

1 **A framework for summarizing chromatin state annotations within**  
2 **and identifying differential annotations across groups of samples**

3

4 Ha Vu <sup>1,2</sup>, Zane Koch <sup>2</sup>, Petko Fizev <sup>1,2,8</sup>, Jason Ernst <sup>1,2,3,4,5,6,7,\*</sup>

5 <sup>1</sup> Bioinformatics Interdepartmental Program, University of California, Los Angeles, CA, 90095, USA.

6 <sup>2</sup> Department of Biological Chemistry, University of California, Los Angeles, Los Angeles, CA 90095,  
7 USA

8 <sup>3</sup> Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at University of  
9 California, Los Angeles, Los Angeles, CA 90095, USA

10 <sup>4</sup> Computer Science Department, University of California, Los Angeles, Los Angeles, CA 90095, USA

11 <sup>5</sup> Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, CA 90095,  
12 USA

13 <sup>6</sup> Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90095, USA

14 <sup>7</sup> Computational Medicine Department, University of California, Los Angeles, Los Angeles, CA 90095,  
15 USA

16 <sup>8</sup> Illumina Inc.

17 \*To whom correspondence should be addressed

## 18 **Abstract**

19 **Motivation:** Genome-wide maps of epigenetic modifications are powerful resources for non-  
20 coding genome annotation. Maps of multiple epigenetics marks have been integrated into cell or  
21 tissue type-specific chromatin state annotations for many cell or tissue types. With the increasing  
22 availability of multiple chromatin state maps for biologically similar samples, there is a need for  
23 methods that can effectively summarize the information about chromatin state annotations within  
24 groups of samples and identify differences across groups of samples at a high resolution.

25 **Results:** We developed CSREP, which takes as input chromatin state annotations for a group of  
26 samples and then probabilistically estimates the state at each genomic position and derives a  
27 representative chromatin state map for the group. CSREP uses an ensemble of multi-class  
28 logistic regression classifiers to predict the chromatin state assignment of each sample given the  
29 state maps from all other samples. The difference of CSREP's probability assignments for two  
30 groups can be used to identify genomic locations with differential chromatin state patterns.

31 Using groups of chromatin state maps of a diverse set of cell and tissue types, we demonstrate  
32 the advantages of using CSREP to summarize chromatin state maps and identify biologically  
33 relevant differences between groups at a high resolution.

34 **Availability and implementation:** The CSREP source code is openly available under  
35 <http://github.com/ernstlab/csrep>.

36 Contact: [jason.ernst@ucla.edu](mailto:jason.ernst@ucla.edu)

## 37 **Introduction**

38 Genome-wide maps of chromatin marks such as histone modifications and variants  
39 provide valuable information for annotating non-coding genome features (Barski *et al.*, 2007;

40 Ernst *et al.*, 2011; Zhu *et al.*, 2013; Xie *et al.*, 2013). Efforts by large consortia and individual  
41 labs have produced chromatin state maps for many cell and tissue types (Roadmap Epigenomics  
42 Consortium *et al.*, 2015; Consortium, 2012; Zhu *et al.*, 2013; Barski *et al.*, 2007). A popular  
43 representation of such data is chromatin states defined by the combinatorial and spatial patterns  
44 of multiple marks, which are generated by methods such as ChromHMM and Segway (Libbrecht  
45 *et al.*, 2021)(Ernst and Kellis, 2010, 2012; Hoffman *et al.*, 2012), and correspond to diverse  
46 classes of genomic elements including various types of enhancers and promoters.

47 Chromatin state maps have been produced for hundreds of different biological samples.  
48 In many cases there are multiple samples representing similar cell and tissue types (Boix *et al.*,  
49 2021; Roadmap Epigenomics Consortium *et al.*, 2015). In such cases, to simplify analyses and  
50 visualizations, it may be desirable to have a single chromatin state annotation that summarizes  
51 the annotations for all samples from each group. A straightforward approach to this task is to  
52 take the most frequent chromatin state assigned at each position across samples. However, when  
53 the number of samples per group is small or the number of states is large, such an approach can  
54 be particularly vulnerable to noise. Furthermore, such an approach does not consider additional  
55 information available about the different chromatin states. For example, if a location was  
56 assigned to three different states in three samples, the summary annotation among these three  
57 states based on the frequency-based method would be arbitrary. However, by leveraging  
58 information about the co-occurrence of state assignments genome-wide, there is additional  
59 information to predict the most likely chromatin state annotation for a new sample from the  
60 group.

61 A related challenge is to identify differences in chromatin state annotations between two  
62 groups at a high resolution and on a per-state basis. Methods such as ChromDiff, chromswitch,

63 and EpiAlign (Yen and Kellis, 2015; Jessa and Kleinman, 2018; Ge *et al.*, 2019) can identify  
64 chromatin state differences between samples, but only calculate a measure of difference for a  
65 broad domain (e.g. a gene body), encompassing a large number of genomic bins for which the  
66 states are defined. Additionally, EpiAlign and Chromswitch are designed to measure the  
67 difference in annotations for one user-input query region in each run, and are not designed to  
68 generate genome-wide output, which is our focus. Another approach, EpiCompare (He and  
69 Wang, 2017) presented an approach for identifying differential enhancer chromatin states across  
70 cell or tissue types, but did not consider other types of chromatin states. SCIDDO (Ebert  
71 and Schulz, 2020) can detect genome-wide significant differential chromatin domains between  
72 two groups of samples while incorporating a measure of similarity among states. However,  
73 SCIDDO only provides a single differential score per position and does not directly answer the  
74 question of what chromatin state switch occurs at each genomic position. Another method, dPCA  
75 (Ji *et al.*, 2013), works directly on chromatin mark signals and does not quantify state differences  
76 across groups of samples.

77 To effectively summarize the chromatin state annotations for a group of samples and  
78 prioritize the chromatin state differences between two groups on a per-state basis, at high  
79 resolution, we introduce CSREP. CSREP leverages both the information about the input  
80 samples' chromatin states at a position being summarized, as well as information of states' co-  
81 occurrences in different samples within the same group across the genome. CSREP does this by  
82 first generating probabilistic estimates of chromatin state annotations by using an ensemble of  
83 multi-class logistic regression classifiers that predict the state assignment in a sample at a  
84 position, given the annotations in other samples, at the corresponding genomic position. From  
85 those predictions, CSREP is then able to produce a single summary state assignment per

86 position. CSREP can also use the difference of summary probabilistic predictions for two groups  
87 of samples to quantify the difference in state assignments between the two groups on a per-state  
88 basis, e.g. one genome-wide score track per chromatin state.

