the procedure  
344 yields higher quality data and is now cheap enough for broad usage; however, almost all of  
345 the training set experiments predate this switch and involve single-end data. The process-  
346 ing of single-end and paired-end data is largely similar, but a crucial difference occurs in  
347 the deduplication step. Specifically, deduplication of single-end reads using PICARD [21]  
348 allows the mapping of only one read start to each position on the genome on each strand.  
349 In contrast, deduplication of paired-end data can result in more than one read-start per po-  
350 sition on each strand because read-pairs are only removed if the read start of *both* ends are  
351 duplicates. Consequently, the number of reads mapping within peaks from paired-end data  
352 can be significantly higher than what one would get using single-end data. Importantly, the  
353 shift is not simply caused by paired-end data being higher quality, as we first explored, but  
354 rather differences in the deduplication step.

355 We confirmed that differences in processing, rather than differences in data quality,  
356 explained the distributional shift by reprocessing the paired-end data sets (except for ATAC-  
357 seq which requires paired-end data) as single-end data. Specifically, for each paired-end  
358 experiment in the test set we concatenated the FASTQ files of reads from both ends and ran  
359 the same single-end processing pipeline that was run on the other single-end experiments  
360 in the challenge. We found that the reprocessed data had distributions of peak signal  
361 values significantly closer to the training set, as measured by the Kolmogorov-Smirnov  
362 (KS) statistic, for four of the histone modification assays including H3K4me3 (Figure 3B).  
363 The remaining two histone modification assays already resembled the training set before  
364 reprocessing. However, we found that the distribution of DNase-seq peak signal values had  
365 a larger KS-statistic after reprocessing than before. This is likely because 21 of the 38  
366 training set experiments contained paired-end data, which would shift the distribution of  
367 signal values in the training set up. Although the most principled next step would be to  
368 reprocess all of the experiments used in the challenge and subsequently re-training and re-  
369 evaluating each submission, this analysis was not possible because we only required that the  
370 three challenge winners submit code that could retrain their models on new data sets. Given  
371 no perfect solution, we chose to continue with the single-end reprocessed test set tracks for  
372 our subsequent analyses.

373 We found that reprocessing the histone modification data significantly reduced the dis-  
374 tributional shift but did not perfectly correct it. The remaining differences are likely related  
375 to small changes in experimental protocol over time, such as improvements in sequencing  
376 technology, antibodies used, and read lengths measured. A general-purpose correction for  
377 the remaining differences is to explicitly quantile normalize the data such that the signal  
378 values in the testing experiments exhibit the same signal distribution as those in the training  
379 experiments. Quantile normalization is powerful because it is a non-linear method, in con-  
380 trast to min-max or z-score scaling, and has been extensively applied to genomics data sets,  
381 including those measuring bulk gene expression [25], single-cell RNA-seq [26], and ChIP-seq



**Figure 3: Distributional shift and quantile normalization.** (A) Experimental signal measuring H3K4me3 in BE2C cells from an unnormalized training set experiment (gray), an unnormalized test set experiment in SJSA1 cells (green), the test set signal after quantile normalization (blue), the test set signal after single-end reprocessing (red), and the test set signal after single-end reprocessing and quantile normalization (purple). (B) Distributions of signal values within peaks in chr16/17 for each reprocessed assay across the unnormalized training set (gray), the unnormalized test set (green), the single-end reprocessed test set (red), and the single-end reprocessed and quantile-normalized test set (purple). The KS statistics between the training set distribution and the test set distributions are shown in the legends and the CDFs are summarized using 25 dots for visualization purposes. (C) An example locus that exhibits a DNase peak in both the training and test sets. (D) A re-scoring of the challenge participants against single-end reprocessed and quantile-normalized test set signal.

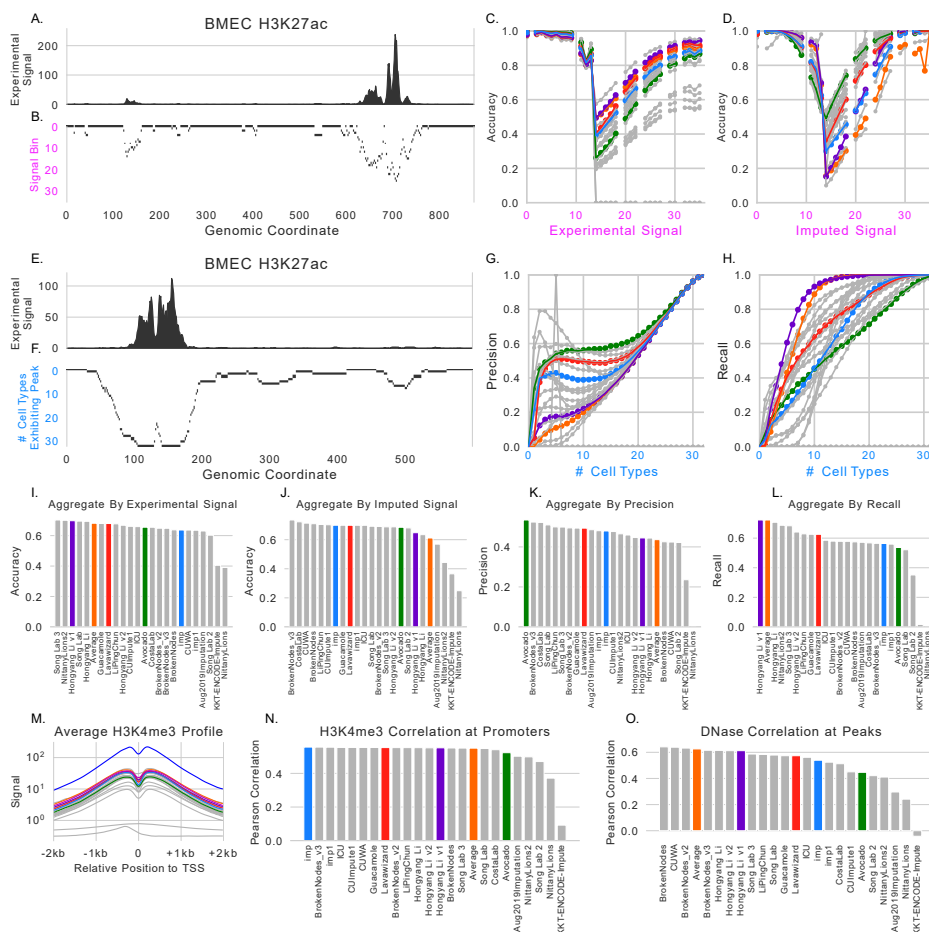
382 data when combined with a spike-in reference [27]. We account for differing proportions of  
383 the genome exhibiting peaks across cell types by separately quantile normalizing the signal  
384 within peaks and the signal in background regions (see Section 2.4 for details). Finally, be-  
385 cause the distribution of signal is significantly different across assays, we apply this quantile  
386 normalization to each assay separately. After normalization, we confirmed that the distribu-  
387 tion of within-peak test signal values was almost identical to the distribution of within-peak  
388 training signal values across all assays (Figure 3B), even for the DNase-seq experiments.

