

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

3

4 Guanjue Xiang<sup>1,2,3</sup>, Xi He<sup>1</sup>, Belinda M. Giardine<sup>4</sup>, Kathryn J. Isaac<sup>5</sup>, Dylan J. Taylor<sup>5</sup>, Rajiv C.  
5 McCoy<sup>5</sup>, Camden Jansen<sup>4</sup>, Cheryl A. Keller<sup>4</sup>, Alexander Q. Wixom<sup>4</sup>, April Cockburn<sup>4</sup>, Amber  
6 Miller<sup>4</sup>, Qian Qi<sup>6</sup>, Yanghua He<sup>6,7</sup>, Yichao Li<sup>6</sup>, Jens Lichtenberg<sup>8</sup>, Elisabeth F. Heuston<sup>8</sup>, Stacie M.  
7 Anderson<sup>9</sup>, Jing Luan<sup>10</sup>, Marit W. Vermunt<sup>10</sup>, Feng Yue<sup>11</sup>, Michael E.G. Sauria<sup>12</sup>, Michael C.  
8 Schatz<sup>12</sup>, James Taylor<sup>5,12</sup>, Berthold Göttgens<sup>13</sup>, Jim R. Hughes<sup>14</sup>, Douglas R. Higgs<sup>14</sup>, Mitchell  
9 J. Weiss<sup>6</sup>, Yong Cheng<sup>6</sup>, Gerd A. Blobel<sup>10</sup>, David M. Bodine<sup>8</sup>, Yu Zhang<sup>15</sup>, Qunhua Li<sup>15,16</sup>, Shaun  
10 Mahony<sup>4,16,17</sup>, Ross C. Hardison<sup>4,16,17</sup> \*

11

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

14 <sup>2</sup>Department of Data Science, Dana-Farber Cancer Institute, Boston, MA 02215

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

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

18 <sup>5</sup>Department of Biology, Johns Hopkins University, Baltimore, MD 21218

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

20 <sup>7</sup>Department of Human Nutrition, Food and Animal Sciences, University of Hawai'i at Mānoa,  
21 Honolulu, HI 96822, USA

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

24 <sup>9</sup>Flow Cytometry Core, National Human Genome Research Institute, Bethesda, MD 20892

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

27 <sup>11</sup>Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine,  
28 Northwestern University, Evanston, IL 60611

29 <sup>12</sup>Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218

30 <sup>13</sup>Wellcome and MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK

31 <sup>14</sup>MRC Weatherall Institute of Molecular Medicine, Oxford University, Oxford, UK

32 <sup>15</sup>Department of Statistics, The Pennsylvania State University, University Park, PA 16802

33 <sup>16</sup>Center 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 <sup>17</sup>Center 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](mailto: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 Systematic Integration**  
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 (Libbrecht 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 granulins 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-](http://www.nealelab.is/uk-biobank/)  
280 [biobank/](http://www.nealelab.is/uk-biobank/)) within the VISION cCREs relative to the rest of the genome (Supplemental Material  
281 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 glycosylated 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., Karličić 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  $\beta$  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  $\beta$  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  $\beta$  coefficient values in Fig. 4A). The  $\beta$  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  $\beta$  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  $\beta$  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 (Dyran 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  
526 positive regulators for the *GATA1/Gata1* gene loci; these regulatory elements were evident in  
527 reconstructed correlation matrices between species (Fig. 6D) and within individual species (Fig.  
528 6E). By applying a Z-score approach to identify peak regions in the factor 3 signal vector (with  
529 FDR < 0.1; Supplemental Material), we pinpointed regions in both species showing an  
530 epigenetic regulatory machinery exhibiting positive regulatory dynamics for *GATA1/Gata1* gene  
531 loci, particularly in the ERY and MK cell types. In contrast, the correlation matrices  
532 reconstructed from the signals for factor 6 (Fig. 6F and G) highlighted regions marked by the  
533 transcription elongation modification H3K36me3 (epigenetic states colored green, Fig. 6G). The  
534 correlations in the factor 6 elongation signature were observed, as expected, between the  
535 human/mouse orthologous gene pairs *GATA1* and *Gata1* as well as between human *HDAC6*  
536 and mouse *Hdac6* (green rectangles in Fig. 6F). The factor 6 correlations were also observed  
537 between the *GATA1/Gata1* and *HDAC6/Hdac6* genes (black rectangles in Fig. 6F and G),  
538 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 revealed that the active elements were highly enriched in genomic bins with  
572 high cross cell-type epigenetic state correlation between species (Supplemental Fig. S35). The  
573 enrichment for active elements was further increased in bins with both high epigenetic state  
574 correlation and interspecies sequence conservation, while enrichment was reduced in bins with  
575 only sequence conservation. These results further support the value of the cross cell-type  
576 epigenetic state correlation between species in predicting and interpreting cCREs.

577

578 The comparison of epigenetic state profiles across cell types also provided a means to  
579 categorize cCREs between species that did not require a match in the underlying genomic DNA  
580 sequence (Supplemental Figs. S36 and S37). Results from that approach indicated that certain  
581 cCREs were potentially involved in regulation of orthologous genes, even for cCREs with DNA  
582 sequences that did not align between species.

583

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

592

## 593 **Discussion**

594

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

601

602 One major resource is the annotation of the epigenetic states across the genomes of progenitor  
603 and mature blood cells of both species. These state maps show the epigenetic landscape in a  
604 compact form, capturing information from the input data on multiple histone modifications, CTCF  
605 occupancy, and chromatin accessibility, and they use a common set of epigenetic states to  
606 reveal the patterns of epigenetic activity associated with gene expression and regulation both  
607 across cell types and between species. A second major resource is a catalog of cCREs  
608 actuated in one or more of the blood cell types in each species. The cCREs are predictions of  
609 discrete DNA segments likely involved in gene regulation, based on the patterns of chromatin  
610 accessibility across cell types, and the epigenetic state annotations suggest the type of activity  
611 for each cCRE in each cell type, such as serving as a promoter or enhancer, participating in  
612 repression, or inactivity. A third major resource is a quantitative estimate of the regulatory  
613 impact of human and mouse cCREs on gene expression in each cell type, i.e., an esRP score,  
614 derived from multivariate regression modeling of the epigenetic states in cCREs as predictors of  
615 gene expression. The esRP scores are a continuous variable capturing not only the integration  
616 of the input epigenetic data, but also the inferred impacts on gene expression. Those impacts



617 may be manifested as activation or repression during regulation or as transcriptional elongation.  
618 They are useful for many downstream analyses, such as determining informative groups of  
619 cCREs by clustering analysis. These resources along with browsers for visualization and tools  
620 for analysis are provided at our project website, <http://usevision.org>. Among these tools is  
621 cCRE\_db, which records the several dimensions of annotation of the cCREs and provides a  
622 query interface to support custom queries from users.

623

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

631

632 The jointly learned state maps and cCRE predictions allowed us to extend previous work on the  
633 evolution of regulatory elements between mouse and human. Several previous studies focused  
634 on transcription factor (TF) occupancy, e.g. examining key TFs in one tissue across multiple  
635 species (Schmidt et al. 2010; Ballester et al. 2014; Villar et al. 2014) or a diverse set of TFs in  
636 multiple cell types and in mouse and human (Cheng et al. 2014; Yue et al. 2014; Denas et al.  
637 2015). Other studies focused on discrete regions of high chromatin accessibility in multiple cell  
638 types and tissues between mouse and human (Stergachis et al. 2014; Vierstra et al. 2014).

639 These previous studies revealed that only a small fraction of elements was conserved both in  
640 genomic sequence and in inferred function. A notable fraction of elements changed  
641 considerably during mammalian diversification, including turnover of TF binding site motifs and  
642 repurposing of elements (Schmidt et al. 2010; Cheng et al. 2014; Stergachis et al. 2014; Denas

643 et al. 2015). These prior studies focused primarily on regions of the genome with sequences  
644 that aligned between human and mouse, with the non-aligning regions used to infer that some  
645 elements were lineage-specific and that many were derived from transposable elements and  
646 endogenous retroviruses (Bourque 2009; Rebollo et al. 2012; Jacques et al. 2013; Sundaram et  
647 al. 2014).

648  
649 Our evolutionary analyses confirmed the previous observations, e.g., finding about one-third of  
650 cCREs are conserved in both sequence and inferred function between human and mouse, and  
651 further showing that this evolutionary category was highly enriched for proximal regulatory  
652 elements. Going beyond the prior studies, our jointly learned epigenetic state maps generated a  
653 representation of multiple epigenetic features, not just TF occupancy or chromatin accessibility,  
654 and they are continuous in bins across genomes of both species. Thus, they provided a basis  
655 for comparisons of the epigenetic profiles between species. These epigenetic comparisons were  
656 a strong complement to genomic sequence alignments, allowing us to find elements with similar  
657 epigenetic profiles even in genomic regions in which the DNA sequence does not align between  
658 species. In the current work, we used both a correlation between profiles of epigenetic states  
659 and joint clusterings of cCREs between species by esRP scores as initial explorations of these  
660 epigenetic comparisons. Previous work compared epigenetic profiles across species, such as  
661 the phylo-HMGP method to find different evolutionary states in multi-species epigenomic data  
662 (Yang et al. 2018) and the LECIF scores to find evidence of conservation from functional  
663 genomic data (Kwon and Ernst 2021). These approaches are powerful but limited to the  
664 genomic regions with DNA sequences that align between the species. In contrast, our approach  
665 of correlating epigenetic states is agnostic to the underlying DNA sequence alignments (or  
666 absence of them), and thus it complements traditional approaches that rely on DNA sequence  
667 alignments to find similar elements. Our inter-species comparisons of loci surrounding pairs of  
668 orthologous genes included both DNA segments that align between human and mouse and

669 those that do not. Our detection, even in segments of DNA that do not align between species, of  
670 epigenetic similarity indicative of a common role in gene regulation suggests that processes or  
671 structures, such as chromatin interactions, chromatin complexes, or molecular condensates,  
672 may be maintained between species in a manner that is not fully revealed by comparisons of  
673 genome sequences. Hence, further studies of this apparent epigenetic dimension of regulatory  
674 conservation may be productive.

675

676 Several innovations were developed to produce the resources introduced here. A major  
677 innovation was to extend the IDEAS framework (Zhang et al. 2016) to jointly learn epigenetic  
678 states and assign them to annotate the epigenomes in human and mouse blood cells. The  
679 IDEAS method employs a Bayesian approach to the modeling to learn the states, which we  
680 utilized to bring in states learned from the data in one species as priors for learning states in the  
681 data from the second species. Another extension of the IDEAS framework was to learn states  
682 based on one feature, specifically ATAC-seq data, defining discrete signal intensity states. This  
683 approach was used for calling cCREs, implemented as the IDEAS-IS method (Xiang et al.  
684 2021). The approach is relatively simple and benefits from joint modeling across the input  
685 datasets. Other methods for predicting cCREs based on chromatin accessibility across many  
686 cell types prevented excessive expansion of the summary calls for overlapping peaks by  
687 employing a centroid determination for the DNase hypersensitive sites (DHS) index (Meuleman  
688 et al. 2020) or by choosing the highest signal peak for the ENCODE cCRE catalog (The  
689 ENCODE Project Consortium et al. 2020). The ENCODE cCRE catalog paired DHS peaks with  
690 individual chromatin modifications or CTCF occupancy, which led to complications when data  
691 on diagnostic features were missing from some cell types. The IDEAS framework used for the  
692 VISION cCRE sets leveraged data in related cell types to ameliorate the impact of missing data.