89 Using CSREP, we generated the summary chromatin state maps for 11 groups of  
90 tissue/cell types from Roadmap Epigenomic Project (Roadmap Epigenomics Consortium et al.,  
91 2015), and for 75 groups from the Epimap Portal (Boix *et al.*, 2021). We show that CSREP can  
92 better predict chromatin state assignments in held-out samples than a counting-based baseline  
93 method. We also verify that the resulting summary chromatin maps show correspondence with  
94 the group's average gene expression profile. Additionally, we show that CSREP's differential  
95 scores can recover differential epigenetic signals on chromosome X between male and female  
96 samples. We also show that CSREP differential scores between samples from two different  
97 tissue groups can predict regions of differential peaks for various chromatin marks. The CSREP  
98 implementation is designed to be user-friendly and includes a detailed tutorial, available at  
99 <https://github.com/ernstlab/csrep>. We expect CSREP will be a useful tool for summarizing  
100 chromatin state maps within groups and finding differences across groups. Additionally, the  
101 summary annotations for different tissue groups that we generated with CSREP are expected to  
102 be useful resource.

## 103 **Results**

### 104 **CSREP method**

105 CSREP takes as input chromatin state maps for a group of samples learned based on a  
106 concatenation approach (Ernst and Kellis, 2010, 2012) to ensure that annotations for different  
107 samples share chromatin state definitions. CSREP then generates as output (1) a summary  
108 probabilistic chromatin state assignment matrix and (2) a summary state map track for the group.

109 The summary state assignment matrix represents the probabilities of each state being present at  
110 each genomic position in a new sample of that group. To generate these, CSREP takes a  
111 supervised learning approach, leveraging information about the co-occurrence of states from the  
112 different samples across the genome. Specifically, for each group of input samples, CSREP  
113 trains an ensemble of multi-class logistic regression classifiers (Hastie *et al.*, 2009) to generate  
114 probabilistic predictions for each chromatin state at each position (**Fig. 1A, Methods**). We used  
115 multi-class logistic regression classifiers since they provide well calibrated probabilities, are  
116 robust, and relatively fast to train. Each classifier is trained with *labels* based on the chromatin  
117 state assignments from one sample and *features* based on the chromatin states in other samples  
118 for the same genomic positions. Each classifier then makes a probabilistic prediction of the  
119 chromatin state assigned at each genomic position in the target sample. The chromatin state input  
120 features to each logistic regression classifier are represented with a one-hot-encoding of the  
121 chromatin states. The classifiers are trained on randomly selected genomic positions that  
122 constitute 10% of the genome, while the predictions are calculated genome-wide. The resolution  
123 of predictions is the same as that of input samples' chromatin state maps (200bp with default  
124 settings from ChromHMM). The prediction result for each sample's chromatin state map are  
125 represented in a matrix with *rows* corresponding to genomic positions and *columns* chromatin  
126 states. The values in each row sum to 1, representing the probabilities of state assignments at a  
127 genomic position. The probabilistic summary of a group is based on averaging the prediction  
128 output matrices for each sample in the group. These probabilistic predictions are then used to  
129 generate a summary chromatin state map for the group of samples by assigning the state with  
130 maximum assignment probability to each genomic position.

131 CSREP's summary probabilistic predictions can be directly used to generate differential  
132 chromatin state maps for two groups with multiple samples. This is achieved by subtracting the  
133 summary chromatin state assignment matrices of one group (first group) from the other's  
134 (second group) (**Fig. 1B, Methods**). At each genomic position, CSREP's chromatin differential  
135 scores for individual chromatin states are bounded between -1 and 1, with a score of 1 in state  $S$   
136 meaning state  $S$  was predicted to be the annotation for the first and second groups with  
137 probabilities 1 and 0 respectively, and vice versa for -1 (**Fig. 1C**). Overall, in addition to  
138 summarizing the state assignments for groups of samples, CSREP can calculate scores of  
139 differential chromatin state assignments for pairs of groups at the resolution of the input  
140 chromatin state maps.

#### 141 **CSREP is predictive of chromatin states on held-out samples**

142 We applied CSREP to a compendium of 18-state chromatin state maps for 64 samples (reference  
143 epigenomes) from 11 tissue groups generated by the Roadmap Epigenomics Project (Roadmap  
144 Epigenomics Consortium et al., 2015). The tissue groups include embryonic stem cells (ESCs),  
145 induced pluripotent stem cells (iPSC), ESC-derived cells, blood & T-cells, HSC & B-cells,  
146 epithelial, brain, muscle, heart, smooth muscle and digestive. The numbers of input samples for  
147 each tissue group range from 3 to 12. We provide the CSREP's genomewide summary  
148 probabilistic and hard state assignments for 11 tissue groups (**Data availability**).

149 We first visualized CSREP's summary chromatin state maps for a group of samples from  
150 digestive and heart tissue groups, which have 10 and 3 samples, respectively (**Fig. 2A, Supp.**  
151 **Fig. 1-4**). For each group, we arbitrarily selected four 500-kb regions and visualized the input  
152 chromatin state maps and CSREP's output probabilistic estimates and summary state map at

153 such genomic windows. We observed expected correspondence between the groups' input and  
154 output chromatin state assignment estimates (**Fig. 2A, Supp. Fig. 1-4**).

155 To quantitatively evaluate CSREP's summary output for a group of samples, we  
156 evaluated the accuracy of CSREP's summary probabilistic chromatin state predictions in a leave-  
157 one-out cross-validation analysis. In particular, for each chromatin state, we calculated Area  
158 Under the Receiver Operating Characteristic (AUROC) curve for predicting genomic locations  
159 assigned to the state in the left-out sample from the group (**Supp. Methods**). We compared the  
160 performance of CSREP against a baseline method, denoted `base_count`, which counts each  
161 state's frequency across input samples at each genomic position (**Supp. Methods**).

162 CSREP showed strong predictive performance for chromatin states in left-out samples  
163 with average AUROCs across 64 samples varying from 0.871 to 0.993 for the 18 states. Across  
164 the 18 states, CSREP consistently had better AUROC in recovering individual states compared  
165 to the baseline method `base_count` (**Fig. 2B**). The average AUROC improvements by CSREP  
166 compared to `base_count` ranged from 0.003 (for state 18\_Quies) to 0.157 (for state 4\_  
167 TSSFlnkD). Larger performance improvements by CSREP relative to `base_count` are observed  
168 for all chromatin states when there are fewer input samples in the group (**Supp. Fig. 5**).