389 In theory, one could apply quantile normalization to the original paired-end test set data  
390 and, by definition, produce signal values with the same distribution without the need for  
391 reprocessing. However, when looking at a representative DNase peak, we found that the  
392 reprocessed data was not a simple monotonic transform of the original data (Figure 3C).  
393 Specifically, the paired-end data exhibited a peak shape unlike that observed in the single-  
394 end data, and simply quantile normalizing the signal does not fix the differences in shape.  
395 More comprehensively, when considering a 10Mbp region of chr1 on each of the 48 repro-  
396 cessed experiments, we clearly observed that paired-end data is not a monotonic transforma-  
397 tion of single-end data (Additional File 2, Supplementary Figure S8). Although the assays  
398 associated with activity, such as H3K4me3 and DNase-seq, exhibit Spearman correlations  
399 up to 0.938 between the paired-end and single-end processed signals, repressive marks ex-  
400 hibit Spearman correlations as low as 0.037, and the average Spearman correlation across  
401 all tracks was only 0.453. Further, even though some assays exhibit high correlation, this  
402 value is inflated by the large number of low-signal values and, indeed, the largest variability  
403 comes at loci with high signal values.

404 Moving forward with our method of reprocessing the test data using single-end settings  
405 and then quantile normalizing to correct the remaining differences, we next re-scored the  
406 originally submitted imputations (Figure 3D, Additional File 5, Additional File 6). We  
407 observed that the number of methods outperforming the average activity baseline increased  
408 from two to 16 and that BrokenNodes\_v3 rose from sixth place to first place in the rankings.  
409 Although HLYGv1 remains within the top three, the other two winners descended in the  
410 rankings. This might be explained by HLYGv1 using quantile normalization, albeit a slightly  
411 different version than the one we used, during training. Interestingly, many of the methods  
412 performed similarly to each other, reinforcing the idea found in the original challenge that  
413 there is not necessarily one way to do imputation. Indeed, the best performing model is a  
414 simple KNN-based approach using arcsinh-transformed data and the second best performing  
415 model uses gradient-boosting trees on quantile transformed data. Critically, we note that  
416 it would not be fair to use these rankings to declare challenge winners because we did not  
417 give the teams an opportunity to retrain or tune their methods on the transformed data.  
418 Rather, our take-away is that the distributional shift is partially responsible for the good  
419 performance of the average activity baseline but does not fully explain it.

### 420 **3.3 Designing more informative performance measures**

421 Although the measures used in the challenge were devised to rank methods independently  
422 for each experiment based on their genome-wide (or across large portions of the genome)  
423 performance, this property meant that they ultimately exhibited a high degree of redundancy  
424 with each other (Supplementary Figure S9). Essentially, by uniformly weighting all positions  
425 along the genome, methods with low genome-wide MSE were likely to have low MSE within  
426 promoters, gene bodies, or the top 1% of signal as well. Exacerbating this issue, MSE-based  
427 measures were disproportionately confounded by the large distributional shift described in



428 the previous section in comparison to the shape-based measures. Illustrating this, we found  
429 that most of the residual—sometimes over 99% in H3K4me3 assays—came at correctly-  
430 predicted peaks (Supplementary Figure S10A). Realizing this weakness, we next designed  
431 three new types of performance measures that, respectively, reweighted genomic bins based  
432 on signal strength, considered multiple experiments simultaneously, and focused on shape  
433 within active areas. All evaluations in this section are done against the reprocessed, quantile-  
434 normalized test set signal.

### 435 3.3.1 Partitioning by signal strength

436 A strategy for measuring performance in a complementary way to uniformly-weighted genome-  
437 wide performance is to explicitly calculate the performance with respect to the magnitude  
438 of either the observed or imputed signal (Figure 4A/B). Rather than being limited by con-  
439 sidering only the top 1% bin of signal, such as by using the `mselobs` or `mseimp` measures,  
440 considering all signal bins provides a finer-grained view of model performance. As an ex-  
441 ample, if the imputations exhibit high accuracy when the imputed signal is high, then one  
442 may be confident that predicted peaks are correct when using imputations for which there  
443 is no corresponding experimental data; in contrast, if the imputations exhibit low accuracy  
444 when the imputed signal is high but higher accuracy when the imputed signal is low, then  
445 one might be more skeptical of imputed peak calls but more trusting of regions not called  
446 as peaks, e.g. facultative peaks that are not active in the studied cell types. Although any  
447 measure can be partitioned by signal magnitude, we focus on accuracy between binarized  
448 imputed signal and peak calls for the experimental signal. Accuracy was excluded from the  
449 original set of performance measures because the sparsity of peaks can make it difficult to  
450 interpret genome-wide; in this setting, we anticipate accuracy to be more valuable in the  
451 signal bins where one might reasonably find a peak. Importantly, we did not use rank-based  
452 classification measures (e.g., AUROC or AUPR) here, because once the signal is partitioned  
453 by strength, applying a rank-based measure to each bin is less meaningful than when applied  
454 genome-wide.