693

694 While the resources introduced here are valuable for many applications, it is prudent to  
695 acknowledge their limitations. First, the quality of the products of integrated analyses are limited  
696 by the quality and completeness of the input, raw data. We endeavored to reduce the impact of  
697 variances in the input data by normalization. The S3V2 procedure (Xiang et al. 2021)  
698 systematically normalized the input data to adjust for differences in signal-to-noise and variance  
699 in signal across the datasets. Some epigenetic features were not determined in some cell types,  
700 and we used the IDEAS method in part because it is able to assign an epigenetic state even in  
701 the context of missing data by learning patterns from local similarities in cell types for which the  
702 data are present (Zhang and Mahony 2019). However, these approaches cannot completely  
703 overcome all issues with variance in input data, and further developments in these directions  
704 (such as Shahraki et al. 2023; Xiang et al. 2023) may help to improve integrative resources.  
705 Second, the resolution of both the epigenetic state assignments and the cCRE inference is  
706 limited to 200 bp, which is the window size we utilized in the IDEAS analyses. Other resources,  
707 such as DHS calls (Meuleman et al. 2020), DNase footprints (Vierstra et al. 2020), and motif  
708 instances (Weirauch et al. 2014), achieve a higher resolution. Indeed, one can use these higher  
709 resolution datasets to derive further information about cCREs, such as families of TFs that are  
710 likely to be binding to them. Regarding esRP scores, a third limitation is that we do not make  
711 explicit assignments for target genes of cCREs. Predictions of a large number of target gene-  
712 cCRE pairs were made in our prior work (Xiang et al. 2020); these assignments cover large  
713 genomic intervals around each gene and are most useful when used with further filtering, such  
714 as restricting cCREs and target genes to the same topologically associated domains. On-going  
715 work is examining other models and approaches for assigning likely target genes to cCREs. A  
716 fourth limitation is that our inference of repression-related cCREs apply only to those with stable  
717 histone modifications. Elements that had been involved in initiation of repression but eventually  
718 were packaged into quiescent chromatin, e.g., via a hit-and-run mechanism (Shah et al. 2019),  
719 would not be detected. A fifth limitation concerns the scale of the studies of epigenetic

720 conservation by correlations of epigenetic states. Our current approach is limited to individual  
721 examination of specific genetic loci, since we used orthologous genes as the initial anchors, and  
722 it is likely that a direct application to whole chromosomes or genomes would generate high false  
723 discovery. Exploring ways to expand the scale of the analytical approach is a goal of future  
724 research. Finally, the work presented here was restricted to blood cell types. In future work,  
725 extension of the approaches developed in this study to a broader spectrum of cell types would  
726 expand the utility of the resulting resources.

727

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

735

## 736 **Methods**

737

### 738 **Data generation, collation, normalization, and integration**

739 The data sets used as input, including the ones generated for the work reported here (with  
740 methods), are described in Supplemental Material section “Data generation and collection” and  
741 Supplemental Tables S1 and S2. The S3V2 approach (Xiang et al. 2021) was used for  
742 normalization and denoising the data sets prior to integration. The data sets were integrated to  
743 find and assign epigenetic states using IDEAS (Zhang et al. 2016; Zhang and Hardison 2017);  
744 the extension of this approach to joint learning and annotation between species is described in

745 Supplemental Material sections “Data normalization” and “Joint systematic integration of human  
746 and mouse blood cell epigenomes by IDEAS”.

747

#### 748 **Prediction, annotation, and estimation of regulatory impact of cCREs**

749 The identification of cCREs as peaks of chromatin accessibility employed IDEAS in the signal  
750 intensity state (IS) mode (Xiang et al. 2021). This approach and comparisons with MACS peaks  
751 (Zhang et al. 2008) are described in Supplemental Material section “Prediction of VISION  
752 cCREs using IDEAS-IS”. The cCREs are provided in Supplemental Table S3. Annotation of  
753 potential cCRE functions used intersections with orthogonal data sets of elements implicated in  
754 regulation or chromatin structure (Supplemental Table S5). Enrichment of genetic variants  
755 associated with blood cell traits used stratified linkage disequilibrium score regression (sLDSC,  
756 Finucane et al. 2015). The impact of epigenetic states in cCREs on regulation of gene  
757 expression used a multivariate linear regression approach like one described previously (Xiang  
758 et al. 2020). Methods and supplementary results on these analyses are presented in detail in  
759 the Supplemental Material.

760

#### 761 **Identification of clusters of cCREs based on epigenetic regulatory potential scores**

762 The sets of human and mouse cCREs were placed jointly into groups based on their epigenetic  
763 regulatory potential (esRP) scores using a series of *k*-means clustering steps, as described in  
764 detail in Supplemental Material and Supplementary Fig. S14. Methods and results for  
765 enrichment of the resulting joint meta-clusters (JmCs) for orthogonal sets of regulatory elements  
766 and SNPs associated with blood cell traits, along with comparisons of clusters based on  
767 chromatin accessibility and H3K27ac signal, are described in Supplemental Material and  
768 Supplementary Figs. S15 - S18. Motifs that were differentially enriched across JmCs were  
769 identified using the Maelstrom tool in the GimmeMotifs suite (v0.17.1) (Bruse and van

770 Heeringen 2018) and SeqUnwinder (Kakumanu et al. 2017), as described in detail in  
771 Supplemental Material and Supplementary Fig. S21.

772

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

775 The human and mouse cCREs were assigned to three evolutionary categories using the  
776 following procedure. The set of human cCREs was mapped to mouse genome assembly mm10  
777 using the liftOver tool at the UCSC Genome Browser (Hinrichs et al. 2006). Human cCREs that  
778 failed to map to mm10 were grouped as N cCREs. Matches to mouse cCREs for the human  
779 cCREs that could be mapped by liftOver to mm10 were determined using the intersect tool in  
780 BEDTools (Quinlan and Hall 2010). Human cCREs that overlapped with mouse cCREs were  
781 labeled as SF cCREs, while human cCREs that mapped to mm10 but did not match mouse  
782 cCREs were labeled as S cCREs. A similar process was performed on the set of mouse cCREs  
783 using liftOver to map to human genome build GRCh38

784

785 **Calculation of pairwise correlation coefficients for epigenetic landscapes between**  
786 **human and mouse**

787 A bin-to-bin pairwise correlation analysis was used to quantify the similarity of epigenetic  
788 landscapes between two DNA regions in human and mouse. For each 200bp bin in one cell  
789 type in one species, the assigned epigenetic state was replaced by a vector of mean signals of  
790 8 epigenetic features in the IDEAS state model. After replacing the states in all 15 matched cell  
791 types (14 analogous cell types and one pseudo-cell type with average values for all cell types)  
792 in the two species, the original two categorical state vectors with 15 elements were converted  
793 into two numeric vectors with 120 numbers (Supplemental Fig. S28). The similarity of cross-cell  
794 type epigenetic landscape between two bins in the two species was defined as the correlation  
795 coefficient between each pair of numeric vectors with 120 numbers. When calculating the

796 correlation coefficients, we added random noise (mean=0, sd=0.2) to the raw values to avoid  
797 high correlation coefficients created between regions with states that have low signals. The  
798 complex correlation matrix was decomposed into distinctive factors using Nonnegative Matrix  
799 Factorization (Lee and Seung 1999). Methods and supplementary results on these analyses are  
800 presented in detail in the Supplemental Material.

801

## 802 **Data access**

803 All raw and processed sequencing data generated in this study have been submitted to the  
804 NCBI Gene Expression Omnibus (GEO; <https://www.ncbi.nlm.nih.gov/geo/>) under accession  
805 number GSE229101 and the NCBI BioProject database  
806 (<https://www.ncbi.nlm.nih.gov/bioproject/>) under accession number PRJNA952902. Resources  
807 developed in the VISION project are available at the website <https://usevision.org>; the data can  
808 be viewed via a track hub at the UCSC Genome Browser or any compatible browser by using  
809 this URL: <https://usevision.org/data/trackHub/hub.txt> or by clicking the track hubs link at  
810 usevision.org. The database cCRE db supports flexible user queries on extensive annotation of  
811 the cCREs, including epigenetic states and esRP scores across cell types, chromatin  
812 accessibility scores across cell types, membership in JmCs, and evolutionary categories. Code  
813 developed for this study is in the Supplemental Material and at these GitHub repositories:  
814 [https://github.com/guanjue/Joint\\_Human\\_Mouse\\_IDEAS\\_State](https://github.com/guanjue/Joint_Human_Mouse_IDEAS_State) for the joint human-mouse  
815 IDEAS pipeline and [https://github.com/usevision/cre\\_heritability](https://github.com/usevision/cre_heritability) for the sLDSC analysis.

816

## 817 **Competing interest statement**

818 The authors declare no competing interests.

819

## 820 **Acknowledgments**



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

827

## 828 **References**

- 829 Agarwal V, Inoue F, Schubach M, Martin BK, Dash PM, Zhang Z, Sohota A, Noble WS,  
830 Yardimci GG, Kircher M et al. 2023. Massively parallel characterization of transcriptional  
831 regulatory elements in three diverse human cell types. *bioRxiv*  
832 doi:10.1101/2023.03.05.531189.
- 833 Ballester B, Medina-Rivera A, Schmidt D, Gonzalez-Porta M, Carlucci M, Chen X, Chessman K,  
834 Faure AJ, Funnell AP, Goncalves A et al. 2014. Multi-species, multi-transcription factor  
835 binding highlights conserved control of tissue-specific biological pathways. *eLife* **3**:  
836 e02626.
- 837 Bauer DE, Kamran SC, Lessard S, Xu J, Fujiwara Y, Lin C, Shao Z, Canver MC, Smith EC,  
838 Pinello L et al. 2013. An erythroid enhancer of BCL11A subject to genetic variation  
839 determines fetal hemoglobin level. *Science* **342**: 253-257.
- 840 Blobel GA, Weiss MJ. 2009. Nuclear Factors that Regulate Erythropoiesis. In *Disorders of*  
841 *Hemoglobin: Genetics, Pathophysiology, and Clinical Management*, (ed. MH Steinberg,  
842 et al.), pp. 62-85. Cambridge University Press, Cambridge.
- 843 Bourque G. 2009. Transposable elements in gene regulation and in the evolution of vertebrate  
844 genomes. *Curr Opin Genet Dev* **19**: 607-612.

