CoRE-ATAC: A Deep Learning model for the functional Classification of Regulatory Elements from single cell and bulk ATAC-seq data

Asa Thibodeau[1], Shubham Khetan[1,2], Alper Eroglu[1], Ryan Tewhey[3], Michael L. Stitzel[1, 4], Duygu Ucar[1, 4*]

[1] The Jackson Laboratory for Genomic Medicine, Farmington, CT, 06032, USA.

[2] Department of Genetics and Genome Sciences, University of Connecticut Health Center, Farmington, CT, 06030, USA.

[3] The Jackson Laboratory, Bar Harbor, ME, 04609, USA.

[4] Institute for Systems Genomics, University of Connecticut Health Center, Farmington, CT, 06030, USA

*To whom correspondence should be addressed.

Abstract

Cis-Regulatory elements (cis-REs) include promoters, enhancers, and insulators that regulate gene expression programs *via* binding of transcription factors. ATAC-seq technology effectively identifies active *cis*-REs in a given cell type (including from single cells) by mapping the accessible chromatin at base-pair resolution. However, these maps are not immediately useful for inferring specific functions of *cis*-REs. For this purpose, we developed a deep learning framework (CoRE-ATAC) with novel data encoders that integrate DNA sequence (reference or personal genotypes) and ATAC-seq read pileups. CoRE-ATAC was trained on 4 cell types (n=6 samples/replicates) and accurately predicted known *cis*-RE functions from 7 cell types (n=40 samples) that were not used in model training (average precision=0.80). CoRE-ATAC enhancer predictions

from 19 human islets coincided with genetically modulated gain/loss of enhancer activity, which was confirmed by massively parallel reporter assays (MPRAs). Finally, CoRE-ATAC effectively inferred functionality of *cis*-REs from aggregate single nucleus ATAC-seq (snATAC) data from human blood-derived immune cells that overlapped well with known functional annotations in sorted immune cells. Performances on snATAC-seq data demonstrate CoRE-ATAC's ability to infer *cis*-RE function in rare cell populations that can be identified by unsupervised clustering of snATAC-seq cells but difficult to capture in bulk ATAC-seq. ATAC-seq maps from primary human cells reveal individual- and cell-specific variation in cis-RE activity. CoRE-ATAC increases the functional resolution of these maps, a critical step for studying regulatory disruptions behind diseases.

Keywords: deep learning, ATAC-seq, snATAC-seq, *cis*-REs, enhancers, insulators

Background

Cis-Regulatory Elements (*cis*-REs) are non-coding DNA sequences that can be bound by transcription factors (TFs) and can take on different functional roles (e.g., promoter, enhancer, or insulator) to regulate gene expression programs. Over 88% of disease associated single nucleotide polymorphisms (SNPs) from genome-wide association studies (GWAS) are within non-coding region of the genome[1]. In particular, these SNPs typically fall within cell-specific enhancer sequences and indirectly disrupt the gene expression programs[2-4]. It is therefore critical to map *cis*-REs and their functions with increased precision to study how GWAS SNPs disrupt gene regulation in different cell types. Furthermore, genetic variation can impact the activity level of *cis*-REs[5-8].

Uncovering the functionality of such genetically-modulated *cis*-REs will guide *in-vivo* and *in-vitro* functional studies. Furthermore, identifying *cis*-RE functions in clinical samples will help uncover disease-associated loci at the individual level and their functional roles in pathogenesis.

ENCODE[9] and Roadmap[10] consortia successfully annotated *cis*-REs for 127 reference human cell/tissue types by profiling their epigenomes and analyzing them using a hidden markov model based approach: ChromHMM[11]. ChromHMM integrates Chromatin Immunoprecipitation with sequencing (ChIP-seq) profiles of multiple histone modification marks and TFs to demarcate promoters, enhancers, insulators and other functional types. These reference epigenomes are valuable resources to uncover and study disease-relevant and cell-type-specific *cis*-REs. However, these maps serve as references and do not capture individual- or condition-specific (e.g., activated cells) *cis*-REs. Moreover, these references do not include *cis*-REs of less frequently studied and/or rarer cells that are gaining attention with the advances in single cell profiling techniques. Although ChromHMM is very effective in inferring *cis*-RE functions, it requires five or more ChIP-seq assays to be generated from the same sample, which is not always feasible due to the cost (both antibody and sequencing) and cell numbers required for ChIP-seq assays (10^4-10^6 cells per experiment).

An alternative strategy for genome-wide interrogation of *cis*-REs is through chromatin accessibility profiling, which identifies open chromatin regions that are accessible for TF binding. One of the most recent and most frequently used methods for identifying open

70 chromatin regions is Assay for Transposase Accessible Chromatin using Sequencing
71 (ATAC-seq)[12, 13]. ATAC-seq utilizes Tn5 transposase to cleave accessible DNA into
72 fragments, which are sequenced to generate a genome-wide map of open chromatin (Fig.
73 1 Top). The low input material needed to generate ATAC-seq libraries allows it to be
74 applied in clinical samples, making it an ideal assay for studying *cis*-REs in individual
75 epigenomes in health and disease. Furthermore, recent developments have enabled the
76 generation of high-quality ATAC-seq maps from single cells (i.e., single nuclei ATAC-seq
77 (snATAC))[14] and generate epigenomic maps from clinical samples of limited quantity.
78 Obtaining snATAC data provides opportunities to 1) study a single cell's chromatin
79 accessibility and 2) analyze rare cell populations that do not have distinct cell markers to
80 identify them using cell sorting methods for bulk ATAC-seq. ATAC-seq is widely used to
81 infer *cis*-REs from diverse cell types and cell states [6, 12, 15-20]. However, functionality
82 of these *cis*-REs (e.g., enhancers, promoters, insulators) cannot be inferred from these
83 assays without the help of computational methods.
84
85 Previously we developed a method for predicting enhancers from bulk ATAC-seq data
86 using a neural network (NN) model on extracted data features (n=24), namely PEAS[21].
87 Although PEAS was effective in predicting enhancers from ATAC-seq maps, it was based
88 on a supervised approach for feature extraction; potentially missing novel predictive
89 features that are hidden within the read count distribution of ATAC-seq data. Furthermore,
90 PEAS did not predict active insulators among ATAC-seq peaks. To overcome these
91 limitations, we turned to deep learning. Deep learning methods have revolutionized the
92 machine learning field due to their ability to learn novel data features, which have also

been successfully applied in genomics to predict chromatin accessibility[22] and enhancers[23, 24] from DNA sequence. We developed a deep learning framework, namely Classification of Regulatory Elements with ATAC-seq (CoRE-ATAC) (Fig. 1 middle and bottom) to harness the power of deep learning for inferring the regulatory function of open chromatin regions. Core-ATAC integrates DNA sequence data with chromatin accessibility data (single cell or bulk) using a novel ATAC-seq data encoder that is designed to be able to integrate an individual's genotype to personalize predictions for *cis*-REs, especially for loci with genetically modulated regulatory activity (e.g., eQTLs, caQTLs).

In this study, we present CoRE-ATAC and evaluate its ability to predict *cis*-RE function (promoters, enhancers, insulators, and others) in 11 different cell types across 46 bulk ATAC-seq samples (Table 1) and demonstrate that CoRE-ATAC is a robust method that consistently predicts *cis*-REs with high average precision (mean micro average precision = 0.8) irrespective of whether the cell type was used in model training. We observed that CoRE-ATAC predictions recapitulated ChromHMM enhancers from the cognate cell type as well as *cis*-REs inferred *via* different assays (CAGE[25] and STARR-seq[26]). We compared CoRE-ATAC *cis*-RE predictions in human islet samples to *cis*-RE activity from massively parallel reporter assays (MPRA)[27] to demonstrate that CoRE-ATAC can predict the loss/gain of *cis*-RE activity linked to genetic variation. Next, we focused on the problem of annotating rare cell populations which can be identified using unsupervised clustering on snATAC-seq data. CoRE-ATAC effectively predicted enhancers from snATAC-seq data in human Peripheral Blood Mononuclear Cells[14] for 7 blood-derived

immune cell type clusters. Enhancers inferred from PBMC snATAC-seq data captured the majority of super enhancers in these immune cell subsets[2, 4] (i.e., cell-type-specific enhancers) and were enriched in SNPs for diseases of the cognate cell/tissue type. These analyses demonstrate the potential of CoRE-ATAC to better annotate open chromatin regions from bulk ATAC-seq and cell clusters from snATAC-seq samples, which will ultimately improve our understanding of how gene expression programs are regulated at the individual level and how this regulation is disrupted in pathologies.

Results

CoRE-ATAC predicts functional annotations for *cis*-REs

Annotation of ATAC-seq peaks using ChromHMM states revealed that between 16-52% are promoters, 18-54% are enhancers, 4-15% are insulators (for samples with insulator states in ChromHMM) and 5-55% are other functional annotations (Fig. 2a). The functional diversity in these annotations establishes a need for functionally annotating ATAC-seq open chromatin maps. Furthermore ChromHMM states are effective as baseline references to assess predictive performances of functional annotations[21] compared to alternatives (e.g., CAGE[25] or P300 binding) since ChromHMM captures a wider array of functional states and detects a larger set of loci (e.g., enhancers) that can be used in model training. For instance, in the third figure of our previous work PEAS, we showed that chromHMM captured the majority of both P300 and CAGE identified enhancers and that these two alternative methods identified smaller but distinct subsets of enhancers between them [21].