455 When we partitioned genomic loci based on experimental signal, we found that model  
456 performance aggregated across all tracks generally falls into three regimes: (1) when the  
457 imputed or experimental signal is low, the accuracy is high, (2) when the imputed or experi-  
458 mental signal is between 1 and 10 the accuracy severely drops, and (3) when the imputed or  
459 experimental signal is high, the accuracy returns to being high (Figure 4C). Although the  
460 second regime includes ambiguous peak calls, it also includes the most difficult to call peaks  
461 (thus, the relatively low accuracy) and should be emphasized by performance measures.  
462 When focusing on H3K27ac signal in brain microvascular endothelial cells we can also see  
463 that rankings flip between the first and second regime (Figures 4C-D); `imp1` and `LavaWiz-  
464 ard` both outperform the average activity and `HLYGv1` when the experimental signal is low,  
465 but perform significantly lower when the experimental signal is higher.

466 Interestingly, the ranking of methods is almost reversed when partitioning genomic loci  
467 using the imputed signal instead of the experimental signal (Figure 4D). `HLYGv1` and the  
468 average activity are among the top performers when partitioning by experimental signal  
469 but are among the worst performers when partitioning by imputed signal. An explanation  
470 for this flip is that these approaches measure notions of precision and recall, respectively,  
471 which have a known trade-off. Because the average activity is essentially a union of peaks  
472 across cell types in the training set, it will have a high recall but a low precision. Methods,  
473 such as `HLYGv1`, that rely too heavily on the average activity will exhibit the same tradeoff

474 (Figure 4C/D).

475 A straightforward way to condense these curves into a single value for a performance  
476 measure is to take the average value across the curve. This value is essentially a re-weighting  
477 of genome-wide accuracy that uniformly values each bin of signal values rather than each  
478 locus, and so will downweight the more common low signal value loci and upweight the  
479 less common higher signal values ones. Notably, the winners of the ENCODE Imputation  
480 Challenge did not perform the best across all test set experiments when partitioning by either  
481 experimental signal or by imputed signal (Figure 4I/J). Indeed, the top two performers when  
482 partitioning by experimental signal (Song Lab 3 and NittanyLions2) came in 12th and 19th  
483 respectively in the original evaluation.

### 484 3.3.2 Prediction of facultative peaks

485 A primary source of error for imputation models comes from loci that exhibit functional  
486 activity in some, but not all, cell types. Evaluating whether the imputations can distinguish  
487 between cell types that do and do not exhibit signal at a given locus is crucial for ensuring  
488 that the imputations are cell type-specific. However, because traditional genome-wide  
489 performance measures treat each experiment independently, they cannot explicitly evaluate  
490 this property. To better understand how well these methods can identify what cell types  
491 loci are active in, for each assay we partitioned genomic positions by the number of experi-  
492 ments that exhibit a peak for that assay and then evaluated each partition separately. For  
493 example, if a locus exhibited a DNase-seq peak in 3 out of 5 cell types, that locus would be  
494 grouped for evaluation with other loci that also exhibited DNase-seq peaks in 3 out of 5 cell  
495 types (Figure 4E/F). This analysis is similar to the one presented by Schreiber et al. [11]

496 We observe trends that are reminiscent of partitioning loci by signal strength. As the  
497 number of cell types that exhibit peaks increases, so too does the precision and recall of the  
498 methods (Figure 4G/H). This indicates that, generally, imputation methods are better at  
499 predicting peaks at facultative peaks than they are at predicting cell type-specific activity.  
500 Interestingly, we noted that several methods had peaks in precision when the number of cell  
501 types the peak was expressed in is low. Given that performance was extremely variable in  
502 this regime, we think that focusing on this measure in future studies will be useful when  
503 comparing models. Consistent with the role that the average activity plays as essentially the  
504 union of peaks across cell types, we see that it has a low aggregate precision score across all  
505 test set tracks but has the second highest aggregate recall score (Figure 4K/L). Put another  
506 way, the average activity is very good at identifying peaks that are common across many  
507 cell types but very poor at identifying the cell types that cell type-specific peaks occur in.  
508 Somewhat surprisingly, the Avocado baseline had the highest aggregate precision score, but  
509 the challenge winners that most resemble it (Lavawizard and imp) did not exhibit the most  
510 similar performance.

### 511 3.3.3 Relative peak shape

512 The performance measures that have been proposed so far predominantly involve genome-  
513 wide calculations, even if they involve re-weighting loci contributions. An alternate form  
514 of performance measure is to focus on specific forms of biochemical activity at loci that  
515 are known to be relevant. The MSEProm, MSEEnh, and MSEGene measures attempt to  
516 quantify this by focusing on promoters, enhancers, and gene bodies respectively, but measure  
517 the performance of all assays at these loci. Next, we investigate two more performance



518 measures that follow the reasoning of Ernst et al. [7] that only specific assays should be  
519 measured at these loci.

