

1 **Interspecies regulatory landscapes and elements revealed by novel joint systematic**
2 **integration of human and mouse blood cell epigenomes**

3
4 Guanjue Xiang^{1,2,3}, Xi He¹, Belinda M. Giardine⁴, Kathryn J. Isaac⁵, Dylan J. Taylor⁵, Rajiv C.
5 McCoy⁵, Camden Jansen⁴, Cheryl A. Keller⁴, Alexander Q. Wixom⁴, April Cockburn⁴, Amber
6 Miller⁴, Qian Qi⁶, Yanghua He^{6,7}, Yichao Li⁶, Jens Lichtenberg⁸, Elisabeth F. Heuston⁸, Stacie M.
7 Anderson⁹, Jing Luan¹⁰, Marit W. Vermunt¹⁰, Feng Yue¹¹, Michael E.G. Sauria¹², Michael C.
8 Schatz¹², James Taylor^{5,12}, Berthold Göttgens¹³, Jim R. Hughes¹⁴, Douglas R. Higgs¹⁴, Mitchell
9 J. Weiss⁶, Yong Cheng⁶, Gerd A. Blobel¹⁰, David M. Bodine⁸, Yu Zhang¹⁵, Qunhua Li^{15,16}, Shaun
10 Mahony^{4,16,17}, Ross C. Hardison^{4,16,17 *}

11

12 ¹Bioinformatics and Genomics Graduate Program, Huck Institutes of the Life Sciences, The
13 Pennsylvania State University, University Park, PA 16802

14 ²Department of Data Science, Dana-Farber Cancer Institute, Boston, MA 02215

15 ³Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02215

16 ⁴Department of Biochemistry and Molecular Biology, The Pennsylvania State University,
17 University Park, PA 16802

18 ⁵Department of Biology, Johns Hopkins University, Baltimore, MD 21218

19 ⁶Department of Hematology, St. Jude Children's Research Hospital, Memphis, TN 38105

20 ⁷Department of Human Nutrition, Food and Animal Sciences, University of Hawai`i at Mānoa,
21 Honolulu, HI 96822, USA

22 ⁸Genetics and Molecular Biology Branch, National Human Genome Research Institute,
23 Bethesda, MD 20892

24 ⁹Flow Cytometry Core, National Human Genome Research Institute, Bethesda, MD 20892

25 ¹⁰Department of Pediatrics, Children's Hospital of Philadelphia, and Perelman School of
26 Medicine, University of Pennsylvania, Philadelphia, PA 19104

27 ¹¹Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine,
28 Northwestern University, Evanston, IL 60611
29 ¹²Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218
30 ¹³Welcome and MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK
31 ¹⁴MRC Weatherall Institute of Molecular Medicine, Oxford University, Oxford, UK
32 ¹⁵Department of Statistics, The Pennsylvania State University, University Park, PA 16802
33 ¹⁶Center for Computational Biology and Bioinformatics, Genome Sciences Institute, Huck
34 Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA
35 16802
36 ¹⁷Center for Eukaryotic Gene Regulation, The Pennsylvania State University, University Park,
37 PA 16802

38

39 * Corresponding author: Ross C. Hardison, Department of Biochemistry and Molecular Biology,
40 The Pennsylvania State University, 304 Wartik Lab, University Park, PA 16802, Phone: 814-
41 863-0113; E-mail: rch8@psu.edu

42 *Running Title:* Integrated epigenomic profiles between species

43 *Key words:* epigenetics, gene regulatory elements, dimensional reduction, regulatory potential

44

45

46

47

48

49

50

51 **Abstract**

52

53 Knowledge of locations and activities of *cis*-regulatory elements (CREs) is needed to decipher
54 basic mechanisms of gene regulation and to understand the impact of genetic variants on
55 complex traits. Previous studies identified candidate CREs (cCREs) using epigenetic features in
56 one species, making comparisons difficult between species. In contrast, we conducted an
57 interspecies study defining epigenetic states and identifying cCREs in blood cell types to
58 generate regulatory maps that are comparable between species, using integrative modeling of
59 eight epigenetic features jointly in human and mouse in our **Validated **S**ystematic **I**ntegration**
60 (VISION) Project. The resulting catalogs of cCREs are useful resources for further studies of
61 gene regulation in blood cells, indicated by high overlap with known functional elements and
62 strong enrichment for human genetic variants associated with blood cell phenotypes. The
63 contribution of each epigenetic state in cCREs to gene regulation, inferred from a multivariate
64 regression, was used to estimate epigenetic state Regulatory Potential (esRP) scores for each
65 cCRE in each cell type, which were used to categorize dynamic changes in cCREs. Groups of
66 cCREs displaying similar patterns of regulatory activity in human and mouse cell types, obtained
67 by joint clustering on esRP scores, harbored distinctive transcription factor binding motifs that
68 were similar between species. An interspecies comparison of cCREs revealed both conserved
69 and species-specific patterns of epigenetic evolution. Finally, we showed that comparisons of
70 the epigenetic landscape between species can reveal elements with similar roles in regulation,
71 even in the absence of genomic sequence alignment.

72

73

74

75

76 **Introduction**

77

78 The morphology and functions of different cell types are determined by the expression of
79 distinctive sets of genes in each. This differential gene expression is regulated by the interplay
80 of transcription factors (TFs) binding to *cis*-regulatory elements (CREs) in the genomic DNA,
81 such as promoters and enhancers, forging interactions among the CREs and components of
82 transcriptional apparatus and ultimately leading to patterns of gene activation and repression
83 characteristic of each cell type (Maston et al. 2006; Hamamoto and Fukaya 2022). Epigenetic
84 features such as accessibility of DNA and modifications of histone tails in chromatin have
85 pronounced impacts on the ability of TFs to bind to CREs, and furthermore, they serve as a
86 molecular memory of transcription and repression (Strahl and Allis 2000; Ringrose and Paro
87 2004). Frequently co-occurring sets of chromatin features define epigenetic states, which are
88 associated with gene regulation and expression (Ernst and Kellis 2010; Hoffman et al. 2013;
89 Zhang et al. 2016). Genome-wide assignment of DNA intervals to epigenetic states (annotation)
90 provides a view of the regulatory landscape that can be compared across cell types, which in
91 turn leads to insights into the processes regulating gene expression (Libbrecht et al. 2021).

92

93 Comprehensive mapping of CREs within the context of the regulatory landscape in different cell
94 types is needed to achieve a broad understanding of differential gene expression. Maps of
95 candidate CREs (cCREs) provide guidance in understanding how changes in cCREs, including
96 single nucleotide variants and indels, can lead to altered expression (Hardison 2012), and they
97 can inform approaches for activation or repression of specific genes in potential strategies for
98 therapies (Bauer et al. 2013). Indeed, most human genetic variants associated with common
99 traits and diseases are localized in or near cCREs (Hindorff et al. 2009; Maurano et al. 2012;
100 The ENCODE Project Consortium 2012). Thus, knowledge of the activity and epigenetic state of

101 cCREs in each cell type can facilitate understanding the impact of trait-associated genetic
102 variants on specific phenotypes. Furthermore, genome editing approaches in somatic cells have
103 recently been demonstrated to have promise as therapeutic modalities (Frangoul et al. 2021),
104 and a full set of cCREs annotated by activity and state can help advance similar applications.

105

106 The different types of blood cells in humans and mice are particularly tractable systems for
107 studying many aspects of gene regulation during differentiation. The striking differences among
108 mature cell types result from progressive differentiation starting from a common hematopoietic
109 stem cell (HSC) (Kondo et al. 2003). While single cell analyses reveal a pattern of ostensibly
110 continuous expression change along each hematopoietic lineage (Laurenti and Göttgens 2018),
111 intermediate populations of multi-lineage progenitor cells with decreasing differentiation
112 potential have been defined, which provide an overall summary and nomenclature for major
113 stages in differentiation. These stem, progenitor, and mature cell populations can be isolated
114 using characteristic cell surface markers (Spangrude et al. 1988; Payne and Crooks 2002),
115 albeit with many fewer cells in progenitor populations. In addition to the primary blood cells,
116 several immortalized cell lines provide amenable systems for intensive study of various aspects
117 of gene regulation during differentiation and maturation of blood cells (Weiss et al. 1997).

118

119 The VISION project aims to produce a Validated Systematic Integration of hematopoietic
120 epigenomes, harvesting extensive epigenetic and transcriptomic datasets from many
121 investigators and large consortia into concise, systematically integrated summaries of regulatory
122 landscapes and cCREs (Hardison et al. 2020). We previously published the results of these
123 analyses for progenitor and mature blood cell types from mouse (Xiang et al. 2020). In the
124 current study, we generated additional epigenetic datasets and compiled data from human
125 blood cells to expand the integrative analyses to include data from both human and mouse. The
126 systematic integrative analysis of epigenetic features across blood cell types was conducted

127 jointly in both species to learn epigenetic states, generate concise views of epigenetic
128 landscapes, and predict regulatory elements that are comparable in both species. This joint
129 modeling enabled further comparisons using approaches that were not dependent on DNA
130 sequence alignments between species, including a demonstration of the role of orthologous
131 transcription factors in cell type-specific regulation in both species. An exploration of
132 comparisons of epigenetic landscapes between species showed that they were informative for
133 inferring regulatory roles of elements in lineage-specific (i.e., non-aligning) DNA. Together, this
134 work provides valuable community resources that enable researchers to leverage the extensive
135 existing epigenomic data into further mechanistic regulatory studies of both individual loci and
136 genome-wide trends in human and mouse blood cells.

137

138 **Results**

139

140 **Extracting and annotating epigenetic states by modeling epigenomic information jointly 141 in human and mouse**

142 A large number of data sets of epigenetic features related to gene regulation and expression
143 (404 data sets, 216 in human and 188 in mouse; Fig. 1, Supplemental Material “Data generation
144 and collection”, Supplemental Tables S1 and S2) served as the input for our joint integrative
145 analysis of human and mouse regulatory landscapes across progenitor and mature blood cell
146 types. The features included chromatin accessibility, which is a hallmark of almost all regulatory
147 elements, occupancy by the structural protein CTCF, and histone modifications associated with
148 gene activation or repression. After normalizing and denoising these diverse data sets
149 (Supplemental Fig. S1), we conducted an iterative joint modeling to discover epigenetic states,
150 i.e., sets of epigenetic features commonly found together, in a consistent manner for both
151 human and mouse blood cells (Fig. 2). The joint modeling took advantage of the Bayesian

152 framework of the Integrative and Discriminative Epigenomic Annotation System, or IDEAS
153 (Zhang et al. 2016; Zhang and Hardison 2017), to iteratively learn states in both species. The
154 joint modeling proceeded in four steps: initial training on randomly selected regions in both
155 species, retaining the 27 epigenetic states that exhibit similar combinatorial patterns of features
156 in both human and mouse, using these 27 states as prior information to sequentially run the
157 IDEAS genome segmentation on the human and mouse data sets, and removal of two
158 heterogenous states (Fig. 2A and Supplemental Figs. S2, S3, S4, and S5). This procedure
159 ensured that the same set of epigenetic states was learned and applied for both species.
160 Previously, the segmentation and genome annotation (Llibrech et al. 2021) method
161 ChromHMM (Ernst and Kellis 2012) was used to combine data between species by
162 concatenating the datasets for both human and mouse cell types (Yue et al. 2014). This earlier
163 approach produced common states between species, but it did not benefit from the positional
164 information and automated approach to handling missing data that are embedded in IDEAS.
165
166 The resulting model with 25 epigenetic states (Fig. 2B) was similar to that obtained from mouse
167 blood cell data (Xiang et al. 2020). The states captured combinations of epigenetic features
168 characteristic of regulatory elements such as promoters and enhancers, transcribed regions,
169 repressed regions marked by either Polycomb (H3K27me3) or heterochromatin (H3K9me3),
170 including states that differ quantitatively in the contribution of specific features to each state. For
171 example, H3K4me1 is the predominant component of states E1 and E, but E1 has a lower
172 contribution of that histone modification. Similar proportions of the genomes of human and
173 mouse were covered by each state (Fig. 2B).
174
175 Assigning all genomic bins in human and mouse to one of the 25 states in each hematopoietic
176 cell type produced an annotation of blood cell epigenomes that gave a concise view of the
177 epigenetic landscape and how it changes across cell types, using labels and color conventions

178 consistently for human and mouse. The value of this concise view can be illustrated in
179 orthologous genomic intervals containing genes expressed preferentially in different cell
180 lineages as well as genes that are uniformly expressed (Fig. 2C, D). For example, the gene
181 *SLC4A1/Slc4a1*, encoding the anion transporter in the erythrocyte plasma membrane, is
182 expressed in the later stages of erythroid maturation (Dore and Crispino 2011). The epigenetic
183 state assignments across cell types matched the differential expression pattern, with genomic
184 intervals in the gene and its flanking regions, including a non-coding gene located upstream (to
185 its right, *Bloodlinc* in mouse), assigned to states indicative of enhancers (yellow and orange)
186 and promoters (red) only in erythroid cell types, with indications of stronger activation in the
187 more mature erythroblasts (region boxed and labeled E in Fig. 2 C, D). A similar pattern was
188 obtained in both human and mouse. Those genomic intervals assigned to the enhancer- or
189 promoter-like states contain candidates for regulatory elements, an inference that was
190 supported by chromatin binding data including occupancy by the transcription factor GATA1 (Xu
191 et al. 2012; Pimkin et al. 2014) and the co-activator EP300 (ENCODE datasets ENCSR000EGE
192 and ENCSR982LJQ) in erythroid cells. Similarly, the gene and flanking regions for *GRN/Grn*,
193 encoding the granulin precursor protein that is produced at high levels in granulocytes and
194 monocytes (Jian et al. 2013), and *ITGA2B/Itga2b*, encoding the alpha 2b subunit of integrin that
195 is abundant in mature megakaryocytes (van Pampus et al. 1992; Pimkin et al. 2014), were
196 assigned to epigenetic states indicative of enhancers and promoters in the expressing cell types
197 (boxed regions labeled G and MK, respectively). In contrast, genes expressed in all the blood
198 cell types, such as *UBTF/Ubtf*, were assigned to active promoter states and transcribed states
199 across the cell types. We conclude that these concise summaries of the epigenetic landscapes
200 across cell types showed the chromatin signatures for differential or uniform gene expression
201 and revealed discrete intervals as potential regulatory elements, with the consistent state
202 assignments often revealing similar epigenetic landscapes of orthologous genes in human and
203 mouse.

204

205 While these resources are useful, some limitations should be kept in mind. For example, IDEAS
206 used data from similar cell types to improve state assignments in cell types with missing data,
207 but the effectiveness of this approach may be impacted by the pattern of missing data. In
208 particular, the epigenetic data on human stem and progenitor cell types were largely limited to
209 ATAC-seq data, whereas histone modification data and CTCF occupancy were available for the
210 analogous cell types in mouse (Fig. 1). Thus, the state assignments for epigenomes in human
211 stem and progenitor cells may be less robust compared to those for similar cell types in mouse.
212 Another limitation is the broad range of quality in the data sets that cannot be completely
213 adjusted by normalization, which leads to over- or under-representation of some epigenetic
214 signals in specific cell types (Supplemental Fig. S5). Despite these limitations, the annotation of
215 blood cell epigenomes after normalization and joint modeling of epigenetic states produced a
216 highly informative painting of the activity and regulatory landscapes across the genomes of
217 human and mouse blood cells.

218

219 **Candidate *cis*-regulatory elements in human and mouse**

220 We define a candidate *cis*-regulatory element, or cCRE, as a DNA interval with a high signal for
221 chromatin accessibility in any cell type (Xiang et al. 2020). We utilized a version of the IDEAS
222 methodology to combine peaks of accessibility across different cell types, running it in the signal
223 intensity state (IS) mode only on chromatin accessibility signals (Xiang et al. 2021), which helps
224 counteract excessive expansion of peak calls when combining them (Supplemental Fig. S6).

225

226 Employing the same peak-calling procedure to data from human and mouse resulted in 200,342
227 peaks of chromatin accessibility for human and 96,084 peaks for mouse blood cell types
228 (Supplemental Table S3). Applying the peak caller MACS3 (Zhang et al. 2008) on the same
229 human ATAC-seq data generated a larger number of peaks, but those additional peaks tended

230 to have low signal and less enrichment for overlap with other function-related genomic datasets
231 (Supplemental Fig. S7).

232

233 The ENCODE Project released regulatory element predictions in a broad spectrum of cell types
234 in the Index of DHSs (Meuleman et al. 2020) and the SCREEN cCRE catalog (The ENCODE
235 Project Consortium et al. 2020), using data that were largely different from those utilized for the
236 VISION analyses. Almost all the VISION cCRE calls in human blood cells were included in the
237 regulatory element predictions from ENCODE (Supplemental Fig. S8A), supporting the quality
238 of the VISION cCRE calls. Furthermore, as expected from its focus on blood cell types, the
239 VISION cCRE catalog shows stronger enrichment for regulatory elements active in blood cells
240 (Supplemental Fig. S8B, Supplemental Table S4).