139 Using ChromHMM annotations in cognate cell types as class labels, CoRE-ATAC was
140 trained on GM12878, K562, HSMM, and $CD14^+$ ATAC-seq samples and model
141 performances were evaluated using held-out test data (i.e., regions within chromosomes
142 3 and 11) (Table 2). After identifying a concordant set of 10 functional states to use as a
143 ground truth based on in-house and Roadmap ChromHMM states (Fig. S1 and S2), we
144 showed that among these, four states could be effectively discriminated from ATAC-seq
145 maps (Fig. S3). We therefore annotated ATAC-seq peaks for three major functional *cis*-
146 RE classes (promoters, enhancers, and insulators) using ChromHMM and combined
147 other states into a fourth class, named "other" which captures states such as repressed,
148 transcribed, and quiescent regions, covering ATAC-seq peaks that do not correspond to
149 *cis*-REs (Methods) (Fig. S4). For the remainder of our analyses, we focus on models
150 utilizing these four functional states.
151
152 CoRE-ATAC models evaluated on 4 cell types (Table 1) had high micro-average precision
153 (0.88-0.96) across all samples with a combined accuracy of 84.20% (Fig. 2b left and Fig.
154 2c). Individual class performances also observed high average precision for all four
155 classes of *cis*-REs (Fig. 2b right, Fig. S5), establishing that high precision/accuracy values
156 from CoRE-ATAC predictions are not driven by predictions of a single functional class.
157 However, we noted that most mis-predictions were between enhancer and "other" classes
158 (Fig. 2c).
159
160 Currently, the only alternatives for inferring *cis*-RE function from ATAC-seq data are
161 PEAS[21] and DNA sequence methods. We therefore compared the performance of

162 CoRE-ATAC with two sequence based methods: DeepSEA[24] and LS-GKM[28], and our
163 previous neural network based approach PEAS[21] using the same training and testing
164 data for enhancer and "other" classes. CoRE-ATAC showed improvement over these
165 alternative methods (Fig. 2d, Fig. S6) (average precision=0.92 for held out test data), with
166 PEAS following second in performance (average precision=0.90). Methods that used only
167 DNA sequence performed very similarly to one another (Fig. S6) (average
168 precision=0.81) but were not as effective as PEAS and CoRE-ATAC. The increased
169 predictive performances of CoRE-ATAC and PEAS compared to sequence-based
170 methods suggest that ATAC-seq signal (e.g., read/insert pileups) contains critical
171 information for functionally classifying *cis*-REs that cannot be captured from DNA
172 sequence alone. This analysis establishes that CoRE-ATAC not only captures similar
173 DNA-sequence methods as previous sequence based methods, but that it is able to
174 leverage this DNA sequence information in conjunction with ATAC-seq signal to improve
175 predictive performances.
176

177 **CoRE-ATAC predicts *cis*-RE functionality across cell types**

178 We evaluated CoRE-ATAC's ability to predict *cis*-RE function in 40 bulk ATAC-seq
179 samples from 7 cell types (i.e., pancreatic islets, MCF7 breast cancer cell line, naïve $CD8^+$
180 T, PBMC, $CD4^+$ T, A549 and EndoC beta cell line) that were not used in model training
181 (Table 1). Functional annotation of ATAC-seq peaks from these 40 samples *via* CoRE-
182 ATAC (Fig. 3a) resembled the state distributions observed with ChromHMM (Fig. 2a).
183 Next, we studied whether functional predictions with CoRE-ATAC can outperform a naïve
184 annotation approach using distance to transcription start site (TSS), MACS2 FDR, and

185 number of CTCF motifs (Methods). CoRE-ATAC annotations outperformed these naïve
186 annotation approaches (Fig. 3b, Fig. S7), observing ~0.58 Matthews correlation
187 coefficient score for CoRE-ATAC compared to ~0.47 on the average across all samples.
188 Matthews correlation coefficient was used to account for the fact that threshold-based
189 approaches produce binary probabilities (i.e. 1.0 or 0.0) that can result in misleading
190 average precision calculations.
191
192 CoRE-ATAC models were highly predictive for cell types that were not utilized in model
193 training, showing ~0.80 micro-average precision across all samples. The highest
194 precision values were detected for promoters (~0.95), followed by enhancers (~0.76) (Fig.
195 3c). Insulator annotations were only available in 4 of the 7 cell types tested (MCF7, A549,
196 $CD4^+$ T and islet), however, insulator states in islets were excluded from these analyses
197 due to the poor quality of CTCF ChIP-seq data in islets (Fig. S8). Among the remaining 3
198 cell types, known insulators (based on ChromHMM states) were predicted with high
199 average precision (~0.78) (Fig. 3c and Fig. S9 for islets). To show the functional relevance
200 of CoRE-ATAC insulators that were either not annotated by ChromHMM (i.e., cell types
201 that are missing insulator states) or were misclassifications compared to ChromHMM
202 insulators, we conducted de novo motif enrichment analysis using HOMER[29]. CTCF
203 motif was by far the most enriched sequence for these loci with a very significant p-value
204 (P-value < 1e-27180) (Fig. 3d), suggesting that CoRE-ATAC predictions capture CTCF
205 insulator sequences.
206

207 A potential pitfall in cross-cell-type model evaluations is the use of the same genomic
208 regions (e.g., same chromosome) in both model training and testing[30]. When the same
209 region (not the same genomic data) is used in training and testing, a model might perform
210 well simply because it "remembers" the specific DNA sequences used during training. To
211 study whether our model suffers from this pitfall, we utilized the two chromosomes that
212 were excluded from model training (chromosomes 3 and 11) and compared cross cell
213 type predictions for these two chromosomes with the rest of the predictions and observed
214 comparable predictive performances (Fig. 3e). These analyses suggest that CoRE-ATAC
215 has learned a function from DNA sequence and chromatin accessibility signals that is
216 transferable across genomic regions for predicting *cis*-RE functionality rather than simply
217 memorizing specific DNA sequences. To further evaluate CoRE-ATAC's cross-cell-type
218 performance, we stratified ATAC-seq peaks by their cell-specificity across 7 cell types
219 and compared prediction performances. CoRE-ATAC predicted common promoters with
220 higher average precision than cell-type-specific promoters. Surprisingly, CoRE-ATAC
221 predicted cell-type-specific enhancers more effectively compared to common enhancers
222 (mean average precision 0.70 versus 0.48), emphasizing the utility of this method to study
223 disease-relevant enhancers that are typically cell-type-specific[2] (Fig. 3f top). The
224 prediction bias observed for cell-specific promoters is potentially due to the fact that the
225 majority of promoters are common among cell types (Fig. 3f bottom), whereas the majority
226 of enhancers are cell-type-specific.
227

228 **Core-ATAC can recapitulate enhancers from different assays**

229 To study whether CoRE-ATAC could recapitulate enhancers identified by alternative
230 methods, we compared ChromHMM annotations and CoRE-ATAC predictions in MCF7,
231 A549, $CD4^+$ T, and PBMC cell types to CAGE[25] enhancers identified by the FANTOM5
232 project[31, 32]. We observed that the majority of FANTOM enhancers overlapped with
233 promoters and enhancers in both ChromHMM and CoRE-ATAC (Fig. 4a). CoRE-ATAC
234 enhancer predictions showed significant overlap with FANTOM enhancers (Fisher's exact
235 test p-values for all cell types < 1.26e-11). The similarity between ChromHMM and CoRE-
236 ATAC further establishes that functional annotations *via* CoRE-ATAC using ATAC-seq
237 data is a cost-effective alternative to annotations *via* ChromHMM using multiple ChIP-seq
238 assays.
239
240 Another alternative method for investigating enhancer activity is massively parallel
241 reporter assays (MPRA)[27], which allows for experimentally testing thousands of
242 sequences in parallel for regulatory activity (i.e., promoters and enhancers)[33]. To
243 understand whether CoRE-ATAC could detect enhancers identified *via* MPRAs, we first
244 compared enhancer predictions in A549 cells to enhancers identified from self-
245 transcribing active regulatory region sequencing (STARR-seq)[26]. The majority of
246 sequences that showed regulatory activity in this assay were predicted as enhancers
247 (overlap significance calculated using Fisher's exact test p-value = 3.247e-61) by CoRE-
248 ATAC (Fig. 4b). We noted that a significant portion of STARR-seq enhancers (~33%)
249 were predicted as promoters in our models and the majority of these regions were close
250 to TSS (Fig. S10a). In contrast, CoRE-ATAC enhancers overlapping STARR-seq active
251 regions were more distal from the TSS (Fig S10b).

252

253 Regulatory activities of certain open chromatin regions are genetically modulated, which
254 can be detected *via* chromatin accessibility quantitative trait loci (caQTL) analyses.
255 Previously, we identified caQTLs from human islet samples (n=19) [6]. More recently, we
256 generated an MPRA library (manuscript in revision)[34], to test and compare the
257 regulatory activity of reference and alternative alleles for islet caQTL and other variants
258 (n=4293 SNPs). These data gave us an opportunity to test whether CoRE-ATAC
259 predictions can detect genetically driven differences in the regulatory activity. Using
260 CoRE-ATAC prediction probabilities from 19 islets (stratified based on genotypes), we
261 identified 237 loci for which a gain or loss of enhancer activity was predicted to be
262 modulated by the individual's genotype using one-tailed point-biserial correlation p-values
263 (Methods). Among these, 110 loci were predicted as losing activity in the alternative allele,
264 whereas 127 were predicted as gaining activity. For these sequences, we compared the
265 regulatory activity from MPRA assays to the predicted gain/loss of activity from CoRE-
266 ATAC and confirmed that the direction of effect coincides with the two analyses (Fig. 4c).
267 More specifically, gain of functions for the alternative allele based on CoRE-ATAC
268 predictions had higher MPRA activity for the alternative allele in comparison to the
269 reference allele (i.e., fold change > 0). A similar concordance was observed for loci
270 associated with loss of function for the alternative allele. For example, islet samples with
271 the alternative allele for SNP rs11205653 (an islet caQTL) had lower enhancer
272 probabilities compared to the samples with the reference allele, in agreement with the
273 activity levels from the MPRA library for this loci (Fig. 4d). In this example, we observed
274 differences in accessibility for individuals with the same genotype, and we suspect that

275 other factors such as disease or a combined effect with additional SNPs may be
276 contributing to this effect, showing the difficulty of identifying enhancer sequences and
277 importance of using methods that incorporate more than DNA sequence. Overall, these
278 results and findings establish CoRE-ATAC's ability to predict changes in *cis*-RE activity
279 due to genetic variation, which can help identify loci important for disease risk at the
280 individual level.
281