- 845 Bruse N, van Heeringen SJ. 2018. GimmeMotifs: an analysis framework for transcription factor  
846 motif analysis. *bioRxiv* doi:<https://doi.org/10.1101/474403>.
- 847 Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A,  
848 Morales J, Mountjoy E, Sollis E et al. 2019. The NHGRI-EBI GWAS Catalog of published  
849 genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic  
850 Acids Res* **47**: D1005-D1012.
- 851 Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of  
852 morphological evolution. *Cell* **134**: 25-36.
- 853 Cheng L, Li Y, Qi Q, Xu P, Feng R, Palmer L, Chen J, Wu R, Yee T, Zhang J et al. 2021.  
854 Single-nucleotide-level mapping of DNA regulatory elements that control fetal  
855 hemoglobin expression. *Nat Genet* **53**: 869-880.
- 856 Cheng Y, Ma Z, Kim BH, Wu W, Cayting P, Boyle AP, Sundaram V, Xing X, Dogan N, Li J et al.  
857 2014. Principles of regulatory information conservation between mouse and human.  
858 *Nature* **515**: 371-375.
- 859 Chi AW, Bell JJ, Zlotoff DA, Bhandoola A. 2009. Untangling the T branch of the hematopoiesis  
860 tree. *Curr Opin Immunol* **21**: 121-126.
- 861 Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, Snyder MP, Pritchard  
862 JK, Kundaje A, Greenleaf WJ et al. 2016. Lineage-specific and single-cell chromatin  
863 accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet* **48**: 1193-  
864 1203.
- 865 Denas O, Sandstrom R, Cheng Y, Beal K, Herrero J, Hardison RC, Taylor J. 2015. Genome-  
866 wide comparative analysis reveals human-mouse regulatory landscape and evolution.  
867 *BMC Genomics* **16**: 87.
- 868 Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, Gingeras TR, Gerstein M,  
869 Guigo R, Birney E et al. 2012. Modeling gene expression using chromatin features in  
870 various cellular contexts. *Genome Biol* **13**: R53.

- 871 Dore LC, Crispino JD. 2011. Transcription factor networks in erythroid cell and megakaryocyte  
872 development. *Blood* **118**: 231-239.
- 873 Dynan WS, Tjian R. 1983. The promoter-specific transcription factor Sp1 binds to upstream  
874 sequences in the SV40 early promoter. *Cell* **35**: 79-87.
- 875 Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for  
876 systematic annotation of the human genome. *Nat Biotechnol* **28**: 817-825.
- 877 Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and  
878 characterization. *Nat Methods* **9**: 215-216.
- 879 Ferreira R, Ohneda K, Yamamoto M, Philipsen S. 2005. GATA1 function, a paradigm for  
880 transcription factors in hematopoiesis. *Mol Cell Biol* **25**: 1215-1227.
- 881 Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, Anttila V, Xu H, Zang C,  
882 Farh K et al. 2015. Partitioning heritability by functional annotation using genome-wide  
883 association summary statistics. *Nat Genet* **47**: 1228-1235.
- 884 Frangoul H, Altshuler D, Cappellini MD, Chen YS, Domm J, Eustace BK, Foell J, de la Fuente J,  
885 Grupp S, Handgretinger R et al. 2021. CRISPR-Cas9 Gene Editing for Sickle Cell  
886 Disease and beta-Thalassemia. *The New England journal of medicine* **384**: 252-260.
- 887 Fujiwara T, O'Geen H, Keles S, Blahnik K, Linnemann AK, Kang YA, Choi K, Farnham PJ,  
888 Bresnick EH. 2009. Discovering hematopoietic mechanisms through genome-wide  
889 analysis of GATA factor chromatin occupancy. *Mol Cell* **36**: 667-681.
- 890 Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, Jackson D, Leith A,  
891 Schreiber J, Noble WS et al. 2019. A Genome-wide Framework for Mapping Gene  
892 Regulation via Cellular Genetic Screens. *Cell* **176**: 377-390 e319.
- 893 Ge T, Chen CY, Neale BM, Sabuncu MR, Smoller JW. 2017. Phenome-wide heritability analysis  
894 of the UK Biobank. *PLoS Genet* **13**: e1006711.
- 895 Graf T, Enver T. 2009. Forcing cells to change lineages. *Nature* **462**: 587-594.

896 Gumucio DL, Heilstedt-Williamson H, Gray TA, Tarle SA, Shelton DA, Tagle D, Slightom J,  
897 Goodman M, Collins FS. 1992. Phylogenetic footprinting reveals a nuclear protein which  
898 binds to silencer sequences in the human g and e globin genes. *Mol Cell Biol* **12**: 4919-  
899 4929.

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

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

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

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

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

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

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

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

- 922 Jacques PE, Jeyakani J, Bourque G. 2013. The majority of primate-specific regulatory  
923 sequences are derived from transposable elements. *PLoS Genet* **9**: e1003504.
- 924 Jansen C, Ramirez RN, El-Ali NC, Gomez-Cabrero D, Tegner J, Merckenschlager M, Conesa A,  
925 Mortazavi A. 2019. Building gene regulatory networks from scATAC-seq and scRNA-seq  
926 using Linked Self Organizing Maps. *PLoS Comput Biol* **15**: e1006555.
- 927 Jian J, Konopka J, Liu C. 2013. Insights into the role of progrenulin in immunity, infection, and  
928 inflammation. *J Leukoc Biol* **93**: 199-208.
- 929 Kaczynski J, Cook T, Urrutia R. 2003. Sp1- and Kruppel-like transcription factors. *Genome Biol*  
930 **4**: 206.
- 931 Kakumanu A, Velasco S, Mazzoni E, Mahony S. 2017. Deconvolving sequence features that  
932 discriminate between overlapping regulatory annotations. *PLoS Comput Biol* **13**:  
933 e1005795.
- 934 Karlić R, Chung HR, Lasserre J, Vlahovicek K, Vingron M. 2010. Histone modification levels are  
935 predictive for gene expression. *Proc Natl Acad Sci U S A* **107**: 2926-2931.
- 936 King DC, Taylor J, Zhang Y, Cheng Y, Lawson HA, Martin J,  
937 ENCODE\_groups\_for\_Transcriptional\_Regulation\_and\_Multispecies\_Sequence\_Analysis  
938 s, Chiaromonte F, Miller W, Hardison RC. 2007. Finding cis-regulatory elements using  
939 comparative genomics: some lessons from ENCODE data. *Genome Res* **17**: 775-786.
- 940 Kondo M, Wagers AJ, Manz MG, Prohaska SS, Scherer DC, Beilhack GF, Shizuru JA,  
941 Weissman IL. 2003. Biology of hematopoietic stem cells and progenitors: implications for  
942 clinical application. *Annu Rev Immunol* **21**: 759-806.
- 943 Kwon SB, Ernst J. 2021. Learning a genome-wide score of human-mouse conservation at the  
944 functional genomics level. *Nature communications* **12**: 2495.
- 945 Laurenti E, Göttgens B. 2018. From haematopoietic stem cells to complex differentiation  
946 landscapes. *Nature* **553**: 418-426.

- 947 Lee DD, Seung HS. 1999. Learning the parts of objects by non-negative matrix factorization.  
948 *Nature* **401**: 788-791.
- 949 Lee DI, Roy S. 2021. GRINCH: simultaneous smoothing and detection of topological units of  
950 genome organization from sparse chromatin contact count matrices with matrix  
951 factorization. *Genome Biol* **22**: 164.
- 952 Libbrecht MW, Chan RCW, Hoffman MM. 2021. Segmentation and genome annotation  
953 algorithms for identifying chromatin state and other genomic patterns. *PLoS Comput Biol*  
954 **17**: e1009423.
- 955 Martens JH, Stunnenberg HG. 2013. BLUEPRINT: mapping human blood cell epigenomes.  
956 *Haematologica* **98**: 1487-1489.
- 957 Maston GA, Evans SK, Green MR. 2006. Transcriptional Regulatory Elements in the Human  
958 Genome. *Annu Rev Genomics Hum Genet* **7**: 29-59.
- 959 Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom  
960 R, Qu H, Brody J et al. 2012. Systematic localization of common disease-associated  
961 variation in regulatory DNA. *Science* **337**: 1190-1195.
- 962 Meuleman W, Muratov A, Rynes E, Halow J, Lee K, Bates D, Diegel M, Dunn D, Neri F,  
963 Teodosiadis A et al. 2020. Index and biological spectrum of human DNase I  
964 hypersensitive sites. *Nature* **584**: 244-251.
- 965 Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, Wolfe SA. 2008.  
966 Analysis of homeodomain specificities allows the family-wide prediction of preferred  
967 recognition sites. *Cell* **133**: 1277-1289.
- 968 Payne KJ, Crooks GM. 2002. Human hematopoietic lineage commitment. *Immunol Rev* **187**:  
969 48-64.
- 970 Pennacchio LA, Rubin EM. 2001. Genomic strategies to identify mammalian regulatory  
971 sequences. *Nat Rev Genet* **2**: 100-109.

- 972 Pimkin M, Kossenkov AV, Mishra T, Morrissey CS, Wu W, Keller CA, Blobel GA, Lee D, Beer  
973 MA, Hardison RC et al. 2014. Divergent functions of hematopoietic transcription factors  
974 in lineage priming and differentiation during erythro-megakaryopoiesis. *Genome Res* **24**:  
975 1932-1944.
- 976 Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution  
977 rates on mammalian phylogenies. *Genome Res* **20**: 110-121.
- 978 Qi Q, Cheng L, Tang X, He Y, Li Y, Yee T, Shrestha D, Feng R, Xu P, Zhou X et al. 2021.  
979 Dynamic CTCF binding directly mediates interactions among cis-regulatory elements  
980 essential for hematopoiesis. *Blood* **137**: 1327-1339.
- 981 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic  
982 features. *Bioinformatics* **26**: 841-842.
- 983 Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and natural  
984 source of regulatory sequences for host genes. *Annu Rev Genet* **46**: 21-42.
- 985 Ringrose L, Paro R. 2004. Epigenetic regulation of cellular memory by the Polycomb and  
986 Trithorax group proteins. *Annu Rev Genet* **38**: 413-443.
- 987 Rothenberg EV, Taghon T. 2005. Molecular genetics of T cell development. *Annu Rev Immunol*  
988 **23**: 601-649.
- 989 Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S,  
990 Martinez-Jimenez CP, Mackay S et al. 2010. Five-vertebrate ChIP-seq reveals the  
991 evolutionary dynamics of transcription factor binding. *Science* **328**: 1036-1040.
- 992 Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W.  
993 2000. PipMaker-A web server for aligning two genomic DNA sequences. *Genome Res*  
994 **10**: 577-586.
- 995 Shah M, Funnell APW, Quinlan KGR, Crossley M. 2019. Hit and Run Transcriptional  
996 Repressors Are Difficult to Catch in the Act. *Bioessays* **41**: e1900041.