241

242 **Enrichment of the cCRE catalog for function-related elements and trait-associated
243 genetic variants**

244 Having generated catalogs of cCREs along with an assignment of their epigenetic states in
245 each cell type, we characterized the human cCREs further by connecting them to orthogonal
246 (not included in VISION predictions) datasets of DNA elements implicated in gene regulation or
247 in chromatin structure and architecture (termed structure-related) (Fig. 3A, Supplemental Fig.
248 S9, Supplemental Table S5). About two-thirds (136,664 or 68%) of the VISION human cCREs
249 overlapped with elements in the broad groups of CRE-related (97,361 cCREs overlapped) and
250 structure-related (83,327 cCREs overlapped) elements, with 44,024 cCREs overlapping
251 elements in both categories (Fig. 3A, B). In contrast, ten sets of randomly chosen DNA intervals,
252 matched in length and GC-content with the human cCRE list, showed much less overlap with
253 the orthogonal sets of elements (Fig. 3B). Of the CRE-related superset, the enhancer-related
254 group of datasets contributed the most overlap with VISION cCREs, followed by SuRE peaks,
255 which measure promoter activity in a massively parallel reporter assay (van Arensbergen et al.

256 2017), and CpG islands (Fig. 3C). Compared to overlaps with the random matched intervals, the
257 VISION cCREs were highly enriched for overlap with each group of CRE-related datasets (Fig.
258 3C). Of the structure-related superset, the set of CTCF occupied segments (OSs) contributed
259 the most overlap, followed by chromatin loop anchors, again with high enrichment relative to
260 overlaps with random matched sets (Fig. 3D). Considering the VISION cCREs that intersected
261 with both structure- and CRE-related elements, major contributors were the cCREs that overlap
262 with enhancers and CTCF OSs or loop anchors (Supplemental Fig. S10). Furthermore, the
263 VISION cCREs captured known blood cell CREs (Supplemental Table S4) and CREs
264 demonstrated to impact a specific target gene in a high throughput analysis (Gasperini et al.
265 2019) (Fig. 3E). We conclude that the intersections with orthogonal, function- or structure-
266 related elements lent strong support for the biological significance of the VISION cCRE calls
267 and added to the annotation of potential functions for each cCRE.

268

269 The catalog of VISION human blood cell cCREs showed a remarkable enrichment for genetic
270 variants associated with blood cell traits, further supporting the utility of the catalog. We initially
271 observed a strong enrichment by overlap with variants from the NHGRI-EBI GWAS Catalog
272 (Buniello et al. 2019) associated with blood cell traits (Supplemental Fig. S11). We then
273 analyzed the enrichments while considering the haplotype structure of human genomes,
274 whereby association signals measured at assayed genetic markers likely reflect an indirect
275 effect driven by linkage disequilibrium (LD) with a causal variant (that may or may not have
276 been genotyped). We employed stratified linkage disequilibrium score regression (sLDSC,
277 Finucane et al. 2015) to account for LD structure and estimate the proportion of heritability of
278 each trait explained by a given genomic annotation, quantifying the enrichment of heritability in
279 587 traits from the UK Biobank (UKBB) GWAS (Ge et al. 2017 and <http://www.nealelab.is/uk-biobank/>) within the VISION cCREs relative to the rest of the genome (Supplemental Material
280 section “Stratified linkage disequilibrium score regression”). These traits encompassed 54

282 “blood count” traits that measure properties including size and counts of specific blood cell
283 types, 60 “blood biochemistry” traits that measure lipid, enzyme, and other molecular
284 concentrations within whole blood samples, and 473 non-blood-related traits, allowing us to
285 assess the specific relevance of the cCREs to regulation of blood-related versus other
286 phenotypes. At a 5% FDR threshold, we discovered 53 traits for which cCREs were significantly
287 enriched in heritability (Fig. 3F). Of these traits, 52 (98%) were blood-related and 50 were blood
288 count traits, representing 93% of all UKBB blood count traits included in our analysis. The
289 remaining 2 significant traits pertained to blood biochemistry, specifically, the male and female
290 glycated hemoglobin concentrations. These metrics and observations together lend support to
291 the VISION cCRE annotation being composed of informative genomic regions associated with
292 regulation of genes involved in development of blood cell traits.

293

294 **Estimates of regulatory impact of cCREs during differentiation**

295 The epigenetic states assigned to cCREs can reveal those that show changes in apparent
296 activity during differentiation. Inferences about the activity of a cCRE in one or more cell types
297 are based on whether the cCRE was actuated, i.e., was found in a peak of chromatin
298 accessibility, and which epigenetic state was assigned to the actuated cCRE. Those states can
299 be associated with activation (e.g., enhancer-like or promoter-like) or repression (e.g.,
300 associated with polycomb or heterochromatin). In addition to these categorical state
301 assignments, quantitative estimates of the impact of epigenetic states on expression of target
302 genes are useful, e.g., to provide an estimate of differences in inferred activity when the states
303 change. Previous work used signals from single or multiple individual features such as
304 chromatin accessibility or histone modifications in regression modeling to explain gene
305 expression (e.g., Karlić et al. 2010; Dong et al. 2012), and we applied a similar regression
306 modeling using epigenetic states as predictor variables to infer estimates of regulatory impact of
307 each state on gene expression (Xiang et al. 2020).

308

309 We used state assignments of cCREs across cell types in a multivariate regression model to
310 estimate the impact of each state on the expression of local genes (Supplemental Material,
311 “Estimation of the impact of epigenetic states and cCREs on gene expression”). That impact
312 was captured as β coefficients, which showed the expected strong positive impact for promoter
313 and enhancer associated states and negative impacts from heterochromatin and polycomb
314 states (Fig. 4A). The β coefficients were then used in further analysis, such as estimating the
315 change in regulatory impact as a cCRE shifts between states during differentiation (difference
316 matrix to the left of the β coefficient values in Fig. 4A). The β coefficient values also were used
317 to generate an epigenetic state Regulatory Potential (esRP) score for each cCRE in each cell
318 type, calculated as the β coefficient values for the epigenetic states assigned to the cCRE
319 weighted by the coverage of the cCRE by each state (Fig. 4B). These esRP scores were the
320 basis for visualizing the collection of cCREs and how their regulatory impact changed across
321 differentiation (Supplemental Fig. S12 and Supplemental movie S1). Comparison of the
322 integrative esRP scores with signal intensities for single features (ATAC-seq and H3K27ac)
323 showed all were informative for visualizations, and esRP performed slightly better than the
324 single features in differentiating cCREs based on locations within gene bodies (Supplemental
325 Fig. S13).

326

327 In addition, we explored the utility of the esRP scores for clustering the cCREs into groups with
328 similar activity profiles across blood cell types in both human and mouse. Focusing on the esRP
329 scores in 12 cell types shared between human and mouse along with the average across cell
330 types, we identified clusters jointly in both species. The clustering proceeded in three steps,
331 specifically finding robust k -means clusters for the combined human and mouse cCREs,
332 identifying the clusters shared by cCREs in both species, and then further grouping those
333 shared k -means clusters hierarchically to define fifteen joint metaclusters (JmCs) (Supplemental

334 Fig. S14). Each cCRE in both mouse and human was assigned to one of the fifteen JmCs, and
335 each JmC was populated with cCREs from both mouse and human.

336

337 These JmCs established discrete categories for the cCREs based on the cell type distribution of
338 their inferred regulatory impact (Fig. 4C). The clusters of cCREs with high esRP scores across
339 cell types were highly enriched for promoter elements (Supplemental Fig. S15A). The cell type-
340 restricted clusters of cCREs showed enrichment both for selected enhancer catalogs and for
341 functional terms associated with those cell types (Supplemental Fig. S15A and B). Furthermore,
342 clustering of human genes by the JmC assignments of cCREs in a 100kb interval centered on
343 their TSS (Supplemental Material section “Enrichment of JmCs assigned to cCREs in gene
344 loci”) revealed a strong enrichment for JmCs with high activity in the cell type(s) in which the
345 genes are expressed (Fig. 4D). Examples include *IFNG* showing enrichment for JmC 12, which
346 has high esRP scores in T and NK cells, *CSF1R* showing enrichment for JmC 15, which has
347 high scores in monocytes, and *GATA1* showing enrichment for JmC 10, which has high scores
348 in erythroid cells and megakaryocytes. Moreover, running sLDSC on cCREs in individual JmCs
349 showed enrichment for heritability of blood cell-related traits in some specific JmCs
350 (Supplemental Fig. S16).

351

352 As expected from previous work (e.g., Heintzman et al. 2009; Meuleman et al. 2020), similar
353 metaclusters of cCREs were generated based on single signals from the histone modification
354 H3K27ac or chromatin accessibility across cell types (Supplemental Fig. S17). Clustering based
355 any of the three features better resolved individual cell types when larger numbers of clusters
356 were considered, prior to collapsing the shared robust clusters into JmCs (Supplemental Fig.
357 S18).

358

359 In summary, we show that the β coefficients and esRP scores provide valuable estimates of
360 regulatory impacts of states and cCREs, respectively. The esRP-driven joint metaclusters
361 provide refined subsets of cCREs that should be informative for investigating cell type-specific
362 and general functions of cCREs. We also built self-organizing maps as a complementary
363 approach to systematic integration of epigenetic features and RNA data across cell types
364 (Supplementary Figure S10, Jansen et al. 2019).

365

366 **Motif enrichment in joint metaclusters of human and mouse cCREs**

367 We examined the sets of cCREs in each JmC to ascertain enrichment for transcription factor
368 binding site (TFBS) motifs because these enriched motifs suggest the families of transcription
369 factors that play a major role in regulation by each category of cCREs. Furthermore, having sets
370 of cCREs determined and clustered for comparable blood cell types in human and mouse
371 provided the opportunity to discover which TFBS motifs were shared between species and
372 whether any were predominant in only one species.

373

374 To find TFBS motifs associated with each JmC, we calculated enrichment for all non-redundant
375 motifs in the Cis-BP database (Weirauch et al. 2014) using Maelstrom from GimmeMotifs
376 (Bruse and van Heeringen 2018) (Supplemental Material “Enrichment for transcription factor
377 binding site motifs in joint metaclusters of cCREs”). The results confirmed previously
378 established roles of specific TFs in cell lineages and showed little evidence for novel motifs (Fig.
379 4E). For example, TFBS motifs for the GATA family of transcription factors were enriched in
380 JmCs 2 and 10, which have high esRP scores in progenitor and mature cells in the erythroid
381 and megakaryocytic lineages, as expected for the known roles of GATA1 and GATA2 in this
382 lineage (Blobel and Weiss 2009; Fujiwara et al. 2009). The GATA motif was also enriched in
383 JmC 14, as expected for the role of GATA3 in natural killer (NK) and T cells (Rothenberg and
384 Taghon 2005). Furthermore, motifs for the known lymphoid transcription factors TBX21,

385 TCF7L1, and LEF1 (Chi et al. 2009) were enriched in cCREs with high esRP scores in NK and
386 T cells (JmCs 9 and 12), and motifs for myeloid-determining transcription factors CEBPA and
387 CEBPB (Graf and Enver 2009) and the myeloid transcription factor PU.1 (Tenen et al. 1997)
388 were enriched in cCREs that are active in progenitor cells and monocytes (JmCs 3 and 15).
389 TFBS motifs for promoter-associated transcription factors such as E2F2 and SP1 (Dynan and
390 Tjian 1983; Kaczynski et al. 2003) were enriched in broadly active cCREs (JmCs 1 and 4).
391 These patterns of motif enrichments in the JmCs fit well with the expectations from previous
392 studies of transcription factor activity across lineages of blood cells, and thus, they lend further
393 credence to the value of the cCRE calls and the JmC groupings for further studies of regulation
394 in the blood cell types.

395
396 The genome-wide collection of cCREs across many blood cell types in human and mouse
397 provided an opportunity for an unbiased and large-scale search for indications of transcription
398 factors that may be active specifically in one species for a shared cell type. Prior studies of
399 transcription factors have shown homologous transcription factors used in analogous cell types
400 across species (e.g., Carroll 2008; Noyes et al. 2008; Schmidt et al. 2010; Cheng et al. 2014;
401 Villar et al. 2014), but it is not clear if there are significant exceptions. In our study, we found that
402 for the most part, the motif enrichments were quite similar between the human and mouse
403 cCREs in each JmC. Note that these similarities were not forced by requiring sequence
404 matches between species; the cCREs were grouped into JmCs based on their pattern of
405 activity, as reflected in the esRP scores, across cell types, not by requiring homologous
406 sequences. This similarity between species indicates that the same transcription factors tend to
407 be active in similar groups of cell types in both mouse and human. An intriguing potential
408 exception to the sharing of motifs between species was the enrichment of TFBS motifs for
409 CTCF and ZBTB7A in some JmCs, suggestive of some species selectivity in their binding in the
410 context of other TFs (Supplemental Figs. S20 and S21). These indications of conditional,

411 preferential usage of these TFs in human or mouse could serve as the basis for more detailed
412 studies in the future.

413

414 In summary, after grouping the cCREs in both human and mouse by their inferred regulatory
415 impact across blood cell in a manner agnostic to DNA sequence or occupancy by TFs, the
416 enrichment for TFBS motifs within those groups recapitulated known activities of TFs both
417 broadly and in specific cell lineages. The results also showed considerable sharing of inferred
418 TF activity in both human and mouse.

419

420 **Evolution of sequence and inferred function of cCREs**

421 The human and mouse cCREs from blood cells were assigned to three distinct evolutionary
422 categories (Fig. 5A). About one-third of the cCREs were present only in the reference species
423 (39% for human, 28% for mouse), as inferred from the failure to find a matching orthologous
424 sequence in whole-genome alignments with the other species. We refer to these as
425 nonconserved (N) cCREs. Of the two-thirds of cCREs with an orthologous sequence in the
426 second species, slightly over 30,000 were also identified as cCREs in the second species. The
427 latter cCREs comprise the set of cCREs conserved in both sequence and inferred function,
428 which we call SF conserved (SF) cCREs. Almost the same number of cCREs in both species
429 fall into the SF category; the small difference resulted from interval splits during the search for
430 orthologous sequences (Supplemental Fig. S22). The degree of chromatin accessibility in
431 orthologous SF cCREs was positively correlated between the two species (Supplemental Fig.
432 S23). The remaining cCREs (91,000 in human and 36,000 in mouse) were conserved in
433 sequence but not in an inferred function as a regulatory element, and we call them S conserved
434 (S) cCREs. The latter group could result from turnover of regulatory motifs or acquisition of
435 different functions in the second species.

436

437 The distributions of epigenetic states assigned to the blood cell cCREs in each of the three
438 evolutionary categories were similar between human and mouse, but those distributions differed
439 among evolutionary categories, with significantly more SF cCREs assigned to promoter-like
440 states than were S or N cCREs (Supplemental Fig. S24). Indeed, the SF cCREs tended to be
441 close to or encompass the TSSs of genes, showing a substantial enrichment in overlap with
442 TSSs compared to the overlap observed for all cCREs (Fig. 5B). Many of the S and N cCREs
443 were assigned to enhancer-like states (Supplemental Fig. S24D), giving a level of enrichment
444 for overlap with enhancer datasets comparable to that observed for the full set of cCREs (Fig.
445 5B).

446

447 For both human and mouse, the level of sequence conservation, estimated by the maximum
448 phyloP score (Pollard et al. 2010), was higher in the collection of cCREs than in sets of
449 randomly chosen genomic intervals matching the cCREs in length and G+C content (Fig. 5C).
450 Among the evolutionary categories of cCREs, the distribution of phyloP scores for SF cCREs
451 was significantly higher than the distribution for S cCREs, which in turn was higher than that for
452 N cCREs, for both species (Fig. 5C). The whole genome alignments underlying the phyloP
453 scores are influenced by proximity to the highly conserved coding exons (King et al. 2007), and
454 the high phyloP scores of the promoter-enriched SF cCREs could reflect both this effect as well
455 as strong constraint on conserved function (Supplemental Fig. S25). In all three evolutionary
456 categories, the distribution of phyloP scores was higher for promoter-proximal cCREs than for
457 distal ones, but the relative levels of inferred conservation were the same for both, i.e., SF>S>N
458 (Supplemental Fig. S26).

459

460 In summary, this partitioning of the cCRE catalogs by conservation of sequence and inferred
461 function revealed informative categories that differed both in evolutionary trajectories and in
462 types of functional enrichment.

463

464 Conservation of non-coding genomic DNA sequences among species has been used
465 extensively to predict regulatory elements (Gumucio et al. 1992; Hardison 2000; Pennacchio
466 and Rubin 2001), but the observation that predicted regulatory elements fall into distinct
467 evolutionary categories (SF, S, and N) raised the question of whether inter-species DNA
468 sequence alignments or annotation of epigenetic states would be more effective in finding
469 elements that were experimentally determined to be active in gene regulation. Recent advances
470 in massively parallel reporter assays have enabled the testing of large sets of candidate
471 elements, approaching comprehensive assessment of the predicted elements (Agarwal et al.
472 2023). We used the set of over 57,000 human genomic elements shown to be active in K562
473 cells to address this question (Supplemental Material), and we found that requiring alignment to
474 the mouse genome would miss about 40% of the active elements, whereas requiring presence
475 in a non-quiescent epigenetic state or one associated with gene activation would cover 87% or
476 82.5%, respectively, of the active elements (Fig. 5D). Thus, the epigenetic state annotation can
477 enable a more comprehensive prediction and examination of gene regulatory elements. This
478 realization motivated a comparison of epigenetic states between human and mouse, as
479 described in the next section.