282 **Predicting disease-relevant enhancers from single nuclei ATAC-seq data**

283 Single nuclei ATAC-seq (snATAC-seq) reveals chromatin accessibility at single nucleus
284 resolution and enables 1) interrogation of chromatin accessibility at the single cell level
285 and 2) identifying epigenomic maps of rare cell types *via* unsupervised clustering
286 methods[14]. We focused on the latter and explored the efficacy of CoRE-ATAC for
287 predicting *cis*-RE function in rare cell types identified by snATAC-seq data. Towards this
288 goal, human PBMC snATAC data[14] were analyzed and grouped into 15 cell clusters
289 based on the similarity of accessibility profiles (Fig. 5a). Comparisons with sorted immune
290 cell bulk ATAC-seq data[15, 19] revealed 7 distinct cell types corresponding to B, natural
291 killer (NK), $CD8^+$ T, $CD4^+$ T, effector $CD4^+$ T, $CD14^+$, and dendritic cells (DCs) (Fig. 5b).
292 CoRE-ATAC models trained on bulk ATAC-seq data (using hg19 reference), predicted
293 *cis*-RE function in snATAC aggregate clusters aligned to hg38 (mean micro-average
294 precision = 0.68) (Fig. 5c), showing the flexibility and robustness of the method. Insulator
295 predictions from snATAC-seq data were significantly enriched for CTCF/BORIS motifs
296 (Fig. S11), confirming the biological relevance of these predictions.
297

298 To understand whether CoRE-ATAC can identify disease-relevant enhancers, we
299 conducted GWAS SNP enrichment analyses using GREGOR[35] for CoRE-ATAC
300 enhancers in the 7 cell types inferred from snATAC-seq data along with enhancers
301 predicted from bulk ATAC-seq data (A549, MCF7, islets, EndoC, $CD4^+$ T and naïve $CD8^+$
302 T). Immune cell enhancers predicted from snATAC data were enriched in diseases
303 related to the immune system (Figure 6d). For example, B cell enhancers were the most
304 significantly enriched for variants linked to Systemic Lupus, a disease characterized by
305 dysfunctions in B cells[36]. Islet enhancers were most significantly enriched for Type 2
306 Diabetes and Fasting Blood Glucose as expected, while T cell were most significantly
307 enriched in immune diseases (e.g., Type 1 Diabetes). Next, we compared CoRE-ATAC
308 enhancers with super enhancers obtained from SEdb[37], which are cell-type-specific
309 enhancers that likely harbor disease causing SNPs[2]. On average, ~79.0% of super
310 enhancers were captured by CoRE-ATAC enhancer predictions within their respective
311 cell types (Figure 5e). Surprisingly, enhancer predictions from snATAC-seq data were at
312 least as effective in capturing cell-specific enhancers compared to bulk ATAC-seq data,
313 indicating the efficacy of single cell epigenomic profiles for detecting noncoding regions
314 associated with complex diseases with the help of advanced computational methods.
315 Chromatin accessibility profiles of snATAC clusters resembles that of bulk data from
316 sorted cells, allowing CoRE-ATAC to effectively predict *cis*-RE function from snATAC
317 data, e.g., the locus around the *IL7R* (an important gene for T cell aging[19]) in $CD4^+$ T
318 cells for both bulk ATAC-seq and snATAC-seq data (Figure 5f). All together, these results
319 demonstrate that CoRE-ATAC is an effective and robust tool for predicting the
320 functionality of *cis*-REs in rare cell types identified from snATAC data.

Discussion

Recent advances in ATAC-seq profiling and single cell genomics revolutionized the epigenomics field by enabling the generation of chromatin accessibility maps from small starting material even at single cell resolution. Due to these advances, epigenomic maps of human cells/tissues from many individuals are being generated at an unprecedented rate, including our own studies, to study how epigenomic landscapes change with age, diseases, and upon in-vivo and in-vitro activation [6, 12, 15-20]. These epigenomic maps are instrumental in inferring *cis*-REs from clinically relevant samples. CoRE-ATAC is an effective method for classifying the functional state of *cis*-REs, which we extensively evaluated to ensure its efficacy in new cell types (i.e., cell types that are missing reference annotations) and in snATAC-seq data.

One of the unique features of CoRE-ATAC is its ability to integrate an individual's genotype with the chromatin accessibility maps by inferring the genotype from ATAC-seq read alignments. CoRE-ATAC accomplishes this with a sequence encoder that uses the frequency of observing a specific base-pair at a genomic position instead of solely using a one hot encoding of the reference genome, separating our method from existing ones that use the commonly used one hot-encoding approach. If genotype data (e.g., SNP arrays) are available for an individual, one can also leverage this information in CoRE-ATAC predictions by providing a reference specific to the individual. Genotype-aware methods can be effective in predicting the annotation and activity of *cis*-REs whose activity is modulated by genetic variation (e.g., caQTLs[5-8]). In alignment with this, we

344 demonstrated that CoRE-ATAC predictions from individual islet epigenomes aligned with
345 enhancer loss/gain between alleles inferred from MPRA assays.
346
347 CoRE-ATAC also established a foundation for inferring functional annotations from
348 snATAC-seq data by enabling the study of rare cell types that can be identified from these
349 assays. For this analysis, we used CoRE-ATAC models trained on bulk data due to
350 current limitations on the availability of snATAC-seq data. Although these models were
351 effective in predicting functionality of *cis*-REs from single cell data, we suspect that
352 models built from snATAC-seq data might further improve its efficacy. More specifically,
353 models can be trained to better learn the background signal specific to snATAC-seq data
354 where chromatin accessibility profiles consist of cells with similar chromatin accessibility
355 profiles due to unsupervised clustering. In future, as more snATAC-seq data becomes
356 available for more samples and more diverse cell types with established ChromHMM and
357 CTCF insulator states available, a CoRE-ATAC model can be trained solely on snATAC-
358 seq data, potentially resulting in better classification performances for the purposes of
359 annotating rare cell populations.
360
361 CoRE-ATAC is widely applicable for inferring *cis*-RE function from ATAC-seq and further
362 improves upon naïve methods as well as existing machine-learning models (Fig 2d, Fig.
363 3b). Although model training is a computationally expensive process, requiring GPUs to
364 train in a timely manner, using these models for predictions is much faster, requiring a
365 little over 2 minutes (excluding the time to load the data into memory) to functionally
366 annotate *cis*-REs for 75000 loci using a 2.3 GHz 8-Core Intel Core i9 processor (Fig.

367 S12). Our in-depth performance analyses suggest that CoRE-ATAC can be widely used,
368 even with limited computational resources, to improve the functional annotations of
369 ATAC-seq *cis*-RE maps.
370

371 **Conclusions**

372 CoRE-ATAC harnesses the power of deep learning to integrate chromatin accessibility
373 maps with DNA sequence and effectively predicts the functionality of *cis*-REs (i.e.,
374 promoters, enhancers, insulators) across diverse cell types and individuals. We
375 extensively evaluated CoRE-ATAC's efficacy to predict *cis*-RE function across multiple
376 cell types including those that are not used in model training. One of the unique features
377 of CoRE-ATAC is its ability to incorporate an individual's own genotype instead of the
378 reference sequence, enabling personalized predictions *via* genotype-aware machine
379 learning models. In alignment with this, our analyses from human islet samples showed
380 that CoRE-ATAC can predict differences in *cis*-RE activity that can be linked to genetic
381 variation in the studied population, which were confirmed experimentally with MPRA
382 assays. Finally, we predicted *cis*-RE function from snATAC-seq data in human blood-
383 derived immune cells. Despite being trained on bulk ATAC-seq data, CoRE-ATAC
384 recapitulated *cis*-REs in snATAC-seq clusters corresponding to specific immune cell
385 subtypes, including the ones that harbor SNPs for diverse immune diseases, enabling the
386 annotation of rare cell populations than can be obtained by snATAC-seq. CoRE-ATAC is
387 a robust model that can be used to infer *cis*-RE function and identify *cis*-RE activity
388 changes/disruptions due to genotype and/or environment, enabling us to further our
389 understanding in human genomics and disease biology. To promote the widespread use

of our predictive model we have made the CoRE-ATAC code and pre-trained model freely available on GitHub (https://github.com/UcarLab/CoRE-ATAC).

Methods

Machine Learning Architecture

CoRE-ATAC utilizes both data encoded in the deep learning framework and features extracted using our previous method PEAS[21]. These two feature sets are provided as two separate inputs into a machine learning model implemented using Keras[38] and Tensorflow[39] libraries. The deep learning component uses four convolutional layer blocks. Each block consists of i) a 1D convolutional neural network (CNN) layer (window size = 19), ii) a batch normalization layer, iii) a second 1D CNN layer (window size = 19) iv) a max pooling layer (pool size =2) and v) a second batch normalization layer, in this order (Figure 1). The first two blocks of convolutional layer units utilize 256 filters while the final two blocks utilize 512 filters. Both the deep learning component and the PEAS component of the model are trained using their own dense neural network layer with n=2048 nodes. The output of these dense layers is then concatenated and provided as input into a combined dense layer with n=512 nodes before classifying *cis*-REs in the final output dense layer using the Adam[40] optimization method. CNN layer window sizes, number of filters, and number of dense layer nodes were selected based on the best parameters observed during model tuning.

ATAC-seq Data Encoders

412 We implemented a novel ATAC-seq data encoder that takes as input: paired-end reads,
413 peaks, and the reference genome (or a personalized genome reflecting an individual's
414 genetic variation) to generate one 10x600 matrix per peak. Rows represent DNA
415 sequence and ATAC-seq signal data, while columns represent the 300 base-pair
416 positions downstream and upstream from the peak center (Figure 1). Details of this matrix
417 are explained below:

418 ***Rows 1-4:*** The 0-1 normalized frequency of observing 1) adenine (A), 2) cytosine (C), 3)
419 guanine (G), or 4) thymine (T) from ATAC-seq read pileups, which is calculated when
420 there is enough read coverage (n >= 10). For low coverage loci; the frequency reduces
421 to a one-hot encoding, representing the corresponding nucleotide in the reference
422 genome.