- 997 Shahraki MF, Farahbod M, Libbrecht MW. 2023. Robust chromatin state annotation. *bioRxiv*  
998 doi:<https://doi.org/10.1101/2023.07.15.549175>.
- 999 Spangrude GJ, Heimfeld S, Weissman IL. 1988. Purification and characterization of mouse  
1000 hematopoietic stem cells. *Science* **241**: 58-62.
- 1001 Stein-O'Brien GL, Arora R, Culhane AC, Favorov AV, Garmire LX, Greene CS, Goff LA, Li Y,  
1002 Ngom A, Ochs MF et al. 2018. Enter the Matrix: Factorization Uncovers Knowledge from  
1003 Omics. *Trends Genet* **34**: 790-805.
- 1004 Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, Byron R, Canfield T,  
1005 Stelhing-Sun S, Lee K et al. 2014. Conservation of trans-acting circuitry during  
1006 mammalian regulatory evolution. *Nature* **515**: 365-370.
- 1007 Strahl BD, Allis CD. 2000. The language of covalent histone modifications. *Nature* **403**: 41-45.
- 1008 Stunnenberg HG, International Human Epigenome C, Hirst M. 2016. The International Human  
1009 Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167**:  
1010 1145-1149.
- 1011 Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. 2014. Widespread  
1012 contribution of transposable elements to the innovation of gene regulatory networks.  
1013 *Genome Res* **24**: 1963-1976.
- 1014 Tenen DG, Hromas R, Licht JD, Zhang DE. 1997. Transcription factors, normal myeloid  
1015 development, and leukemia. *Blood* **90**: 489-519.
- 1016 The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the  
1017 human genome. *Nature* **489**: 57-74.
- 1018 The ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shores N,  
1019 Adrian J, Kawli T, Davis CA, Dobin A et al. 2020. Expanded encyclopaedias of DNA  
1020 elements in the human and mouse genomes. *Nature* **583**: 699-710.



- 1021 Valverde-Garduno V, Guyot B, Anguita E, Hamlett I, Porcher C, Vyas P. 2004. Differences in  
1022 the chromatin structure and cis-element organization of the human and mouse GATA1  
1023 loci: implications for cis-element identification. *Blood* **104**: 3106-3116.
- 1024 van Arensbergen J, FitzPatrick VD, de Haas M, Pagie L, Sluimer J, Bussemaker HJ, van  
1025 Steensel B. 2017. Genome-wide mapping of autonomous promoter activity in human  
1026 cells. *Nat Biotechnol* **35**: 145-153.
- 1027 van Pampus EC, Denkers IA, van Geel BJ, Huijgens PC, Zevenbergen A, Ossenkoppele GJ,  
1028 Langenhuijsen MM. 1992. Expression of adhesion antigens of human bone marrow  
1029 megakaryocytes, circulating megakaryocytes and blood platelets. *Eur J Haematol* **49**:  
1030 122-127.
- 1031 Vierstra J, Lazar J, Sandstrom R, Halow J, Lee K, Bates D, Diegel M, Dunn D, Neri F, Haugen  
1032 E et al. 2020. Global reference mapping of human transcription factor footprints. *Nature*  
1033 **583**: 729-736.
- 1034 Vierstra J, Rynes E, Sandstrom R, Zhang M, Canfield T, Hansen RS, Stehling-Sun S, Sabo PJ,  
1035 Byron R, Humbert R et al. 2014. Mouse regulatory DNA landscapes reveal global  
1036 principles of cis-regulatory evolution. *Science* **346**: 1007-1012.
- 1037 Villar D, Flicek P, Odom DT. 2014. Evolution of transcription factor binding in metazoans -  
1038 mechanisms and functional implications. *Nat Rev Genet* **15**: 221-233.
- 1039 Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS,  
1040 Lambert SA, Mann I, Cook K et al. 2014. Determination and inference of eukaryotic  
1041 transcription factor sequence specificity. *Cell* **158**: 1431-1443.
- 1042 Weiss MJ, Yu C, Orkin SH. 1997. Erythroid-cell-specific properties of transcription factor GATA-  
1043 1 revealed by phenotypic rescue of a gene-targeted cell line. *Mol Cell Biol* **17**: 1642-  
1044 1651.

- 1045 Xiang G, Giardine BM, Mahony S, Zhang Y, Hardison RC. 2021. S3V2-IDEAS: a package for  
1046 normalizing, denoising and integrating epigenomic datasets across different cell types.  
1047 *Bioinformatics* **37**: 3011-3013.
- 1048 Xiang G, Guo Y, Bumcrot D, Sigova A. 2023. JMnorm: a novel Joint Multi-feature normalization  
1049 method for integrative and comparative epigenomics. *bioRxiv*  
1050 doi:<https://doi.org/10.1101/2023.06.14.545004>.
- 1051 Xiang G, Keller CA, Heuston E, Giardine BM, An L, Wixom AQ, Miller A, Cockburn A, Sauria  
1052 MEG, Weaver K et al. 2020. An integrative view of the regulatory and transcriptional  
1053 landscapes in mouse hematopoiesis. *Genome Res* **30**: 472-484.
- 1054 Xu J, Shao Z, Glass K, Bauer DE, Pinello L, Van Handel B, Hou S, Stamatoyannopoulos JA,  
1055 Mikkola HK, Yuan GC et al. 2012. Combinatorial assembly of developmental stage-  
1056 specific enhancers controls gene expression programs during human erythropoiesis.  
1057 *Dev Cell* **23**: 796-811.
- 1058 Yang Y, Gu Q, Zhang Y, Sasaki T, Crivello J, O'Neill RJ, Gilbert DM, Ma J. 2018. Continuous-  
1059 Trait Probabilistic Model for Comparing Multi-species Functional Genomic Data. *Cell*  
1060 *Syst* **7**: 208-218 e211.
- 1061 Yue F Cheng Y Breschi A Vierstra J Wu W Ryba T Sandstrom R Ma Z Davis C Pope BD et al.  
1062 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**:  
1063 355-364.
- 1064 Zhang Y, An L, Yue F, Hardison RC. 2016. Jointly characterizing epigenetic dynamics across  
1065 multiple human cell types. *Nucleic Acids Res* **44**: 6721-6731.
- 1066 Zhang Y, Hardison RC. 2017. Accurate and reproducible functional maps in 127 human cell  
1067 types via 2D genome segmentation. *Nucleic Acids Res* **45**: 9823-9836.
- 1068 Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM,  
1069 Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**:  
1070 R137.

1071 Zhang Y, Mahony S. 2019. Direct prediction of regulatory elements from partial data without  
1072 imputation. *PLoS Comput Biol* **15**: e1007399.

1073

## 1074 **Figure Legends**

### 1075 **Figure 1. Cell types and data sets used for systematic integration of epigenetic features**

1076 **of blood cells. (A)** The tree on the left shows the populations of stem, progenitor, and mature

1077 blood cells and cell lines in human. The diagram on the right indicates the epigenetic features

1078 and transcriptomes for which genome-wide data sets were generated or collected, with

1079 distinctive icons for the major sources of data, specifically the Blueprint project (Martens and

1080 Stunnenberg 2013; Stunnenberg et al. 2016), Corces et al. (2016), abbreviated CMB, and St.

1081 Jude Children's Research Hospital (SJCRH, Cheng et al. 2021; Qi et al. 2021). (B) Cell types

1082 and epigenetic data sets in mouse, diagrammed as for panel A. Sources were described in

1083 Xiang et al. (2020) and Supplemental Table S1. Abbreviations for blood cells and lines are: HSC

1084 = hematopoietic stem cell, MPP = multipotent progenitor cell, LMPP = lymphoid-myeloid primed

1085 progenitor cell, CMP = common myeloid progenitor cell, MEP = megakaryocyte-erythrocyte

1086 progenitor cell, K562 = a human cancer cell line with some features of early megakaryocytic and

1087 erythroid cells, HUDEP = immortalized human umbilical cord blood-derived erythroid progenitor

1088 cell lines expressing fetal globin genes (HUDEP1) or adult globin genes (HUDEP2), CD34\_E =

1089 human erythroid cells generated by differentiation from CD34+ blood cells, ERY = erythroblast,

1090 RBC = mature red blood cell, MK = megakaryocyte, GMP = granulocyte monocyte progenitor

1091 cell, EOS = eosinophil, MON = monocyte, MONp = primary monocyte, MONc = classical

1092 monocyte, NEU = neutrophil, CLP = common lymphoid progenitor cell, B = B cell, NK = natural

1093 killer cell, TCD4 = CD4+ T cell, TCD8 = CD8+ T cell, LSK = Lin-Sca1+Kit+ cells from mouse

1094 bone marrow containing hematopoietic stem and progenitor cells, HPC7 = immortalized mouse

1095 cell line capable of differentiation in vitro into more mature myeloid cells, G1E = immortalized

1096 mouse cell line blocked in erythroid maturation by a knockout of the *Gata1* gene and its subline

1097 ER4 that will further differentiate after restoration of *Gata1* function in an estrogen inducible  
1098 manner (Weiss et al. 1997), MEL = murine erythroleukemia cell line that can undergo further  
1099 maturation upon induction (designated iMEL), CFUE = colony forming unit erythroid, FL =  
1100 designates ERY derived from fetal liver, BM = designates ERY derived from adult bone marrow,  
1101 CFUMK = colony forming unit megakaryocyte, iMK = immature megakaryocyte, MK\_fl =  
1102 megakaryocyte derived from fetal liver.

1103

1104 **Figure 2. Genome segmentation and annotation jointly between human and mouse using**

1105 **IDEAS. (A)** Workflow for joint modeling. (1) Initial epigenetic states from 100 randomly selected  
1106 regions separately in human and mouse hematopoietic cell types were identified in IDEAS runs.  
1107 (2) States that were reproducible and shared in both species were retained. (3a and 3b) The  
1108 profile of epigenetic feature contribution to each of the reproducible states was sequentially  
1109 refined by applying IDEAS across the full genomes of human and of mouse, updating the state  
1110 model after each IDEAS run. (4) Two heterogeneous states were removed to generate the final  
1111 joint epigenetic states in the two species. **(B)** The 25 joint epigenetic states for human and  
1112 mouse hematopoietic cell types. The average signal of the epigenetic features for each state  
1113 are shown in the heatmap. The corresponding state colors, the state labels based on the  
1114 function, and the average proportions of the genome covered by each state across cell types  
1115 are listed on the right-side of the heatmap. **(C)** Annotation of epigenetic states in a large  
1116 genomic interval containing *SLC4A1* and surrounding genes across human blood cell types.  
1117 The genomic interval is 210kb, GRCh38 Chr17:44,192,001-44,402,000, with gene annotations  
1118 from GENCODE V38. Binding patterns for selected transcription factors are from the VISION  
1119 project ChIP-seq tracks (CTCF and GATA1 in adult erythroblasts, signal tracks from MACS,  
1120 track heights 100 and 80, respectively) or from the ENCODE data portal (EP300 in K562 cells,  
1121 experiment ENCSR000EGE, signal track is fold change over background, track height is 50).  
1122 The epigenetic state assigned to each genomic bin in the different cell types is designated by

1123 the color coding shown in panel (B). The replicates in each cell type examined in Blueprint are  
1124 labeled by the id for the donor of biosamples. Genes and regulatory regions active primarily in  
1125 erythroid (E), granulocytes (G), and megakaryocytes (MK) are marked by gray rectangles. **(D)**  
1126 Annotation of epigenetic states in a large genomic interval containing *Slc4a1* and surrounding  
1127 genes across mouse blood cell types. The genomic interval is 198kb, mm10  
1128 Chr11:102,290,001-102,488,000, with gene annotations from GENCODE VM23. Binding  
1129 patterns for selected transcription factors are from the VISION project CHIP-seq tracks (CTCF in  
1130 adult erythroblasts, GATA1 and EP300 from the highly erythroid fetal liver, signal tracks from  
1131 MACS, track heights 200, 200, and 150, respectively; the EP300 track was made by re-mapping  
1132 reads from ENCODE experiment ENCSR982LJQ). The tracks of epigenetic states and  
1133 highlighted regions are indicated as in panel (C).

