Early epigenetic influence on EGR1, KLF2 and KLF4 transcriptional network associates with alteration of HSCs quiescence

Running title: Early epigenetic influence on HSPCs

Alexandre Pelletier[1,3,*], Arnaud Carrier[1,3,*], Yong Mei Zhao[2], Mickaël Canouil[1,3], Mehdi Derhourhi[1], Emmanuelle Durand[1], Lionel Berberian-Ferrato[1], John Greally[4], Francine Hughes[5], Philippe Froguel[1,3,6], Amélie Bonnefond[1,3,6], Fabien Delahaye[1,3].

1) Inserm U1283, CNRS UMR 8199, European Genomic Institute for Diabetes, Institut Pasteur de Lille, Lille, France.
2) Department of Obstetrics & Gynecology and Women's Health, Albert Einstein College of Medicine, 1300 Morris Park Ave, Bronx, NY, 10461, USA.
3) Lille University Hospital, University of Lille, Lille, France.
4) Department of Genetics, Albert Einstein College of Medicine, 1301 Morris Park Avenue, Price Building, Room 322, Bronx, New York 10461, USA.
5) Obstetrics & Gynecology and Women's Health, Division of Maternal-Fetal Medicine, Montefiore Medical Center/Albert Einstein College of Medicine, Bronx, NY 10461, USA.
6) Department of Metabolism, Digestion and Reproduction, Imperial College London, London, U.K.

* These authors contributed equally to this work

Corresponding authors:

Fabien Delahaye; Institut Pasteur de Lille-UMR1283, 1 rue du Prof Calmette, 59000 Lille, France ; fabien.delahaye@pasteur-lille.fr; (+33) 320877258

Philippe Froguel; EGID-UMR1283 1 place de Verdun 59045 Lille, France ; philippe.froguel@cnrs.fr; (+33) 374008101

Summary

Excessive fetal growth is associated with DNA methylation alterations in human hematopoietic stem and progenitor cells (HSPC), but their functional impact remains elusive. We implemented an integrative analysis combining single-cell epigenomics, single-cell transcriptomics, and *in vitro* analyses to functionally link DNA methylation changes to putative alterations of HSPC functions. We showed in hematopoietic stem cells (HSC) from large for gestational age neonates that both DNA hypermethylation and chromatin rearrangement target a specific network of transcription factors known to sustain stem cell quiescence. In parallel, we found a decrease expression of key genes regulating HSC differentiation including *EGR1*, *KLF2, SOCS3,* and *JUNB*. Our functional analyses showed that this epigenetic programming was associated with a decreased ability for HSCs to stay quiescent. Taken together, our multimodal approach using single-cell (epi)genomics showed that human fetal overgrowth affects hematopoietic stem cells quiescence maintenance via epigenetic programming.

Key Words: Epigenomic / Single-cell / Stem-cells / Fetal Programming / Hematopoiesis

Introduction

The hematopoietic stem cells (HSC) are involved in essential processes such as inflammation, cardiovascular repair, and immunity throughout the entire life.[1, 2] Thus, alterations in HSC ability to self-renew and to adequately produce differentiated progeny have been suggested to contribute to the onset and progression of age-related diseases such as cancer and cardiovascular diseases.[3, 4] Systemic alterations or the action of various stressors like aging[5, 6] can result in alteration of HSC destiny, and ultimately of hematopoietic functions. The early mechanisms that control their long-term functions in human are not well understood, in part due to the diversity of phenotypes and behaviors of HSCs.[7]

In mice, maternal high-fat diet during gestation limits fetal hematopoietic stem and progenitor cells (HSPC) expansion and repopulation ability while inducing myeloid-biased differentiation.[18] In humans, a limited number of studies have been conducted. Fetal growth was shown to alter the number of circulating CD34+ HSCs[16, 17]. We previously described a global increase of DNA methylation in cord blood derived CD34+ HSPCs from large for gestational age (LGA) infants compared to normal size neonates.[19] Still, the functional impacts of these early epigenetic alterations remain to be elucidated. Such an effort is essential to determine how these epigenetic modifications could mediate the association between early-life exposures and induce persistent life-long functional changes within the hematopoietic system.