423 ***Row 5:*** The number of insert pileups within the 600 base-pair window, where each
424 position is z-standardized with respect to pileups across all peaks for that position.

425 ***Rows 6 and 7:*** The number of 5' (row 6) and 3' (row 7) cuts at a given position (z-
426 standardized). Savitzky-Golay smoothing filter[41] is applied with a window size of 15 to
427 account for an increase in the number of cuts across multiple positions in close proximity
428 due to sequencing depth.

429 ***Rows 8 and 9:*** The median fragment length for 5' and 3' cut site, applying the same
430 standardization and smoothing used for rows 6/7.

431 ***Row 10:*** The original peak region, setting values to 1 if a base-pair position is within the
432 peak and 0 otherwise. This allows for distinguishing between peaks that are in close
433 proximity to one another and providing an indicator of the most relevant positions.

434

435 In all analyses, peaks used for data encoding were called with MACS2[42] using
436 parameters "-f BAMPE --nomodel". Duplicates were kept only in the snATAC-seq data
437 analyses using the "--keep-dup all" option.
438

439 **Ground Truth Selection & Model Training**

440 CoRE-ATAC was trained on ATAC-seq data from 4 different cell types: GM12878[12]
441 (GSE47753), K562[20] (GSE121993), HSMM[18] (GSE109828, donor "54-1" at time 0),
442 and $CD14^+$ monocytes[19] (EGAS00001002605) (Table 1). Class labels were assigned
443 by co-analyzing 15-state ChromHMM[11] models generated in these cell types with
444 corresponding 18-state Roadmap[10] ChromHMM annotations. 15 state ChromHMM
445 models were trained for each cell type independently using H3K4me1, H3K4me3,
446 H3K27ac, H3K27me3, H3K36me3, H3K9me3 and CTCF ChIP-seq data from
447 ENCODE[9]. Hierarchical clustering of emission probabilities revealed 10 distinct clusters
448 (Fig. S1) corresponding to active promoters, promoters, flanking enhancers, active
449 enhancers, enhancers, genic enhancers, transcribed, insulator, repressed, and low signal
450 states. Preliminary machine learning models were trained using the 10 clusters identified
451 and peaks were selected based on their concordance with Roadmap 18 state models
452 (Fig. S2) in order to select the best examples for model training. Performance evaluations
453 on the 10-state models revealed that only 4 major classes of *cis*-REs could be effectively
454 predicted from such models across all cell types: promoter, enhancer, insulator and other
455 classes (Fig. S3). Based on these results, we focused on training models with these 4
456 major classes and selected peaks using a wider range of concordance with Roadmap
457 states (Fig. S4).

458

459 To make CoRE-ATAC predictions robust to instances of low chromatin accessibility levels
460 or false positives in ATAC-seq peak calls, non-peak locations were also included for a
461 subset of samples. We selected non-peak locations by taking the union of all peaks
462 across all 6 samples and selecting additional genomic regions for HSMM, $CD14^+$ sample
463 2 (S2), and K562 replicate 2 (R2) samples that were called as peaks in other samples but
464 were not called within the sample itself. Only half of the samples were selected to ensure
465 that CoRE-ATAC could be robust to differences in normalization steps when provided
466 either a list of genomic regions that could contain no ATAC-seq reads or when provided
467 ATAC-seq peaks where every region must contain read information to call the peak at the
468 given location. We further filtered these regions to ensure that ChromHMM of these
469 additional annotations were annotated as “other” in the respective cell types. Finally, we
470 randomly selected non-peak regions so that only up to 25% of the “other” category for the
471 sample was represented by these non-peak regions. This additional filtering was applied
472 to ensure that non-peak regions did not dominate the “other” class to the point of CoRE-
473 ATAC only classifying peaks as “other” if the genomic regions have reduced or no ATAC-
474 seq read counts. In total, 153311 loci across 4 cell types were used in model training for
475 4-state classification, with 2551 of these loci representing non-peak regions.

476

477 Deep learning and PEAS components of CoRE-ATAC were initially trained separately to
478 handle different rates of overfitting observed between the two models (overfitting
479 evaluated using training and validation loss). Parameter tuning was performed in this
480 stage of training to adjust i) the number of filters, ii) convolutional layer window size, iii)

481 number of convolutional blocks, and iv) number of nodes within dense layers (where
482 applicable) using held out validation data (i.e., ATAC-seq peaks in chromosomes 2 and
483 10) (Table 2). Individually each component achieved 82.56% and 79.25% accuracy on
484 held out validation data for deep learning and PEAS components respectively using the
485 best observed parameters for minimizing the validation loss. We then combined these
486 models into a single unified model by concatenating deep learning and PEAS
487 components with a concatenation layer, for a final round of model training.

488 **Model Tuning**

489 Hyperparameter selection was performed on i) the 1D convolution kernel size {9, 11, 19, 21}, ii)
490 the number of convolutional layer block {3,4,5}, iii) intermediate dense layer size {1024, 2048,
491 4096} and iv) and the number of convolutional layer filters {32, 128, 256} where for each block
492 the first two convolutional layers were set to the initial filter size and then doubled for the
493 remaining two convolutional layers. Due to the long training time on these models, parameter,
494 not every combination (108 total) was tested, instead different parameters were tested one by
495 one, and the best performing value for the parameter was kept.

496

497 **Model Evaluation**

498 To test the performance of CoRE-ATAC on datasets and cell types that were not used in
499 model training, we predicted *cis*-REs in ATAC-seq data obtained for 7 different cell types:
500 MCF7[16] (GSE97583) (n=2), Naïve $CD8^+$ T[15, 19] (GSE118189 & EGAS00001002605)
501 (n=10), Peripheral Blood Mononuclear Cells (PBMCs) [15, 19] (EGAS00001002605)
502 (n=6), $CD4^+$ T[12] (GSE47753) (n=1), A549[43] (GSE117089) (n=1), pancreatic islets[6]

503 (SRP117935) (n=19) and EndoC beta cell line[17] (GSE118588) (n=1) (Table 1). Model
504 performances were evaluated using ChromHMM states in each cell type.
505

506 **Annotations for Model Evaluation**

507 ChromHMM 18 state model annotations from Roadmap[10] were used for naïve $CD8^+$ T
508 and A549, where annotations were converted to 4 major classes (promoter, enhancer,
509 insulator and other), mapping states to the most relevant class label. "TSSA" and
510 "TssFlnk" were mapped to promoter, "EnhG1", "EnhG2", "EnhA1", "EnhA2", and "EnhWk"
511 were mapped to enhancer and "Tx", "TxWk", "ZNF/Rpts", "Het", "ReprPC", "ReprPCWk",
512 and "Quies" were mapped to "other". Peaks that could not be mapped to these major
513 classes (i.e., "TssFlnkU", "TSSFlnkD", "TSSBiv", and "EnhBiv") were excluded. For PBMC
514 and $CD4^+$ T samples, 15 state ChromHMM models were trained using H3K4me1,
515 H3K4me3, H3K27ac, H3K27me3, H3K36me3, and H3K9me3, ChIP-seq from Roadmap.
516 ChromHMM states for MCF7, Islet, and EndoC were obtained from the following
517 studies[3, 17] and were converted to the 4 major class labels used by CoRE-ATAC. Only
518 ATAC-seq peaks with greater than 90% overlap with a single class label were kept to filter
519 out ambiguous regions.
520

521 **Comparison with alternative methods/assays**

522 We compared CoRE-ATAC to two sequence-based methods (DeepSEA[24] and LS-
523 GKM[28]) and our previous neural network (NN) based method (PEAS[21]). For this
524 comparison we focused on enhancer versus "other" predictions, the most difficult
525 discrimination task in our models (Figure 2c). For this, we chose all enhancers and "other"

526 annotated regions from our ground truth test set (i.e., regions within chromosomes 3 and
527 11 for GM12878, K562, HSMM, and $CD14^+$ samples) (Table 1 and Table 2). The same
528 training and testing set was used for CoRE-ATAC, PEAS, and LS-GKM. For DeepSEA,
529 we used the web annotation tool (http://deepsea.princeton.edu). DeepSEA makes
530 multiple predictions of activity for a wide array histone marks and transcription factors
531 (TFs) across multiple cell types. We therefore selected enhancer probabilities by taking
532 the maximum probability score for H3K4me1 and H3K27ac across all cell types predicted
533 by DeepSEA. The area under the receiver operating characteristic curve (ROC AUC) and
534 average precision metrics were used to compare the four methods (Fig. 2d, Fig. S2).
535
536 Naïve method comparisons focused on the 40 samples not used in model training to fairly
537 assess CoRE-ATAC's performance with respect to these methods. Multiple thresholds
538 were applied for each Naïve method to identify the best threshold to set for predicting
539 promoters, enhancers, and insulators. Promoters were tested using 1kb, 2kb, and 5kb
540 distances to the nearest TSS, (distances calculated using HOMER[29]). Enhancers were
541 tested using MACS2[42] FDR qval of 0.01, 0.001, and 0.0001. Insulators were tested
542 using the number of CTCF motifs greater than 0, 2, and 4. Promoters, enhancers, and
543 insulators were selected using these thresholds, assigning probabilities 1.0 when the
544 threshold requirement is met, and 0.0 otherwise. Finally, combined naïve approach
545 performance was calculated selecting the best threshold (i.e., 2kb, 0.001, and 0 for
546 promoter TSS, enhancer MACS2 qval, and insulator number of CTCF motifs
547 respectively). Due to the nature of selecting enhancers using MACS2 qval, priority was
548 given to promoters and then insulators. For the remaining regions that were not annotated