### 169 **CSREP summary chromatin state maps' association with gene expression**

170 Transcription start sites (TSS) are marked by different histone modifications and variants  
171 that can correlate gene transcription (Kimura, 2013; Soboleva *et al.*, 2014). Here, we evaluated  
172 how CSREP's summary state map for a tissue group is predictive of the group's gene expression  
173 profiles at transcription start sites (TSS) of genes. First, we obtained gene expression data for  
174 available samples for the 11 tissue groups as above, and calculated the average protein-coding  
175 gene expression for each group (**Supp. Methods**). We then calculated the Pearson correlation



176 between (1) the group's average expression for protein coding genes and (2) CSREP's summary  
177 state assignment probabilities for state 1\_TssA (active TSS state) at the corresponding genes'  
178 TSSs. We did the same evaluation for base\_count. CSREP had significantly higher correlations  
179 than base\_count (paired t-test p-value: 0.009, average 0.550 vs. 0.534, **Supp. Methods**). We next  
180 extended this analysis for a larger dataset for 552 samples in 75 groups from EpiMap repository  
181 based on state 1\_TssA from the same 18-state annotations (Boix *et al.*, 2021) (**Supp. Methods**).  
182 The 75 groups were previously formed based on tissue types and developmental stages with the  
183 number of samples per group ranging from 3 to 38 (**Supp. Methods, Data Availability**). Of the  
184 75 groups, 65 also had gene expression data available for at least one sample. Across these 65  
185 groups, again CSREP's had significantly higher correlations than base\_count (paired t-test p-val:  
186 5.5e-08, average 0.545 vs. 0.538, **Supp. Methods**). Overall, CSREP's summary chromatin state  
187 maps at TSS for the TssA state show significantly higher correspondence with gene expression  
188 levels compared to the base\_count method.

### 189 **CSREP detects differential chromatin regions associated with different sexes**

190 We next investigated the performance of CSREP at identifying biologically meaningful  
191 chromatin state changes between groups of male and female samples based on its ability to  
192 prioritize chromatin state differences on chromosome X (chrX) relative to autosomal  
193 chromosomes. Specifically, we applied CSREP to calculate differential chromatin state  
194 scores between 25 female and 44 male samples from Roadmap Epigenomics (**Supp. Methods**)  
195 (Yen and Kellis, 2015; Ge *et al.*, 2019) by subtracting CSREP's summary state probability  
196 matrix for the female samples from the corresponding matrix for the male samples.

197 We analyzed CSREP's differential scores for all chromatin states across autosomal  
198 chromosomes and chrX (**Fig. 3A, Supp. Fig 6-7**). Three states with the largest magnitude of

199 difference in mean scores between the sex chrX and autosomes were states 13\_Het  
200 (heterochromatin, marked by H3K9me3), 17\_ReprPCWk (weak polycomb repressed complex)  
201 and 18\_Quies (quiescent). In chrX, compared to autosomal chromosomes, the distribution of  
202 differential scores for states 13\_Het and 17\_ReprPCWk showed a larger tail of negative. ChrX's  
203 average score minus the autosomes' average score values for states 13\_Het and 17\_ReprPCWk  
204 were -0.039 and -0.054, respectively (**Supp. Fig. 7**), implying that on chromosome X, female  
205 samples are more often assigned to these states compared to male samples. State 18\_Quies  
206 showed the opposite trend with a difference of 0.11(**Fig. 3A, Supp. Fig. 7**). These results are  
207 consistent with sex-specific chrX inactivation, which is used in female mammals to achieve  
208 dosage compensation between the two sexes (Wutz, 2011; Yen and Kellis, 2015).

209 We next compared the performance of CSREP and other methods in recovering  
210 annotated transcription start sites (TSSs) on chrX, using the above-mentioned states, given  
211 varying numbers of input samples (**Supp. methods**) (**Fig. 3B**). To do this, we randomly selected  
212 30 subsets of size  $n$  male and  $n$  female samples from the set of available 44 male and 25 female  
213 samples, where  $n$  is varied within the set of 3, 5, 9, 12 or 15 samples. Given each set of input  
214 male and female samples, we calculated the receiver operating characteristic (ROC) curve when  
215 using differential chromatin scores between male and female groups to predict locations  
216 overlapping annotated TSSs on chrX, against the background of those overlapping all annotated  
217 TSSs in the genome (**Supp. Methods**). We observed that CSREP showed the largest advantage  
218 over base\_count, as measured by AUROCs, when the number of input samples from Male and  
219 Female groups is relatively small, e.g. 3 samples in each group (**Fig. 3B**). As the number of input  
220 samples from each group increases sufficiently, the overall gap of performance between CSREP  
221 and base\_count goes away. In all cases, CSREP and base\_count show better performance

222 compared to SCIDDO (Ebert and Schulz, 2020) (**Fig. 3B**). Overall, CSREP showed the greatest  
223 advantage over other approaches when the number of samples is relatively small, which occurs  
224 frequently in practice.

### 225 **CSREP differential scores recover differential chromatin mark peaks**

226 We next analyzed how well CSREP's, base-count's and SCIDDO's differential  
227 chromatin state scores can predict genomic regions overlapping differential signals of DNase I  
228 hypersensitivity (DNase), H3K9ac and H3K27ac between samples from embryonic stem cell  
229 (ESC) and brain. DNase and H3K9ac signals were not used for learning the 18-state model used  
230 to annotate the two groups' input samples, providing an independent validation. While H3K27ac  
231 was used in learning the input chromatin state maps, since all the methods being compared  
232 (CSREP, base\_count, SCIDDO) had access to the same maps as input, and H3K27ac is a well-  
233 established mark of cell-type specific activity (Creyghton *et al.*, 2010), we still considered  
234 H3K27ac in the evaluations of methods' performance.

235 For each of the three chromatin marks, we first obtained a set of bases that are present in  
236 peaks in all samples from ESC but not in any from the Brain group and vice versa (**Supp.**  
237 **Methods**). We then calculated CSREP and base\_count differential chromatin scores by  
238 subtracting the summary chromatin state map of Brain from that of the ESC. Additionally, we  
239 applied SCIDDO to the same set of input data (**Supp. Methods**). We evaluated, in terms of  
240 AUROC, how well the methods prioritize regions overlapping bases in the ESC-/brain-specific  
241 sets of peaks (**Supp. Methods**). For CSREP and base\_count, we conducted separate evaluations  
242 for each chromatin state, but did *not* for SCIDDO since it outputs one score track that measures  
243 the overall difference across the chromatin state landscape between the two groups.

244 Across the different marks being evaluated, the highest AUROCs were consistently from  
245 CSREP based on its scores for either from promoter or enhancer associated states (**Fig. 4**). For  
246 example, for identifying brain specific H3K9ac peaks, CSREP had an AUROC of 0.717 based  
247 on the evaluation with state 9\_EnhA1, an active enhancer state, while the maximum AUROC  
248 achieved for base\_count was 0.617 and SCIDDO's AUROC was 0.564. These analyses suggest  
249 that CSREP differential scores better correspond to locations of individual mark differences  
250 between two groups of samples genomewide, compared to other approaches that also aim to  
251 identify chromatin state differences between two groups. The advantage of CSREP over  
252 SCIDDO may in part be due to CSREP producing scores with respect to specific chromatin  
253 states and including the direction of change (with positive/negative scores implying one group's  
254 higher state assignment probabilities compared to the other's).