480

481 **Comparison of epigenetic states around orthologous genes in human and mouse**

482 The consistent state assignments from the joint modeling facilitated epigenetic comparisons
483 between species. Such comparisons are particularly informative for orthologous genes with
484 similar expression patterns but some differences in their regulatory landscapes. For example,
485 the orthologous genes *GATA1* in human and *Gata1* in mouse each encode a transcription factor
486 with a major role in regulating gene expression in erythroid cells, megakaryocytes, and
487 eosinophils (Ferreira et al. 2005), with a similar pattern of gene expression across blood cell
488 types in both species (Supplemental Fig. S27). The human and mouse genomic DNA

489 sequences aligned around these orthologous genes, including their promoters and proximal
490 enhancers; the alignments continued through the genes downstream of *GATA1/Gata1* (Fig. 6A).
491 An additional, distal regulatory element located upstream of the mouse *Gata1* gene, which was
492 bound by GATA1 and EP300 (Fig. 6A), was found only in mouse (Valverde-Garduno et al.
493 2004). The DNA sequences of the upstream interval harboring the mouse regulatory element
494 did not align between mouse and human except in portions of the *GLOD5/Glod5* genes (Fig.
495 6A). Thus, the interspecies sequence alignments provide limited information about this distal
496 regulatory element.

497

498 This limitation to sequence alignments led us to explore whether comparisons of epigenetic
499 information would be more informative, utilizing the consistent assignment of epigenetic states
500 in both human and mouse, which do not rely on DNA sequence alignment. In the large genomic
501 regions (76kb and 101kb in the two species) encompassing the orthologous human *GATA1* and
502 mouse *Gata1* genes and surrounding genes, we computed the correlation for each genomic bin
503 between the epigenetic state assignments across cell types in one species and that in the other
504 species for all the bins (Supplemental Fig. S28). This local, all-versus-all comparison of the two
505 loci yielded a matrix of correlation values showing similarities and differences in profiles of
506 epigenetic states in the two species (Fig. 6B). The conserved promoter and proximal enhancers
507 of the *GATA1/Gata1* genes were highly correlated in epigenetic states across cell types
508 between the two species, in a region of the matrix that encompassed the aligning DNA
509 sequences (labeled Px in Fig. 6B). In contrast, whereas the mouse-specific distal regulatory
510 element did not align with the human DNA sequence, the epigenetic states annotating it
511 presented high correlations with active epigenetic states in the human *GATA1* locus (labeled D
512 in Fig. 6B).

513

514 The complexity of the correlation matrix (Fig. 6B) indicated that multiple epigenetic trends could
515 be contributing to the patterns. To systematically reduce the high dimensionality of the matrix to
516 a set of simpler matrices, we employed nonnegative matrix factorization (NMF) because of its
517 interpretability (Stein-O'Brien et al. 2018; Lee and Roy 2021). The decomposed matrices from
518 NMF revealed a set of factors, each of which (represented by each column in the mouse matrix
519 and each row in the human matrix in Fig. 6C) captures a group of highly correlated elements in
520 the original matrix that show a pattern distinct from the rest of the elements. The complex
521 correlation matrix was decomposed into six distinct factors, as determined by the number of
522 factors at which an “elbow” was found in the BIC score (Supplemental Fig. S29). Each factor
523 encapsulated a specific epigenetic regulatory machinery or process exhibiting consistent cross-
524 cell type patterns in both humans and mice (Supplemental Fig. S30). For example, the
525 correlation matrices reconstructed by using signals from factor 3 exclusively highlighted the cell
526 type-specific positive regulators for the *GATA1/Gata1* gene loci; these regulatory elements were
527 evident in reconstructed correlation matrices between species (Fig. 6D) and within individual
528 species (Fig. 6E). By applying a Z-score approach to identify peak regions in the factor 3 signal
529 vector (with FDR < 0.1; Supplemental Material), we pinpointed regions in both species showing
530 an epigenetic regulatory machinery exhibiting positive regulatory dynamics for the orthologous
531 *GATA1/Gata1* gene loci, particularly in the ERY and MK cell types. In contrast, the correlation
532 matrices reconstructed from the signals for factor 6 (Fig. 6F and G) highlighted regions marked
533 by the transcription elongation modification H3K36me3 (epigenetic states colored green, Fig.
534 6G). The correlations in the factor 6 elongation signature were observed, as expected, between
535 the human/mouse orthologous gene pairs *GATA1* and *Gata1* as well as between human
536 *HDAC6* and mouse *Hdac6* (green rectangles in Fig. 6F). The factor 6 correlations were also
537 observed between the *GATA1/Gata1* and *HDAC6/Hdac6* genes (black rectangles in Fig. 6F and
538 G), showing a common process, specifically transcriptional elongation, at both loci. A similar
539 analysis for other factors revealed distinct regulatory processes or elements, such as active

540 promoters (factor 2), exhibiting unique cross-cell type patterns (Supplemental Fig. 30). The
541 genomic bins with high scores for a given NMF factor in human showed high correlation with
542 bins with high scores for that same factor in mouse, indicating that the NMF factors capture a
543 similar set of epigenetic state patterns in each species (Supplemental Fig. S31). The patterns
544 captured by NMF factors 3 and 6 were robust to the choice of k in the NMF (Supplemental Fig.
545 S32). Overall, these results underscore this method's capability to objectively highlight
546 regulatory regions with analogous epigenetic patterns across cell types in both species. This
547 method could aid in extracting additional information about similar epigenetic patterns between
548 human and model organisms such as mice, for which only a portion of their genome aligns with
549 human.

550

551 Because some of the NMF factors reflected processes in gene expression and regulation that
552 occur in many genes, some of the highly correlated regions across species could reflect false
553 positives. Thus, it is prudent to restrict the current approach to genomic intervals around
554 orthologous genes to reduce the impact of false discovery. We examined patterns of epigenetic
555 state correlations across cell types between the human *GATA1* gene locus and three non-
556 orthologous loci in mouse to investigate the scope of this issue (Supplemental Material). While
557 genomic bins of high epigenetic state correlation were observed between non-orthologous loci,
558 the discovery of bins implicated in a cell type-specific process, such as erythroid or
559 megakaryocytic regulation, could be enhanced by utilizing a broader background model for
560 computing peaks of NMF signal (Supplemental Fig. S33). With this refined approach to peak
561 identification, the false discovery rate estimated for epigenetic state comparison between the
562 human *GATA1* locus and the mouse *Cd4* locus was reduced to 0.1 or less (Supplemental Fig.
563 S33R). Furthermore, the epigenetic state comparisons between the human *GATA1* locus and
564 the mouse *Rps19* locus revealed a previously unreported region with hallmarks of erythroid
565 regulatory elements (Supplemental Fig. S34). These initial results suggest that the genomic

566 scale of the epigenetic state correlations could be expanded in future work with judicious
567 attention to reducing false discovery, e.g., by linking the discovered elements to evidence of
568 conserved synteny between species.

569
570 Examination of human genomic elements shown to be active in a lentiMPRA assay (Agarwal et
571 al. 2023) at 30 loci (Supplemental Table S6) revealed that the active elements were enriched in
572 genomic bins with high cross cell-type epigenetic state correlation between species
573 (Supplemental Fig. S35). The enrichment for active elements was further increased in bins with
574 both high epigenetic state correlation and interspecies sequence conservation, while enrichment
575 was reduced or comparable (depending on approaches used for false discovery thresholds) in
576 bins with only sequence conservation. These results further support the value of the cross cell-
577 type epigenetic state correlation between species in predicting and interpreting cCREs
578 (Supplemental Fig. S36).

579
580 The comparison of epigenetic state profiles across cell types also provided a means to
581 categorize cCREs between species that did not require a match in the underlying genomic DNA
582 sequence (Supplemental Figs. S37 and S38). Results from that approach indicated that certain
583 cCREs were potentially involved in regulation of orthologous genes, even for cCREs with DNA
584 sequences that did not align between species.

585
586 In summary, the IDEAS joint modeling on the input data compiled here and the consistent state
587 assignments in both mouse and human confirmed and extended previous observations on
588 known regulatory elements, and they revealed both shared and distinctive candidate regulatory
589 elements and states between species. Correlations of state profiles between species provided a
590 comparison of chromatin landscapes even in regions with DNA sequences that were not
591 conserved between species. Our initial results reported here support continuing the

592 development of this approach of comparing cross cell-type epigenetic state profiles between
593 species for functional prediction and interpretation of cCREs.

594

595 **Discussion**

596

597 In this paper, the VISION consortium introduces a set of resources describing the regulatory
598 landscapes of both human and mouse blood cell epigenomes. A key, novel aspect of our work
599 is that the systematic integrative modeling that generated these resources was conducted jointly
600 across the data from both species, which enabled robust comparisons between species without
601 being limited by sequence alignments, allowing comparisons in non-conserved and lineage-
602 specific genomic regions.

603

604 One major resource is the annotation of the epigenetic states across the genomes of progenitor
605 and mature blood cells of both species. These state maps show the epigenetic landscape in a
606 compact form, capturing information from the input data on multiple histone modifications, CTCF
607 occupancy, and chromatin accessibility, and they use a common set of epigenetic states to
608 reveal the patterns of epigenetic activity associated with gene expression and regulation both
609 across cell types and between species. A second major resource is a catalog of cCREs
610 actuated in one or more of the blood cell types in each species. The cCREs are predictions of
611 discrete DNA segments likely involved in gene regulation, based on the patterns of chromatin
612 accessibility across cell types, and the epigenetic state annotations suggest the type of activity
613 for each cCRE in each cell type, such as serving as a promoter or enhancer, participating in
614 repression, or inactivity. A third major resource is a quantitative estimate of the regulatory
615 impact of human and mouse cCREs on gene expression in each cell type, i.e., an esRP score,
616 derived from multivariate regression modeling of the epigenetic states in cCREs as predictors of

617 gene expression. The esRP scores are a continuous variable capturing not only the integration
618 of the input epigenetic data, but also the inferred impacts on gene expression. Those impacts
619 may be manifested as activation or repression during regulation or as transcriptional elongation.
620 They are useful for many downstream analyses, such as determining informative groups of
621 cCREs by clustering analysis. These resources along with browsers for visualization and tools
622 for analysis are provided at our project website, <http://usevision.org>. Among these tools is
623 cCRE_db, which records the several dimensions of annotation of the cCREs and provides a
624 query interface to support custom queries from users.

625

626 Our human blood cell cCRE catalog should be valuable for mechanistic interpretations of trait-
627 related human genetic variants. Human genetic variants associated with traits intrinsic to blood
628 cells were significantly enriched in the VISION cCRE catalog, whereas variants associated with
629 a broad diversity of other traits were not enriched. We expect that the extensive annotations in
630 our cCRE catalog combined with information about TFBS motifs and TF occupancy should lead
631 to specific, refined hypotheses for mechanisms by which a variant impacts expression, such as
632 alterations in TF binding, which can be tested experimentally in further work.

633

634 The jointly learned state maps and cCRE predictions allowed us to extend previous work on the
635 evolution of regulatory elements between mouse and human. Several previous studies focused
636 on transcription factor (TF) occupancy, e.g. examining key TFs in one tissue across multiple
637 species (Schmidt et al. 2010; Ballester et al. 2014; Villar et al. 2014) or a diverse set of TFs in
638 multiple cell types and in mouse and human (Cheng et al. 2014; Yue et al. 2014; Denas et al.
639 2015). Other studies focused on discrete regions of high chromatin accessibility in multiple cell
640 types and tissues between mouse and human (Stergachis et al. 2014; Vierstra et al. 2014).
641 These previous studies revealed that only a small fraction of elements was conserved both in
642 genomic sequence and in inferred function. A notable fraction of elements changed

643 considerably during mammalian diversification, including turnover of TF binding site motifs and
644 repurposing of elements (Schmidt et al. 2010; Cheng et al. 2014; Stergachis et al. 2014; Denas
645 et al. 2015). These prior studies focused primarily on regions of the genome with sequences
646 that aligned between human and mouse, with the non-aligning regions used to infer that some
647 elements were lineage-specific and that many were derived from transposable elements and
648 endogenous retroviruses (Bourque 2009; Rebollo et al. 2012; Jacques et al. 2013; Sundaram et
649 al. 2014). Our evolutionary analyses confirmed the previous observations, e.g., finding about
650 one-third of cCREs are conserved in both sequence and inferred function between human and
651 mouse, and further showing that this evolutionary category was highly enriched for proximal
652 regulatory elements.

653
654 Going beyond the prior comparative epigenetic studies, our jointly learned epigenetic state
655 maps generated a representation of multiple epigenetic features, not just TF occupancy or
656 chromatin accessibility, and they are continuous in bins across genomes of both species. Using
657 the same set of epigenetic states for annotation of both the human and mouse genomes gave a
658 common “alphabet” (set of states) for both species, which enabled comparisons of the
659 epigenetic profiles between species. In the current work, we explored the utility of these
660 epigenetic comparisons in several ways. For example, the joint clusterings of cCREs between
661 species by esRP scores (derived from the epigenetic state annotations) enabled an analysis
662 that was agnostic to DNA sequence or occupancy by TFs to show considerable sharing of
663 inferred TF activity in both human and mouse. Furthermore, the common alphabet of states
664 allowed us to compare the cross-cell type epigenetic state patterns in large genomic intervals of
665 both species containing orthologous genes, again in a manner agnostic to underlying DNA
666 sequence similarities or differences. These epigenetic comparisons were a strong complement
667 to genomic sequence alignments, revealing regulatory elements with similar epigenetic profiles
668 even in genomic regions in which the DNA sequence does not align between species. Our

669 detection, even in segments of DNA that do not align between species, of epigenetic similarity
670 indicative of a common role in gene regulation suggests that processes or structures, such as
671 chromatin interactions, chromatin complexes, or molecular condensates, may be maintained
672 between species in a manner that is not fully revealed by comparisons of genome sequences.
673 Hence, further studies of this apparent epigenetic dimension of regulatory conservation may be
674 productive. For example, the complex interspecies epigenetic state correlation matrices were
675 decomposed into NMF factors that represented major types of regulatory mechanisms, some
676 that were common across cell types and others that were specific to certain cell types. Further
677 investigation indicated the potential for judicious use of the cell type specific NMF factors in a
678 context of conserved synteny for expanding the scale of the state correlation analysis in future
679 studies.

680
681 Previous work compared epigenetic profiles across species, such as the phylo-HMGP method
682 to find different evolutionary states in multi-species epigenomic data (Yang et al. 2018) and the
683 LECIF scores to find evidence of conservation from functional genomic data (Kwon and Ernst
684 2021). These approaches are powerful but limited to the genomic regions with DNA sequences
685 that align between the species, and thus they will miss the approximately 40% of experimentally
686 demonstrated CREs that are not in aligning regions (Fig. 5D). In contrast, our approach of
687 correlating epigenetic states included both DNA segments that align between human and
688 mouse and those that do not, and it captures more of the experimentally verified cCREs. For
689 comparisons between species, both genomic sequence alignment and epigenetic state
690 annotation across cell types provide important sources of information. Combining both types of
691 data into joint models for predicting CREs could be a productive avenue for future work, not only
692 for improved accuracy but also to allow the contributions of each type of information to
693 determined systematically.