549 as either promoter or insulator, enhancers and “other” were classified using MACS2 qval.
550 Performances were evaluated using Matthews correlation coefficient.
551
552 Alternative enhancer definitions were explored to understand how well CoRE-ATAC can
553 predict active regulatory elements (i.e., promoter and enhancers) from FANTOM[31, 32]
554 and STARR-seq[26]. FANTOM enhancers identified using Cap Analysis of Gene
555 Expression (CAGE) technology[25] were obtained for MCF7, A549, $CD4^+$ T cells, and
556 PBMCs. We obtained STARR-seq active regulatory sites for A549[44] from GEO[45, 46]
557 (GSE114063). These regions were then compared with CoRE-ATAC predictions,
558 counting the number of predictions for each class within FANTOM5 enhancers and
559 STARR-seq sites. Fisher’s exact test was used to calculate the significance of CoRE-
560 ATAC enhancers overlapping with Starr-seq enhancers.
561
562 Massively Parallel Reporter Assay (MPRA) data were generated in MIN6 pancreatic beta
563 cell line to study the regulatory activity of variants associated with Type-2-Diabetes
564 (manuscript in revision) [34]. Briefly, 4293 variants within islet ATAC-seq peaks were
565 profiled for regulatory element activity using MPRA. Taking the union of islet peaks called
566 across all 19 samples, we predicted *cis*-RE functions for each islet sample to obtain class
567 probabilities for each region and islet sample. We then putatively identified *cis*-REs
568 showing loss or gain of *cis*-RE activity using one-tailed point-biserial correlation p-values,
569 identifying loci with probabilities that were significantly lower or higher between reference
570 and alternative genotypes. The maximum absolute value correlation was obtained for
571 each peak with genotype information by calculating the correlations for all comparisons

572 among ref/ref, ref/alt, and alt/alt genotypes, using the sum of promoter and enhancer
573 probabilities. Peaks with significant point-biserial correlation coefficients (p-value < 0.01)
574 were separated into two groups corresponding to the loss or gain of *cis*-RE activity
575 (negative and positive correlations respectively) based on CoRE-ATAC predictions for
576 different alleles. Finally, MPRA activity differences between alternative and reference
577 alleles (log fold change in MPRA analyses) were compared with the activity differences
578 inferred from CoRE-ATAC *cis*-RE class probabilities. Student t-test, and Mann Whitney
579 U test were used to calculate the significance of MPRA log fold change values observed
580 for predicted loss and gain of *cis*-RE activity both individually as well as comparatively.
581

582 **snATAC seq data analyses and clustering**

583 Single nuclei ATAC-seq (snATAC) PBMC data[14] was obtained from GEO[45, 46]
584 (GSE129785). Sequence reads were processed using Cell Ranger and cells were
585 clustered using our own implementation of a recently described snATAC clustering
586 method[14], which uses two passes of clustering to identify cell type clusters. In the first
587 pass, genomic regions were binned into 2500bp windows, counting the number of paired-
588 end reads within each bin. The top 50,000 bins with the greatest number of reads were
589 selected for clustering using Seurat[47, 48], requiring a minimum of 185 cells per cluster.
590 Peaks were called for each cluster independently using MACS2 ("-f BAMPE –nomodel, -
591 -keep-dup all") and were used to perform a second pass of cell clustering as before, using
592 peaks instead of bins, identifying a total of 15 clusters.
593

594 Cell type annotations for these clusters were obtained by comparing snATAC profiles with
595 ATAC-seq profiles of sorted immune cell types *via* flow cytometry. To identify marker
596 peaks for each cell type, we selected 2-3 representative and high quality ATAC-seq
597 samples per cell type[15, 19] for 19 different immune cell types based on ATAC-seq
598 library quality (FRIP score and read depth) and similarity between biological replicates
599 (Spearman r). Marker peaks were identified from the signature profile generated by
600 CIBERSORT[49], using its data pre-processing step. Hierarchical clustering was
601 performed for read pileups within marker peaks for snATAC-seq clusters and sorted bulk
602 data, which enabled annotating 15 clusters into 7 cell types. Cells that belong to the same
603 cell type were pooled together and the functionality of *cis*-REs were predicted using
604 CoRE-ATAC for 7 cell types by allowing duplicate reads. Predictions were evaluated
605 using either ChromHMM annotations from Roadmap, or in-house ChromHMM states
606 ($CD4^+$ T and $CD14^+$ monocytes).
607

608 **SNP enrichments**

609 GWAS SNP enrichments were performed with GREGOR[35] software using index SNPs
610 as well as linked SNPs (linkage disequilibrium threshold of R^2 0.7 for the European (EUR)
611 population). NHGRI-EBI GWAS Catalog SNPs[50] (Obtained January 8th 2020) for 3981
612 traits/diseases were used in enrichment analyses for different enhancer sets inferred by
613 CoRE-ATAC.
614

615 **Super enhancer analysis**

616 Super enhancer annotations were obtained for A549, MCF7, Islets, $CD14^+$, $CD4^+$ T, $CD8^+$
617 T, monocyte-derived dendritic cells, NK, and B cells from SEdb[37]. We then calculated
618 the percent of super enhancers found among CoRE-ATAC enhancer predictions to
619 measure how well CoRE-ATAC predictions recapitulate these cell-type-specific
620 enhancers.
621
622 **Declarations**
623 **Ethics approval and consent to participate**
624 Not applicable
625
626 **Consent for publication**
627 Not applicable
628
629 **Availability of data and materials**
630 The publicly available datasets analyzed during the current study are available from the
631 GEO and EGA repositories and websites:
632
633 GM12878 & CD4T ATAC-seq:
634 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE47753
635 K562 ATAC-seq: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121993
636 HSMM ATAC-seq: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE109828
637 MCF7 ATAC-seq: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE97583

Naive CD8 ATAC-seq:

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118189

https://ega-archive.org/studies/EGAS00001002605

PBMC ATAC-seq: https://ega-archive.org/studies/EGAS00001002605

A549 ATAC-seq: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117089

EndoC ATAC-seq: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118588

A549 Starr-seq: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE114063

snATAC PBMC data: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE129785

FANTOM enhancers: https://fantom.gsc.riken.jp/5/datafiles/latest/extra/Enhancers/

Encode ChIP-seq: https://www.encodeproject.org/

Roadmap ChromHMM states:

https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html

Islet MIN6 MPRA data analyzed during the current study are not publicly available yet due as the publication for these data is currently under revision, but are available from the corresponding author on reasonable request.

CoRE-ATAC code and pretrained models are available on our GitHub page:

https://github.com/UcarLab/CoRE-ATAC

Competing Interests

The authors declare that they have no competing interests

661 **Funding**

662 This research was supported by the PhRMA Foundation postdoctoral fellowship award
663 in bioinformatics (to AT) and National Institute of General Medical Sciences (NIGMS)
664 under award number GM124922 (to DU).

665

666 **Author's Contributions**

667 AT and DU designed the study. AT implemented the CoRE-ATAC method, trained the
668 deep learning model and conducted all downstream analyses. SK, RT, and MLS
669 provided the MIN6 MPRA data and additional support for analyses conducted using
670 these data. AT and AE processed the snATAC-seq data performed snATAC clustering
671 of PBMCs. AT and DU annotated the final snATAC clusters. AT and DU wrote the
672 manuscript. All authors read and approved the final manuscript.

673

674 **Acknowledgements**

675 We thank members of the Ucar, Stitzel, Beck, and Lee labs at JAX-GM for their insightful
676 feedback throughout this project.

677

678 **References:**

679 1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al.
680 Potential etiologic and functional implications of genome-wide association loci for human
681 diseases and traits. Proceedings of the National Academy of Sciences.
682 2009;106(23):9362-7.

683 2. Hnisz D, Abraham Brian J, Lee Tong I, Lau A, Saint-André V, Sigova Alla A, et al.
684 Super-Enhancers in the Control of Cell Identity and Disease. Cell. 2013;155(4):934-47.
685 3. Parker SCJ, Stitzel ML, Taylor DL, Orozco JM, Erdos MR, Akiyama JA, et al.
686 Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human
687 disease risk variants. Proceedings of the National Academy of Sciences.
688 2013;110(44):17921-6.
689 4. Whyte Warren A, Orlando David A, Hnisz D, Abraham Brian J, Lin Charles Y,
690 Kagey Michael H, et al. Master Transcription Factors and Mediator Establish Super-
691 Enhancers at Key Cell Identity Genes. Cell. 2013;153(2):307-19.
692 5. Alasoo K, Rodrigues J, Mukhopadhyay S, Knights AJ, Mann AL, Kundu K, et al.
693 Shared genetic effects on chromatin and gene expression indicate a role for enhancer
694 priming in immune response. Nature Genetics. 2018;50(3):424-31.
695 6. Khetan S, Kursawe R, Youn A, Lawlor N, Jillette A, Marquez EJ, et al. Type 2
696 Diabetes–Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets.
697 Diabetes. 2018;67(11):2466-77.
698 7. Kumasaka N, Knights AJ, Gaffney DJ. Fine-mapping cellular QTLs with RASQUAL
699 and ATAC-seq. Nature Genetics. 2015;48(2):206-13.
700 8. Gaffney DJ, McVicker G, Pai AA, Fondufe-Mittendorf YN, Lewellen N, Michelini K,
701 et al. Controls of Nucleosome Positioning in the Human Genome. PLoS Genetics.
702 2012;8(11):e1003036.
703 9. Consortium EP. An integrated encyclopedia of DNA elements in the human
704 genome. Nature. 2012;489(7414):57-74.