## 255 **Discussion**

256 Here, we proposed CSREP, a method for probabilistically summarizing the chromatin  
257 state maps from a group of samples. CSREP achieves this by training multi-class logistic  
258 regression models to predict the chromatin state annotations of one sample using data from  
259 others, and then averaging the prediction probabilities across all samples in the group. CSREP  
260 outputs the probabilities of each chromatin state being assigned to each genomic position, at the  
261 same resolution that chromatin states are annotated. We applied CSREP to generate summary  
262 18-state chromatin state assignment probability matrices for 11 groups of cell and tissue types  
263 from Roadmap Epigenomics Project (Roadmap Epigenomics Consortium et al., 2015), and 75  
264 groups of samples stratified by cell and tissue types and developmental phases from EpiMap  
265 (Boix *et al.*, 2021), and have made them publicly available (**Data Availability**).

266 Our analyses reveal that CSREP's probabilistic summary of state assignments better  
267 predicts the chromatin states of held out samples compared to the counting-based baseline  
268 approach. We also showed that CSREP's summary assignment probabilities of state 1\_TssA at  
269 TSS was well correlated with the average gene expression of the group, and significantly higher  
270 than those achieved by the counting-based baseline.

271 CSREP can also be used to directly quantify the difference in chromatin state maps  
272 between two groups with multiple samples, at the resolution of the input annotations. CSREP  
273 produces differential scores for each chromatin state at each genomic position, which represent  
274 the difference in probabilities that samples from two input groups are assigned to each specific  
275 state. Therefore, CSREP differential scores are bounded (-1 to 1), interpretable with respect to  
276 specific chromatin state changes, and indicative of the direction of change, which contrasts it  
277 with other approaches that provide a single score showing magnitude of difference per genomic  
278 position. We used CSREP to compare the chromatin state annotations between male and female  
279 samples from Roadmap Epigenomics (Roadmap Epigenomics Consortium et al., 2015), and  
280 showed that CSREP can better predict regions overlapping genes' TSS on chrX, particularly  
281 when there are few samples in each group. CSREP's differential scores for states associated with  
282 active enhancers and promoters better recovered tissue-group-specific peaks of  
283 DNase/H3K27ac/H3K9ac signals compared to alternative approaches, suggesting that CSREP  
284 provides useful additional information for analyzing epigenomic changes across tissue types.  
285 Future work could apply CSREP to compare additional biological conditions or disease state  
286 (e.g. cancer vs non-cancer).

287 CSREP works directly off of chromatin state annotations, which makes CSREP agnostic  
288 to the specific methods used to produce those annotations. Some methods for learning chromatin

289 state annotations have the option to expose posterior probability estimates of annotations, which  
290 could potentially be used in an extended version of CSREP. However, assuming accurately  
291 determined posterior probability estimates are available as input would also make CSREP less  
292 generally applicable.

293 To facilitate the use of CSREP, we provide an implementation of CSREP as a snakemake  
294 pipeline (Mölder *et al.*, 2021) with a detailed tutorial that only requires users to modify  
295 parameters in a yaml file. The program can be run either on local computers or on computing  
296 clusters, in which case snakemake will optimize the workflow for execution.

297 We expect CSREP to be a useful tool and the output we have provided from it a valuable  
298 resource for summarizing summarize chromatin state maps from groups of samples and  
299 prioritizing regions with differential chromatin state changes across pairs of groups of samples.

## 300 **Methods**

### 301 **CSREP's summarization of a group of samples**

302 Let  $G$  denote the number of genomic bins across the genome,  $S$  the number of chromatin  
303 states, and  $N$  the number of samples in the target group of samples. Let  $C_{i,n}$  denote the chromatin  
304 state assigned to sample  $n$  at genomic position  $i$ , which can take one value of  $1, 2, \dots, S$ . Let  
305  $N_n$  denote the set of samples not including  $n$ , i.e.  $N_n = \{1, \dots, N\} - \{n\}$ . In general, CSREP is an  
306 ensemble of  $N$  multi-class logistic regression classifiers such that for each sample  $n$ , CSREP  
307 trains a classifier to predict the chromatin state map of this sample based on features in the  
308 remaining samples ( $N_n$ ). The predictor variables for such a model include one-hot encoding  
309 chromatin state maps of the  $N - 1$  samples (all samples in the group except  $n$ ) and an intercept  
310 term, resulting in  $(N - 1) * S + 1$  predictor variables. The response variable is the chromatin  
311 state of the target sample  $n$ , which can take one value of  $1, 2, \dots, S$ .

312 In the multi-class logistic regression model, let  $X_i$  denote the vector of predictor variables  
313 at position  $i$ , which has length  $(N-1)*S + 1$  and takes values  $\{0,1\}$ . The last entry of  $X_i$  is 1,  
314 corresponding to the intercept term. Let  $Y_i$  denote the value of the response variable at position  $i$ ,  
315 which takes values  $\{1,2,\dots,S\}$ . Since the input chromatin state maps segment the genome into  
316 200-bp bins, we refer to each genomic position as one 200-bp window in the genome. We  
317 randomly selected genomic positions for the training data set, such that these positions constitute  
318 10% of the genome. Given the training data set, for each state  $s \in \{1, \dots, S - 1\}$ , the multi-class  
319 logistic regression model learns a coefficient vector  $\beta_s$  with length  $(N - 1) * S + 1$ ,  
320 corresponding to the number of predictor variables. The probability of sample  $n$ 's chromatin  
321 state  $s$  being assigned at position  $i$  is calculated as:

$$P(Y_i = s) = \frac{e^{\beta_s * X_i}}{1 + \sum_{j=1}^{S-1} e^{\beta_j * X_i}}$$

322 for  $s \in \{1, \dots, S - 1\}$ , and as the following when  $s = S$ :

$$P(Y_i = S) = \frac{1}{1 + \sum_{j=1}^{S-1} e^{\beta_j * X_i}}$$

323 After CSREP trains the multi-class logistic regression model on training data that constitute 10%  
324 of the genome, and  $l_2$ -norm penalty. The model is implemented using Python's sklearn,  
325 pybedtools package and snakemake (Dale *et al.*, 2011; Quinlan and Hall, 2010; Mölder *et al.*,  
326 2021). CSREP applies the model to generate predictions of genome-wide probabilistic chromatin  
327 state map for sample  $n$ , which is presented in a matrix of size  $G * S$ . The output matrices from  
328  $N$  predictions for  $N$  samples are then averaged, so at each genomic bin, the sum of state  
329 assignment probabilities across  $S$  states is 1. In addition, the chromatin state with the maximum  
330 probability in each row is recorded to produce a single representative chromatin state map for the  
331 entire group of samples.