1134

1135 **Figure. 3. Overlaps of VISION cCREs with other catalogs and enrichment for variants**  
1136 **associated with blood cell traits. (A)** Venn diagram showing intersections of human VISION  
1137 cCREs with a combined superset of elements associated with nuclear structure (CTCF OSs,  
1138 loop anchors, and TAD boundaries) and with a combined superset of DNA intervals associated  
1139 with *cis*-regulatory elements (CREs), including TSSs, CpG islands, peaks from a massively  
1140 parallel promoter and enhancer assay, and enhancers predicted from enhancer RNAs, peaks of  
1141 binding by EP300, and histone modifications in erythroblasts (see Supplemental Material,  
1142 Supplemental Fig. S9, and Supplemental Table S5). **(B)** The proportions of cCREs and  
1143 randomly selected, matched sets of intervals in the overlap categories are compared in the bar  
1144 graph. For the random sets, the bar shows the mean, and the dots show the values for each of  
1145 ten random sets. **(C)** The UpSet plot provides a higher resolution view of intersections of  
1146 VISION cCREs with the four groups of CRE-related elements, specifically enhancer-related  
1147 (Enh), transcription start sites (TSS), Survey of Regulatory Elements (SuRE), and CpG islands  
1148 (CpG). The enrichment for the cCRE overlaps compared to those in randomly selected,

1149 matched sets of intervals are shown in the boxplots below each overlap subset, with dots for the  
1150 enrichment relative to individual random sets. **(D)** Overlaps and enrichments of VISION cCREs  
1151 for three sets of structure-related elements, specifically CTCF OSs (CT), loop anchors (LA), and  
1152 TAD boundary elements. **(E)** Overlaps of VISION cCREs with two sets of experimentally  
1153 determined blood cell cCREs. **(F)** Enrichment of SNPs associated with blood cell traits from UK  
1154 Biobank in VISION cCREs. Results of the sLDSC analysis of all cCREs are plotted with  
1155 enrichment of the cCRE annotation in heritability of each trait on the x-axis, and the significance  
1156 of the enrichment on the y-axis. The analysis covers 292 unique traits with GWAS results from  
1157 both males and females and 3 traits with results only from males. The vertical dotted line  
1158 indicates an enrichment of 1, and the horizontal dotted line delineates the 5% FDR significance  
1159 threshold. Points and labels in red represent traits for which there was significant enrichment of  
1160 SNPs associated with the VISION cCREs. Traits with a negative enrichment were assigned an  
1161 arbitrary enrichment of 0.1 for plotting and appear as the column of points at the bottom left of  
1162 the plot. The shape of the point indicates the sex in which the GWAS analysis was performed  
1163 for each trait.

1164

1165 **Figure 4. Beta coefficients of states, esRP scores of cCREs, joint human-mouse**  
1166 **metaclusters of cCREs based on esRP scores, and enrichment for TFBS motifs. (A)** Beta  
1167 coefficients and the difference of beta coefficients of the 25 epigenetic states. The vertical  
1168 columns on the right show the beta coefficients along with the ID, color, and labels for the 25  
1169 joint epigenetic states. The triangular heatmap shows the difference of the beta coefficients  
1170 between two states in the right columns. Each value in the triangle heatmap shows the  
1171 difference in beta coefficients between the state on top and the state below based on the order  
1172 of states in the right columns. **(B)** An example of calculating esRP score for a cCRE in a cell  
1173 type based on the beta coefficients of states. For a cCRE covering more than one 200bp bin,  
1174 the esRP equals the weighted sum of beta coefficients of states that covers the cCRE, where

1175 the weights are the region covered by different states. **(C)** The average esRP score of all  
1176 cCREs in JmCs across blood cell types shared by human and mouse. The right column shows  
1177 the number of human cCREs in each JmC. **(D)** The average enrichment of JmCs in 15  
1178 homologous gene clusters. The genes are clustered based on the JmCs' enrichments by *k*-  
1179 means. **(E)** Motifs enriched in joint metaclusters. The top heatmap shows the enrichment of  
1180 motifs in the cCREs in each JmC in human (H) and mouse (M) as a Z-score. The logo for each  
1181 motif is given to the right of the heat map, labeled by the family of transcription factors that  
1182 recognize that motif. The heatmap below is aligned with the motif enrichment heatmap, showing  
1183 the mean esRP score for the cCREs in each JmC for all the common cell types examined  
1184 between human and mouse. A summary description of the cell types in which the cCREs in  
1185 each JmC are more active is given at the bottom.

1186

1187 **Figure. 5. Evolutionary and epigenetic comparisons of cCREs.** **(A)** Workflow to partition  
1188 blood cell cCREs in human and mouse into three evolutionary categories. N=nonconserved,  
1189 S=conserved in sequence but not inferred function, SF=conserved in both sequence and  
1190 inferred function as a cCRE, y=yes, n=no. **(B)** Enrichment of SF-conserved human cCREs for  
1191 TSSs. The number of elements in seven sets of function-related DNA intervals that overlap with  
1192 the 32,422 SF human cCREs was determined, along with the number that overlap with three  
1193 subsets (32,422 each) randomly selected from the full set of 200,342 human cCREs. The ratio  
1194 of the number of function-related elements overlapping SF-cCREs to the number overlapping a  
1195 randomly chosen subset of all cCREs gave the estimate of enrichment plotted in the graph. The  
1196 mean for the three determinations of enrichment is indicated by the horizontal line for each set.  
1197 Results are also shown for a similar analysis for the S and N cCREs. **(C)** Distribution of phyloP  
1198 scores for three evolutionary categories of cCREs in human and mouse. The maximum phyloP  
1199 score for each genomic interval was used to represent the score for each cCRE, using genome  
1200 sequence alignments of 100 species with human as the reference (phyloP100) and alignments

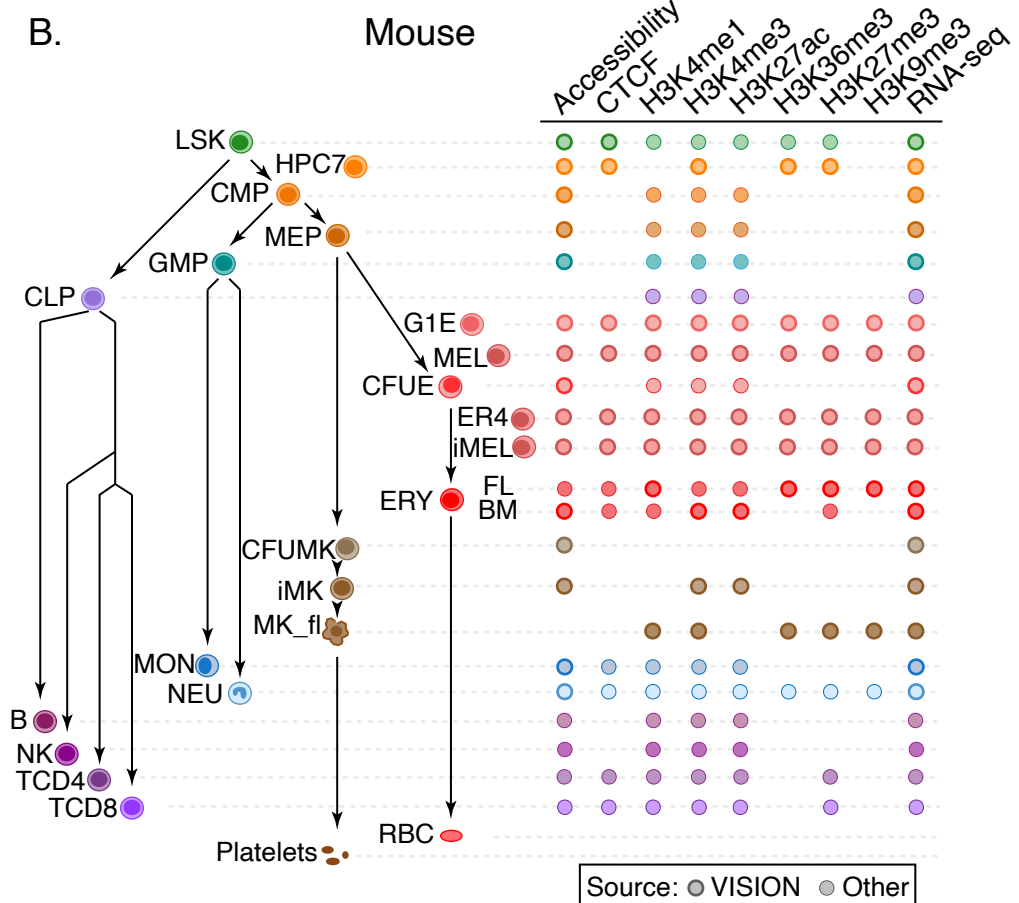
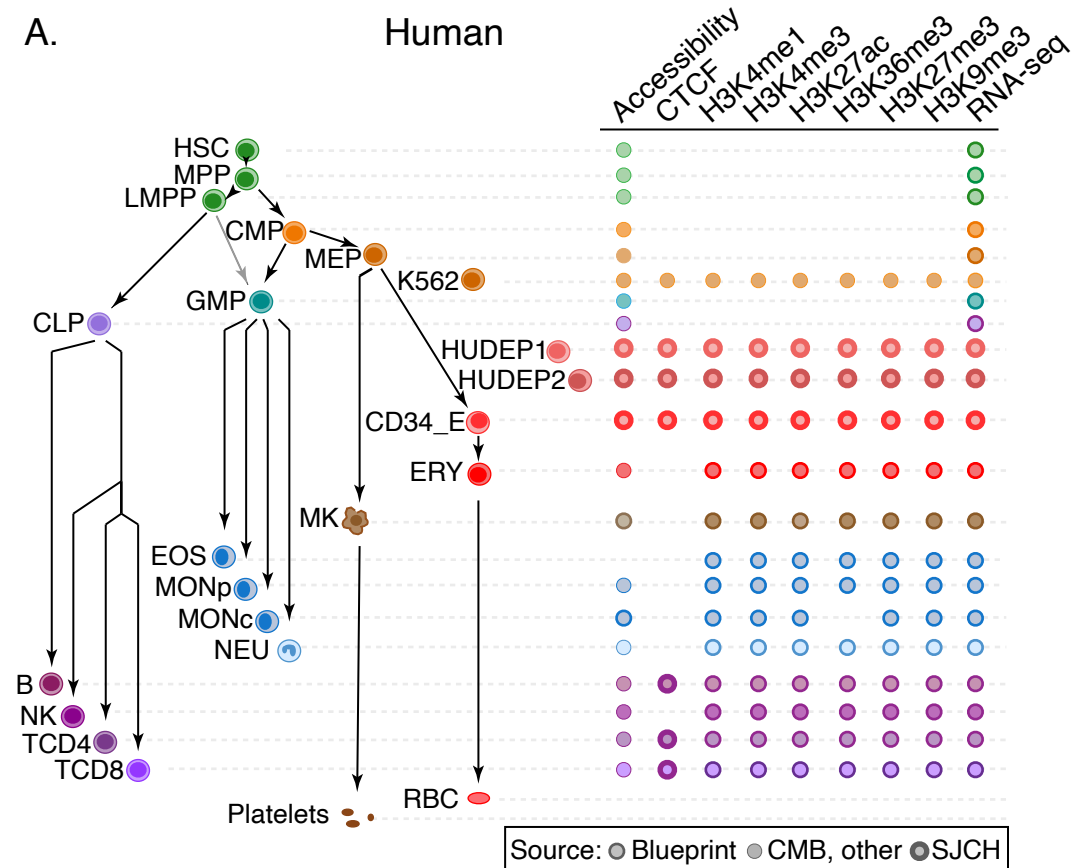
1201 of 60 species with mouse as the reference (phyloP60). The distribution of phyloP scores for  
1202 each group are displayed as a violin plot. All ten random sets had distributions similar to the one  
1203 shown. The asterisk (\*) over brackets indicates comparison for which the P values for Welch's *t*-  
1204 test is less than  $2.2 \times 10^{-16}$ . **(D)** Proportion of human genomic elements active in a massively  
1205 parallel reporter assay (MPRA) that align with mouse or are in a state reflecting dynamic  
1206 chromatin. A set of 57,061 genomic elements found to be active in a lentivirus MPRA that tested  
1207 a close to comprehensive set of predicted regulatory elements in K562 cells (Agarwal et al.  
1208 2023) were assessed for their ability to align with the mouse genome (blue bar) or whether the  
1209 IDEAS epigenetic state assigned in K562 cells was not quiescent or was in a set of states  
1210 associated with gene activation (magenta bars). The results are plotted as percentages of the  
1211 total number of MPRA-active elements.