705 10. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et
706 al. Integrative analysis of 111 reference human epigenomes. Nature.
707 2015;518(7539):317-30.
708 11. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and
709 characterization. Nature Methods. 2012;9(3):215-6.
710 12. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of
711 native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-
712 binding proteins and nucleosome position. Nature Methods. 2013;10(12):1213-8.
713 13. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for
714 Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol. 2015;109:21 9 1-
715 9.
716 14. Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, et al. Massively
717 parallel single-cell chromatin landscapes of human immune cell development and
718 intratumoral T cell exhaustion. Nature Biotechnology. 2019;37(8):925-36.
719 15. Calderon D, Nguyen MLT, Mezger A, Kathiria A, Müller F, Nguyen V, et al.
720 Landscape of stimulation-responsive chromatin across diverse human immune cells.
721 Nature Genetics. 2019;51(10):1494-505.
722 16. Cho SW, Xu J, Sun R, Mumbach MR, Carter AC, Chen YG, et al. Promoter of
723 lncRNA Gene PVT1 Is a Tumor-Suppressor DNA Boundary Element. Cell.
724 2018;173(6):1398-412.e22.
725 17. Lawlor N, Márquez EJ, Orchard P, Narisu N, Shamim MS, Thibodeau A, et al.
726 Multiomic Profiling Identifies cis-Regulatory Networks Underlying Human Pancreatic β
727 Cell Identity and Function. Cell Reports. 2019;26(3):788-801.e6.

728 18. Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM,
729 Aghamirzaie D, et al. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell
730 Chromatin Accessibility Data. Molecular Cell. 2018;71(5):858-71.e8.
731 19. Ucar D, Márquez EJ, Chung C-H, Marches R, Rossi RJ, Uyar A, et al. The
732 chromatin accessibility signature of human immune aging stems from CD8+ T cells. The
733 Journal of Experimental Medicine. 2017;214(10):3123-44.
734 20. Wang Z, Chu T, Choate LA, Danko CG. Identification of regulatory elements from
735 nascent transcription using dREG. Genome Research. 2019;29(2):293-303.
736 21. Thibodeau A, Uyar A, Khetan S, Stitzel ML, Ucar D. A neural network based model
737 effectively predicts enhancers from clinical ATAC-seq samples. Scientific Reports.
738 2018;8(1).
739 22. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible
740 genome with deep convolutional neural networks. Genome Research. 2016;26(7):990-9.
741 23. Yang B, Liu F, Ren C, Ouyang Z, Xie Z, Bo X, et al. BiRen: predicting enhancers
742 with a deep-learning-based model using the DNA sequence alone. Bioinformatics.
743 2017;33(13):1930-6.
744 24. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep
745 learning–based sequence model. Nature Methods. 2015;12(10):931-4.
746 25. de Hoon M, Shin JW, Carninci P. Paradigm shifts in genomics through the
747 FANTOM projects. Mammalian Genome. 2015;26(9-10):391-402.
748 26. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. Genome-Wide
749 Quantitative Enhancer Activity Maps Identified by STARR-seq. Science.
750 2013;339(6123):1074-7.

751 27. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, et al. Systematic
752 dissection and optimization of inducible enhancers in human cells using a massively
753 parallel reporter assay. Nature Biotechnology. 2012;30(3):271-7.
754 28. Lee D. LS-GKM: a new gkm-SVM for large-scale datasets. Bioinformatics.
755 2016;32(14):2196-8.
756 29. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple
757 Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory
758 Elements Required for Macrophage and B Cell Identities. Molecular Cell. 2010;38(4):576-
759 89.
760 30. Schreiber J, Singh R, Bilmes J, Noble WS. A pitfall for machine learning methods
761 aiming to predict across cell types. bioRxiv 512434v2 [Preprint]. 2019.
762 31. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al.
763 An atlas of active enhancers across human cell types and tissues. Nature.
764 2014;507(7493):455-61.
765 32. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al.
766 Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome
767 Biology. 2015;16(1).
768 33. Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, et al. Direct
769 Identification of Hundreds of Expression-Modulating Variants using a Multiplexed
770 Reporter Assay. Cell. 2016;165(6):1519-29.
771 34. Khetan S, Kales S, Kursawe R, Jillette A, Reilly SK, Ucar D, et al. Functional
772 characterization of thousands of type 2 diabetes-associated and chromatin-modulating

773 variants under steady state and endoplasmic reticulum stress. bioRxiv
774 20200212939348v1 [Preprint]. 2020.
775 35. Schmidt EM, Zhang J, Zhou W, Chen J, Mohlke KL, Chen YE, et al. GREGOR:
776 evaluating global enrichment of trait-associated variants in epigenomic features using a
777 systematic, data-driven approach. Bioinformatics. 2015;31(16):2601-6.
778 36. Chan OTM, Madaio MR, Shlomchik MJ, Chan OIM, Madaio MP, Shlomchik MJ.
779 The central and multiple roles of B cells in lupus pathogenesis. Immunological Reviews.
780 1999;169(1):107-21.
781 37. Jiang Y, Qian F, Bai X, Liu Y, Wang Q, Ai B, et al. SEdb: a comprehensive human
782 super-enhancer database. Nucleic Acids Research. 2019;47(D1):D235-D43.
783 38. Chollet F. Keras. 2015.
784 39. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow:
785 Large-scale machine learning on heterogeneous systems. 2015.
786 40. Kingma D, Ba J. ADAM: A Method for Stochastic Optimization. arXiv:14126980
787 [Preprint]. 2014.
788 41. Savitzky A, Golay MJE. Smoothing and Differentiation of Data by Simplified Least
789 Squares Procedures. Analytical Chemistry. 1964;36(8):1627-39.
790 42. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-
791 based Analysis of ChIP-Seq (MACS). Genome Biology. 2008;9(9):R137.
792 43. Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, et al. Joint
793 profiling of chromatin accessibility and gene expression in thousands of single cells.
794 Science. 2018;361(6409):1380-5.

795 44. Johnson GD, Barrera A, McDowell IC, D'Ippolito AM, Majoros WH, Vockley CM,
796 et al. Human genome-wide measurement of drug-responsive regulatory activity. Nature
797 Communications. 2018;9(1).
798 45. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al.
799 NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Research.
800 2012;41(D1):D991-D5.
801 46. Edgar R. Gene Expression Omnibus: NCBI gene expression and hybridization
802 array data repository. Nucleic Acids Research. 2002;30(1):207-10.
803 47. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell
804 transcriptomic data across different conditions, technologies, and species. Nature
805 Biotechnology. 2018;36(5):411-20.
806 48. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al.
807 Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888-902.e21.
808 49. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust
809 enumeration of cell subsets from tissue expression profiles. Nature Methods.
810 2015;12(5):453-7.
811 50. Buniello A, MacArthur JA L, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al.
812 The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted
813 arrays and summary statistics 2019. Nucleic Acids Research. 2019;47(D1):D1005-D12.
814
815
816
817

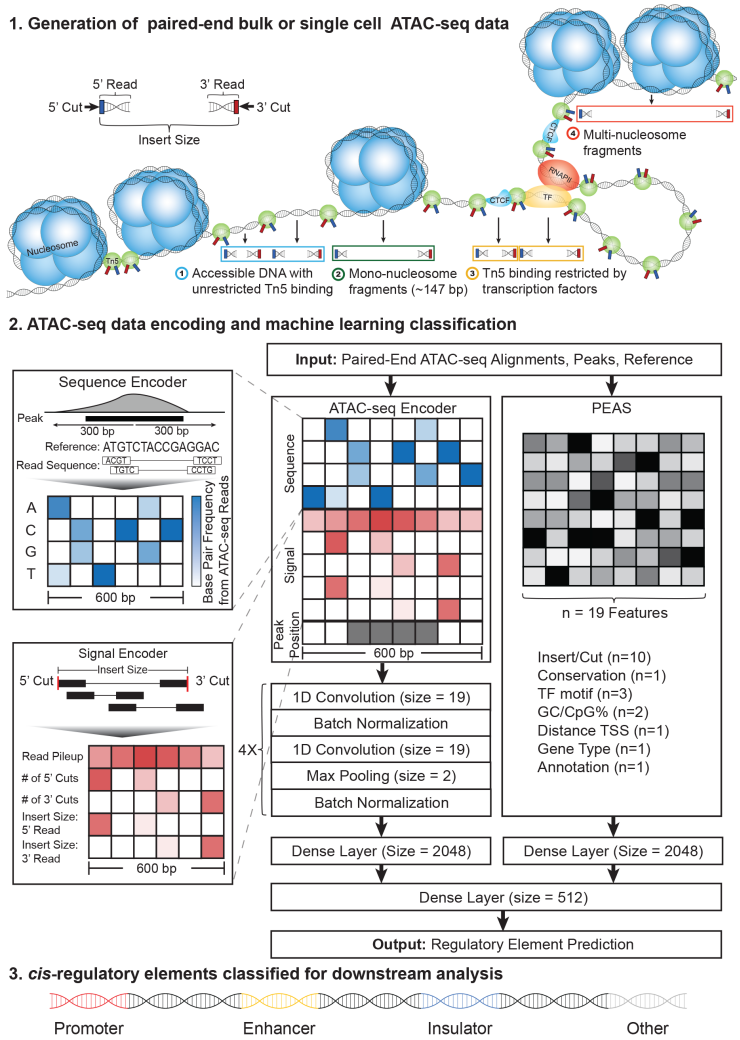

Fig. 1. Overview of the CoRE-ATAC framework. Paired-end ATAC-seq data captures different insert size distributions corresponding to the presence or absence of nucleosomes or TFs. In CoRE-ATAC, ATAC-seq data is encoded into a 10x600 matrix and 19 data features from PEAS algorithm to predict the functionality of an open chromatin region using both novel and manually selected features. In the final step, CoRE-ATAC classifies REs into 4 functional classes: promoter, enhancer, insulator, and other.