We conducted a multimodal analysis combining single-cell epigenomics, single-cell transcriptomics, and *in vitro* analyses to link the DNA methylation alterations observed in LGA neonates with functional alterations in human cord blood derived HSPCs. We developed novel analytical approaches to improve integration of the single-cell

71 epigenomic and transcriptomic data. We found that the DNA methylation alterations
72 observed in LGA neonates are associated with an alteration of the expression of key
73 genes and transcription factors involved in the immediate early gene signalling
74 especially in hematopoietic stem cells as well as in the balance between quiescence
75 and proliferation in HSPCs.

76

77 Results

78 Optimized methylation gene set analysis reveals association between LGA
79 DNA hypermethylation and stem cell differentiation pathways

80 To confirm the LGA-associated DNA hypermethylation we previously observed we
81 significantly increased the power of our analysis. We expanded our original study
82 trough additional patient inclusions, thereby doubling the size of our cohort.[19] Using,
83 HELP-tagging assay (HpaII tiny fragment Enrichment by Ligation-mediated PCR), we
84 generated genome-wide DNA methylation data on 16 CTRL and 16 LGA cord-blood
85 derived human CD34+ HSPC samples. We independently retrieved in this new
86 dataset the global DNA hypermethylation initially found in LGA compared to controls[19]
87 (**Figure 1A**). Then, to increase our detection power, we pooled both dataset and
88 detected a total of 4815 differentially methylated CpGs (DMC) with 4787 CpGs
89 hypermethylated and 28 CpGs hypomethylated in LGA (n=36) compared to CTRL
90 (n=34, p-value<0.001 and |methylation difference|>25%; **Figure 1A, Supplemental**
91 **Table 1**). This new set of DMCs was then used throughout following analysis.

92 As the functional interpretation is performed at the gene level, each CpG (or DMC)
93 must be linked to a specific gene. Thus, our ability to adequately infer the regulatory

effect of a CpG and its target gene will affect our ability to identify relevant pathways. Standard analytical approaches usually rely on distance between CpG and transcription start site (TSS) of the targeted gene and often only consider the top candidate DMC per gene, not taking into account the cell specific genomic context. Therefore, we refined CpG-gene association to optimally assess the influence of DNA methylation changes on gene expression and enhance functional interpretation. We built a novel gene-methylation score considering 1) the distance between TSS and CpG; 2) the CpG overlap with expression quantitative trait loci (eQTL) annotation, as eQTL information allows us to identify tissue specific genomic region link to gene expression changes; and 3) the regulatory annotation (e.g. Promoter, Enhancer) based on cell-specific histone marks[25] and on the Ensembl Regulatory database, as we know that the relationship between change in DNA methylation and change in gene expression will depend on a cell specific genomic context (**Figure 1B**). We established 756470 CpG-gene associations including 34% of them found thanks to eQTL annotation. We then summarized the CpG information at gene level, generating a gene-methylation score for each gene (n=24857, **Supplemental Table 2**). We first confirmed that the gene-methylation score properly recapitulates the influence of key parameters in DMC analysis such as significance and effect size of the methylation change, number of DMCs per gene, distance from TSS, as well as promoter and enhancer localization (**Figure 1C**). We also confirmed that while preserving key information from standard methylation metrics, the gene-methylation score presented better association with DEGs than significance or methylation change alone. Thus, the gene-methylation score appears as a better predictor of the methylation influence on gene expression (**Figure 1C**). We then used our gene-methylation score to perform

118 pathway enrichment analysis and data integration especially considering integration
119 with gene expression data.