## 332 **CSREP's application to prioritizing differential chromatin state changes between two** 333 **groups of samples**

334 To calculate differential chromatin state maps between two groups of samples, group1 and  
335 group2, CSREP first calculates the probabilistic chromatin state map matrices for each group as  
336 described above, denoted as  $R_1$  and  $R_2$ , respectively. After this, CSREP subtracts the two  
337 matrices to represent the differential chromatin state map between group1 and group2 (denoted  
338  $D_{12}$ ), i.e.  $D_{12} = R_1 - R_2$ . We note that we used signed and not absolute difference here and  
339 thus the score range from  $-1$  to  $1$ . A score on row  $i$  and column  $s$  of  $D_{12}$ , denoted  $D_{12,i,s}$ , being  
340  $-1$  means group2 is estimated to have probability 1 of being assigned to state  $s$  at position  $i$   
341 while group1 has probability of 0. Additionally, since CSREP assigns  $S$  scores of differential  
342 chromatin maps to each genomic position  $i$ , corresponding to  $S$  states, CSREP can uncover  
343 specific chromatin states switch. For example, if  $D_{12,i,s} = 0.8$  when  $s = 1$  while  $D_{12,i,s} =$   
344  $-0.8$  when  $s = 2$ , we can say it is likely that at position  $i$ , group1 is more likely to be in state 1  
345 while group2 is likely to be in state 2.

## 346 **Data availability**

347 The summary chromatin state maps (the chromatin state assignment matrices and the  
348 corresponding state annotation) for 11 tissue groups in Roadmap Project and 75 groups in  
349 Epimap Portal are available for download at <https://github.com/ernstlab/csrep>. The summary  
350 state maps for samples in Roadmap Epigenomics and EpiMap are provided both in hg38 and in  
351 hg19.

## 352 **Acknowledgements**

353 We thank all members of the Ernst lab for their helpful suggestions on the manuscript.

## 354 **Funding**

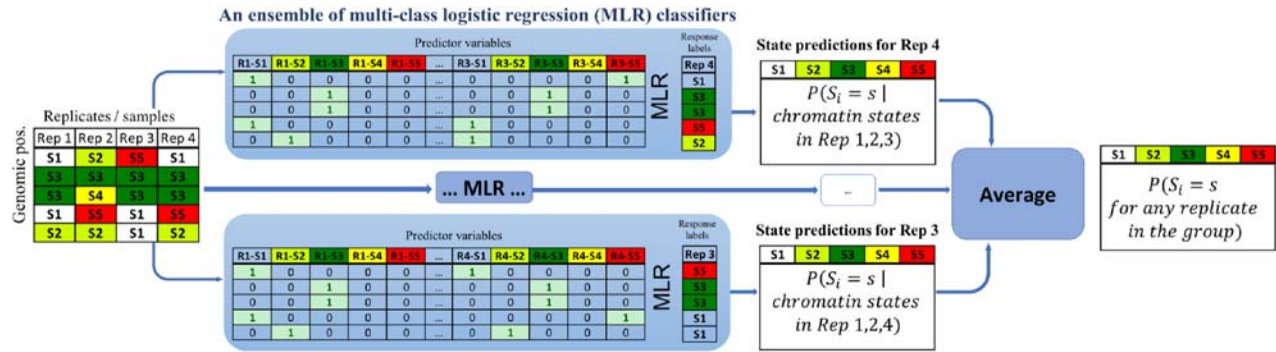


355 US National Institute of Health (DP1DA044371, U01MH105578, UH3NS104095, U01HG012079); US  
356 National Science Foundation (1254200, 1705121, 2125664); a Rose Hills Innovator Award, and the  
357 UCLA Jonsson Comprehensive Cancer Center and Eli and Edythe Broad Center of Regenerative  
358 Medicine and Stem Cell Research Ablon Scholars Program.

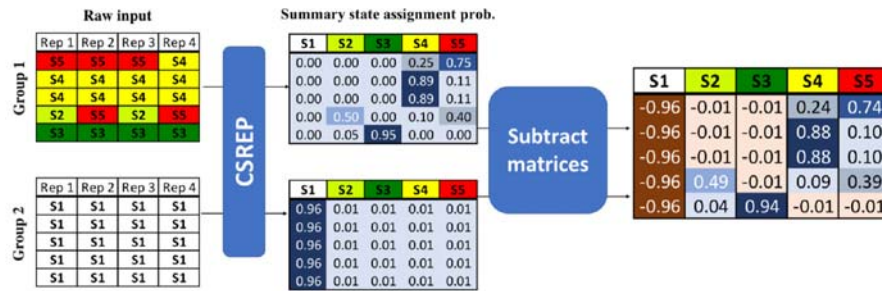
## 359 **References**

- 360 Barski,A. *et al.* (2007) High-resolution profiling of histone methylations in the human genome.  
361 *Cell*, **129**, 823–837.
- 362 Boix,C.A. *et al.* (2021) Regulatory genomic circuitry of human disease loci by integrative  
363 epigenomics. *Nature*, **590**, 300–307.
- 364 Consortium,E.P. (2012) An integrated encyclopedia of DNA elements in the human genome.  
365 *Nature*, **489**, 57–74.
- 366 Creighton,M.P. *et al.* (2010) Histone H3K27ac separates active from poised enhancers and  
367 predicts developmental state. *Proc. Natl. Acad. Sci.*, **107**, 21931–21936.
- 368 Dale,R.K. *et al.* (2011) Pybedtools: a flexible Python library for manipulating genomic datasets  
369 and annotations. *Bioinformatics*, **27**, 3423–3424.
- 370 Ebert,P. and Schulz,M.H. (2020) Fast detection of differential chromatin domains with SCIDDO.  
371 *Bioinformatics*.
- 372 Ernst,J. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell  
373 types. *Nature*, **473**, 43–49.
- 374 Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and  
375 characterization. *Nat. Methods*, **9**, 215–216.
- 376 Ernst,J. and Kellis,M. (2010) Discovery and characterization of chromatin states for systematic  
377 annotation of the human genome. *Nat. Biotechnol.*, **28**, 817–825.
- 378 Ge,X. *et al.* (2019) EpiAlign: an alignment-based bioinformatic tool for comparing chromatin  
379 state sequences. *Nucleic Acids Res.*, **47**, e77–e77.
- 380 Hastie,T. *et al.* (2009) The elements of statistical learning: data mining, inference, and prediction  
381 Springer.
- 382 He,Y. and Wang,T. (2017) EpiCompare: an online tool to define and explore genomic regions  
383 with tissue or cell type-specific epigenomic features. *Bioinformatics*, **33**, 3268–3275.
- 384 Hoffman,M.M. *et al.* (2012) Unsupervised pattern discovery in human chromatin structure  
385 through genomic segmentation. *Nat. Methods*, **9**, 473.
- 386 Jessa,S. and Kleinman,C.L. (2018) Chromswitch: a flexible method to detect chromatin state  
387 switches. *Bioinformatics*, **34**, 2286–2288.
- 388 Ji,H. *et al.* (2013) Differential principal component analysis of ChIP-seq. *Proc. Natl. Acad. Sci.*,  
389 **110**, 6789–6794.
- 390 Kimura,H. (2013) Histone modifications for human epigenome analysis. *J. Hum. Genet.*, **58**,  
391 439–445.
- 392 Roadmap Epigenomics Consortium *et al.* (2015) Integrative analysis of 111 reference human  
393 epigenomes. *Nature*, **518**, 317–330.
- 394 Libbrecht,M.W. *et al.* (2021) Segmentation and genome annotation algorithms for identifying  
395 chromatin state and other genomic patterns. *PLoS Comput. Biol.*, **17**, e1009423.
- 396 Mölder,F. *et al.* (2021) Sustainable data analysis with Snakemake. *F1000Research*, **10**.