520 The first measure evaluates the shape of H3K4me3 signal at promoter regions. This hi-  
521 stone modification is known to be enriched at promoter elements and is indicative of active  
522 transcription. Further, after correcting for the strand of the promoter, the mark exhibits a  
523 distinctive bimodal pattern (Figure 4M). We reasoned that focusing on the ability to recap-  
524 ture this shape would provide an orthogonal evaluation to the other performance measures  
525 proposed so far. We calculated the average Pearson correlation between the imputed signal  
526 and the quantile-normalized experimental signal across all gene promoters for all test set  
527 tracks measuring H3K4me3. Most of the methods outperformed the average activity base-  
528 line but only one of the challenge winners were in the top five according to this measure  
529 (Figure 4N).

530 The second measure evaluates the shape of DNase signal at observed DNase peaks. We  
531 anticipated that recapturing the shape of DNase signal would be more challenging because  
532 DNase does not exhibit a pattern that is as consistent as H3K4me3 at promoter regions.  
533 Further, the subtle patterns encoded in DNase signal can be useful for deciphering the  
534 precise regulatory role that the underlying nucleotide sequence is playing. Consistent with  
535 predicting DNase signal being a more challenging task, we found that methods exhibited  
536 a wider range of performances than they did with H3K4me3 prediction (Figure 4O). We  
537 also found that only three methods outperformed the average activity baseline. This might  
538 initially be counterintuitive, because chromatin accessibility is fairly cell type-specific. How-  
539 ever, because this evaluation is limited to observed DNase peaks, methods are not being  
540 penalized for incorrectly predicting that non-peak regions are exhibiting peaks. This obser-  
541 vation indicates that accessible loci largely retain the shape of their peaks across cell types  
542 when binned at 25 bp resolution.

## 543 4 Discussion

544 A central theme of this work is that evaluating models that rely on large collections of  
545 genomic data sets can be more difficult than one might initially expect and, consequently,  
546 that results can be confounded even when one does not make any obvious mistakes. In our  
547 analysis, we identified three issues that made analysis of imputation models more difficult  
548 than we initially thought: distributional differences in the underlying data, previous eval-  
549 uation focusing on well-characterized cell types and in larger compendia, and performance  
550 measures that were either redundant or sensitive to the first two issues. We addressed these  
551 issues by proposing a quantile normalization approach that treats peak and background  
552 signal separately, and proposing new performance measures that were less redundant with  
553 each other and covered more aspects of performance than the original measures.

554 When the challenge was originally designed, the participants were not required to submit  
555 working code in order to lower the barrier to entry and allow participants to use their own  
556 custom hardware. Although this likely increased participation, it also caused a recurring  
557 problem in our later analyses because we could not retrain models on reprocessed data, or  
558 on different subsets of data. For example, reprocessing all the data using the single-end  
559 settings would likely have been the correct thing to do from a theoretical point of view, but  
560 was impossible as a practical matter because we did not have the required code. Likewise,  
561 we had hypothesized that part of the reason for changes in rankings between the first and  
562 second stages (including in our baselines) was because the first stage involved evaluation

563 on a randomly selected held-out test set of experiments, which are biased towards well-  
564 characterized cell types, and the second stage explicitly evaluated only poorly characterized  
565 cell types. Because we could not re-train the models and evaluate them on cell types giving  
566 variable amounts of information, we could not comprehensively pursue this line of inquiry  
567 using the challenge data.

568 Based on our experience running this challenge, we have several recommendations for the  
569 organizers of future challenges involving genomic data sets. First, ensure that participants  
570 are compared against naive baselines such as the average activity. Without this baseline,  
571 we might not have identified as easily the distributional shift or the worse performance  
572 on sparsely characterized cell types. Second, participants should be required to submit  
573 code that can reproduce the training of their models so that more in-depth analysis can be  
574 done later. Potentially, the organizers should provide a scaffold that the participants fill  
575 in with their own code so that the organizers do not need to decipher each submission to  
576 use it properly. Third, organizers should explicitly look for distributional shifts across data  
577 splits, and even between pairs of data sets, as a quality control step. For example, paired end  
578 datasets from cancer cell lines can often contain large regional distribution shifts and outliers  
579 driven by cell line-specific copy number variation. Even when these shifts are explained by  
580 biological processes rather than experimental biases, tailoring an analysis that accounts for  
581 these shifts can be an important aspect of a fair evaluation. Finally, organizers should  
582 design performance measures that have minimal redundancy with each other, potentially  
583 as measured using the average activity before the challenge begins. Naturally, without a  
584 singular end-goal in mind it can be difficult to balance the various aspects of performance in  
585 a manner that will satisfy everyone, but having redundant performance measures is clearly  
586 not helpful.

587 An unaddressed, but important, issue is determining the most informative target for  
588 imputation methods to predict. The most common target in imputation literature has been  
589 the statistical significance from a peak-calling algorithm. Predicting the statistical signifi-  
590 cance can be more informative than predicting read counts directly because read counts  
591 can suffer from unwanted experimental biases and the peak-calling algorithm can explicitly  
592 consider a control track. Our challenge setting is consistent with that literature. However,  
593 an issue with predicting p-values is that fewer tools take those as input than take read  
594 counts as input. In fact, performing peak calling using imputations is not obvious because  
595 it is unclear that simply thresholding the uncalibrated p-values is the correct approach.  
596 Potentially, future iterations of the imputation work could involve imputing read counts  
597 but allowing models to directly incorporate the control tracks and other covariates such  
598 as sequencing depth, single-end or paired-end status and data quality metrics as well [28].  
599 Although there would be some engineering challenges with such a task, such as designing  
600 alternate loss functions or performance measures based on counts, imputation of read counts  
601 might be more readily adopted.