120 Using the gene ontology (GO) reference database, we performed methylation gene-
121 set enrichment analysis (GSEA) based on the gene-methylation score. We found that
122 change in DNA methylation in LGA HSPC samples targeted genes involved in
123 signaling regulating fetal development as well as in key stem cell pathways such as
124 Wnt signaling, cell fate specification, and cell fate commitment pathways (adjusted *p*-
125 value<0.01, **Figure 1D**) confirming previous findings[19].

126 **Single-cell transcriptomic analysis confirms alteration of hyper-methylated**
127 **genes in pathways regulating stem cell differentiation among LGA HSCs**

128 To identify genes altered in LGA and get further biological insight into the functional
129 consequences of the DNA methylation modifications observed in LGA, we performed
130 single-cell transcriptomic analysis comparing CTRL and LGA HSPCs.

131 To enable lineage specific transcriptomic analysis, we created a hematopoietic
132 reference map (*i.e.* hematomap) by integrating data generated from cord blood
133 derived CD34+ HSPC cells (*n*=18520) from 7 control neonates (**Figure 2A**). Based on
134 cluster specific gene expression, we identified 18 distinct clusters representative of
135 major lineages (Long-Term HSC, HSC, Multi-Potent progenitor, Lymphoid, Myeloid,
136 and Erythroid) of the hematopoietic compartment (**Figure 2B**, **Supplemental Figure**
137 **2**). Each cluster was annotated using cell type specific markers. Markers were then
138 ranked based on their expression fold change and the specificity for the cluster. Top
139 cluster-specific markers were compared with published cell type-specific genes[27-30]
140 (**Supplemental Table 3**). Candidate cell subpopulations were distributed as follows:
141 1% LT-HSC (*ID1*); 24% HSC (*AVP*); 45% MPP/LMPP (*CDK6*); Lymphoid (*CD99,*

142 *LTB*); 1% B cell (*IGHM*); 1% T cell (*CD7*); 14% Erythro-Mas (*GATA1*); <1%Mk/Er
143 (*PLEK, HBD*); 8% Myeloid (*MPO*); <1% DC (*CST3*, *CD83*).

144 To identify differentially expressed genes (DEG) between CTRL and LGA samples,
145 we implemented the Hash Tag Oligonucleotide (HTO) multiplexing strategy[31] allowing
146 simultaneous processing of CTRL and LGA samples. Multiplexing is a means to limit
147 the influence of technique-drive batch effects at every stage of the analysis to improve
148 biological relevance of the finding. We generated multiplexed single-cell transcriptomic
149 data from 6 LGA (*n*=6861 cells) and 7 CTRL (*n*=5823 cells) samples. We observed in
150 LGA a shift toward downregulated genes (**Supplemental Figure 3**) especially in the
151 HSC subpopulation (*n*=285 downregulated genes over 373 DEGs, adjusted *p*-
152 value<0.05 and $\log_2$FC<(-0.5), **Figure 2C; Supplemental Table 4**). Notably, the well-
153 known *EGR1*, *JUNB,* and *KLF2* genes were among the top affected genes. Using GO
154 enrichment analysis, we found that downregulated genes were enriched in growth
155 related pathway (e.g. regulation of growth) as well as in stress related biological
156 process (e.g. response to temperature stimulus, cellular response to chemical stress;
157 **Figure 2D,** adjusted p-value<0.05).

158 To assess if these HSC specific transcriptomic changes may be associated with
159 epigenetic changes, we integrated bulk DNA methylation with single-cell gene
160 expression data using the gene-methylation score. We found that DEGs, and
161 particularly the down-regulated genes, mostly showed high gene-methylation score
162 (**Figure 3A, Figure 3B**). We then assessed the association between changes in DNA
163 methylation and gene expression at pathway level. We looked for enrichment for
164 differentially methylated genes considering pathways that were identified based on
165 DEGs. We found a significant overlap between GO terms enriched in LGA HSC

166 downregulated genes and GO terms enriched in hypermethylated genes (10 out of 46;
167 *p*-value <0.05, hypergeometric test). Among the overlapping terms, we found the
168 module "regulation of growth" including notably *SOCS3*, *SIRT1* and *SESN2* genes that
169 are both downregulated and within the top 10% of hypermethylated genes (**Figure**
170 **3C**).