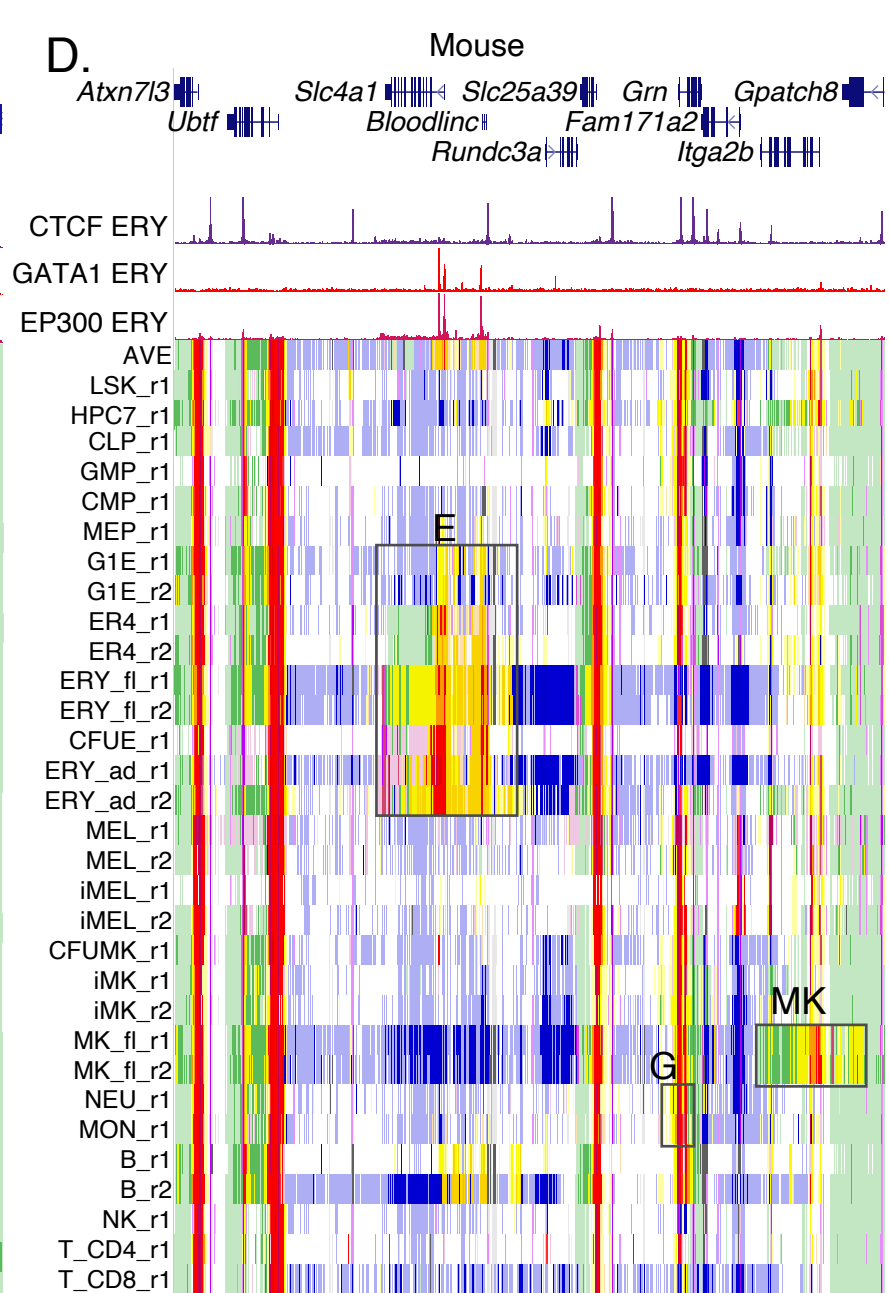
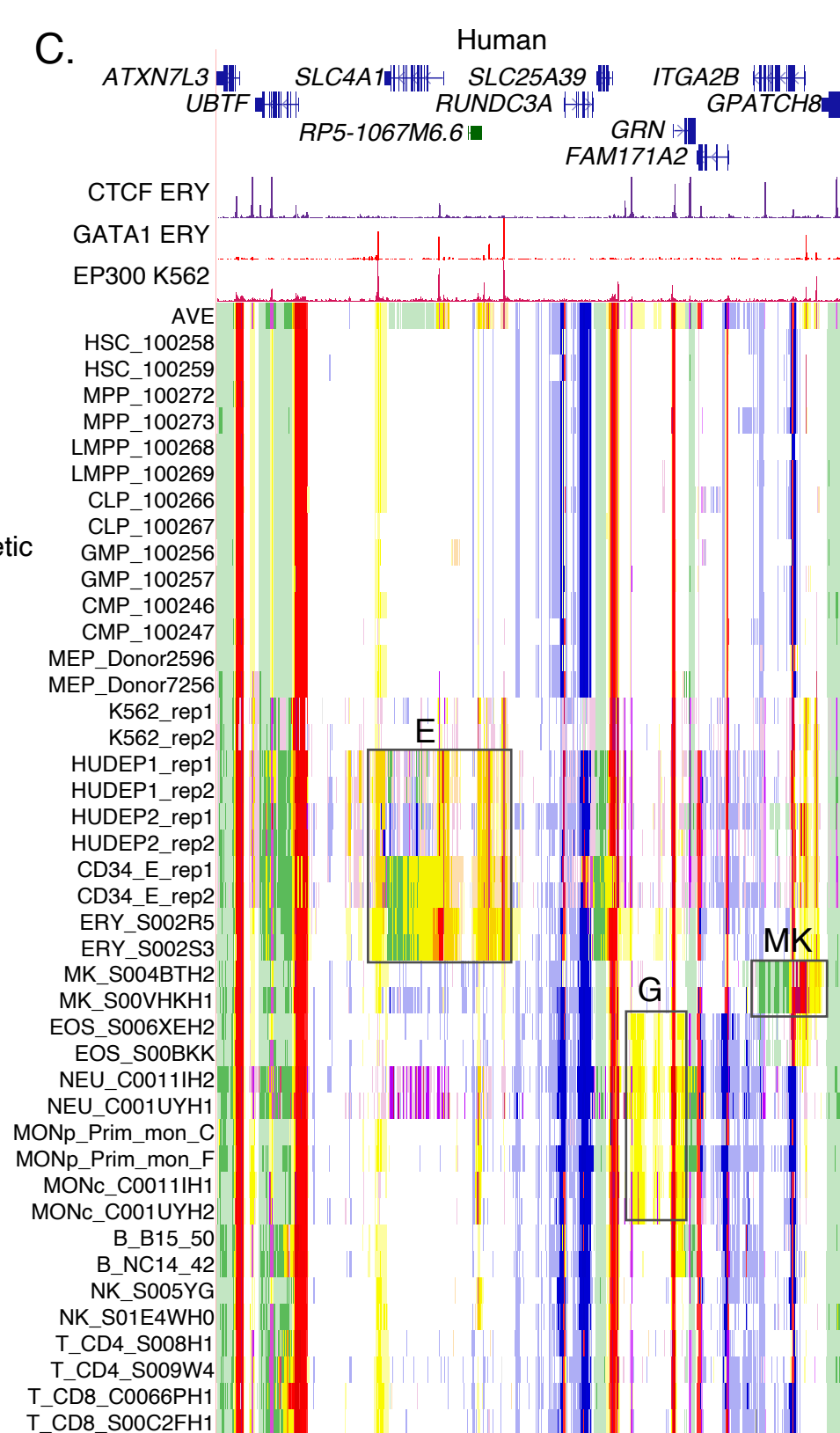
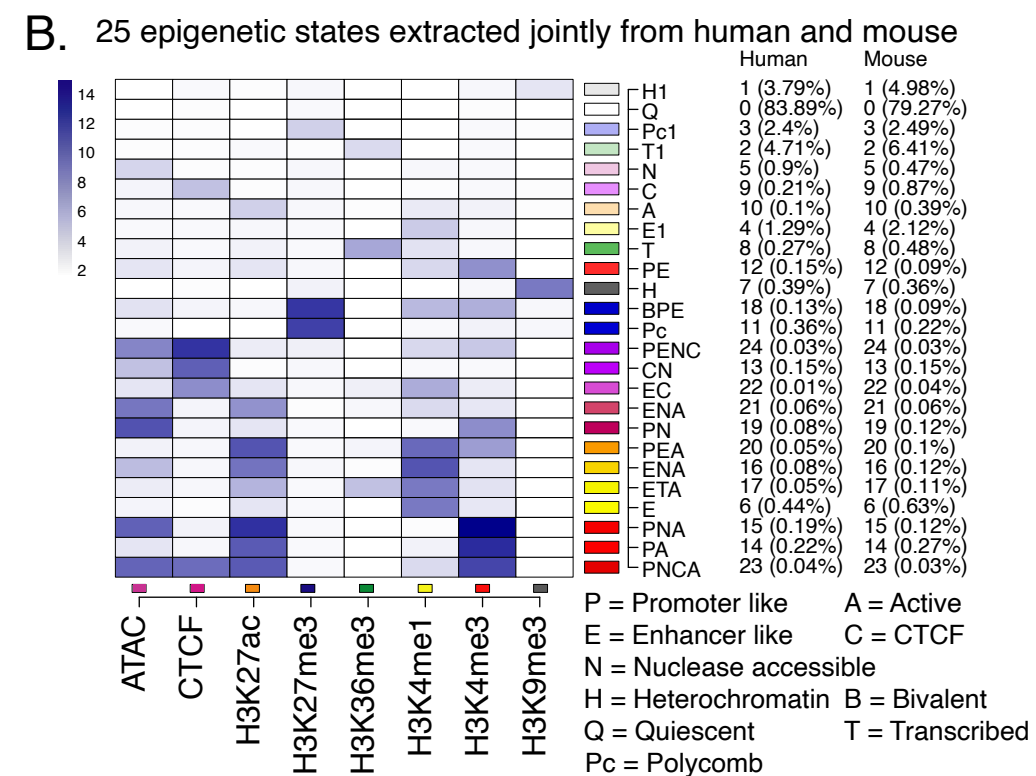
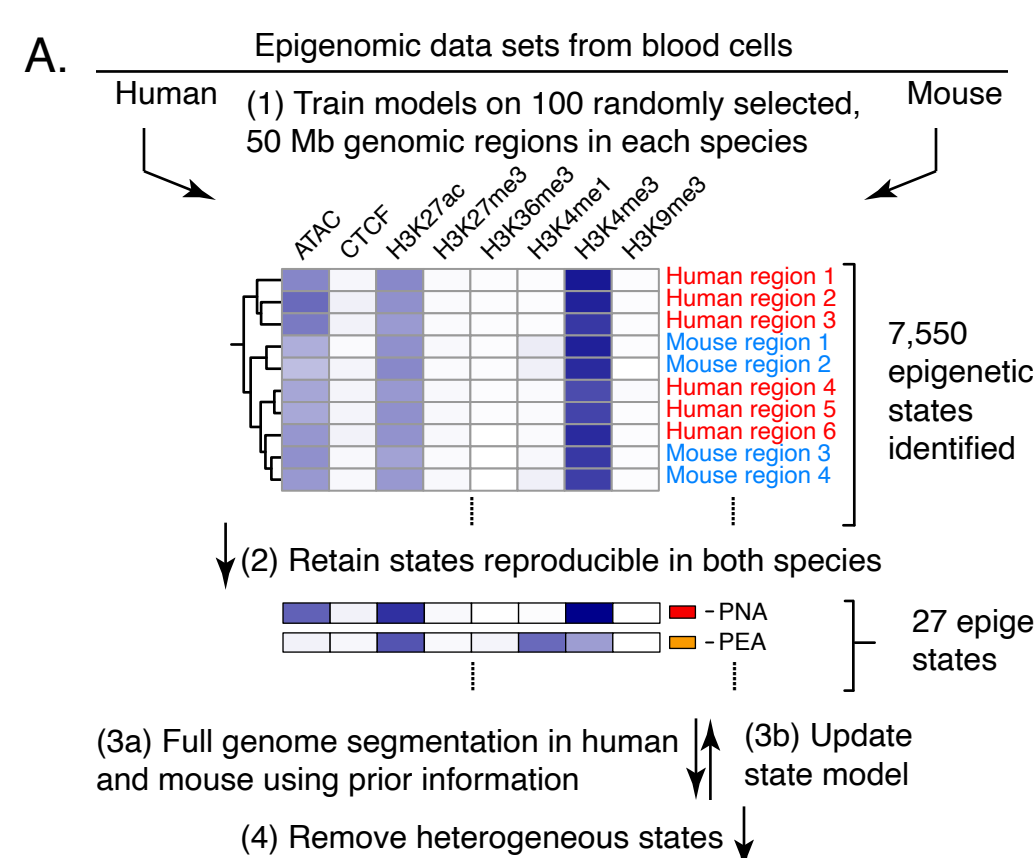
1212