694

695 Several innovations were developed to produce the resources introduced here. A major
696 innovation was to extend the IDEAS framework (Zhang et al. 2016) to jointly learn epigenetic
697 states and assign them to annotate the epigenomes in human and mouse blood cells. The
698 IDEAS method employs a Bayesian approach to the modeling to learn the states, which we
699 utilized to bring in states learned from the data in one species as priors for learning states in the
700 data from the second species. Another extension of the IDEAS framework was to learn states
701 based on one feature, specifically ATAC-seq data, defining discrete signal intensity states. This
702 approach was used for calling cCREs, implemented as the IDEAS-IS method (Xiang et al.
703 2021). The approach is relatively simple and benefits from joint modeling across the input
704 datasets. Other methods for predicting cCREs based on chromatin accessibility across many
705 cell types prevented excessive expansion of the summary calls for overlapping peaks by
706 employing a centroid determination for the DNase hypersensitive sites (DHS) index (Meuleman
707 et al. 2020) or by choosing the highest signal peak for the ENCODE cCRE catalog (The
708 ENCODE Project Consortium et al. 2020). The ENCODE cCRE catalog paired DHS peaks with
709 individual chromatin modifications or CTCF occupancy, which led to complications when data
710 on diagnostic features were missing from some cell types. The IDEAS framework used for the
711 VISION cCRE sets leveraged data in related cell types to ameliorate the impact of missing data.
712

713 While the resources introduced here are valuable for many applications, it is prudent to
714 acknowledge their limitations. First, the quality of the products of integrated analyses are limited
715 by the quality and completeness of the input, raw data. We endeavored to reduce the impact of
716 variances in the input data by normalization. The S3V2 procedure (Xiang et al. 2021)
717 systematically normalized the input data to adjust for differences in signal-to-noise and variance
718 in signal across the datasets. Some epigenetic features were not determined in some cell types,
719 and we used the IDEAS method in part because it is able to assign an epigenetic state even in
720 the context of missing data by learning patterns from local similarities in cell types for which the

721 data are present (Zhang and Mahony 2019). However, these approaches cannot completely
722 overcome all issues with variance in input data, and further developments in these directions
723 (such as Shahraki et al. 2023; Xiang et al. 2024) may help to improve integrative resources.
724 Second, the resolution of both the epigenetic state assignments and the cCRE inference is
725 limited to 200 bp, which is the window size we utilized in the IDEAS analyses. Other resources,
726 such as DHS calls (Meuleman et al. 2020), DNase footprints (Vierstra et al. 2020), and motif
727 instances (Weirauch et al. 2014), achieve a higher resolution. Indeed, one can use these higher
728 resolution datasets to derive further information about cCREs, such as families of TFs that are
729 likely to be binding to them. Regarding esRP scores, a third limitation is that we do not make
730 explicit assignments for target genes of cCREs. Predictions of a large number of target gene-
731 cCRE pairs were made in our prior work (Xiang et al. 2020); these assignments cover large
732 genomic intervals around each gene and are most useful when used with further filtering, such
733 as restricting cCREs and target genes to the same topologically associated domains. On-going
734 work is examining other models and approaches for assigning likely target genes to cCREs. A
735 fourth limitation is that our inference of repression-related cCREs applies only to those with
736 stable histone modifications. Elements that had been involved in initiation of repression but
737 eventually were packaged into quiescent chromatin, e.g., via a hit-and-run mechanism (Shah et
738 al. 2019), would not be detected. A fifth limitation concerns the scale of the studies of epigenetic
739 conservation by correlations of epigenetic states. Our current approach is limited to individual
740 examination of specific genetic loci since we used orthologous genes as the initial anchors.
741 Exploring ways to expand the scale of the analytical approach is a goal of future research.
742 Finally, the work presented here was restricted to blood cell types. In future work, extension of
743 the approaches developed in this study to a broader spectrum of cell types would expand the
744 utility of the resulting resources.

745

746 In conclusion, we present several important new resources to enable further and more detailed
747 studies of gene regulation in human and mouse blood cells both during normal differentiation
748 and in pathological contexts. The patterns of epigenetic states in cCREs across cell types show
749 value in developing an understanding of how genetic variants impact blood cell traits and
750 diseases. Furthermore, the joint modeling between species opens avenues for further
751 exploration of comparisons of epigenetic landscapes in addition to sequence alignments for
752 insights into evolution and function of regulatory elements between species.

753

754 **Methods**

755

756 **Data generation, collation, normalization, and integration**

757 The data sets used as input, including the ones generated for the work reported here (with
758 methods), are described in Supplemental Material section “Data generation and collection” and
759 Supplemental Tables S1 and S2. The S3V2 approach (Xiang et al. 2021) was used for
760 normalization and denoising the data sets prior to integration. The data sets were integrated to
761 find and assign epigenetic states using IDEAS (Zhang et al. 2016; Zhang and Hardison 2017);
762 the extension of this approach to joint learning and annotation between species is described in
763 Supplemental Material sections “Data normalization” and “Joint systematic integration of human
764 and mouse blood cell epigenomes by IDEAS”.

765

766 **Prediction, annotation, and estimation of regulatory impact of cCREs**

767 The identification of cCREs as peaks of chromatin accessibility employed IDEAS in the signal
768 intensity state (IS) mode (Xiang et al. 2021). This approach and comparisons with MACS peaks
769 (Zhang et al. 2008) are described in Supplemental Material section “Prediction of VISION
770 cCREs using IDEAS-IS”. The cCREs are provided in Supplemental Table S3. Annotation of

771 potential cCRE functions used intersections with orthogonal data sets of elements implicated in
772 regulation or chromatin structure (Supplemental Table S5). Enrichment of genetic variants
773 associated with blood cell traits used stratified linkage disequilibrium score regression (sLDSC,
774 Finucane et al. 2015). The impact of epigenetic states in cCREs on regulation of gene
775 expression used a multivariate linear regression approach like one described previously (Xiang
776 et al. 2020). Methods and supplementary results on these analyses are presented in detail in
777 the Supplemental Material.

778

779 **Identification of clusters of cCREs based on epigenetic regulatory potential scores**

780 The sets of human and mouse cCREs were placed jointly into groups based on their epigenetic
781 regulatory potential (esRP) scores using a series of *k*-means clustering steps, as described in
782 detail in Supplemental Material and Supplementary Fig. S14. Methods and results for
783 enrichment of the resulting joint meta-clusters (JmCs) for orthogonal sets of regulatory elements
784 and SNPs associated with blood cell traits, along with comparisons of clusters based on
785 chromatin accessibility and H3K27ac signal, are described in Supplemental Material and
786 Supplementary Figs. S15 - S18. Motifs that were differentially enriched across JmCs were
787 identified using the Maelstrom tool in the GimmeMotifs suite (v0.17.1) (Bruse and van
788 Heeringen 2018) and SeqUnwinder (Kakumanu et al. 2017), as described in detail in
789 Supplemental Material and Supplementary Fig. S21.

790

791 **Partitioning cCREs to evolutionary categories based on DNA sequence alignments and**
792 **cCRE calls between species**

793 The human and mouse cCREs were assigned to three evolutionary categories using the
794 following procedure. The set of human cCREs was mapped to mouse genome assembly mm10
795 using the liftOver tool at the UCSC Genome Browser (Hinrichs et al. 2006). Human cCREs that
796 failed to map to mm10 were grouped as N cCREs. Matches to mouse cCREs for the human

797 cCREs that could be mapped by liftOver to mm10 were determined using the intersect tool in
798 BEDTools (Quinlan and Hall 2010). Human cCREs that overlapped with mouse cCREs were
799 labeled as SF cCREs, while human cCREs that mapped to mm10 but did not match mouse
800 cCREs were labeled as S cCREs. A similar process was performed on the set of mouse cCREs
801 using liftOver to map to human genome build GRCh38

802

803 **Calculation of pairwise correlation coefficients for epigenetic landscapes between
804 human and mouse**

805 A bin-to-bin pairwise correlation analysis was used to quantify the similarity of epigenetic
806 landscapes between two DNA regions in human and mouse. For each 200bp bin in one cell
807 type in one species, the assigned epigenetic state was replaced by a vector of mean signals of
808 8 epigenetic features in the IDEAS state model. After replacing the states in all 15 matched cell
809 types (14 analogous cell types and one pseudo-cell type with average values for all cell types)
810 in the two species, the original two categorical state vectors with 15 elements were converted
811 into two numeric vectors with 120 numbers (Supplemental Fig. S28). The similarity of cross-cell
812 type epigenetic landscape between two bins in the two species was defined as the correlation
813 coefficient between each pair of numeric vectors with 120 numbers. When calculating the
814 correlation coefficients, we added random noise (mean=0, sd=0.2) to the raw values to avoid
815 high correlation coefficients created between regions with states that have low signals. The
816 complex correlation matrix was decomposed into distinctive factors using Nonnegative Matrix
817 Factorization (Lee and Seung 1999). Methods and supplementary results on these analyses are
818 presented in detail in the Supplemental Material.

819

820 **Data access**

821 All raw and processed sequencing data generated in this study have been submitted to the
822 NCBI Gene Expression Omnibus (GEO; <https://www.ncbi.nlm.nih.gov/geo/>) under accession
823 number GSE229101 and the NCBI BioProject database
824 (<https://www.ncbi.nlm.nih.gov/bioproject/>) under accession number PRJNA952902. Resources
825 developed in the VISION project are available at the website <https://usevision.org>; the data can
826 be viewed via a track hub at the UCSC Genome Browser or any compatible browser by using
827 this URL: <https://usevision.org/data/trackHub/hub.txt> or by clicking the track hubs link at
828 usevision.org. The database cCRE db supports flexible user queries on extensive annotation of
829 the cCREs, including epigenetic states and esRP scores across cell types, chromatin
830 accessibility scores across cell types, membership in JmCs, and evolutionary categories. Code
831 developed for this study is in the Supplemental Material and at these GitHub repositories:
832 https://github.com/guanjue/Joint_Human_Mouse_IDEAS_State for the joint human-mouse
833 IDEAS pipeline and https://github.com/usevision/cre_heritability for the sLDSC analysis.
834

835 **Competing interest statement**

836 The authors declare no competing interests.
837

838 **Acknowledgments**

839 This work was supported by grants from the National Institutes of Health:
840 R24DK106766 to RCH, GAB, MJW, YZ, FY, JT, MS, DB, DH, JRH, BG; R01DK054937 to GAB;
841 R01GM121613 to YZ and SM; R01GM109453 to QL; R35GM133747 to RCM; F31HG012900 to
842 DJT; R01HG011139; National Science Foundation DBI CAREER 2045500 to SM, and
843 intramural funds from the National Human Genome Research Institute. We dedicate this paper
844 to the memory of JT.

845

846 References

847 Agarwal V, Inoue F, Schubach M, Martin BK, Dash PM, Zhang Z, Sohota A, Noble WS,
848 Yardimci GG, Kircher M et al. 2023. Massively parallel characterization of transcriptional
849 regulatory elements in three diverse human cell types. *bioRxiv*
850 doi:10.1101/2023.03.05.531189.

851 Ballester B, Medina-Rivera A, Schmidt D, Gonzalez-Porta M, Carlucci M, Chen X, Chessman K,
852 Faure AJ, Funnell AP, Goncalves A et al. 2014. Multi-species, multi-transcription factor
853 binding highlights conserved control of tissue-specific biological pathways. *eLife* **3**:
854 e02626.

855 Bauer DE, Kamran SC, Lessard S, Xu J, Fujiwara Y, Lin C, Shao Z, Canver MC, Smith EC,
856 Pinello L et al. 2013. An erythroid enhancer of BCL11A subject to genetic variation
857 determines fetal hemoglobin level. *Science* **342**: 253-257.

858 Blobel GA, Weiss MJ. 2009. Nuclear Factors that Regulate Erythropoiesis. In *Disorders of*
859 *Hemoglobin: Genetics, Pathophysiology, and Clinical Management*, (ed. MH Steinberg,
860 et al.), pp. 62-85. Cambridge University Press, Cambridge.

861 Bourque G. 2009. Transposable elements in gene regulation and in the evolution of vertebrate
862 genomes. *Curr Opin Genet Dev* **19**: 607-612.

863 Bruse N, van Heeringen SJ. 2018. GimmeMotifs: an analysis framework for transcription factor
864 motif analysis. *bioRxiv* doi:<https://doi.org/10.1101/474403>.

865 Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A,
866 Morales J, Mountjoy E, Sollis E et al. 2019. The NHGRI-EBI GWAS Catalog of published
867 genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic
868 Acids Res* **47**: D1005-D1012.

869 Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of
870 morphological evolution. *Cell* **134**: 25-36.

871 Cheng L, Li Y, Qi Q, Xu P, Feng R, Palmer L, Chen J, Wu R, Yee T, Zhang J et al. 2021.

872 Single-nucleotide-level mapping of DNA regulatory elements that control fetal

873 hemoglobin expression. *Nat Genet* **53**: 869-880.

874 Cheng Y, Ma Z, Kim BH, Wu W, Cayting P, Boyle AP, Sundaram V, Xing X, Dogan N, Li J et al.

875 2014. Principles of regulatory information conservation between mouse and human.

876 *Nature* **515**: 371-375.

877 Chi AW, Bell JJ, Zlotoff DA, Bhandoola A. 2009. Untangling the T branch of the hematopoiesis

878 tree. *Curr Opin Immunol* **21**: 121-126.

879 Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, Snyder MP, Pritchard

880 JK, Kundaje A, Greenleaf WJ et al. 2016. Lineage-specific and single-cell chromatin

881 accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet* **48**: 1193-

882 1203.

883 Denas O, Sandstrom R, Cheng Y, Beal K, Herrero J, Hardison RC, Taylor J. 2015. Genome-

884 wide comparative analysis reveals human-mouse regulatory landscape and evolution.

885 *BMC Genomics* **16**: 87.

886 Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, Gingeras TR, Gerstein M,

887 Guigo R, Birney E et al. 2012. Modeling gene expression using chromatin features in

888 various cellular contexts. *Genome Biol* **13**: R53.

889 Dore LC, Crispino JD. 2011. Transcription factor networks in erythroid cell and megakaryocyte

890 development. *Blood* **118**: 231-239.

891 Dynan WS, Tjian R. 1983. The promoter-specific transcription factor Sp1 binds to upstream

892 sequences in the SV40 early promoter. *Cell* **35**: 79-87.

893 Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for

894 systematic annotation of the human genome. *Nat Biotechnol* **28**: 817-825.

895 Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and
896 characterization. *Nat Methods* **9**: 215-216.

897 Ferreira R, Ohneda K, Yamamoto M, Philipsen S. 2005. GATA1 function, a paradigm for
898 transcription factors in hematopoiesis. *Mol Cell Biol* **25**: 1215-1227.

899 Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, Anttila V, Xu H, Zang C,
900 Farh K et al. 2015. Partitioning heritability by functional annotation using genome-wide
901 association summary statistics. *Nat Genet* **47**: 1228-1235.

902 Frangoul H, Altshuler D, Cappellini MD, Chen YS, Domm J, Eustace BK, Foell J, de la Fuente J,
903 Grupp S, Handgretinger R et al. 2021. CRISPR-Cas9 Gene Editing for Sickle Cell
904 Disease and beta-Thalassemia. *The New England journal of medicine* **384**: 252-260.

905 Fujiwara T, O'Geen H, Keles S, Blahnik K, Linnemann AK, Kang YA, Choi K, Farnham PJ,
906 Bresnick EH. 2009. Discovering hematopoietic mechanisms through genome-wide
907 analysis of GATA factor chromatin occupancy. *Mol Cell* **36**: 667-681.

908 Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, Jackson D, Leith A,
909 Schreiber J, Noble WS et al. 2019. A Genome-wide Framework for Mapping Gene
910 Regulation via Cellular Genetic Screens. *Cell* **176**: 377-390 e319.

911 Ge T, Chen CY, Neale BM, Sabuncu MR, Smoller JW. 2017. Phenome-wide heritability analysis
912 of the UK Biobank. *PLoS Genet* **13**: e1006711.

913 Graf T, Enver T. 2009. Forcing cells to change lineages. *Nature* **462**: 587-594.

914 Gumucio DL, Heilstedt-Williamson H, Gray TA, Tarle SA, Shelton DA, Tagle D, Slightom J,
915 Goodman M, Collins FS. 1992. Phylogenetic footprinting reveals a nuclear protein which
916 binds to silencer sequences in the human g and e globin genes. *Mol Cell Biol* **12**: 4919-
917 4929.

918 Hamamoto K, Fukaya T. 2022. Molecular architecture of enhancer-promoter interaction. *Curr
919 Opin Cell Biol* **74**: 62-70.

920 Hardison RC. 2000. Conserved noncoding sequences are reliable guides to regulatory
921 elements. *Trends in Genetics* **16**: 369-372.

922 Hardison RC. 2012. Genome-wide epigenetic data facilitate understanding of disease
923 susceptibility association studies. *J Biol Chem* **287**: 30932-30940.

924 Hardison RC, Zhang Y, Keller CA, Xiang G, Heuston EF, An L, Lichtenberg J, Giardine BM,
925 Bodine D, Mahony S et al. 2020. Systematic integration of GATA transcription factors
926 and epigenomes via IDEAS paints the regulatory landscape of hematopoietic cells.
927 *IUBMB Life* **72**: 27-38.

928 Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart
929 RK, Ching CW et al. 2009. Histone modifications at human enhancers reflect global cell-
930 type-specific gene expression. *Nature* **459**: 108-112.

931 Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009.
932 Potential etiologic and functional implications of genome-wide association loci for human
933 diseases and traits. *Proc Natl Acad Sci U S A* **106**: 9362-9367.

934 Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey
935 TS, Harte RA, Hsu F et al. 2006. The UCSC Genome Browser Database: update 2006.
936 *Nucleic Acids Res* **34**: D590-598.

937 Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen
938 PM, Bilmes JA, Birney E et al. 2013. Integrative annotation of chromatin elements from
939 ENCODE data. *Nucleic Acids Res* **41**: 827-841.

940 Jacques PE, Jeyakani J, Bourque G. 2013. The majority of primate-specific regulatory
941 sequences are derived from transposable elements. *PLoS Genet* **9**: e1003504.