171 **DNA methylation changes occurs in HSCs and DEGs associated open**
172 **chromatin regions**

173 We profiled chromatin accessibility at single-cell level (i.e. single-cell ATAC-seq) to
174 optimize the resolution of our epigenomic analyses. We generated open chromatin
175 data across 8733 cells in HSPCs from 6 CTRL and 5 LGA neonates. We first
176 annotated subpopulations using label transfer approach between ATAC-seq data and
177 the lineage labels from the Hematomap (**Figure 4A, Supplemental Figure 4A**). To
178 validate the relevance of our lineage annotation, we performed TF motif enrichment
179 and observed that lineage specific peaks were effectively associated with well-known
180 lineage specific TF (**Supplemental Figure 4B**).

181 We then integrated our bulk DNA methylation data with our single-cell ATAC-seq data
182 to assess DMCs distribution within open chromatin regions (OCRs). Overall, 31% of
183 the 211479 peaks contain CpGs queried by our genome wide methylation assay. We
184 first observed a strong enrichment for DMCs in OCRs with 74% of them located in
185 OCRs compare to only 34% of overall queried CpGs (pvalue < 0.001, hypergeometric
186 test). Such enrichment supports the putative regulatory influence of our DMCs. By
187 performing lineage specific analysis, we observed a DMCs enrichment in HSC-specific
188 open chromatin region with a total of 11% of HSC specific peaks containing DMCs
189 (adjusted *p*-value<0.001, **Figure 4B**), while no enrichment was observed for the others

190 lineages. This result corroborates the HSC specific transcriptional impact of the DNA
191 methylation change observed in LGA. Furthermore, we observed that DEGs in LGA
192 HSC and especially down regulated genes were enriched for OCRs containing DMCs
193 (**Figure 4C**).

194 Not limiting our analysis to the regulatory role of DMCs within open chromatin regions,
195 we then assessed the interaction between change in DNA methylation, change in
196 expression, and change in chromatin accessibility among LGA samples. We identified
197 278 open chromatin regions that significantly differ between LGA and CTRL in HSCs,
198 with 215 showing decreased and 63 showing increased accessibility in LGA versus
199 CTRL (adjusted *p*-value<0.001 and $|\log_2 FC|>0.25$, **Supplemental Figure 4C**). Motif
200 enrichment analysis identified EGR1, KLF2 and KLF4 among the top motif enriched
201 in regions with decreased accessibility (**Figure 4D, Supplemental Figure 4D**).
202 Regions with decreased accessibility were also strongly enriched in peaks including
203 DMCs and peaks associated with DEGs (**Figure 4E**), with 3-fold and a 2.5-fold
204 enrichment respectively. Finally, these regions were strongly enriched for peak
205 containing both DMCs and associated to DEGs (23-fold enrichment) illustrating that
206 early epigenetic programming is actually not limited to changes in DNA methylation
207 but also involves chromatin rearrangement targeting altered genes.

208 **EGR1, KLF2 and KLF4 are key upstream regulators influenced by early**
209 **epigenetic programming in LGA**

210 To further characterize the molecular mechanisms affected in LGA HSCs and
211 identified master regulators, we leveraged the single-cell resolution of our approaches
212 to perform a co-regulatory network analysis. This approach allowed us to model the
213 influence of upstream transcription factors (TF) on expression changes of downstream

214 target genes. We performed co-expression analysis to identify co-regulated genes, *i.e.*
215 regulons, and performed analysis of cis-regulatory motif to assign TF to each regulon
216 (SCENIC). We identified a total of 250 regulons but only considered for further
217 analyses the 106 regulons identified based on high confidence cis-regulatory motif.
218 These regulons only rely on associations for which the presence of TF motif was
219 validated. We scored the activity of regulons in each cell and identified lineage specific
220 regulons associated with key hematopoietic TFs such as GATA2, GATA3, MEIS1,
221 TAL1, TCF3, EGR1, CEBPB, HOXB4, SPI1 and STAT1/3 further supporting our
222 subpopulation annotation and the SCENIC approach (**Supplemental Figure 5A,**
223 **Supplemental Table 5**)[39].