602 Although the issues we described made the analysis of the results of this challenge more  
603 difficult, we made several important findings that we hope will guide the design and analysis  
604 of predictive models that rely on genomics data in the future. Specifically, even outside the  
605 context of a challenge, being aware of distributional shifts and evaluating a newly proposed  
606 model with a wide set of performance measures can help ensure that the model is robust in  
607 practice. Further, the difficulties that we faced are not unique to the setting of imputation.  
608 Indeed, these issues can affect any model that is trained or evaluated using large collections  
609 of publicly available data sets.

## 610 4.1 Author Contributions

611 A.K., C.B., J.S.S., M.K. and W.S.N. designed the challenge. J.L. processed the data for the  
612 challenge. J.L. and J.S.S. provided technical support for the challenge. A.K., W.S.N., J.S.,  
613 and C.B. ran the challenge. J.S. and C.B. manually validated the challenge winners. J.S.  
614 designed and performed the subsequent analyses after the challenge concluded and wrote  
615 the manuscript. C.B., A.K. and W.S.N. edited the manuscript.

616 H.L. and Y.G. participated in the challenge under the team name “Hongyang Li and  
617 Yuanfang Guan.” C.C. and J.C. participated in the challenge under the team names “LiP-  
618 ingChun,” “Guacamole,” and “Lavawizard.” A.H., B.S, G.S., and M.R.C. participated in  
619 the challenge under the team names “imp” and “imp1.” A.C., F.G., L.N., M.M., M.J.C., and  
620 P.P. participated in the challenge under the team name “BrokenNodes.” C.H. and K.Y.Y.  
621 participated in the challenge under the team names “CUImpute1,” “CUWA,” and “ICU.”  
622 J.P.S., S.S.B., and Y.S.S. participated in the challenge under the team name “Song Lab.”  
623 S.M. and Z.Z. participated in the challenge under the team name “NittanyLions.” W.T.,  
624 Y.S., Y.S., and Y.S. participated in the challenge under the team name “KKT-ENCODE.”

625 M.S. and J.A. and R.S. and N.F. and J.H. and K.L. and L.J. and X.Y. and M.C.  
626 performed experiments to create the blind test set used to evaluate the methods.

### 627 4.1.1 Acknowledgements

628 We would like to thank Alan Min for providing feedback on a draft of the manuscript,  
629 Oana Ursu for suggesting analyses, and Jason Ernst for providing manuscript feedback and  
630 ChromImpute imputations on the challenge data.

## 631 References

- 632 [1] Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst,  
633 Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo  
634 Zhang, Jianrong Wang, Michael J Ziller, Viren Amin, John W Whitaker, Matthew D  
635 Schultz, Lucas D Ward, Abhishek Sarkar, Gerald Quon, Richard S Sandstrom,  
636 Matthew L Eaton, Yi-Chieh Wu, Andreas R Pfening, Xinchun Wang, Melina Clauss-  
637 nitzer, Yaping Liu, Cristian Coarfa, R Alan Harris, Noam Shores, Charles B Epstein,  
638 Elizabeta Gjoneska, Danny Leung, Wei Xie, R David Hawkins, Ryan Lister, Chibo  
639 Hong, Philippe Gascard, Andrew J Mungall, Richard Moore, Eric Chuah, Angela Tam,  
640 Theresa K Canfield, R Scott Hansen, Rajinder Kaul, Peter J Sabo, Mukul S Bansal,  
641 Annaick Carles, Jesse R Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim,  
642 Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, Ginell Elliott, Tim R Mercer,  
643 Shane J Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C  
644 Sallari, Kyle T Siebenthall, Nicholas A Sinnott-Armstrong, Michael Stevens, Robert E  
645 Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E Beaudet, Laurie A Boyer, Philip L  
646 De Jager, Peggy J Farnham, Susan J Fisher, David Haussler, Steven J M Jones, Wei  
647 Li, Marco A Marra, Michael T McManus, Shamil Sunyaev, James A Thomson, Thea D  
648 Tlsty, Li-Huei Tsai, Wei Wang, Robert A Waterland, Michael Q Zhang, Lisa H Chad-  
649 wick, Bradley E Bernstein, Joseph F Costello, Joseph R Ecker, Martin Hirst, Alexan-  
650 der Meissner, Aleksandar Milosavljevic, Bing Ren, John A Stamatoyannopoulos, Ting  
651 Wang, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes.  
652 *Nature*, 518(7539):317–330, February 2015.