942 Jansen C, Ramirez RN, El-Ali NC, Gomez-Cabrero D, Tegner J, Merkenschlager M, Conesa A,
943 Mortazavi A. 2019. Building gene regulatory networks from scATAC-seq and scRNA-seq
944 using Linked Self Organizing Maps. *PLoS Comput Biol* **15**: e1006555.

945 Jian J, Konopka J, Liu C. 2013. Insights into the role of progranulin in immunity, infection, and
946 inflammation. *J Leukoc Biol* **93**: 199-208.

947 Kaczynski J, Cook T, Urrutia R. 2003. Sp1- and Kruppel-like transcription factors. *Genome Biol*
948 **4**: 206.

949 Kakumanu A, Velasco S, Mazzoni E, Mahony S. 2017. Deconvolving sequence features that
950 discriminate between overlapping regulatory annotations. *PLoS Comput Biol* **13**:
951 e1005795.

952 Karlić R, Chung HR, Lasserre J, Vlahovicek K, Vingron M. 2010. Histone modification levels are
953 predictive for gene expression. *Proc Natl Acad Sci U S A* **107**: 2926-2931.

954 King DC, Taylor J, Zhang Y, Cheng Y, Lawson HA, Martin J, ENCODE groups for
955 Transcriptional Regulation and Multispecies Sequence Analysis, Chiaromonte F, Miller
956 W, Hardison RC. 2007. Finding cis-regulatory elements using comparative genomics:
957 some lessons from ENCODE data. *Genome Res* **17**: 775-786.

958 Kondo M, Wagers AJ, Manz MG, Prohaska SS, Scherer DC, Beilhack GF, Shizuru JA,
959 Weissman IL. 2003. Biology of hematopoietic stem cells and progenitors: implications for
960 clinical application. *Annu Rev Immunol* **21**: 759-806.

961 Kwon SB, Ernst J. 2021. Learning a genome-wide score of human-mouse conservation at the
962 functional genomics level. *Nature communications* **12**: 2495.

963 Laurenti E, Göttgens B. 2018. From haematopoietic stem cells to complex differentiation
964 landscapes. *Nature* **553**: 418-426.

965 Lee DD, Seung HS. 1999. Learning the parts of objects by non-negative matrix factorization.
966 *Nature* **401**: 788-791.

967 Lee DI, Roy S. 2021. GRiNCH: simultaneous smoothing and detection of topological units of
968 genome organization from sparse chromatin contact count matrices with matrix
969 factorization. *Genome Biol* **22**: 164.

970 Libbrecht MW, Chan RCW, Hoffman MM. 2021. Segmentation and genome annotation
971 algorithms for identifying chromatin state and other genomic patterns. *PLoS Comput Biol*
972 17: e1009423.

973 Martens JH, Stunnenberg HG. 2013. BLUEPRINT: mapping human blood cell epigenomes.
974 *Haematologica* 98: 1487-1489.

975 Maston GA, Evans SK, Green MR. 2006. Transcriptional Regulatory Elements in the Human
976 Genome. *Annu Rev Genomics Hum Genet* 7: 29-59.

977 Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom
978 R, Qu H, Brody J et al. 2012. Systematic localization of common disease-associated
979 variation in regulatory DNA. *Science* 337: 1190-1195.

980 Meuleman W, Muratov A, Rynes E, Halow J, Lee K, Bates D, Diegel M, Dunn D, Neri F,
981 Teodosiadis A et al. 2020. Index and biological spectrum of human DNase I
982 hypersensitive sites. *Nature* 584: 244-251.

983 Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, Wolfe SA. 2008.
984 Analysis of homeodomain specificities allows the family-wide prediction of preferred
985 recognition sites. *Cell* 133: 1277-1289.

986 Payne KJ, Crooks GM. 2002. Human hematopoietic lineage commitment. *Immunol Rev* 187:
987 48-64.

988 Pennacchio LA, Rubin EM. 2001. Genomic strategies to identify mammalian regulatory
989 sequences. *Nat Rev Genet* 2: 100-109.

990 Pimkin M, Kossenkov AV, Mishra T, Morrissey CS, Wu W, Keller CA, Blobel GA, Lee D, Beer
991 MA, Hardison RC et al. 2014. Divergent functions of hematopoietic transcription factors
992 in lineage priming and differentiation during erythro-megakaryopoiesis. *Genome Res* 24:
993 1932-1944.

994 Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution
995 rates on mammalian phylogenies. *Genome Res* 20: 110-121.

996 Qi Q, Cheng L, Tang X, He Y, Li Y, Yee T, Shrestha D, Feng R, Xu P, Zhou X et al. 2021.

997 Dynamic CTCF binding directly mediates interactions among cis-regulatory elements

998 essential for hematopoiesis. *Blood* **137**: 1327-1339.

999 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic

1000 features. *Bioinformatics* **26**: 841-842.

1001 Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and natural

1002 source of regulatory sequences for host genes. *Annu Rev Genet* **46**: 21-42.

1003 Ringrose L, Paro R. 2004. Epigenetic regulation of cellular memory by the Polycomb and

1004 Trithorax group proteins. *Annu Rev Genet* **38**: 413-443.

1005 Rothenberg EV, Taghon T. 2005. Molecular genetics of T cell development. *Annu Rev Immunol*

1006 **23**: 601-649.

1007 Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S,

1008 Martinez-Jimenez CP, Mackay S et al. 2010. Five-vertebrate ChIP-seq reveals the

1009 evolutionary dynamics of transcription factor binding. *Science* **328**: 1036-1040.

1010 Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W.

1011 2000. PipMaker-A web server for aligning two genomic DNA sequences. *Genome Res*

1012 **10**: 577-586.

1013 Shah M, Funnell APW, Quinlan KGR, Crossley M. 2019. Hit and Run Transcriptional

1014 Repressors Are Difficult to Catch in the Act. *Bioessays* **41**: e1900041.

1015 Shahraki MF, Farahbod M, Libbrecht MW. 2023. Robust chromatin state annotation. *bioRxiv*

1016 doi:<https://doi.org/10.1101/2023.07.15.549175>.

1017 Spangrude GJ, Heimfeld S, Weissman IL. 1988. Purification and characterization of mouse

1018 hematopoietic stem cells. *Science* **241**: 58-62.

1019 Stein-O'Brien GL, Arora R, Culhane AC, Favorov AV, Garmire LX, Greene CS, Goff LA, Li Y,

1020 Ngom A, Ochs MF et al. 2018. Enter the Matrix: Factorization Uncovers Knowledge from

1021 Omics. *Trends Genet* **34**: 790-805.

1022 Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, Byron R, Canfield T,
1023 Stelhing-Sun S, Lee K et al. 2014. Conservation of trans-acting circuitry during
1024 mammalian regulatory evolution. *Nature* **515**: 365-370.

1025 Strahl BD, Allis CD. 2000. The language of covalent histone modifications. *Nature* **403**: 41-45.

1026 Stunnenberg HG, International Human Epigenome C, Hirst M. 2016. The International Human
1027 Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167**:
1028 1145-1149.

1029 Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. 2014. Widespread
1030 contribution of transposable elements to the innovation of gene regulatory networks.
1031 *Genome Res* **24**: 1963-1976.

1032 Tenen DG, Hromas R, Licht JD, Zhang DE. 1997. Transcription factors, normal myeloid
1033 development, and leukemia. *Blood* **90**: 489-519.

1034 The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the
1035 human genome. *Nature* **489**: 57-74.

1036 The ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N,
1037 Adrian J, Kawli T, Davis CA, Dobin A et al. 2020. Expanded encyclopaedias of DNA
1038 elements in the human and mouse genomes. *Nature* **583**: 699-710.

1039 Valverde-Garduno V, Guyot B, Anguita E, Hamlett I, Porcher C, Vyas P. 2004. Differences in
1040 the chromatin structure and cis-element organization of the human and mouse GATA1
1041 loci: implications for cis-element identification. *Blood* **104**: 3106-3116.

1042 van Arensbergen J, FitzPatrick VD, de Haas M, Pagie L, Sluimer J, Bussemaker HJ, van
1043 Steensel B. 2017. Genome-wide mapping of autonomous promoter activity in human
1044 cells. *Nat Biotechnol* **35**: 145-153.

1045 van Pampus EC, Denkers IA, van Geel BJ, Huijgens PC, Zevenbergen A, Ossenkoppele GJ,
1046 Langenhuijsen MM. 1992. Expression of adhesion antigens of human bone marrow

1047 megakaryocytes, circulating megakaryocytes and blood platelets. *Eur J Haematol* **49**:
1048 122-127.

1049 Vierstra J, Lazar J, Sandstrom R, Halow J, Lee K, Bates D, Diegel M, Dunn D, Neri F, Haugen
1050 E et al. 2020. Global reference mapping of human transcription factor footprints. *Nature*
1051 **583**: 729-736.

1052 Vierstra J, Rynes E, Sandstrom R, Zhang M, Canfield T, Hansen RS, Stehling-Sun S, Sabo PJ,
1053 Byron R, Humbert R et al. 2014. Mouse regulatory DNA landscapes reveal global
1054 principles of cis-regulatory evolution. *Science* **346**: 1007-1012.

1055 Villar D, Flieck P, Odom DT. 2014. Evolution of transcription factor binding in metazoans -
1056 mechanisms and functional implications. *Nat Rev Genet* **15**: 221-233.

1057 Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS,
1058 Lambert SA, Mann I, Cook K et al. 2014. Determination and inference of eukaryotic
1059 transcription factor sequence specificity. *Cell* **158**: 1431-1443.

1060 Weiss MJ, Yu C, Orkin SH. 1997. Erythroid-cell-specific properties of transcription factor GATA-
1061 1 revealed by phenotypic rescue of a gene-targeted cell line. *Mol Cell Biol* **17**: 1642-
1062 1651.

1063 Xiang G, Giardine BM, Mahony S, Zhang Y, Hardison RC. 2021. S3V2-IDEAS: a package for
1064 normalizing, denoising and integrating epigenomic datasets across different cell types.
1065 *Bioinformatics* **37**: 3011-3013.

1066 Xiang G, Guo Y, Bumcrot D, Sigova A. 2024. JMnorm: a novel joint multi-feature normalization
1067 method for integrative and comparative epigenomics. *Nucleic Acids Res* **52**: e11.

1068 Xiang G, Keller CA, Heuston E, Giardine BM, An L, Wixom AQ, Miller A, Cockburn A, Sauria
1069 MEG, Weaver K et al. 2020. An integrative view of the regulatory and transcriptional
1070 landscapes in mouse hematopoiesis. *Genome Res* **30**: 472-484.

1071 Xu J, Shao Z, Glass K, Bauer DE, Pinello L, Van Handel B, Hou S, Stamatoyannopoulos JA,
1072 Mikkola HK, Yuan GC et al. 2012. Combinatorial assembly of developmental stage-

1073 specific enhancers controls gene expression programs during human erythropoiesis.

1074 *Dev Cell* **23**: 796-811.

1075 Yang Y, Gu Q, Zhang Y, Sasaki T, Crivello J, O'Neill RJ, Gilbert DM, Ma J. 2018. Continuous-

1076 Trait Probabilistic Model for Comparing Multi-species Functional Genomic Data. *Cell Syst* **7**: 208-218 e211.

1077

1078 Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD et al.

1079 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**:

1080 355-364.

1081 Zhang Y, An L, Yue F, Hardison RC. 2016. Jointly characterizing epigenetic dynamics across

1082 multiple human cell types. *Nucleic Acids Res* **44**: 6721-6731.

1083 Zhang Y, Hardison RC. 2017. Accurate and reproducible functional maps in 127 human cell

1084 types via 2D genome segmentation. *Nucleic Acids Res* **45**: 9823-9836.

1085 Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM,

1086 Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**:

1087 R137.

1088 Zhang Y, Mahony S. 2019. Direct prediction of regulatory elements from partial data without

1089 imputation. *PLoS Comput Biol* **15**: e1007399.

1090

1091 **Figure Legends**

1092 **Figure 1. Cell types and data sets used for systematic integration of epigenetic features**

1093 **of blood cells. (A)** The tree on the left shows the populations of stem, progenitor, and mature

1094 blood cells and cell lines in human. The diagram on the right indicates the epigenetic features

1095 and transcriptomes for which genome-wide data sets were generated or collected, with

1096 distinctive icons for the major sources of data, specifically the Blueprint project (Martens and

1097 Stunnenberg 2013; Stunnenberg et al. 2016), Corces et al. (2016), abbreviated CMB, and St.

1098 Jude Children's Research Hospital (SJCRH, Cheng et al. 2021; Qi et al. 2021). (B) Cell types

1099 and epigenetic data sets in mouse, diagrammed as for panel A. Sources were described in
1100 Xiang et al. (2020) and Supplemental Table S1. Abbreviations for blood cells and lines are: HSC
1101 = hematopoietic stem cell, MPP = multipotent progenitor cell, LMPP = lymphoid-myeloid primed
1102 progenitor cell, CMP = common myeloid progenitor cell, MEP = megakaryocyte-erythrocyte
1103 progenitor cell, K562 = a human cancer cell line with some features of early megakaryocytic and
1104 erythroid cells, HUDEP = immortalized human umbilical cord blood-derived erythroid progenitor
1105 cell lines expressing fetal globin genes (HUDEP1) or adult globin genes (HUDEP2), CD34_E =
1106 human erythroid cells generated by differentiation from CD34+ blood cells, ERY = erythroblast,
1107 RBC = mature red blood cell, MK = megakaryocyte, GMP = granulocyte monocyte progenitor
1108 cell, EOS = eosinophil, MON = monocyte, MONp = primary monocyte, MONc = classical
1109 monocyte, NEU = neutrophil, CLP = common lymphoid progenitor cell, B = B cell, NK = natural
1110 killer cell, TCD4 = CD4+ T cell, TCD8 = CD8+ T cell, LSK = Lin-Sca1+Kit+ cells from mouse
1111 bone marrow containing hematopoietic stem and progenitor cells, HPC7 = immortalized mouse
1112 cell line capable of differentiation in vitro into more mature myeloid cells, G1E = immortalized
1113 mouse cell line blocked in erythroid maturation by a knockout of the *Gata1* gene and its subline
1114 ER4 that will further differentiate after restoration of *Gata1* function in an estrogen inducible
1115 manner (Weiss et al. 1997), MEL = murine erythroleukemia cell line that can undergo further
1116 maturation upon induction (designated iMEL), CFUE = colony forming unit erythroid, FL =
1117 designates ERY derived from fetal liver, BM = designates ERY derived from adult bone marrow,
1118 CFUMK = colony forming unit megakaryocyte, iMK = immature megakaryocyte, MK_fl =
1119 megakaryocyte derived from fetal liver.

1120
1121 **Figure. 2. Genome segmentation and annotation jointly between human and mouse using**
1122 **IDEAS. (A)** Workflow for joint modeling. (1) Initial epigenetic states from 100 randomly selected
1123 regions separately in human and mouse hematopoietic cell types were identified in IDEAS runs.
1124 (2) States that were reproducible and shared in both species were retained. (3a and 3b) The

1125 profile of epigenetic feature contribution to each of the reproducible states was sequentially
1126 refined by applying IDEAS across the full genomes of human and of mouse, updating the state
1127 model after each IDEAS run. (4) Two heterogeneous states were removed to generate the final
1128 joint epigenetic states in the two species. **(B)** The 25 joint epigenetic states for human and
1129 mouse hematopoietic cell types. The average signal of the epigenetic features for each state
1130 are shown in the heatmap. The corresponding state colors, the state labels based on the
1131 function, and the average proportions of the genome covered by each state across cell types
1132 are listed on the right-side of the heatmap. **(C)** Annotation of epigenetic states in a large
1133 genomic interval containing *SLC4A1* and surrounding genes across human blood cell types.
1134 The genomic interval is 210kb, GRCh38 Chr17:44,192,001-44,402,000, with gene annotations
1135 from GENCODE V38. Binding patterns for selected transcription factors are from the VISION
1136 project ChIP-seq tracks (CTCF and GATA1 in adult erythroblasts, signal tracks from MACS,
1137 track heights 100 and 80, respectively) or from the ENCODE data portal (EP300 in K562 cells,
1138 experiment ENCSR000EGE, signal track is fold change over background, track height is 50).
1139 The epigenetic state assigned to each genomic bin in the different cell types is designated by
1140 the color coding shown in panel (B). The replicates in each cell type examined in Blueprint are
1141 labeled by the id for the donor of biosamples. Genes and regulatory regions active primarily in
1142 erythroid (E), granulocytes (G), and megakaryocytes (MK) are marked by gray rectangles. **(D)**
1143 Annotation of epigenetic states in a large genomic interval containing *Slc4a1* and surrounding
1144 genes across mouse blood cell types. The genomic interval is 198kb, mm10
1145 Chr11:102,290,001-102,488,000, with gene annotations from GENCODE VM23. Binding
1146 patterns for selected transcription factors are from the VISION project ChIP-seq tracks (CTCF in
1147 adult erythroblasts, GATA1 and EP300 from the highly erythroid fetal liver, signal tracks from
1148 MACS, track heights 200, 200, and 150, respectively; the EP300 track was made by re-mapping
1149 reads from ENCODE experiment ENCSR982LJQ). The tracks of epigenetic states and
1150 highlighted regions are indicated as in panel (C).