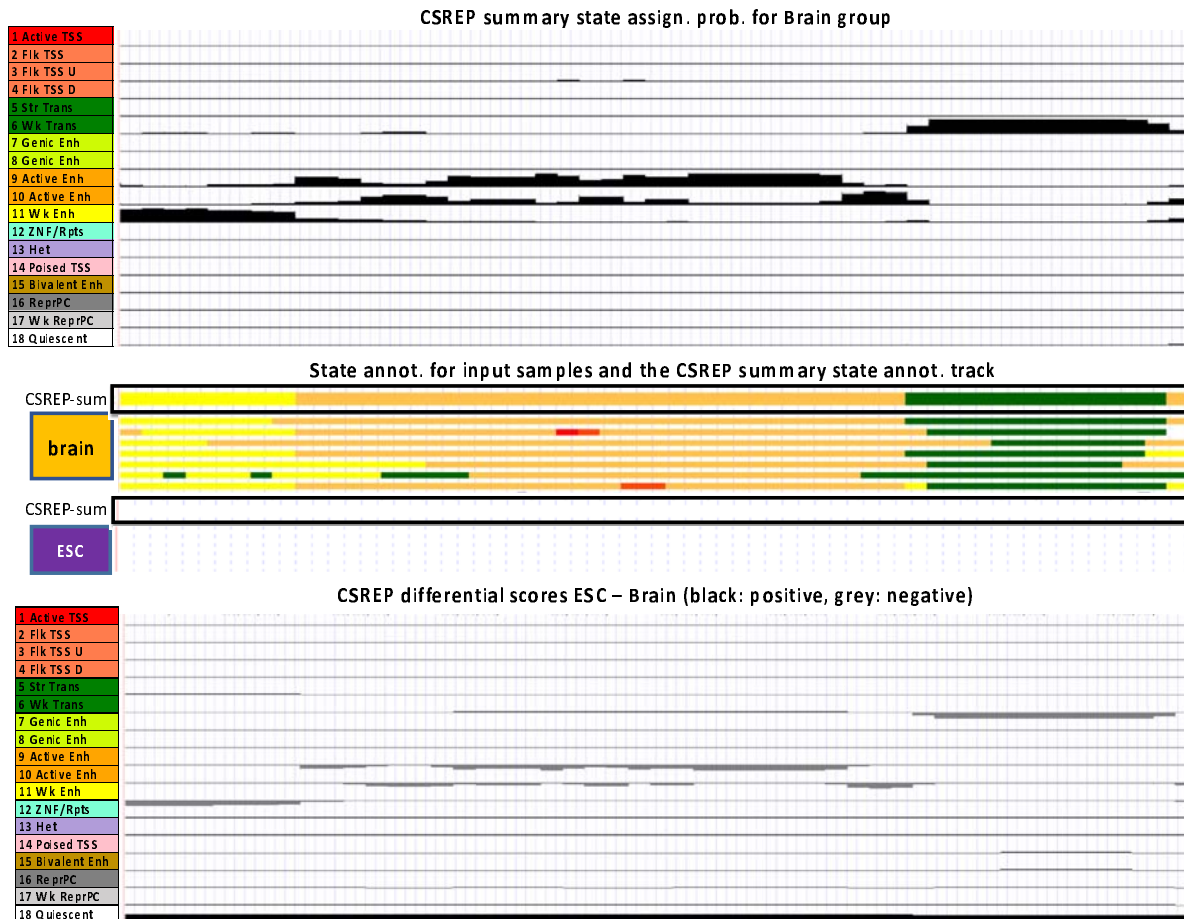
- 397 Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic  
398 features. *Bioinformatics*, **26**, 841–842.
- 399 Soboleva,T.A. *et al.* (2014) Histone variants at the transcription start-site. *Trends Genet.*, **30**,  
400 199–209.
- 401 Wutz,A. (2011) Gene silencing in X-chromosome inactivation: advances in understanding  
402 facultative heterochromatin formation. *Nat. Rev. Genet.*, **12**, 542–553.
- 403 Xie,W. *et al.* (2013) Epigenomic analysis of multilineage differentiation of human embryonic  
404 stem cells. *Cell*, **153**, 1134–1148.
- 405 Yen,A. and Kellis,M. (2015) Systematic chromatin state comparison of epigenomes associated  
406 with diverse properties including sex and tissue type. *Nat. Commun.*, **6**, 1–13.
- 407 Zhu,J. *et al.* (2013) Genome-wide chromatin state transitions associated with developmental and  
408 environmental cues. *Cell*, **152**, 642–654.
- 409