- 653 [2] ENCODE Project Consortium, Jill E Moore, Michael J Purcaro, Henry E Pratt,  
654 Charles B Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A Davis,  
655 Alexander Dobin, Rajinder Kaul, Jessica Halow, Eric L Van Nostrand, Peter Freese,  
656 David U Gorkin, Yin Shen, Yupeng He, Mark Mackiewicz, Florencia Pauli-Behn,  
657 Brian A Williams, Ali Mortazavi, Cheryl A Keller, Xiao-Ou Zhang, Shaimae I El-  
658 hajjajy, Jack Huey, Diane E Dickel, Valentina Snetkova, Xintao Wei, Xiaofeng Wang,  
659 Juan Carlos Rivera-Mulia, Joel Rozowsky, Jing Zhang, Surya B Chhetri, Jialing Zhang,  
660 Alec Victorsen, Kevin P White, Axel Visel, Gene W Yeo, Christopher B Burge, Eric  
661 Lécuyer, David M Gilbert, Job Dekker, John Rinn, Eric M Mendenhall, Joseph R  
662 Ecker, Manolis Kellis, Robert J Klein, William S Noble, Anshul Kundaje, Roderic  
663 Guigó, Peggy J Farnham, J Michael Cherry, Richard M Myers, Bing Ren, Brenton R  
664 Graveley, Mark B Gerstein, Len A Pennacchio, Michael P Snyder, Bradley E Bernstein,  
665 Barbara Wold, Ross C Hardison, Thomas R Gingeras, John A Stamatoyannopoulos,  
666 and Zhiping Weng. Expanded encyclopaedias of DNA elements in the human and  
667 mouse genomes. *Nature*, 583(7818):699–710, July 2020.
- 668 [3] Hendrik G Stunnenberg, International Human Epigenome Consortium, and Martin  
669 Hirst. The international human epigenome consortium: A blueprint for scientific col-  
670 laboration and discovery. *Cell*, 167(5):1145–1149, November 2016.
- 671 [4] Jordan A Ramilowski, Chi Wai Yip, Saumya Agrawal, Jen-Chien Chang, Yari Ciani,  
672 Ivan V Kulakovskiy, Mickaël Mendez, Jasmine Li Ching Ooi, John F Ouyang, Nick  
673 Parkinson, Andreas Petri, Leonie Roos, Jessica Severin, Kayoko Yasuzawa, Imad  
674 Abugessaisa, Altuna Akalin, Ivan V Antonov, Erik Arner, Alessandro Bonetti, Hide-  
675 masa Bono, Beatrice Borsari, Frank Brombacher, Christopher J F Cameron, Carlo Vit-  
676 torio Cannistraci, Ryan Cardenas, Melissa Cardon, Howard Chang, Josée Dostie,  
677 Luca Ducoli, Alexander Favorov, Alexandre Fort, Diego Garrido, Noa Gil, Juliette  
678 Gimenez, Reto Guler, Lusy Handoko, Jayson Harshbarger, Akira Hasegawa, Yuki  
679 Hasegawa, Kosuke Hashimoto, Norihito Hayatsu, Peter Heutink, Tetsuro Hirose, Ed-  
680 die L Imada, Masayoshi Itoh, Bogumil Kaczkowski, Aditi Kanhere, Emily Kawabata,  
681 Hideya Kawaji, Tsugumi Kawashima, S Thomas Kelly, Miki Kojima, Naoto Kondo,  
682 Haruhiko Koseki, Tsukasa Kouno, Anton Kratz, Mariola Kurowska-Stolarska, Andrew  
683 Tae Jun Kwon, Jeffrey Leek, Andreas Lennartsson, Marina Lizio, Fernando López-  
684 Redondo, Joachim Luginbühl, Shiori Maeda, Vsevolod J Makeev, Luigi Marchionni,  
685 Yulia A Medvedeva, Aki Minoda, Ferenc Müller, Manuel Muñoz-Aguirre, Mitsuyoshi  
686 Murata, Hiromi Nishiyori, Kazuhiro R Nitta, Shuhei Noguchi, Yukihiko Noro, Ramil  
687 Nurtdinov, Yasushi Okazaki, Valerio Orlando, Denis Paquette, Callum J C Parr, Owen  
688 J L Rackham, Patrizia Rizzu, Diego Fernando Sánchez Martínez, Albin Sandelin, Pil-  
689 lay Sanjana, Colin A M Semple, Youtaro Shibayama, Divya M Sivaraman, Takahiro  
690 Suzuki, Suzannah C Szumowski, Michihira Tagami, Martin S Taylor, Chikashi Terao,  
691 Malte Thodberg, Supat Thongjuea, Vidisha Tripathi, Igor Ulitsky, Roberto Verardo,  
692 Ilya E Vorontsov, Chinatsu Yamamoto, Robert S Young, J Kenneth Baillie, Alistair  
693 R R Forrest, Roderic Guigó, Michael M Hoffman, Chung Chau Hon, Takeya Kasukawa,  
694 Sakari Kauppinen, Juha Kere, Boris Lenhard, Claudio Schneider, Harukazu Suzuki,  
695 Ken Yagi, Michiel J L de Hoon, Jay W Shin, and Piero Carninci. Functional an-  
696 notation of human long noncoding RNAs via molecular phenotyping. *Genome Res.*,  
697 30(7):1060–1072, July 2020.
- 698 [5] GTEx Consortium, Laboratory, Data Analysis & Coordinating Center (LDACC)—

- 699 Analysis Working Group, Statistical Methods groups—Analysis Working Group,  
700 Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI,  
701 NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen  
702 Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank  
703 Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project  
704 Management, ELSI Study, Genome Browser Data Integration & Visualization—EBI,  
705 Genome Browser Data Integration & Visualization—UCSC Genomics Institute, Univer-  
706 sity of California Santa Cruz, Lead analysts:, Laboratory, Data Analysis & Coordinating  
707 Center (LDACC):, NIH program management:, Biospecimen collection:, Pathology:,  
708 eQTL manuscript working group:, Alexis Battle, Christopher D Brown, Barbara E En-  
709 gelhardt, and Stephen B Montgomery. Genetic effects on gene expression across human  
710 tissues. *Nature*, 550(7675):204–213, October 2017.
- 711 [6] Rik G H Lindeboom, Aviv Regev, and Sarah A Teichmann. Towards a human cell  
712 atlas: Taking notes from the past. *Trends Genet.*, April 2021.
- 713 [7] Jason Ernst and Manolis Kellis. Large-scale imputation of epigenomic datasets for  
714 systematic annotation of diverse human tissues. *Nat. Biotechnol.*, 33(4):364–376, April  
715 2015.
- 716 [8] Wei-Li Guo and De-Shuang Huang. An efficient method to transcription factor bind-  
717 ing sites imputation via simultaneous completion of multiple matrices with positional  
718 consistency. *Mol. Biosyst.*, 13(9):1827–1837, August 2017.
- 719 [9] Qian Qin and Jianxing Feng. Imputation for transcription factor binding predictions  
720 based on deep learning. *PLoS Comput. Biol.*, 13(2):e1005403, February 2017.
- 721 [10] Timothy J Durham, Maxwell W Libbrecht, J Jeffrey Howbert, Jeff Bilmes, and  
722 William Stafford Noble. PREDICTD PaRallel epigenomics data imputation with cloud-  
723 based tensor decomposition. *Nat. Commun.*, 9(1):1402, April 2018.
- 724 [11] Jacob Schreiber, Timothy Durham, Jeffrey Bilmes, and William Stafford Noble. Av-  
725ocado: a multi-scale deep tensor factorization method learns a latent representation of  
726 the human epigenome. *Genome Biol.*, 21(1):81, March 2020.
- 727 [12] Jacob Schreiber, Jeffrey Bilmes, and William Stafford Noble. Completing the EN-  
728CODE3 compendium yields accurate imputations across a variety of assays and human  
729 biosamples. *Genome Biol.*, 21(1):82, March 2020.
- 730 [13] Carles A Boix, Benjamin T James, Yongjin P Park, Wouter Meuleman, and Manolis  
731 Kellis. Regulatory genomic circuitry of human disease loci by integrative epigenomics.  
732 *Nature*, 590(7845):300–307, February 2021.
- 733 [14] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans,  
734 Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle,  
735 If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab  
736 Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward,  
737 Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacque-  
738 line Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress,  
739 Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje van Baren, Michael Brent, David Haus-  
740 sler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic

- 741 Guigó, and Tim J Hubbard. GENCODE: the reference human genome annotation for  
742 the ENCODE project. *Genome Res.*, 22(9):1760–1774, September 2012.
- 743 [15] FANTOM Consortium and the RIKEN PMI and CLST (DGT), Alistair R R Forrest,  
744 Hideya Kawaji, Michael Rehli, J Kenneth Baillie, Michiel J L de Hoon, Vanja Haberle,  
745 Timo Lassmann, Ivan V Kulakovskiy, Marina Lizio, Masayoshi Itoh, Robin Anders-  
746 son, Christopher J Mungall, Terrence F Meehan, Sebastian Schmeier, Nicolas Bertin,  
747 Mette Jørgensen, Emmanuel Dimont, Erik Arner, Christian Schmidl, Ulf Schaefer,  
748 Yulia A Medvedeva, Charles Plessy, Morana Vitezic, Jessica Severin, Colin A Sem-  
749 ple, Yuri Ishizu, Robert S Young, Margherita Francescato, Intikhab Alam, Davide  
750 Albanese, Gabriel M Altschuler, Takahiro Arakawa, John A C Archer, Peter Arner,  
751 Magda Babina, Sarah Rennie, Piotr J Balwierz, Anthony G Beckhouse, Swati Pradhan-  
752 Bhatt, Judith A Blake, Antje Blumenthal, Beatrice Bodega, Alessandro Bonetti, James  
753 Briggs, Frank Brombacher, A Maxwell Burroughs, Andrea Califano, Carlo V Cannis-  
754 traci, Daniel Carbajo, Yun Chen, Marco Chierici, Yari Ciani, Hans C Clevers, Emiliano  
755 Dalla, Carrie A Davis, Michael Detmar, Alexander D Diehl, Taeko Dohi, Finn Drabløs,  
756 Albert S B Edge, Matthias Edinger, Karl Ekwall, Mitsuhiro Endoh, Hideki Enomoto,  
757 Michela Fagiolini, Lynsey Fairbairn, Hai Fang, Mary C Farach-Carson, Geoffrey J  
758 Faulkner, Alexander V Favorov, Malcolm E Fisher, Martin C Frith, Rie Fujita, Shiro  
759 Fukuda, Cesare Furlanello, Masaaki Furino, Jun-Ichi Furusawa, Teunis B Geijtenbeek,  
760 Andrew P Gibson, Thomas Gingeras, Daniel Goldowitz, Julian Gough, Sven Guhl,  
761 Reto Guler, Stefano Gustincich, Thomas J Ha, Masahide Hamaguchi, Mitsuko Hara,  
762 Matthias Harbers, Jayson Harshbarger, Akira Hasegawa, Yuki Hasegawa, Takehiro  
763 Hashimoto, Meenhard Herlyn, Kelly J Hitchens, Shannan J Ho Sui, Oliver M Hof-  
764 mann, Ilka Hoof, Furni Hori, Lukasz Huminiecki, Kei Iida, Tomokatsu Ikawa, Boris R  
765 Jankovic, Hui Jia, Anagha Joshi, Giuseppe Jurman, Bogumil Kaczkowski, Chieko Kai,  
766 Kaoru Kaida, Ai Kaiho, Kazuhiro Kajiyama, Mutsumi Kanamori-Katayama, Artem S  
767 Kasianov, Takeya Kasukawa, Shintaro Katayama, Sachi Kato, Shuji Kawaguchi, Hi-  
768 roshi Kawamoto, Yuki I Kawamura, Tsugumi Kawashima, Judith S Kempfle, Tony J  
769 Kenna, Juha Kere, Levon M Khachigian, Toshio Kitamura, S Peter Klinken, Alan J  
770 Knox, Miki Kojima, Soichi Kojima, Naoto Kondo, Haruhiko Koseki, Shigeo Koyasu,  
771 Sarah Krampitz, Atsutaka Kubosaki, Andrew T Kwon, Jeroen F J Laros, Weonju Lee,  
772 Andreas Lennartsson, Kang Li, Berit Lilje, Leonard Lipovich, Alan Mackay-Sim, Ri-  
773 Ichiroh Manabe, Jessica C Mar, Benoit Marchand, Anthony Mathelier, Niklas Mejhert,  
774 Alison Meynert, Yosuke Mizuno, David A de Lima Morais, Hiromasa Morikawa, Mit-  
775 suru Morimoto, Kazuyo Moro, Efthymios Motakis, Hozumi Motohashi, Christine L  
776 Mummery, Mitsuyoshi Murata, Sayaka Nagao-Sato, Yutaka Nakachi, Fumio Nakahara,  
777 Toshiyuki Nakamura, Yukio Nakamura, Kenichi Nakazato, Erik van Nimwegen, Noriko  
778 Ninomiya, Hiromi Nishiyori, Shohei Noma, Shohei Noma, Tadasuke Noazaki, Soichi  
779 Ogishima, Naganari Ohkura, Hiroko Ohimiya, Hiroshi Ohno, Mitsuhiro Ohshima,  
780 Mariko Okada-Hatakeyama, Yasushi Okazaki, Valerio Orlando, Dmitry A Ovchinnikov,  
781 Arnab Pain, Robert Passier, Margaret Patrikakis, Helena Persson, Silvano Piazza,  
782 James G D Prendergast, Owen J L Rackham, Jordan A Ramilowski, Mamoon Rashid,  
783 Timothy Ravasi, Patrizia Rizzu, Marco Roncador, Sugata Roy, Morten B Rye, Eri Sai-  
784 jyo, Antti Sajantila, Akiko Saka, Shimon Sakaguchi, Mizuho Sakai, Hiroki Sato, Suzana  
785 Savvi, Alka Saxena, Claudio Schneider, Erik A Schultes, Gundula G Schulze-Tanzil,  
786 Anita Schwegmann, Thierry Sengstag, Guojun Sheng, Hisashi Shimoji, Yishai Shimoni,  
787 Jay W Shin, Christophe Simon, Daisuke Sugiyama, Takaai Sugiyama, Masanori Suzuki,