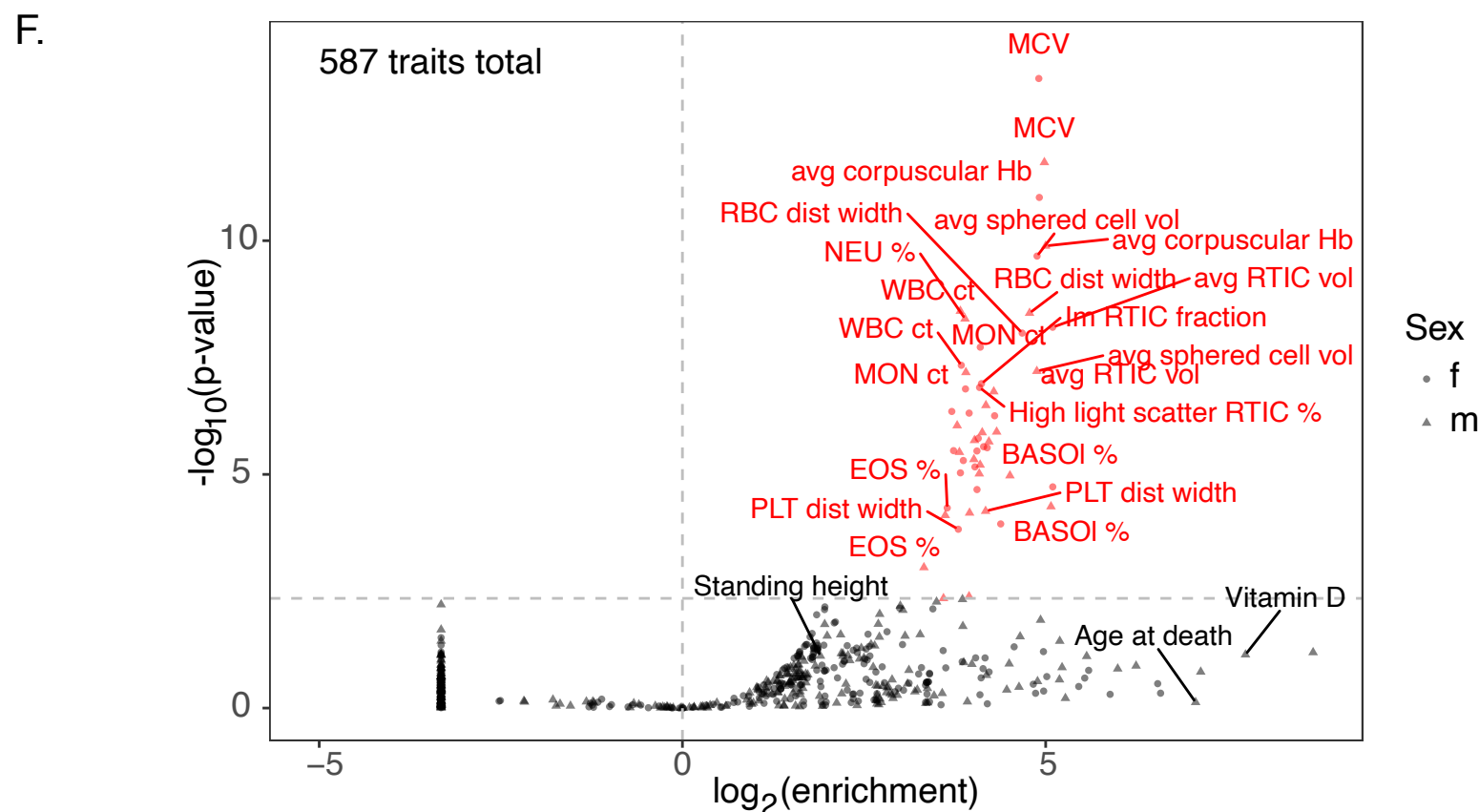
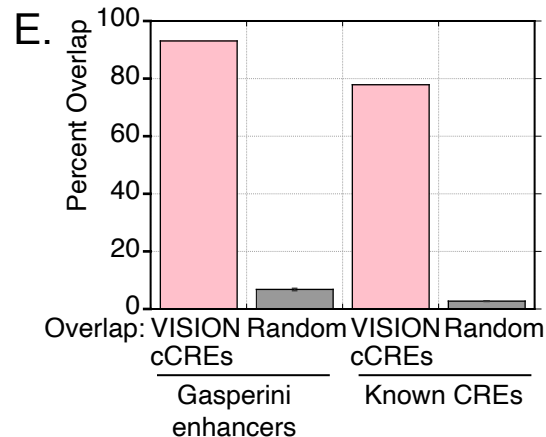
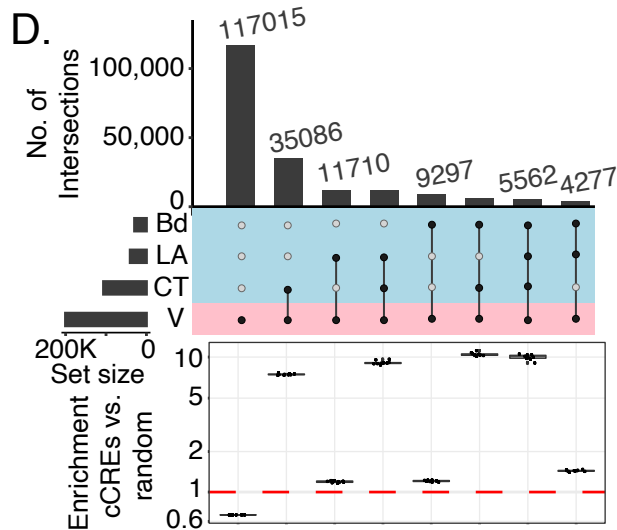
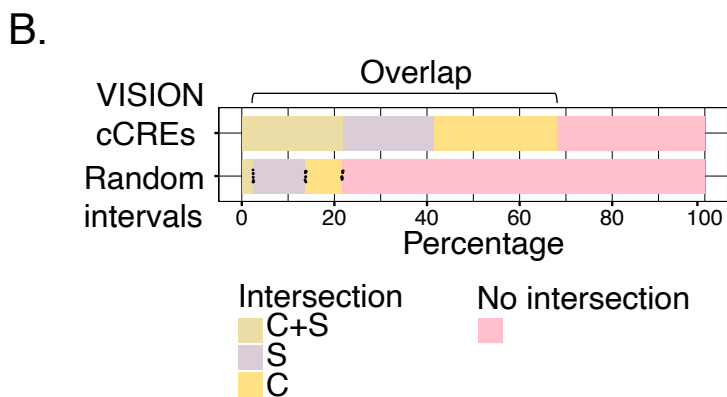
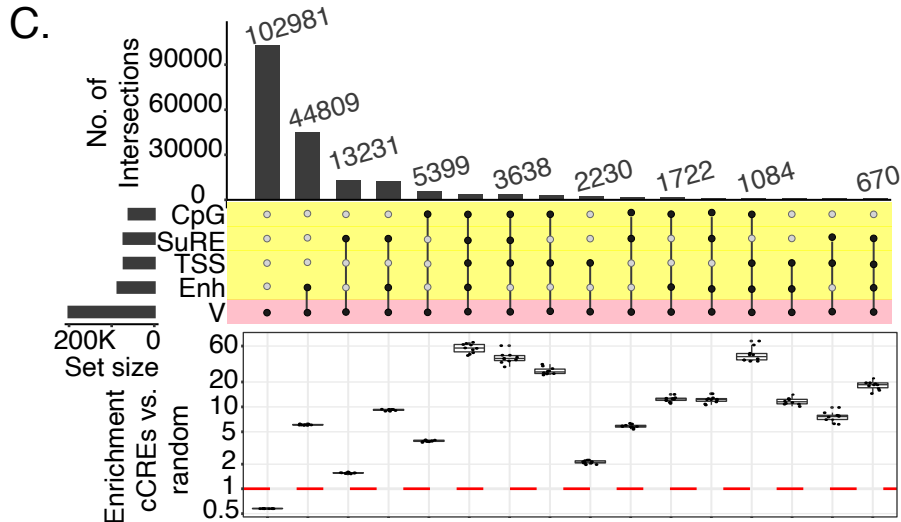
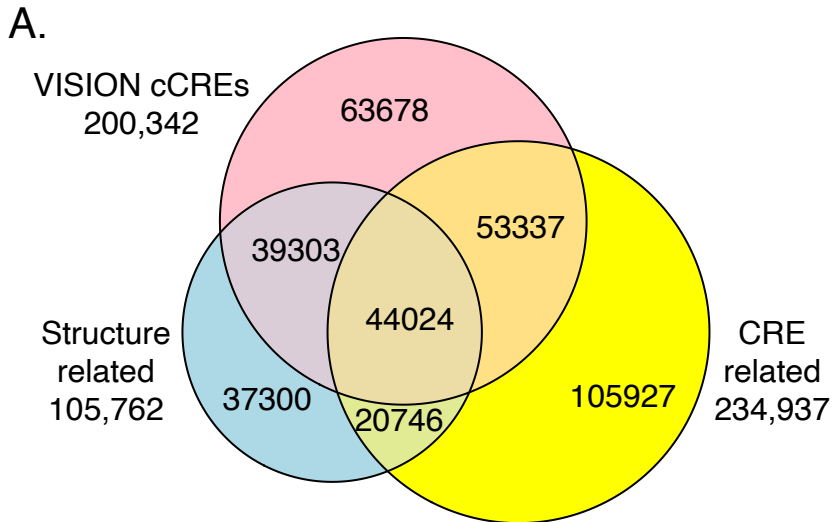
1213 **Figure. 6. Epigenetic comparisons of regulatory landscapes and cCREs. (A and B)** DNA  
1214 sequence alignments and correlations of epigenetic states in human *GATA1* and mouse *Gata1*  
1215 genes and flanking genes. **(A)** Dot-plot view of chained blastZ alignments by PipMaker  
1216 (Schwartz et al. 2000) between genomic intervals encompassing and surrounding the human  
1217 *GATA1* (GRCh38 ChrX:48,760,001-48,836,000; 76kb) and mouse *Gata1* (mm10  
1218 ChrX:7,919,401-8,020,800; 101.4kb, reverse complement of reference genome) genes. The  
1219 axes are annotated with gene locations (GENCODE), predicted *cis*-regulatory elements  
1220 (cCREs), and binding patterns for GATA1 and EP300 in erythroid cells. **(B)** Matrix of Pearson's  
1221 correlation values between epigenetic states (quantitative contributions of each epigenetic  
1222 feature to the assigned state) across 15 cell types analogous for human and mouse. The  
1223 correlation is shown for each 200bp bin in one species with all the bins in the other species,  
1224 using a red-blue heat map to indicate the value of the correlation. Axes are annotated with  
1225 genes and cCREs in each species. **(C)** Decomposition of the correlation matrix (panel **B**) into  
1226 six component parts or factors using nonnegative matrix factorization. **(D-G)** Correlation