224 To identify TF associated with the changes in gene expression observed in LGA HSC,
225 we compared the regulon activity between CTRL and LGA. We found seven regulons
226 with a significant decrease in activity in the LGA HSC population (adjusted *p*-
227 value<0.001 and |activity score fold change|>10%, **Supplemental Table 6**). No
228 regulons were upregulated. These regulons were associated to ARID5A, EGR1,
229 KLF2, KLF4, KLF10, FOSB, and JUN (**Figure 5A**). Among them, ARID5A, EGR1,
230 KLF2, FOSB and JUN were part of the 10 top active regulons in HSCs (**Supplemental**
231 **Table 7**). Based on functional enrichment analysis using as reference GO:BP gene
232 sets, and HSC signatures of quiescence or proliferative state[40], we showed that these
233 regulons were enriched in genes regulating stress response, proliferation, and HSC
234 differentiation (**Figure 5B**).

235 In parallel and to further support the association between change in DNA methylation
236 and change in gene expression previously identified at gene level, we performed
237 GSEA analysis to identify regulon enriched for both differentially methylated and

238 differentially expressed genes. We found 9 regulons enriched in both hypermethylated
239 and downregulated genes (adjusted *p*-value<0.01 and NES < -1.6), including the
240 differentially active regulons ARID5A, EGR1, FOSB, JUN, KLF2, and KLF4 (**Figure**
241 **5C**). We also found 9 regulons enriched in hypermethylated and upregulated genes
242 (adjusted *p*-value<0.01 and NES >1.6) with key HSPC specific regulons such as SPI1
243 promoting myeloid differentiation;[41] and HOX family (HOXA9, HOXA10, HOXB4)
244 promoting HSPC expansion (**Figure 5C**).[42-44]

245 To confirm the putative influence of methylation change on TF activity, we performed
246 TF motif analysis considering the proximal regions surrounding each DMCs (±20bp).
247 We found significant enrichment for 23 TF motifs (adjusted *p*-value<0.05, **Figure 5D**).
248 Among them, we found EGR1 and several members of the Kruppel-like factors (KLF)
249 family: KLF14, KLF5, KLF1, KLF6. Furthermore, by taking advantage of our single-cell
250 ATAC-seq data, we looked at the enrichment for TF motif in open chromatin region of
251 HSC containing DMCs. We found a strong enrichment in EGR1, KLF2, and KLF4
252 motifs indicating that DNA methylation change occurred in active regions of the
253 EGR1/KLF2/KLF4 TF network (**Figure 5E**).

254 **Multimodal co-regulatory network recapitulating TF-gene interactions**
255 **influenced by early epigenetic programming in LGA**

256 Based on the integration of the DNA methylation, single-cell ATAC-seq, and single-
257 cell RNA-seq data, we built a network recapitulating interaction between mains TFs
258 and downstream target genes within the principal regulons altered in LGA neonates:
259 EGR1, KLF2, and KLF4 (**Figure 6**). EGR1, KLF2, and KLF4 relies on highly
260 interconnected genes (**Figure 6A**). For each target gene, we confirmed the presence
261 of unique or shared upstream TF binding motif within the open chromatin regions. We