- 788 Naoko Suzuki, Rolf K Swoboda, Peter A C 't Hoen, Michihira Tagami, Naoko Taka-  
789 hashi, Jun Takai, Hiroshi Tanaka, Hideki Tatsukawa, Zuotian Tatum, Mark Thompson,  
790 Hiroo Toyodo, Tetsuro Toyoda, Elvind Valen, Marc van de Wetering, Linda M van den  
791 Berg, Roberto Verado, Dipti Vijayan, Ilya E Vorontsov, Wyeth W Wasserman, Shoko  
792 Watanabe, Christine A Wells, Louise N Winteringham, Ernst Wolvetang, Emily J  
793 Wood, Yoko Yamaguchi, Masayuki Yamamoto, Misako Yoneda, Yohei Yonekura, Shige-  
794 hiro Yoshida, Susan E Zabierowski, Peter G Zhang, Xiaobei Zhao, Silvia Zucchelli,  
795 Kim M Summers, Harukazu Suzuki, Carsten O Daub, Jun Kawai, Peter Heutink,  
796 Winston Hide, Tom C Freeman, Boris Lenhard, Vladimir B Bajic, Martin S Taylor,  
797 Vsevolod J Makeev, Albin Sandelin, David A Hume, Piero Carninci, and Yoshihide  
798 Hayashizaki. A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–  
799 470, March 2014.
- 800 [16] Sage Bionetworks. [no title]. <https://www.synapse.org/!Synapse:syn6131484/wiki/>.  
801 Accessed: 2021-5-12.
- 802 [17] Jacob Schreiber, Ritambhara Singh, Jeffrey Bilmes, and William Stafford Noble. A  
803 pitfall for machine learning methods aiming to predict across cell types. *Genome Biol.*,  
804 21(1):282, November 2020.
- 805 [18] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoutte, David S Johnson, Bradley E  
806 Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu.  
807 Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9(9):R137, September 2008.
- 808 [19] Jinwook Lee, Daniel Kim, Grey Cristoforo, Chuan-Sheng Foo, Chris Probert, Nathan  
809 Beley, and Anshul Kundaje. ENCODE ATAC-seq pipeline, December 2019.
- 810 [20] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nat.*  
811 *Methods*, 9(4):357–359, March 2012.
- 812 [21] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis,  
813 Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and  
814 Mark A DePristo. The genome analysis toolkit: a MapReduce framework for analyzing  
815 next-generation DNA sequencing data. *Genome Res.*, 20(9):1297–1303, September  
816 2010.
- 817 [22] Haley M Amemiya, Anshul Kundaje, and Alan P Boyle. The ENCODE blacklist:  
818 Identification of problematic regions of the genome. *Sci. Rep.*, 9(1):9354, June 2019.
- 819 [23] Jin Lee, J Seth Strattan, annashcherbina, Karl Sebby, Meenakshi Kagda, and Paul L  
820 Maurizio. ENCODE-DCC/chip-seq-pipeline2: v1.9.0, May 2021.
- 821 [24] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-  
822 Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009.
- 823 [25] Yaxing Zhao, Limsoon Wong, and Wilson Wen Bin Goh. How to do quantile normal-  
824 ization correctly for gene expression data analyses. *Sci. Rep.*, 10(1):15534, September  
825 2020.
- 826 [26] F William Townes and Rafael A Irizarry. Quantile normalization of single-cell RNA-seq  
827 read counts without unique molecular identifiers. *Genome Biol.*, 21(1):160, July 2020.

- 828 [27] Nicolas Bonhoure, Gergana Bounova, David Bernasconi, Viviane Praz, Fabienne Lam-  
829 mers, Donatella Canella, Ian M Willis, Winship Herr, Nouria Hernandez, Mauro De-  
830 lorentzi, and CycliX Consortium. Quantifying ChIP-seq data: a spiking method provid-  
831 ing an internal reference for sample-to-sample normalization. *Genome Res.*, 24(7):1157–  
832 1168, July 2014.
- 833 [28] Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari,  
834 Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, and  
835 Julia Zeitlinger. Base-resolution models of transcription-factor binding reveal soft motif  
836 syntax. *Nat. Genet.*, 53(3):354–366, March 2021.