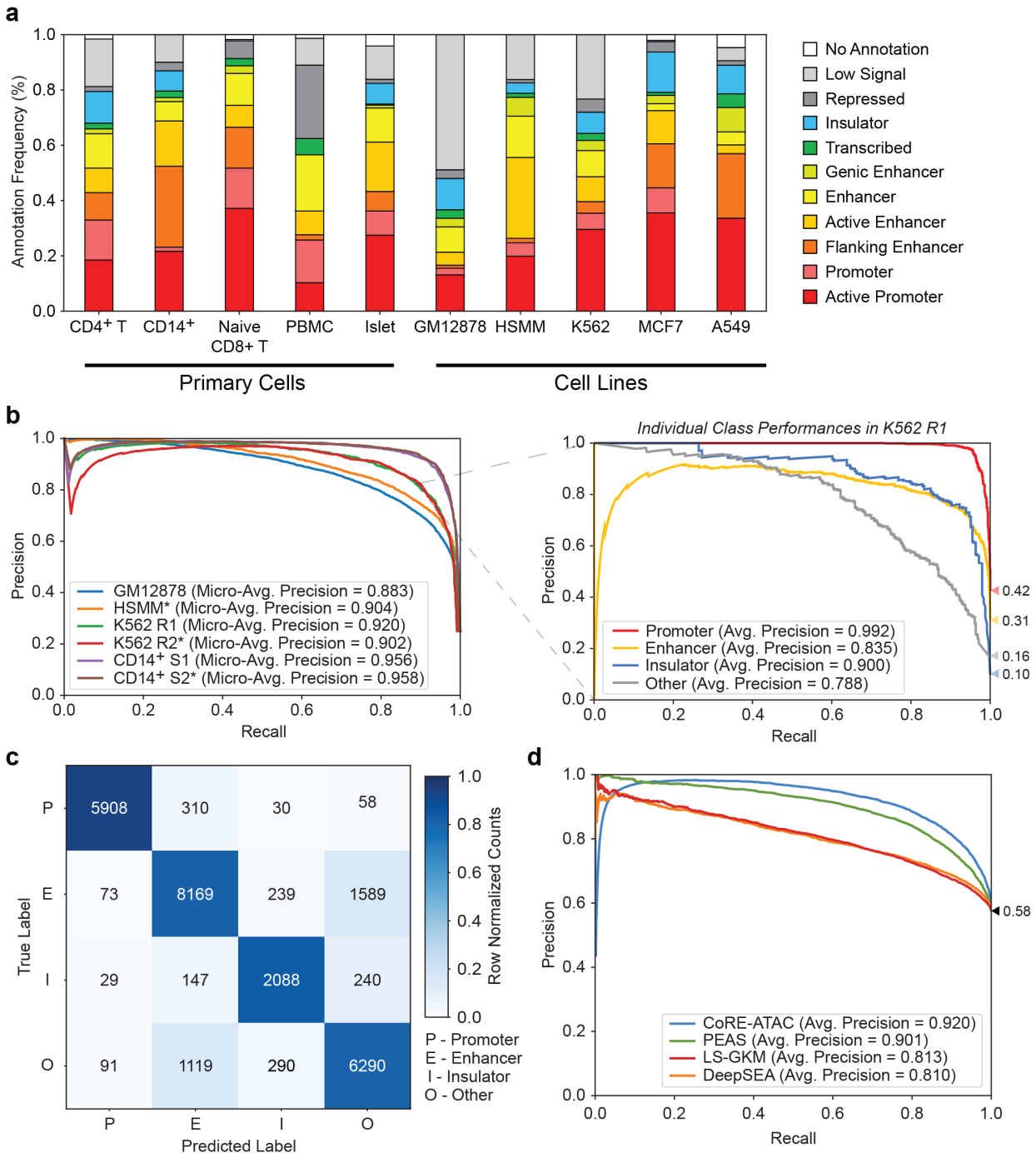

Fig. 2. CoRE-ATAC outperforms alternative enhancer prediction methods. CoRE-ATAC predictions were evaluated using held out test data (chromosomes 3 and 11). (**a**) ChromHMM state distributions for different cell types used in this study. ATAC-seq open chromatin maps correspond to a multitude of *cis*-RE functional states, corresponding to promoters, enhancers, insulators, and other functional annotations. (**b**) Micro-average precision values (left) were calculated, summarizing the average precision values for individual class predictions for all cell types used in model training. A breakdown of individual class average precision scores is shown for K562 (right). (**c**) Combined confusion matrix of model predictions across all cell types used in model training. Note that models are predictive for all class labels. However, mispredictions were more frequently observed between enhancer and other functional classes. (**d**) Receiver operating characteristic (ROC) curves for different enhancer prediction models: CoRE-ATAC, PEAS, DeepSEA and LS-GKM. Note that CoRE-ATAC outperforms alternative methods.

821

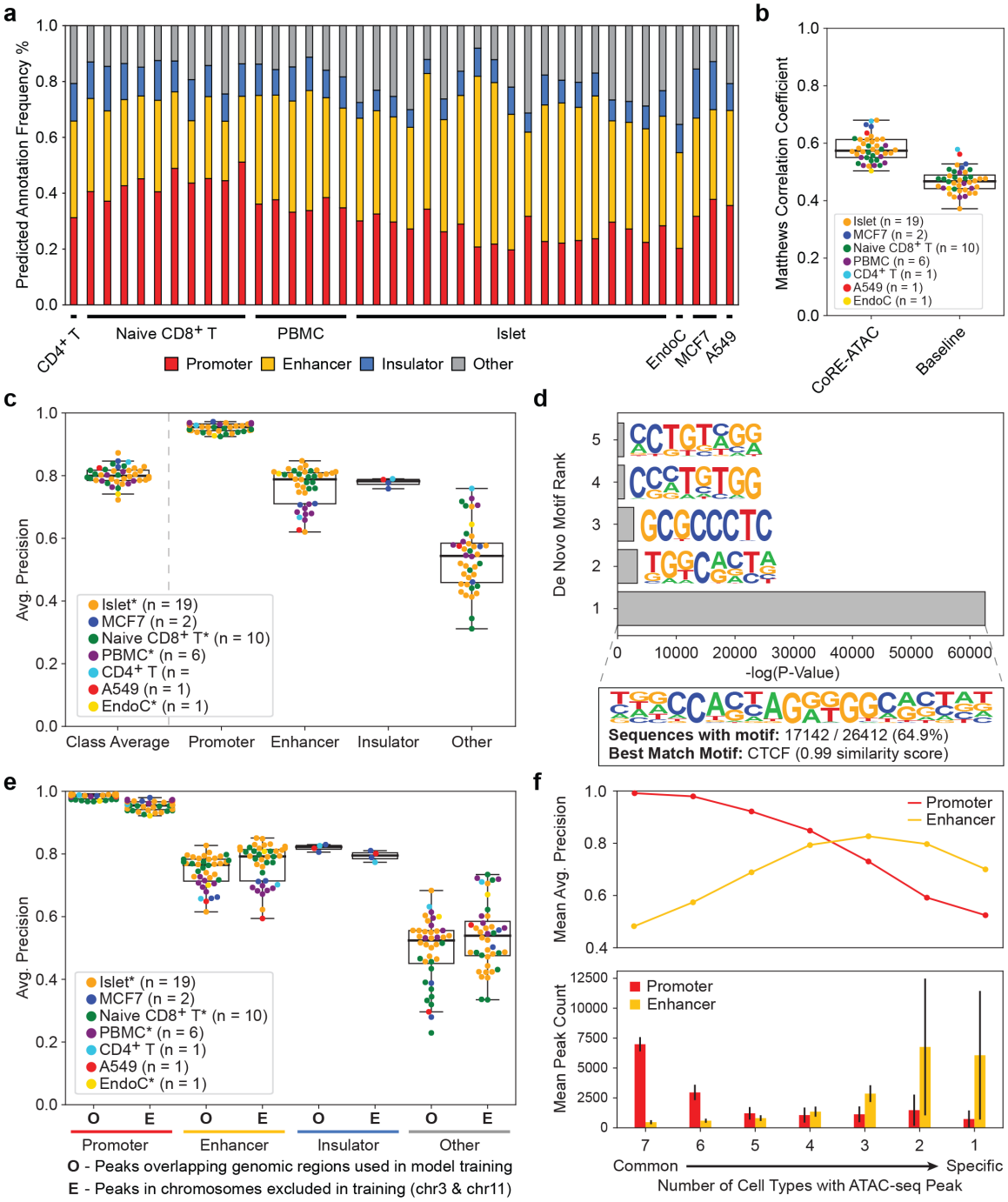

Fig. 3. CoRE-ATAC can predict REs across cell-types. CoRE-ATAC was evaluated in 7 cell types using 40 samples that are not used in model training. (**a**) Distribution of CoRE-ATAC predictions. Prediction distributions are similar to those observed by ChromHMM state annotations. (**b**) Comparison of CoRE-ATAC to baseline/naive predictions based on thresholds for distance to TSS, MACS2 FDR, and number of CTCF motifs. CoRE-ATAC improves upon baseline performances. (**c**) Boxplot of Average precision scores for predicting REs. Micro-average precision was used to calculate class average scores. Co-RE ATAC is predictive across cell types and different functional classes with an exception of insulators in islets, which is due to CTCF ChIP-seq quality in islets. (**d**) De novo motif enrichment results for regions predicted as insulators by CoRE-ATAC but were not annotated as insulators by ChromHMM. Note that these regions are significantly enriched for the CTCF motif (0.983 similarity), suggesting that CoRE-ATAC insulator predictions are functionally relevant. (**e**) CoRE-ATAC performances for i) predictions overlapping regions used in model training (O), and ii) predictions within regions that are on held-out test chromosomes (E). Note the performance similarity between these two prediction categories across all classes. (**f**) CoRE-ATAC model performances (top) and the average number of promoters and enhancers observed (bottom) by cell-type-specificity. We observed that CoRE-ATAC was more effective in predicting common promoters and cell-type-specific enhancers, for which we had more examples represented in the data. CoRE-ATAC's ability to predict cell-type-specific enhancers demonstrates its usefulness for interrogating individual and cell-type-specific enhancers.