1151

1152 **Figure. 3. Overlaps of VISION cCREs with other catalogs and enrichment for variants**
1153 **associated with blood cell traits. (A)** Venn diagram showing intersections of human VISION
1154 cCREs with a combined superset of elements associated with nuclear structure (CTCF OSs,
1155 loop anchors, and TAD boundaries) and with a combined superset of DNA intervals associated
1156 with *cis*-regulatory elements (CREs), including TSSs, CpG islands, peaks from a massively
1157 parallel promoter and enhancer assay, and enhancers predicted from enhancer RNAs, peaks of
1158 binding by EP300, and histone modifications in erythroblasts (see Supplemental Material,
1159 Supplemental Fig. S9, and Supplemental Table S5). **(B)** The proportions of cCREs and
1160 randomly selected, matched sets of intervals in the overlap categories are compared in the bar
1161 graph. For the random sets, the bar shows the mean, and the dots show the values for each of
1162 ten random sets. **(C)** The UpSet plot provides a higher resolution view of intersections of
1163 VISION cCREs with the four groups of CRE-related elements, specifically enhancer-related
1164 (Enh), transcription start sites (TSS), Survey of Regulatory Elements (SuRE), and CpG islands
1165 (CpG). The enrichment for the cCRE overlaps compared to those in randomly selected,
1166 matched sets of intervals are shown in the boxplots below each overlap subset, with dots for the
1167 enrichment relative to individual random sets. **(D)** Overlaps and enrichments of VISION cCREs
1168 for three sets of structure-related elements, specifically CTCF OSs (CT), loop anchors (LA), and
1169 TAD boundary elements. **(E)** Overlaps of VISION cCREs with two sets of experimentally
1170 determined blood cell cCREs. **(F)** Enrichment of SNPs associated with blood cell traits from UK
1171 Biobank in VISION cCREs. Results of the sLDSC analysis of all cCREs are plotted with
1172 enrichment of the cCRE annotation in heritability of each trait on the x-axis, and the significance
1173 of the enrichment on the y-axis. The analysis covers 292 unique traits with GWAS results from
1174 both males and females and 3 traits with results only from males. The vertical dotted line
1175 indicates an enrichment of 1, and the horizontal dotted line delineates the 5% FDR significance
1176 threshold. Points and labels in red represent traits for which there was significant enrichment of

1177 SNPs associated with the VISION cCREs. Traits with a negative enrichment were assigned an
1178 arbitrary enrichment of 0.1 for plotting and appear as the column of points at the bottom left of
1179 the plot. The shape of the point indicates the sex in which the GWAS analysis was performed
1180 for each trait.

1181

1182 **Figure 4. Beta coefficients of states, esRP scores of cCREs, joint human-mouse**
1183 **metaclusters of cCREs based on esRP scores, and enrichment for TFBS motifs. (A)** Beta
1184 coefficients and the difference of beta coefficients of the 25 epigenetic states. The vertical
1185 columns on the right show the beta coefficients along with the ID, color, and labels for the 25
1186 joint epigenetic states. The triangular heatmap shows the difference of the beta coefficients
1187 between two states in the right columns. Each value in the triangle heatmap shows the
1188 difference in beta coefficients between the state on top and the state below based on the order
1189 of states in the right columns. **(B)** An example of calculating esRP score for a cCRE in a cell
1190 type based on the beta coefficients of states. For a cCRE covering more than one 200bp bin,
1191 the esRP equals the weighted sum of beta coefficients of states that covers the cCRE, where
1192 the weights are the region covered by different states. **(C)** The average esRP score of all
1193 cCREs in JmCs across blood cell types shared by human and mouse. The right column shows
1194 the number of human cCREs in each JmC. **(D)** The average enrichment of JmCs in 15
1195 homologous gene clusters. The genes are clustered based on the JmCs' enrichments by k -
1196 means. **(E)** Motifs enriched in joint metaclusters. The top heatmap shows the enrichment of
1197 motifs in the cCREs in each JmC in human (H) and mouse (M) as a Z-score. The logo for each
1198 motif is given to the right of the heat map, labeled by the family of transcription factors that
1199 recognize that motif. The heatmap below is aligned with the motif enrichment heatmap, showing
1200 the mean esRP score for the cCREs in each JmC for all the common cell types examined
1201 between human and mouse. A summary description of the cell types in which the cCREs in
1202 each JmC are more active is given at the bottom.

1203

1204 **Figure. 5. Evolutionary and epigenetic comparisons of cCREs. (A)** Workflow to partition
1205 blood cell cCREs in human and mouse into three evolutionary categories. N=nonconserved,
1206 S=conserved in sequence but not inferred function, SF=conserved in both sequence and
1207 inferred function as a cCRE, y=yes, n=no. **(B)** Enrichment of SF-conserved human cCREs for
1208 TSSs. The number of elements in seven sets of function-related DNA intervals that overlap with
1209 the 32,422 SF human cCREs was determined, along with the number that overlap with three
1210 subsets (32,422 each) randomly selected from the full set of 200,342 human cCREs. The ratio
1211 of the number of function-related elements overlapping SF-cCREs to the number overlapping a
1212 randomly chosen subset of all cCREs gave the estimate of enrichment plotted in the graph. The
1213 mean for the three determinations of enrichment is indicated by the horizontal line for each set.
1214 Results are also shown for a similar analysis for the S and N cCREs. **(C)** Distribution of phyloP
1215 scores for three evolutionary categories of cCREs in human and mouse. The maximum phyloP
1216 score for each genomic interval was used to represent the score for each cCRE, using genome
1217 sequence alignments of 100 species with human as the reference (phyloP100) and alignments
1218 of 60 species with mouse as the reference (phyloP60). The distribution of phyloP scores for
1219 each group are displayed as a violin plot. All ten random sets had distributions similar to the one
1220 shown. The asterisk (*) over brackets indicates comparison for which the P values for Welch's *t*-
1221 test is less than 2.2×10^{-16} . **(D)** Proportion of human genomic elements active in a massively
1222 parallel reporter assay (MPRA) that align with mouse or are in a state reflecting dynamic
1223 chromatin. A set of 57,061 genomic elements found to be active in a lentivirus MPRA that tested
1224 a close to comprehensive set of predicted regulatory elements in K562 cells (Agarwal et al.
1225 2023) were assessed for their ability to align with the mouse genome (blue bar) or whether the
1226 IDEAS epigenetic state assigned in K562 cells was not quiescent or was in a set of states
1227 associated with gene activation (magenta bars). The results are plotted as percentages of the
1228 total number of MPRA-active elements.

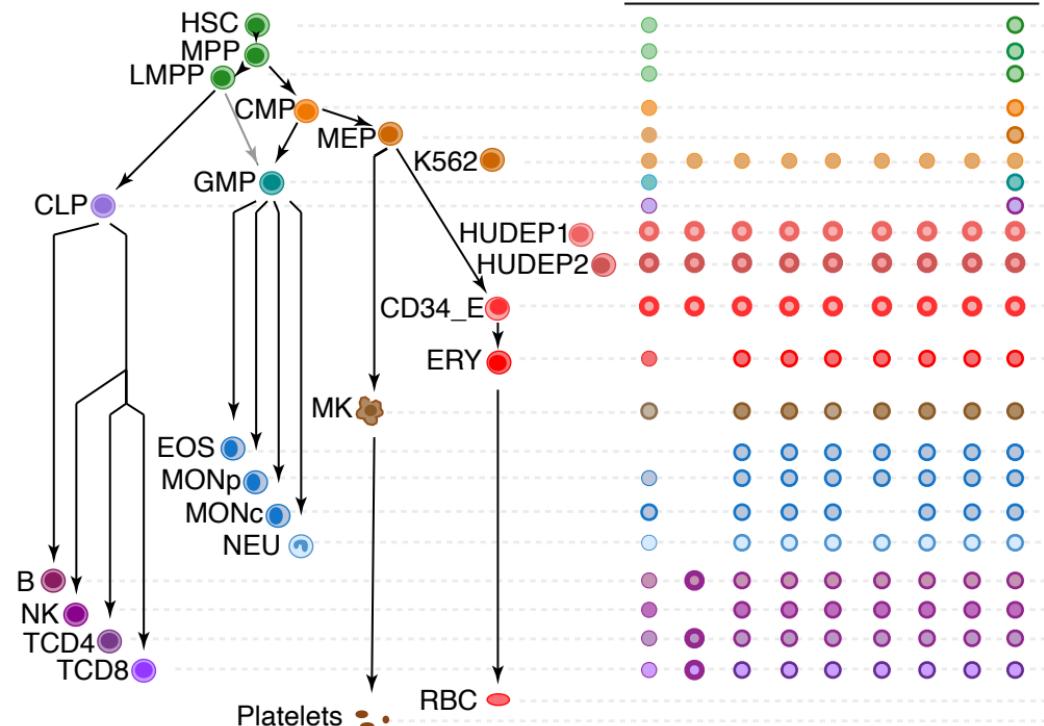
1229

1230 **Figure. 6. Epigenetic comparisons of regulatory landscapes and cCREs. (A and B)** DNA
1231 sequence alignments and correlations of epigenetic states in human *GATA1* and mouse *Gata1*
1232 genes and flanking genes. **(A)** Dot-plot view of chained blastZ alignments by PipMaker
1233 (Schwartz et al. 2000) between genomic intervals encompassing and surrounding the human
1234 *GATA1* (GRCh38 ChrX:48,760,001-48,836,000; 76kb) and mouse *Gata1* (mm10
1235 ChrX:7,919,401-8,020,800; 101.4kb, reverse complement of reference genome) genes. The
1236 axes are annotated with gene locations (GENCODE), predicted *cis*-regulatory elements
1237 (cCREs), and binding patterns for GATA1 and EP300 in erythroid cells. **(B)** Matrix of Pearson's
1238 correlation values between epigenetic states (quantitative contributions of each epigenetic
1239 feature to the assigned state) across 15 cell types analogous for human and mouse. The
1240 correlation is shown for each 200bp bin in one species with all the bins in the other species,
1241 using a red-blue heat map to indicate the value of the correlation. Axes are annotated with
1242 genes and cCREs in each species. **(C)** Decomposition of the correlation matrix (panel **B**) into
1243 six component parts or factors using nonnegative matrix factorization. **(D-G)** Correlation
1244 matrices for genomic intervals encompassing *GATA1/Gata1* and flanking genes, reconstructed
1245 using values from NMF factors. **(D and E)** Correlation matrices using values of NMF factor 3
1246 between human and mouse (panel **D**) or within human and within mouse (panel **E**). The red
1247 rectangles highlight the positive regulatory patterns in the *GATA1/Gata1* genes (labeled Px),
1248 which exhibit conservation of both DNA sequence and epigenetic state pattern. The orange
1249 rectangles denote the distal positive regulatory region present only in mouse (labeled D), which
1250 shows conservation of epigenetic state pattern without corresponding sequence conservation.
1251 Beneath the correlation matrices in panel **E** are maps of IDEAS epigenetic states across 15 cell
1252 types, followed by a graph of the score and peak calls for NMF factor 3 and annotation of
1253 cCREs (thin black rectangles) and genes. **(F and G)** Correlation matrices using values of NMF
1254 factor 6 between human and mouse (panel **F**) or within human and within mouse (panel **G**). The

1255 green rectangles highlight the correlation of epigenetic state patterns within the same gene,
1256 both across the two species and within each species individually, while the black rectangles
1257 highlight the high correlation observed between the two genes *GATA1* and *HDAC6*.

A.

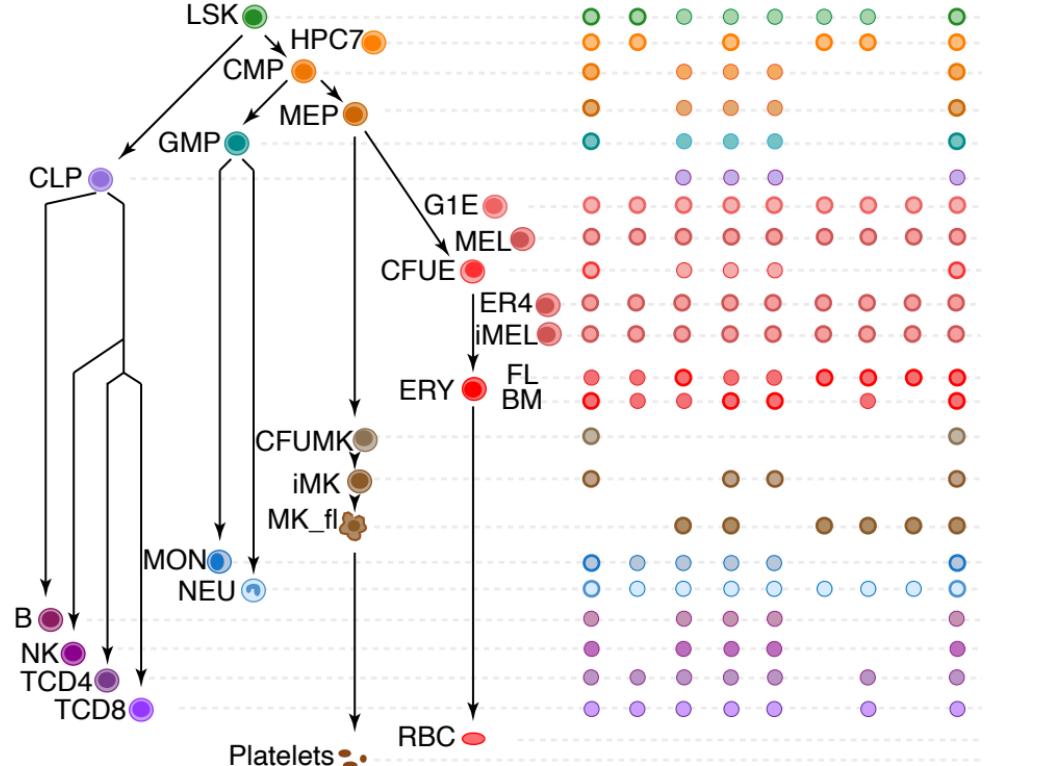
Human



Source: ● Blueprint ● CMB, other ● SJCH

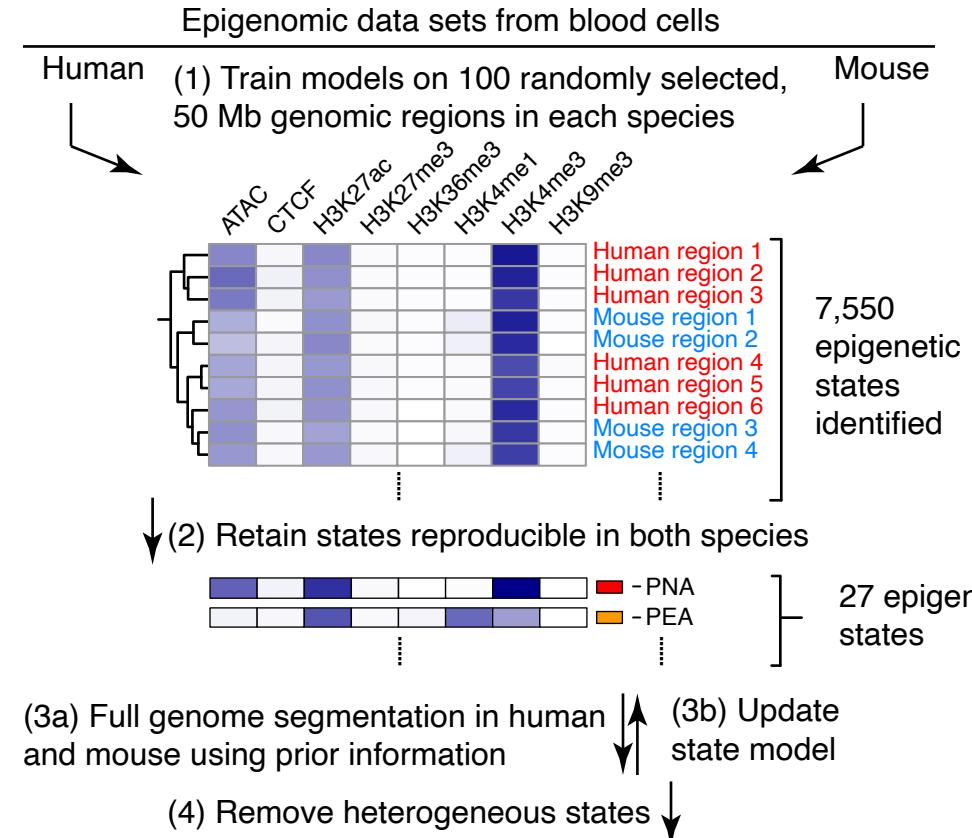
B.