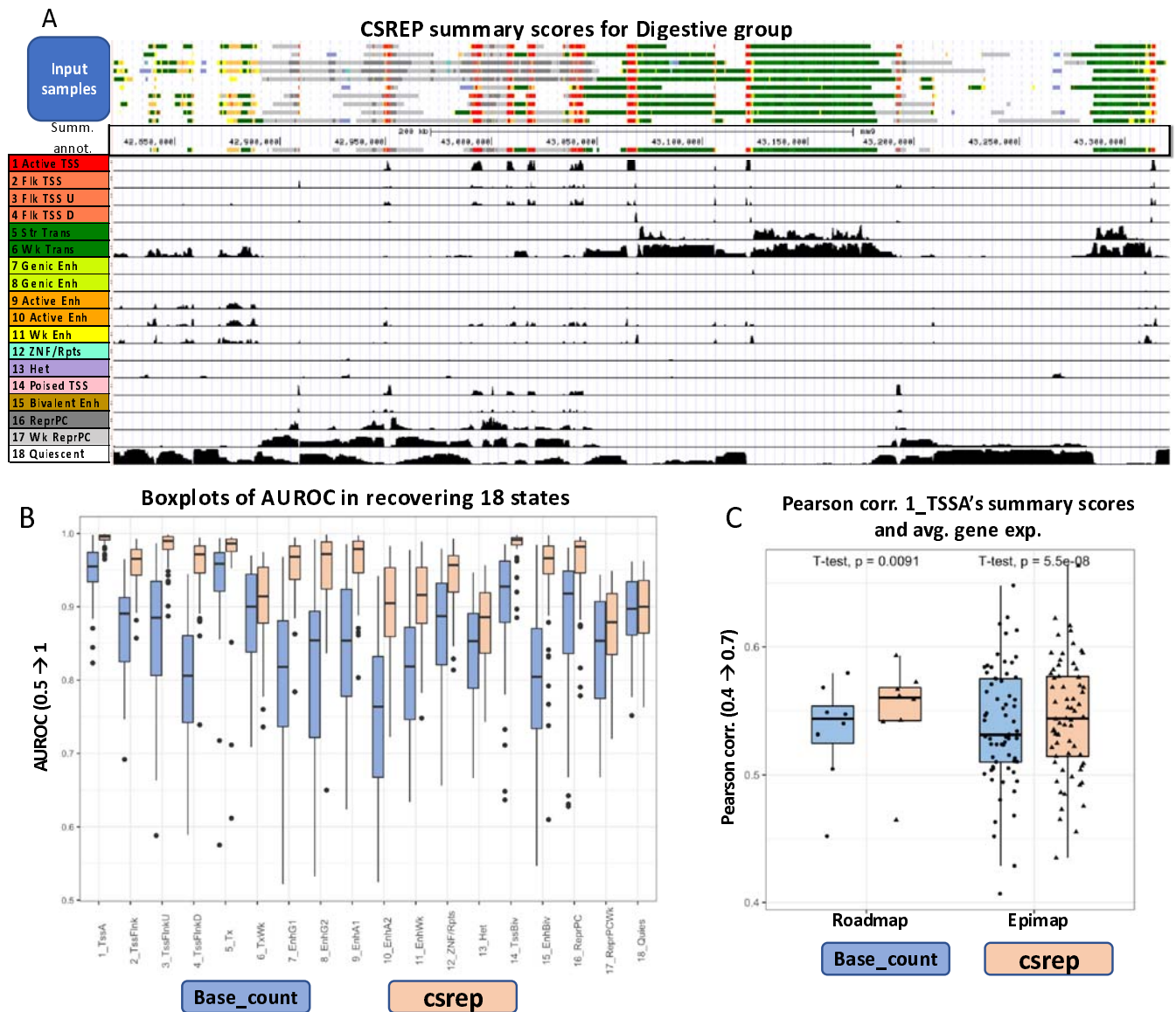
**B Procedure to calculate the differential chromatin state scores**



**C ESC and Brain samples' input chromatin state maps and CSREP's output for regions chr5:156012600-156022400**



411 **Fig. 1: Overview of CSREP.** (A) CSREP uses an ensemble of multi-class logistic regression models. In  
412 each model, the chromatin state map at the target sample is predicted based on the one-hot encoding of  
413 chromatin state assignments at the corresponding genomic positions in other samples. Multi-class logistic  
414 regression outputs the probabilities that each genomic position (row) in the target sample will be assigned  
415 to each state (column). CSREP averages the prediction matrices for target samples, to output the summary  
416 state assignment probability matrix. (B) The operations to obtain differential chromatin state assignment  
417 scores between two groups with multiple samples. CSREP calculates the summary chromatin state  
418 assignment matrices for two groups, and subtracts one group's summary matrix from the other's to obtain  
419 differential chromatin scores. Different chromatin scores are bounded between -1 (brown) and 1 (blue).  
420 (C) Visualization of CSREP's output in a genomic region (hg19, chr5:156,012,600-156,022,400). The top  
421 of the subpanel shows the CSREP's summary chromatin state probabilities for 18 states across seven  
422 Brain reference epigenomes. Each track shows the probabilities of assignment for one state, as named and  
423 colored on the left. The middle subpanel shows the 18-state chromatin state maps for 7 Brain samples and  
424 5 ESC samples from Roadmap Epigenomics (Roadmap Epigenomics Consortium et al., 2015), and the  
425 CSREP's output summary chromatin state maps for each group, outlined in black. States are colored as in  
426 legend as at the top of this subpanel. The last subpanel shows the differential chromatin scores when  
427 ESC's summary state probabilities are subtracted from Brain's. Each track shows one state's differential  
428 scores. Scores between 0 and 1 are colored black, while those between -1 and 0 are colored grey.  
429



430

431 **Fig. 2: Performance of CSREP in summarizing a group with multiple samples' chromatin state**

432 **maps. (A)** Visualization of one arbitrarily selected 500-kb region (chr5: 42,821,109-43,321,109, hg19).

433 The first 10 tracks show chromatin state maps of 10 input samples from the Roadmap Epigenomics

434 Consortium of the Digestive group, which were input to CSREP. The following track shows the summary

435 chromatin state map from CSREP, which shows strong agreement with the input. States are colored based

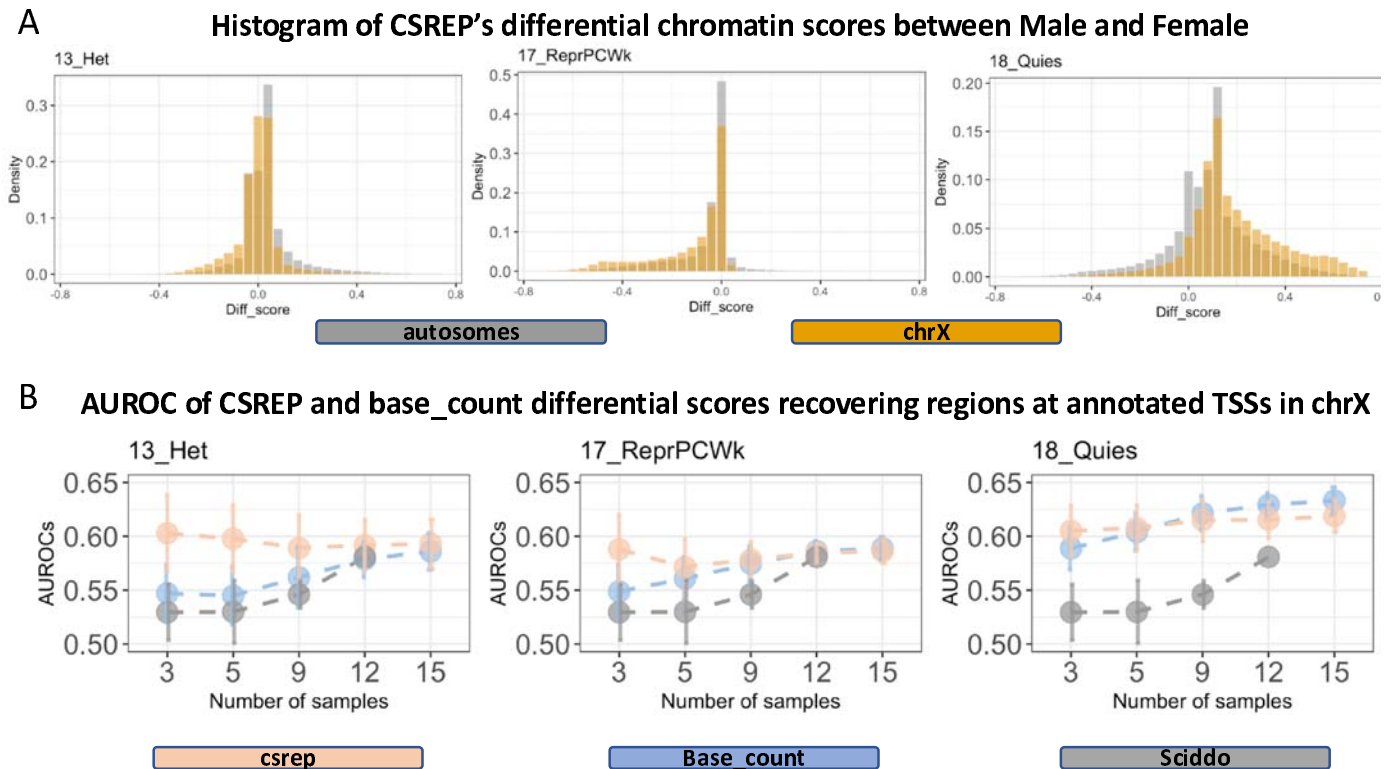
436 on the legend on the lower left. In the following 18 tracks, each track shows CSREP's probabilities of

437 assignment for each of 18 states, with state annotations in legend on left.