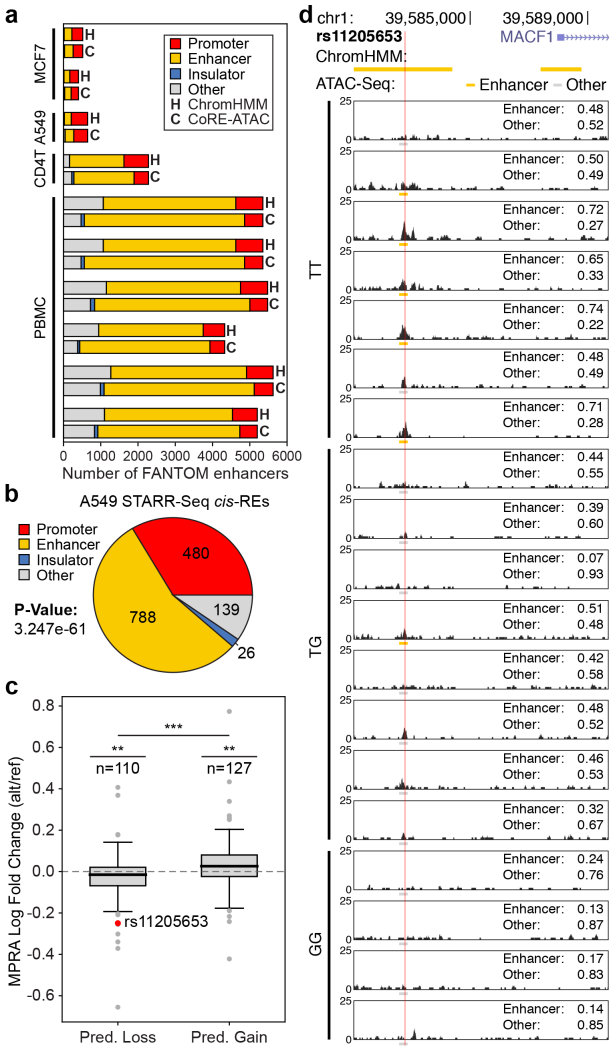

Fig. 4. CoRE-ATAC predictions compare favorably to alternative assays. (**a**) CoRE-ATAC (C) and ChromHMM (H) predictions and annotations of FANTOM enhancers in MCF7, A549, CD4$^+$ T and PBMC samples. CoRE-ATAC predicted the majority of FANTOM enhancers as enhancers or promoters, recapitulating these experimentally identified enhancers. CoRE-ATAC annotations were similar to ChromHMM annotations. (**b**) CoRE-ATAC predictions for active regulatory regions identified by STARR-seq in A549 cell line. The majority of active enhancers identified by STARR-seq were predicted as promoter or enhancer by CoRE-ATAC. (**c**) MIN6 MPRA log fold change values for genomic regions predicted as losing or gaining RE function based on CoRE-ATAC probabilities for reference and alternative alleles. Significance for predicted loss and predicted gain categories was calculated using student's t-test for MPRA log fold change values being less than or greater than 0 respectively. Significance comparing the predicted loss and predicted gain of MPRA fold change distributions was calculated using Mann-Whitney U test. We observed concordant direction of effect both for CoRE-ATAC predictions and MPRA acitivity levels when alternative and reference alleles are compared. (**d**) Genome browsers of 19 islet samples highlighting a loss of enhancer activity for rs11205653 (also highlighted in (**c**)) for the alternative allele (G). Values for enhancer and other represent the probability assigned to those classes of REs by CoRE-ATAC. We observe that for 5 out of 7 individuals with the reference allele (TT) CoRE-ATAC predicted enhancer activity, reflecting ChromHMM reference annotations, while for the individuals with GT or GG alleles, we observed an enhancer activity loss for all but one individual based on CoRE-ATAC predictions.

823

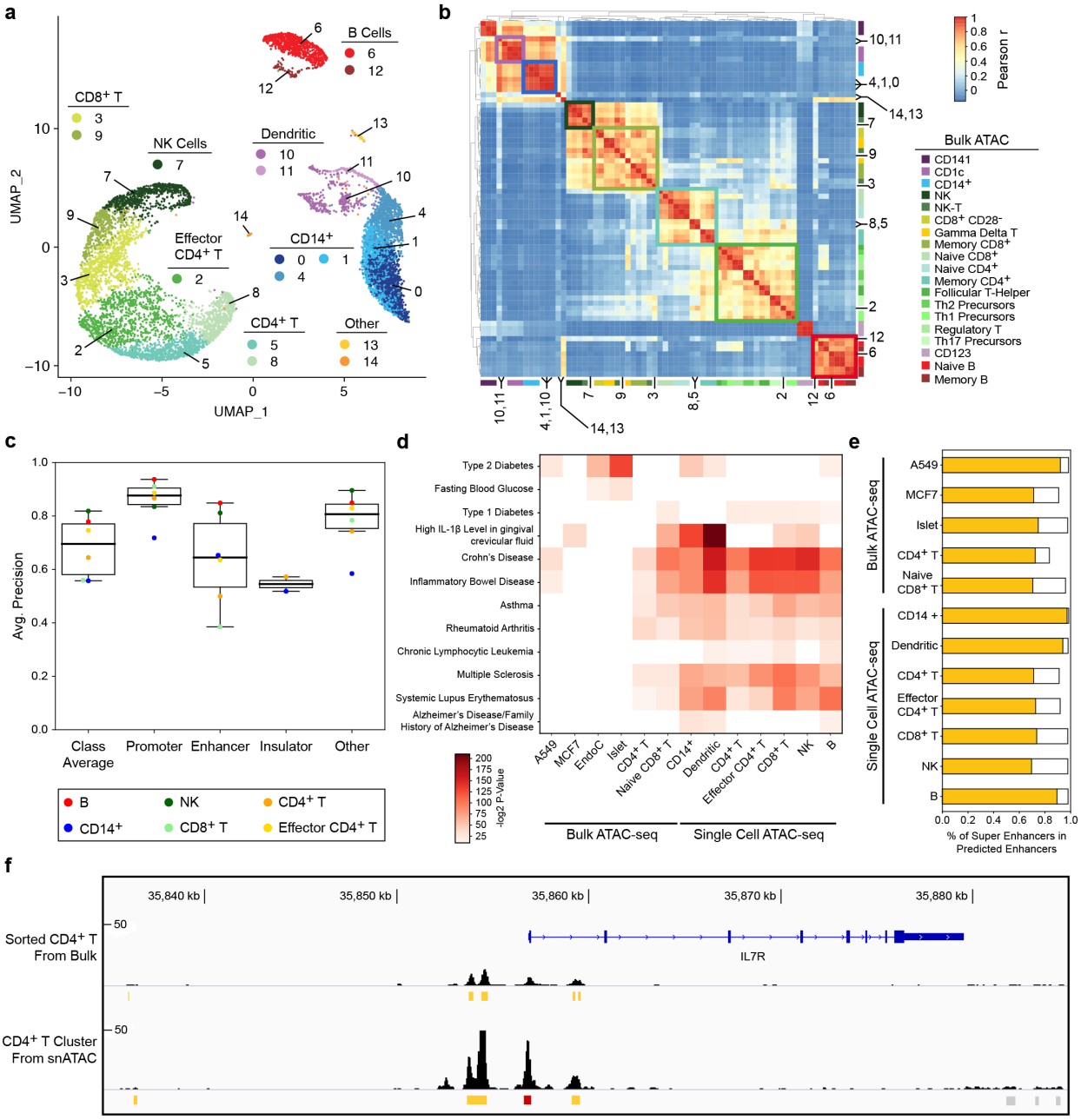

Fig. 5. Predicting functionality of REs from PBMC snATAC data. (**a**) Single cell clusters annotated for 7 immune cell types. Two-pass clustering identified a total of 15 cell clusters which we annotated using hierarchical clustering with sorted bulk ATAC-seq data (shown in (**b**)) to identify 7 different immune cells corresponding to these clusters. (**b**) Hierarchical clustering of scATAC clusters with bulk ATAC-seq data. Numbers and highlighted regions within the heatmap correspond to cell clusters and annotations in (**a**). 7 immune cell types were observed with both scATAC and bulk ATAC-seq samples. (**c**) Average precision values for predicting RE function in scATAC for the 7 annotated clusters. Model performances suggest that CoRE-ATAC is an effective tool for interrogating *cis*-RE activity from snATAC data. (**d**) GREGOR SNP enrichment analysis highlighting selected diseases whose SNPs were significantly enriched within the enhancer elements predicted by CoRE-ATAC. Enhancers from scATAC-seq were significantly enriched for SNPs associated with immune diseases. (**e**) Percent of super enhancers detected among CoRE-ATAC enhancers. Note CoRE-ATAC's ability to identify cell-type-specific enhancers that are most relevant to disease. (**f**) Genome browser view of *IL7R* for bulk ATAC and snATAC samples for CD4$^+$ T cells. ATAC-seq read profiles and CoRE-ATAC predictions between snATAC and bulk ATAC were found to be similar to one another, demonstrating CoRE-ATAC as a robust method for *cis*-RE predictions. Red represents promoter predictions, yellow represent enhancer predictions , and gray represent "other" predictions from CoRE-ATAC.

824

825

826 **Table 1.** ATAC-seq samples used in model training and evaluation.

Cell Type	Usage	#	Data Accession Id
$CD14^+$	Model Training	2	EGAS00001002605
K562	Model Training	2	GSE121993
HSMM	Model Training	1	GSE109828
GM12878	Model Training	1	GSE47753
Pancreatic Islets	Cross-cell Predictions	19	SRP117935
Naïve $CD8^+$ T	Cross-cell Predictions	6	EGAS00001002605
PBMC	Cross-cell Predictions	6	EGAS00001002605
Naïve $CD8^+$ T	Cross-cell Predictions	4	GSE118189
MCF7	Cross-cell Predictions	2	GSE97583
A549	Cross-cell Predictions	1	GSE117089
$CD4^+$ T	Cross-cell Predictions	1	GSE47753
EndoC	Cross-cell Predictions	1	GSE118588
PBMC	snATAC Predictions	1	GSE129785

827

828 **Table 2.** Chromosomes selected for training, validation, and test data.

Category	Selected Chromosomes
Training	1, 4, 5, 6, 7, 8, 9, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22
Validation	2, 10
Testing	3, 11

829

830