Mouse

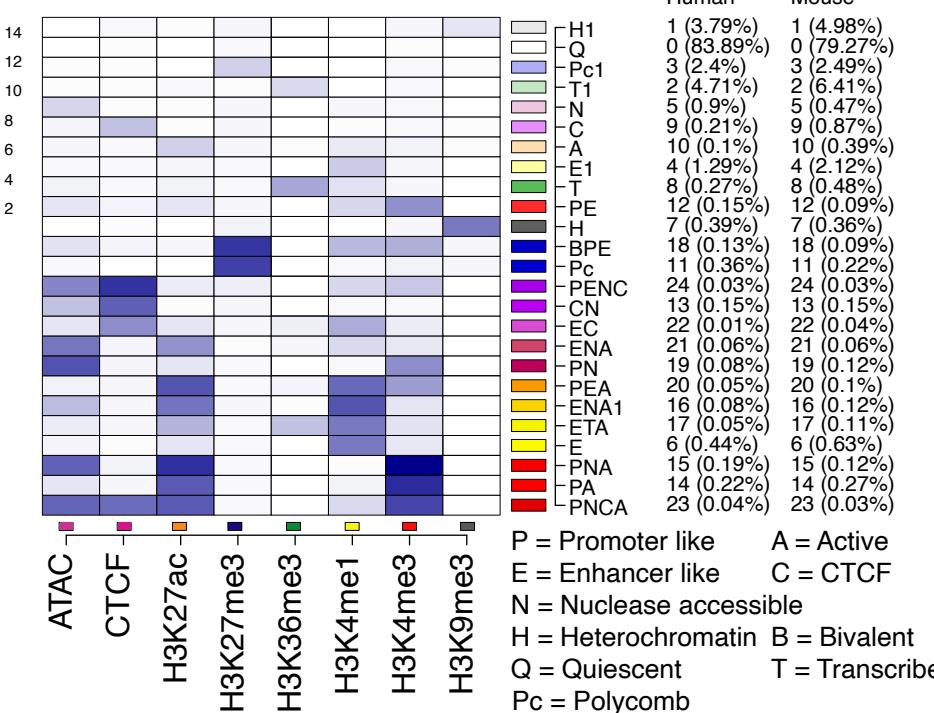


Source: ● VISION ● Other

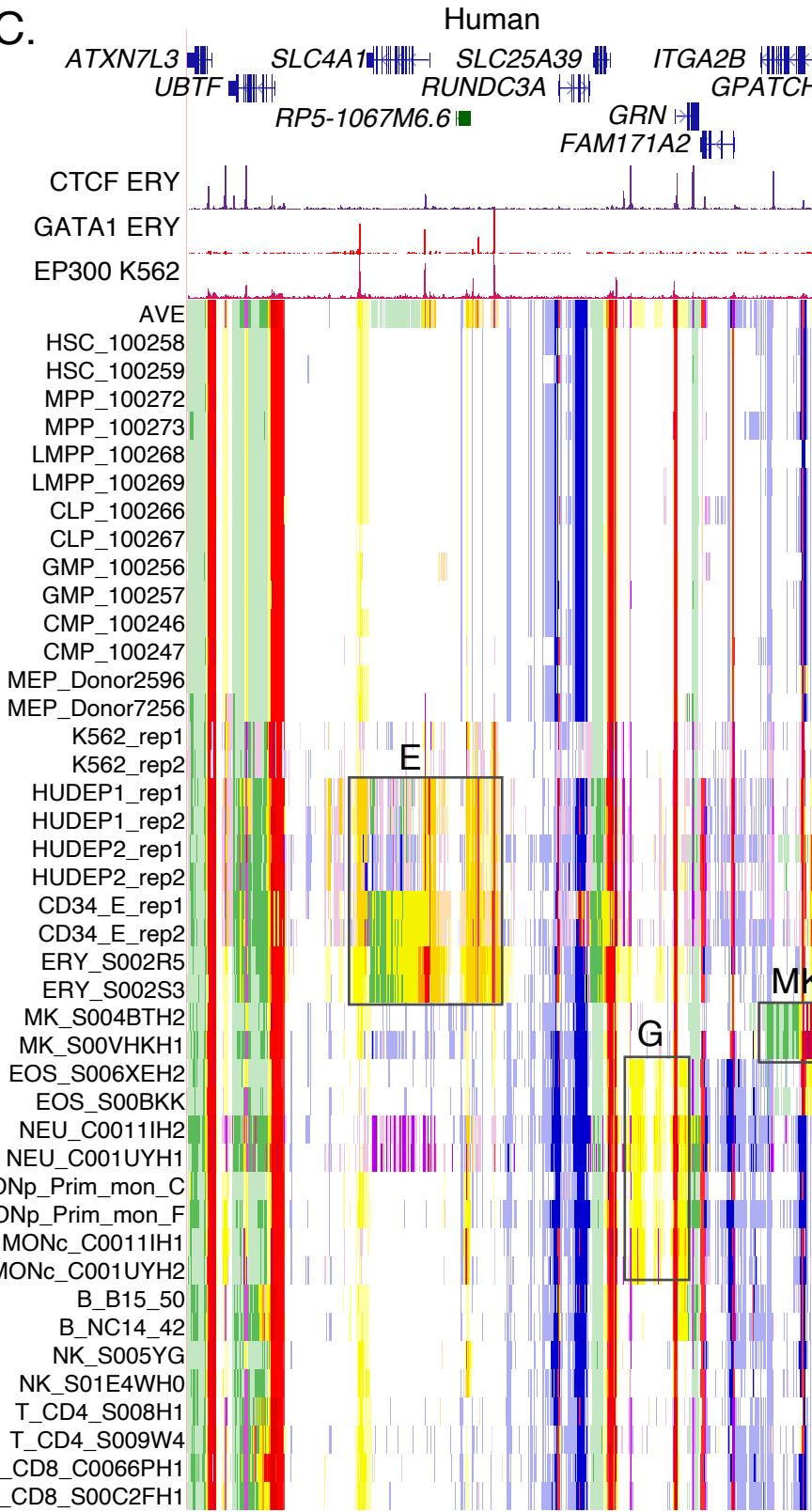
A.



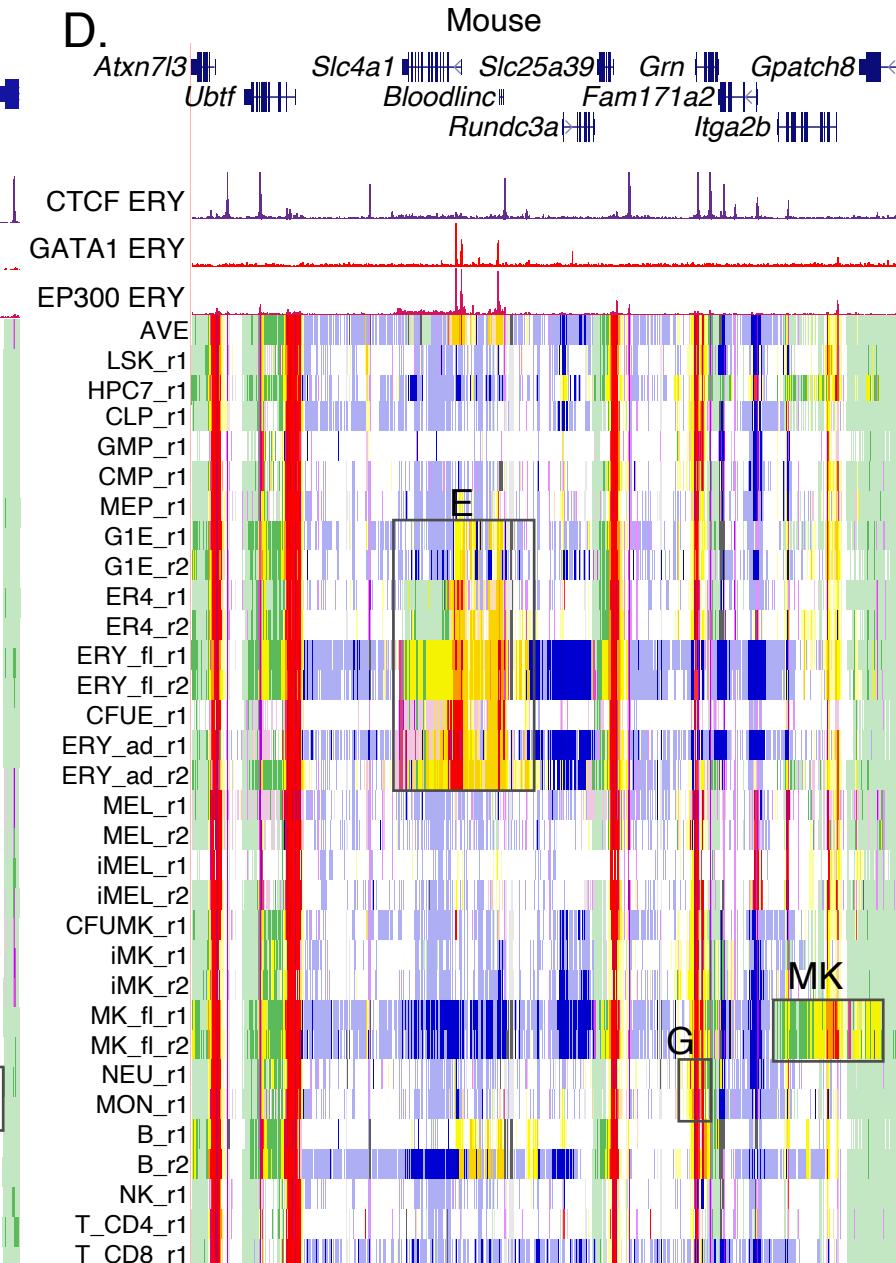
B.

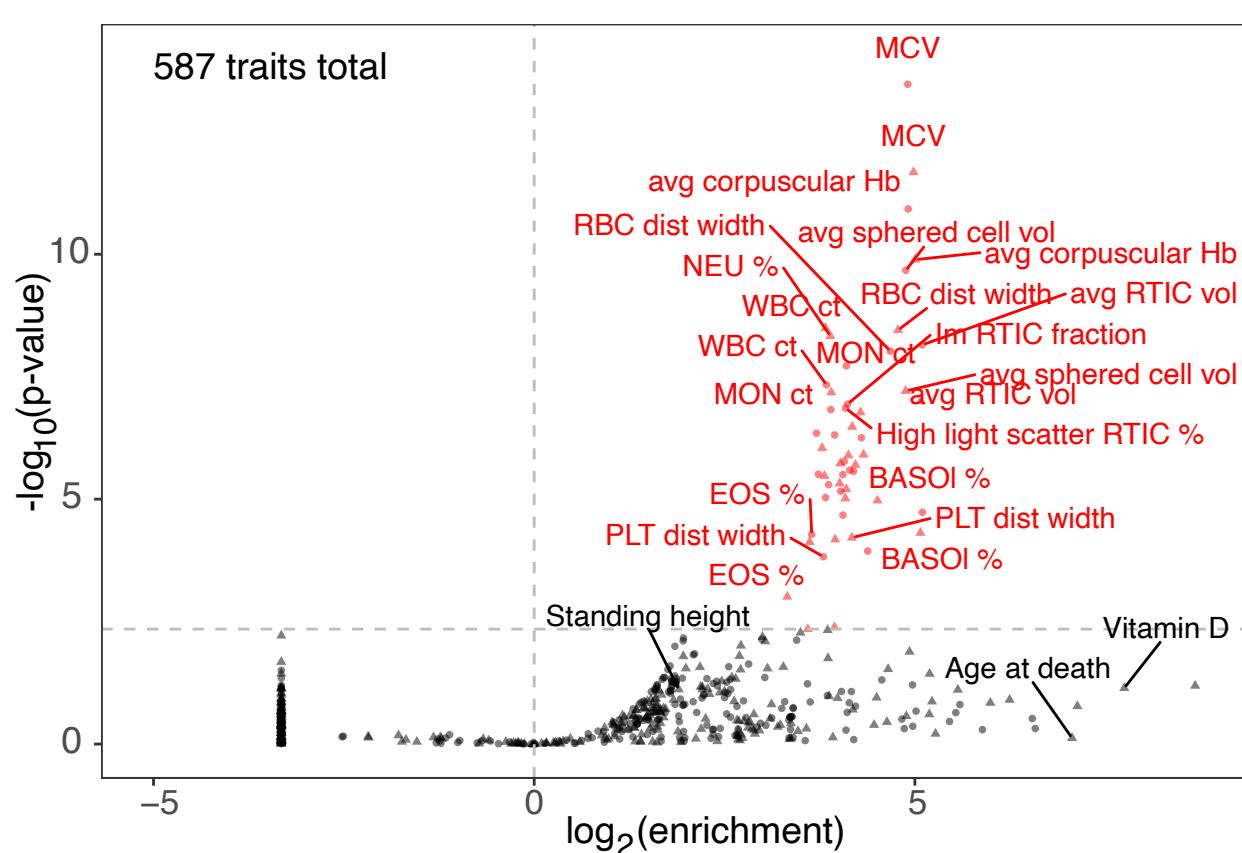
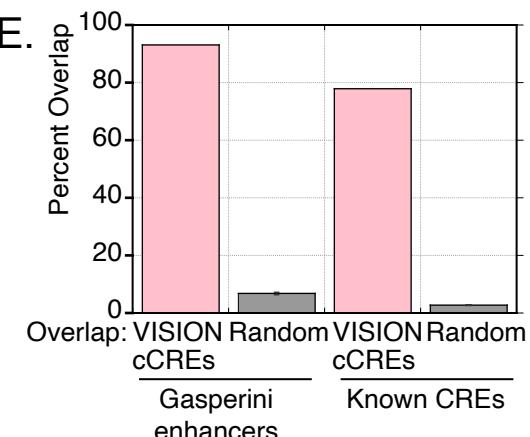
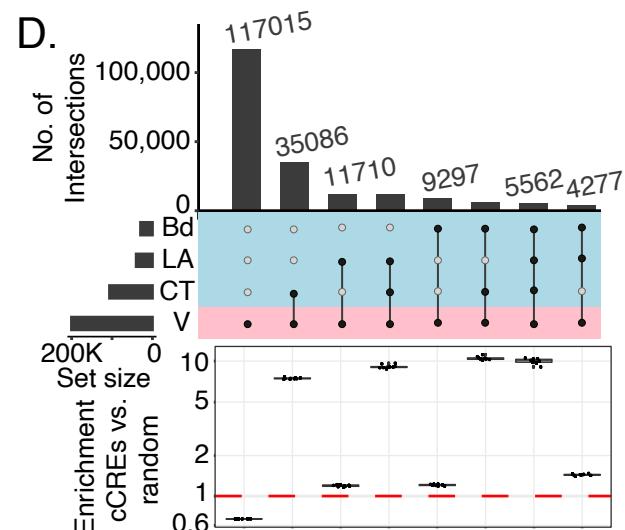
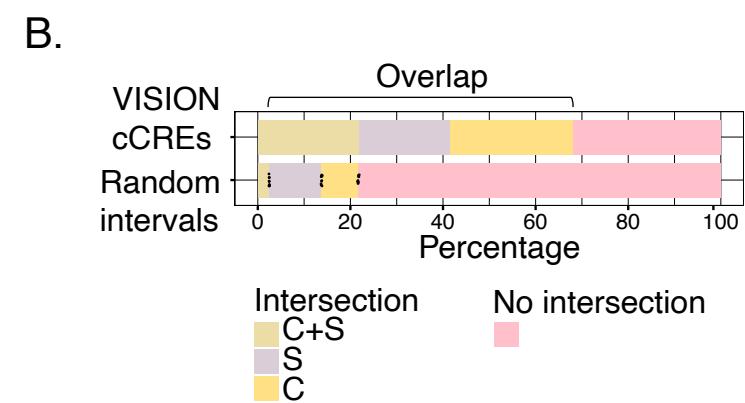
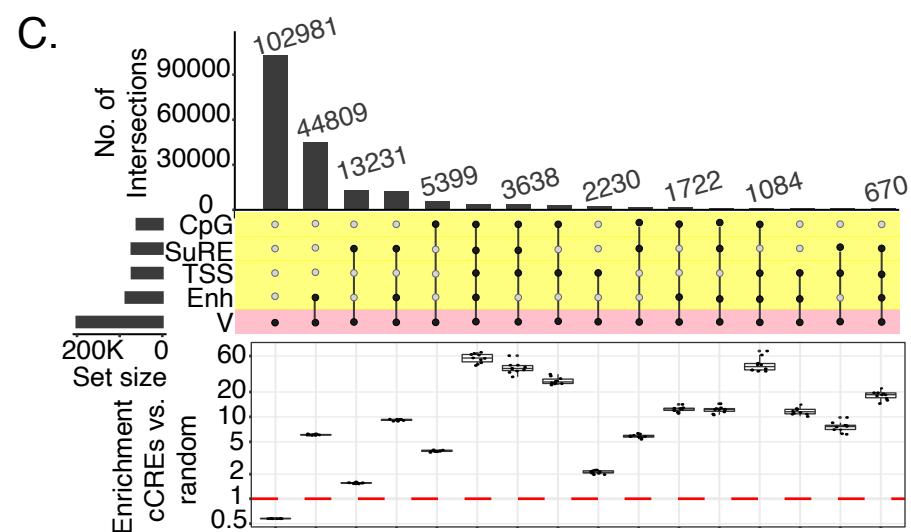
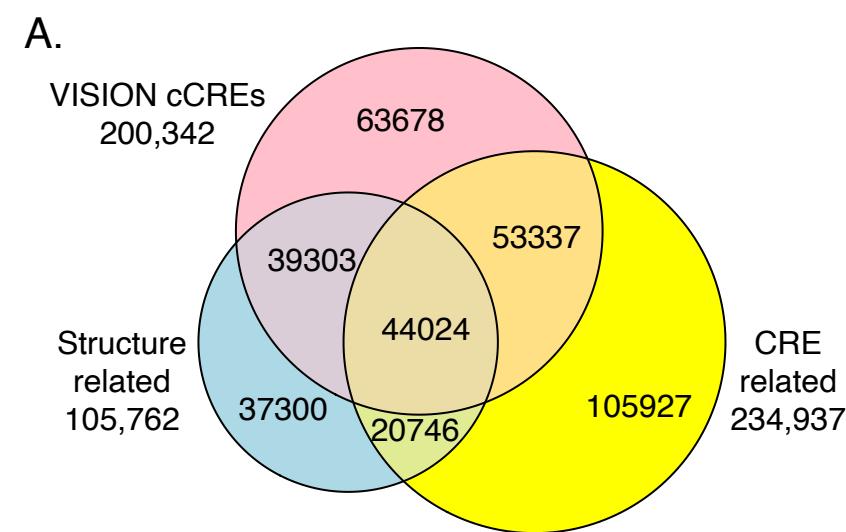


C.