438 **(B)** Boxplots showing the CSREP and base\_count methods' average, range and 25, 75% quantiles of the  
439 AUROCs across 64 samples, for each of the 18 chromatin states. The AUROCs were calculated in leave-  
440 one-out cross validation analysis where we used a group's summary probabilistic chromatin state map to  
441 predict genomic locations of individual chromatin states in a left-out sample from the same cell/tissue  
442 group (**Supp. Methods**). States 1-18 (x-axis) are annotated as in **(A)**.

443 **(C)** Boxplots showing the Pearson correlations between a group of samples' (1) summary probabilities of  
444 state 1\_TssA (active TSS) at annotated TSSs, and (2) the corresponding groups' average gene expression  
445 (**Supp. Methods**). We obtained the correlations for 11 groups of cell types from the Roadmap  
446 Epigenomics Project, and 65 groups from EpiMap. Each dot shows the Pearson correlation for data from  
447 a group of samples.

448



449

450 **Fig. 3: CSREP show signals of differential chromatin state scores in chromosome X when**

451 **comparing male and female samples.** (A) Each subpanel shows the histogram of CSREP's differential

452 scores in autosomes and chromosome X, for states associated with heterochromatin (13\_Het), weak

453 polycomb repressed domains (17\_ReprPCWk), and quiescent regions (18\_Quies). The x-axis shows

454 differential scores, with positive values implying male samples have higher probabilities of being in the

455 state compared to female samples, and vice versa for negative values. Histograms of scores for all states

456 are in **Supp. Fig 6.** (B) AUROCs of recovering regions overlapping annotated TSSs on chromosome X,

457 using differential chromatin scores of three states as in (A), outputted by CSREP and base\_count for Male

458 and Female groups (**Supp. Methods**). We calculated the AUROCs using different sets of input male and

459 female samples, with varying number of samples in each group (x-axis). For each number of samples (x-

460 axis), we conducted the analysis for 30 sets of male and female input samples (**Supp. Methods**). The

461 plots show the average (dots) and standard deviation (error bars) of the AUROCs across the 30 sets of

462 input samples. SCIDDO did not successfully generate output for the case of 15 input samples so no  
 463 results are reported for that.

**AUROC for predicting Brain-/ESC- specific peaks of chromatin marks**

state	DNase peaks				H3K27ac peaks				H3K9ac peaks			
	Brain-spec		ESC-spec		Brain-spec		ESC-spec		Brain-spec		ESC-spec	
	csrep	base	csrep	base	csrep	base	csrep	base	csrep	base	csrep	base
1 Active TSS	0.540	0.506	0.595	0.502	0.584	0.524	0.702	0.513	0.587	0.527	0.659	0.548
2 Flk TSS	0.536	0.510	0.601	0.506	0.552	0.512	0.664	0.504	0.532	0.519	0.648	0.547
3 Flk TSS U	0.549	0.509	0.615	0.509	0.613	0.521	0.725	0.526	0.633	0.530	0.684	0.544
4 Flk TSS D	0.540	0.509	0.645	0.512	0.578	0.514	0.726	0.514	0.570	0.517	0.693	0.523
5 Str Trans	0.512	0.498	0.452	0.486	0.579	0.490	0.494	0.501	0.511	0.478	0.413	0.481
6 Wk Trans	0.490	0.483	0.542	0.496	0.557	0.527	0.579	0.512	0.436	0.440	0.449	0.463
7 Genic Enh	0.547	0.501	0.497	0.504	0.722	0.534	0.554	0.521	0.680	0.529	0.497	0.509
8 Genic Enh	0.566	0.502	0.498	0.501	0.720	0.507	0.563	0.507	0.710	0.516	0.529	0.511
9 Active Enh	0.568	0.529	0.584	0.546	0.740	0.582	0.712	0.623	0.717	0.617	0.606	0.584
10 Active Enh	0.564	0.537	0.594	0.521	0.729	0.644	0.738	0.615	0.700	0.613	0.623	0.530
11 Wk Enh	0.539	0.524	0.636	0.614	0.689	0.583	0.729	0.663	0.621	0.540	0.577	0.567
12 ZNF/Rpts	0.516	0.509	0.465	0.496	0.517	0.508	0.457	0.491	0.517	0.508	0.429	0.490
13 Het	0.474	0.504	0.399	0.484	0.414	0.504	0.310	0.477	0.447	0.507	0.341	0.478
14 Poised TSS	0.560	0.501	0.527	0.514	0.617	0.485	0.533	0.500	0.625	0.491	0.583	0.538
15 Bivalent Enh	0.556	0.500	0.529	0.513	0.618	0.496	0.531	0.501	0.610	0.496	0.484	0.489
16 ReprPC	0.521	0.499	0.469	0.490	0.505	0.487	0.469	0.491	0.517	0.492	0.395	0.453
17 Wk ReprPC	0.492	0.504	0.464	0.467	0.344	0.445	0.406	0.473	0.344	0.455	0.389	0.442
18 Quiescent	0.441	0.447	0.431	0.417	0.316	0.350	0.296	0.294	0.338	0.390	0.386	0.384
sciddo	0.517		0.549		0.562		0.535		0.564		0.585	

464  
 465 **Fig. 4: CSREP better recovers differential chromatin marks signals between ESC and Brain.** The  
 466 table shows AUROCs for differential scores' predictions of genomic regions associated with differential  
 467 peak signals for one chromatin mark, from left to right: DNase, H3K27ac and H3K9ac. For each  
 468 chromatin mark, it shows the AUROCs of predicting signal peaks observed in Brain and ESC exclusively  
 469 (Brain-spec and ESC-spec). Differential scores outputted by CSREP or baseline are shown for each  
 470 chromatin state (rows). In each category of comparisons, the top three scores that show highest AUROCs  
 471 are highlighted in green. Along the bottom is the AUROC for SCIDDO.

472  
 473

474  
 475

476  
 477

478



479

480