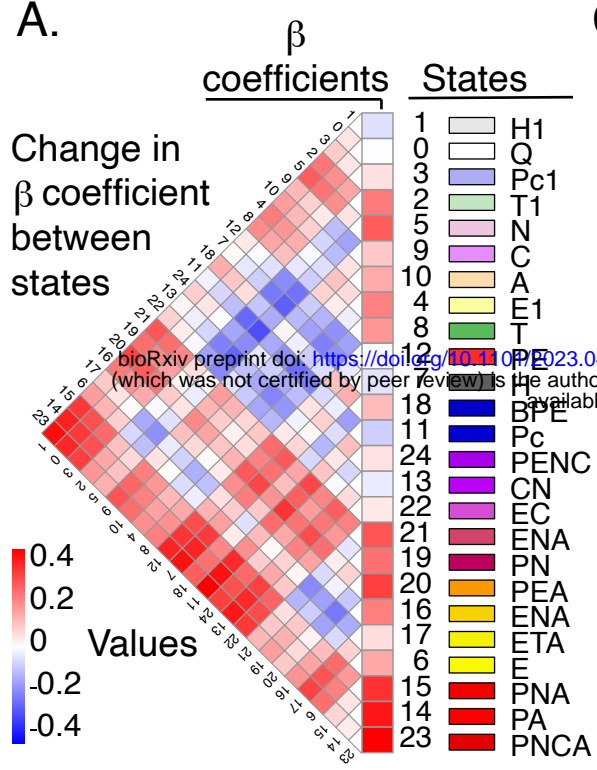
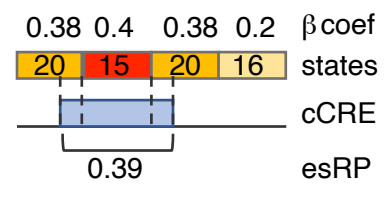
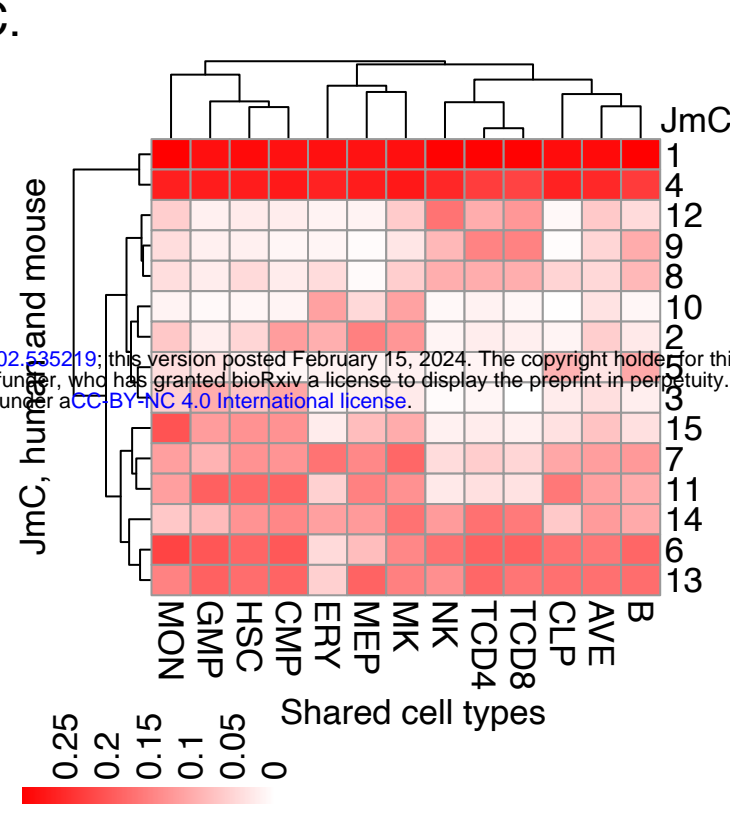


1227 matrices for genomic intervals encompassing *GATA1/Gata1* and flanking genes, reconstructed  
1228 using values from NMF factors. **(D and E)** Correlation matrices using values of NMF factor 3  
1229 between human and mouse (panel **D**) or within human and within mouse (panel **E**). The red  
1230 rectangles highlight the positive regulatory patterns in the *GATA1/Gata1* genes (labeled Px),  
1231 which exhibit conservation of both DNA sequence and epigenetic state pattern. The orange  
1232 rectangles denote the distal positive regulatory region present only in mouse (labeled D), which  
1233 shows conservation of epigenetic state pattern without corresponding sequence conservation.  
1234 Beneath the correlation matrices in panel **E** are maps of IDEAS epigenetic states across 15 cell  
1235 types, followed by a graph of the score and peak calls for NMF factor 3 and annotation of  
1236 cCREs (thin black rectangles) and genes. **(F and G)** Correlation matrices using values of NMF  
1237 factor 6 between human and mouse (panel **F**) or within human and within mouse (panel **G**). The  
1238 green rectangles highlight the correlation of epigenetic state patterns within the same gene,  
1239 both across the two species and within each species individually, while the black rectangles  
1240 highlight the high correlation observed between the two genes *GATA1* and *HDAC6*.



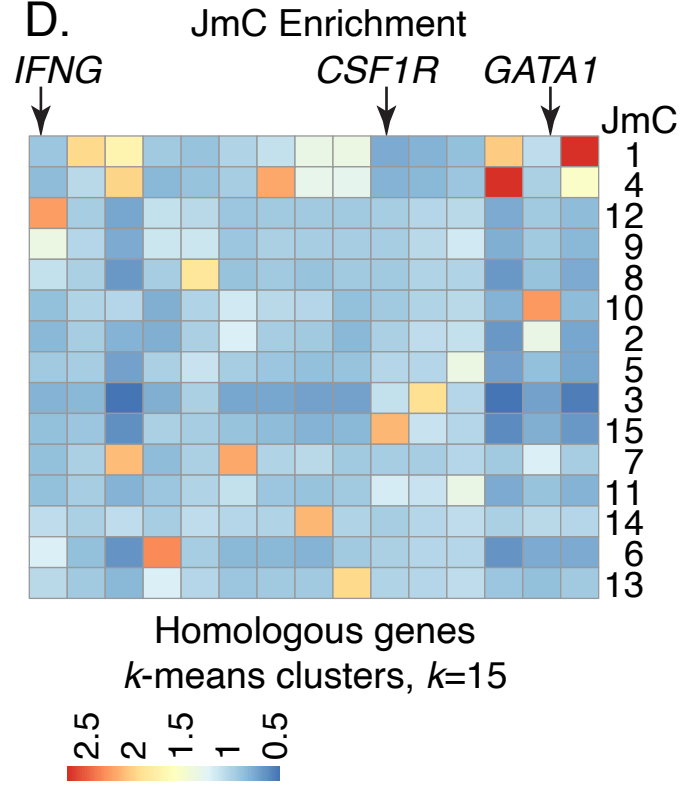




**A.****B.****C.**

**Nmbr of cCREs**

27814
17138
8272
11352
18078
15416
15461
15402
18416
17031
9249
5503
8015
9612
6583

**D.****E.**