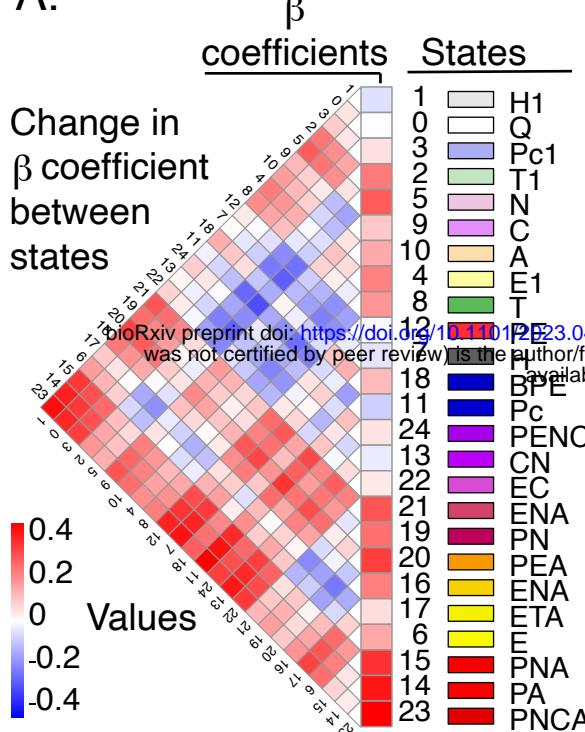


D.

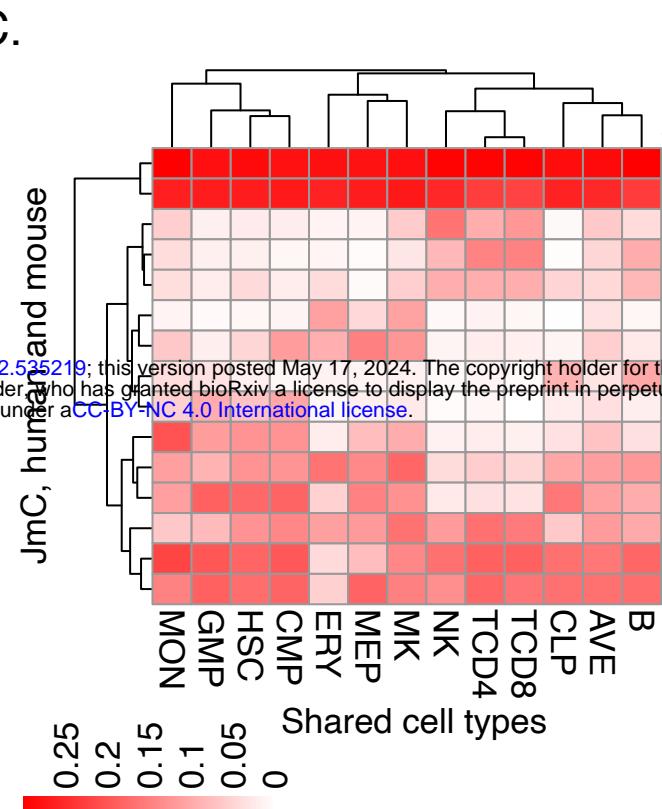




A.



C.



Nmrbr of cCREs

JmC

cCREs

1

4

12

9

8

10

15416

15461

18416

17031

9249

5503

8015

9612

6583

13

JmC Enrichment

CSF1R

GATA1

JmC

IFNG

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

↓

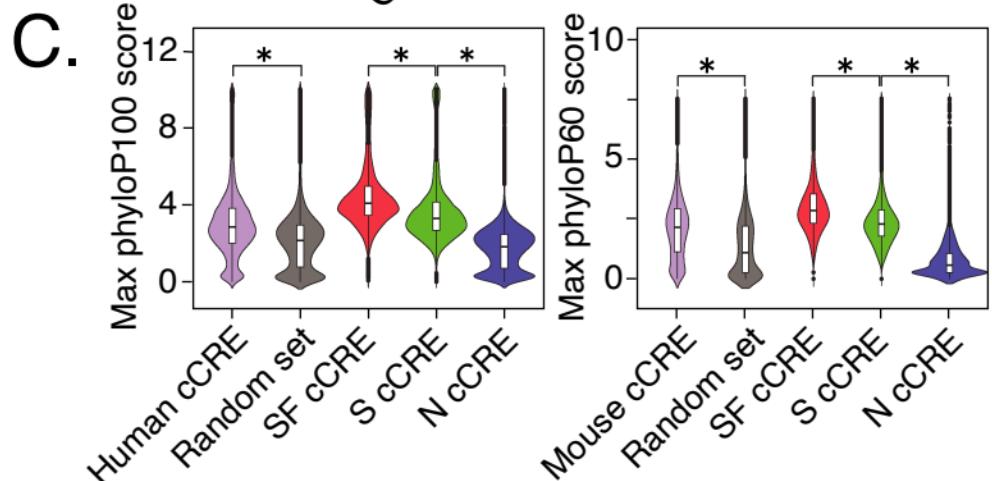
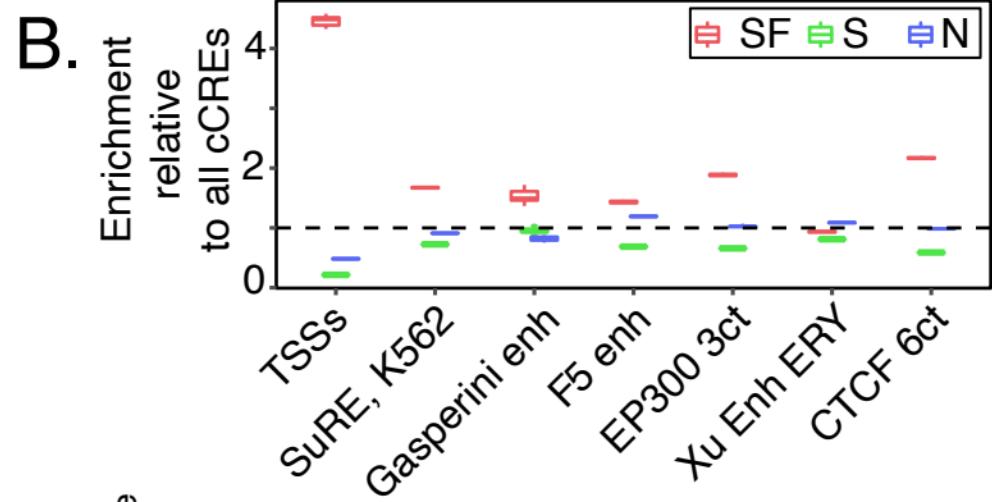
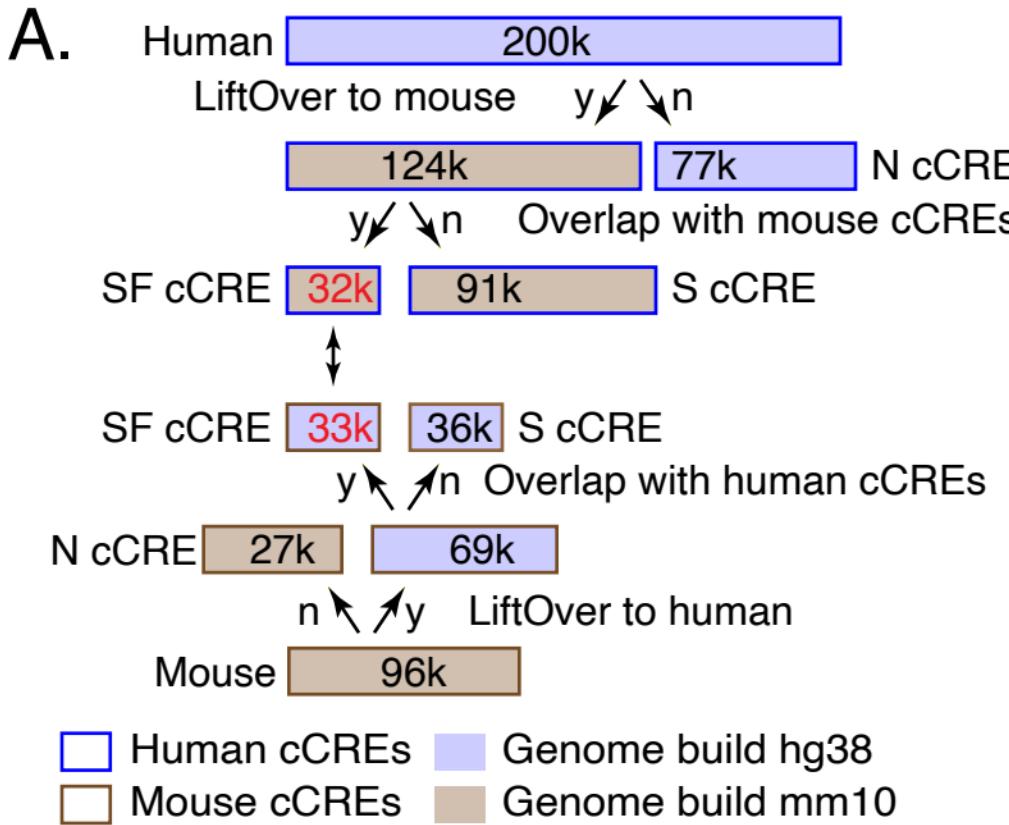
↓

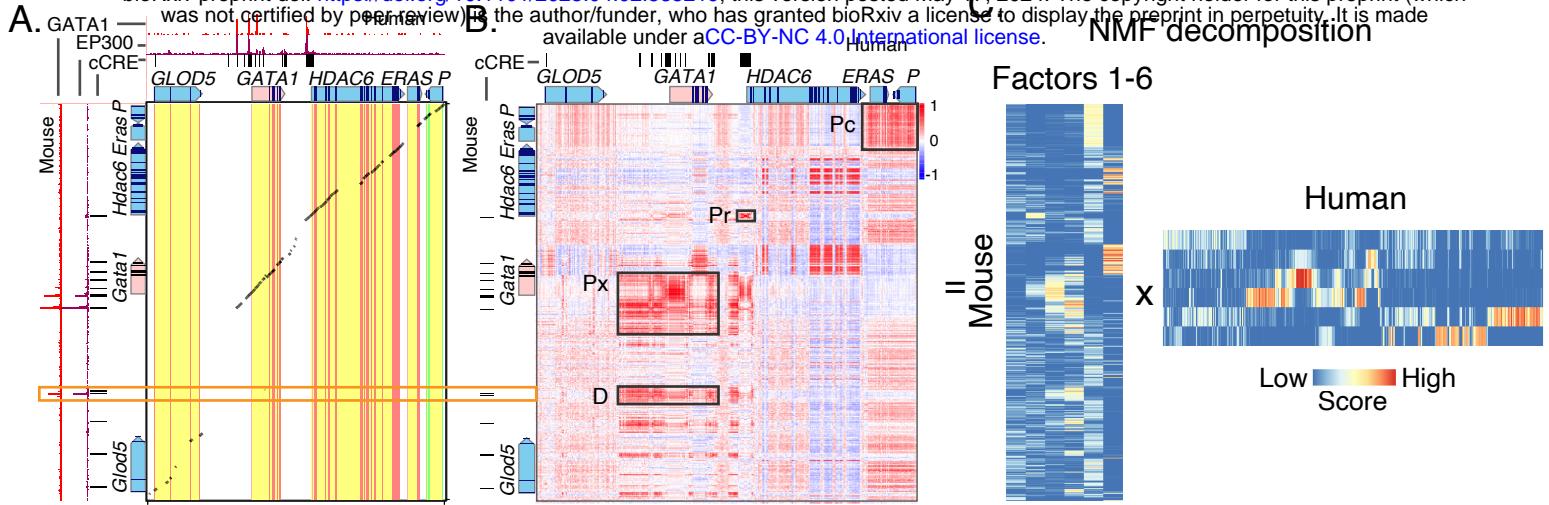
↓

↓

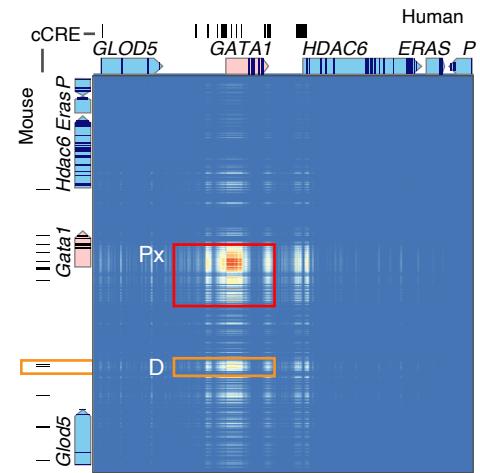
↓

↓

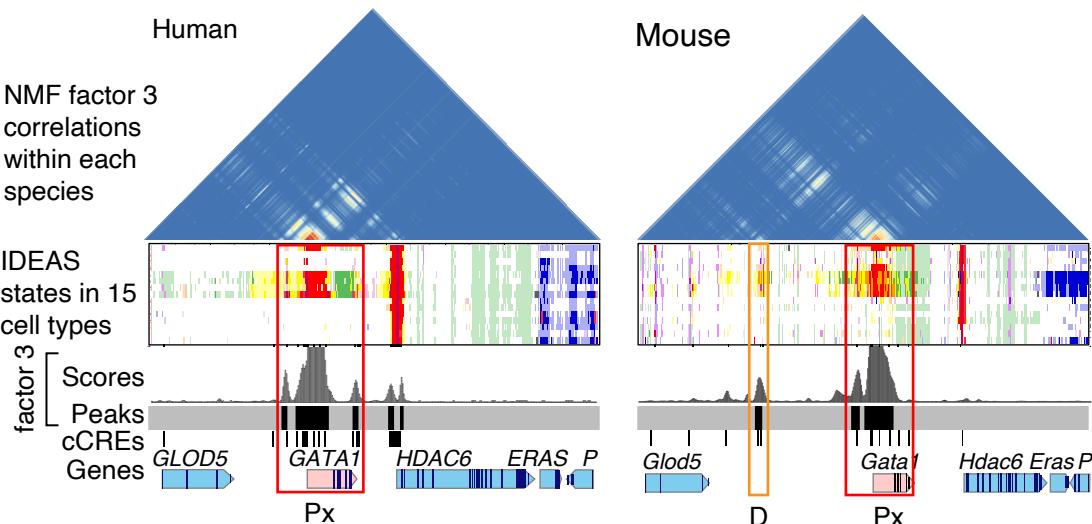




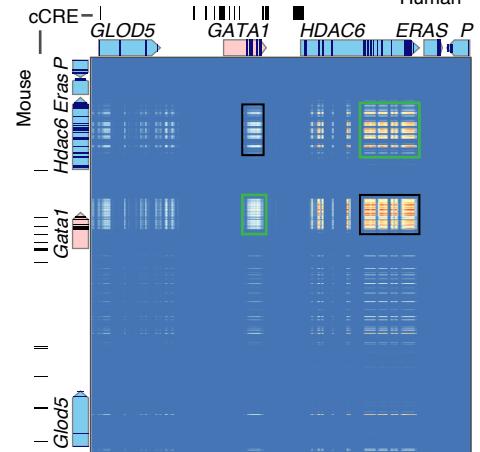
D. NMF factor 3 correlation



E. NMF factor 3: positive regulation ERY & MK



F. NMF factor 6 correlation



G. NMF factor 6: transcription elongation