262 observed a high concordance between the regulons and open chromatin motif
263 analysis: 96%, 91% and 95% of genes included in EGR1, KLF2, and KLF4 regulons
264 respectively were associated to at least one peak containing the corresponding TF
265 motif supporting the association between genes and TFs. We then looked for evidence
266 of epigenetic modifications that may alter TF-target interactions. We annotated genes
267 with associated open chromatin regions containing at least one DMC (middle area) or
268 identified as differentially accessible between CTRL and LGA (inside area) (**Figure**
269 **6B)**. Overall, 23% (n=27) of genes targeted by these TFs networks have epigenetic
270 alteration (DMCs or decrease accessibility) in open chromatin regions while 22%
271 (n=26 genes) appear downregulated in LGA. Finally, we highlight KLF2 as possible
272 master regulators influenced by early programming. Indeed, we identified KLF2 as
273 hypermethylated and down regulated gene that interacts directly with EGR1 and KLF4
274 suggesting downstream influence of KLF2 on these TFs. Conversely, KLF2 was not
275 identified as part of EGR1 and KLF4 regulons suggesting that KLF2 is not a target of
276 these TFs. This network also further validated *JUNB and SOCS3* being highly
277 epigenetically altered in cis-regulatory regions **(Figure 6C**)*,* as well as *ID1, CDKN1A,*
278 *IER2, IER3*, and *IER5* as key downstream altered targets of KLF2, EGR1, and/or
279 KLF4, again highlighting how early programming alters signaling involved in the
280 regulation of cell proliferation and differentiation.

281 ***In vitro* analysis confirms the alteration of HSPCs differentiation capacities in**
282 **LGA**

283 Our integrative analyses highlighted epigenetic and transcriptomic alterations
284 targeting signaling pathways involved in regulation of HSC differentiation and
285 proliferation. Thus, we decided to challenge HSPC differentiation and proliferation

286 potential *in vitro* using colony forming unit (CFU) assays. After 14 days of expansion,
287 colonies from 4 CTRL and 4 LGA samples were classified in 3 categories: those
288 derived from common myeloid progenitors (CFU_GEMM), erythroid progenitors (BFU-
289 E), and granulocyte-macrophage progenitors (CFU_GM) based on the morphology of
290 each colony. We observed a significant decrease in the number of common myeloid
291 progenitor colonies in LGA samples (p-value<0.05; **Figure 7A**) as well as striking
292 differences in shape and size of more differentiated colonies (**Figure 7B**).
293 CFU_GEMM colonies are the product of a non-committed hematopoietic progenitor
294 able to self-renew and differentiate in both erythroid and myeloid lineage. In our
295 samples, only HSC have these features, suggesting that the decreased CFU GEMM
296 proportion in LGA reflects either fewer HSC in starting cell subpopulation composition
297 or a decreased proliferation and differentiation capacity. To evaluate these two
298 possibilities, we monitored cell populations distribution across conditions at molecular
299 resolution using our single-cell expression dataset. We observed a decrease in HSC
300 cells (p-value=0.015) and a trend toward increased MPP cells (p-value=0.13, **Figure**
301 **7C**) in LGA compared to CTRL.

302 Another way to look at population shift is to use pseudotime, *i.e.* a measure that
303 reflects how far an individual cell is in a differentiation process. Indeed, cord-blood
304 derived CD34+ HSPCs represent a heterogeneous population of cells ranging from
305 progenitors to progressively restricted cells of the erythroid, myeloid, or lymphoid
306 lineages as confirmed by our single-cell transcriptomic analysis. To follow cell
307 distribution through these levels of differentiation and assess the influence of the LGA
308 environment we used pseudotime tool from Monocle.[32] Collecting the pseudotimes
309 across our different cell populations, we observed a positive correlation between
310 pseudotime and lineage differentiation as expected (r=0.99, Pearson correlation,

311 **Figure 7D**, **Supplemental Figure 6A**). We then compared the distribution of the
312 pseudotime between LGA and CTRL using the least differentiated cells as roots, i.e.
313 the long-term HSCs. At the population level, we observed an increase in pseudotime
314 in LGA (*p*-value<0.001, **Figure 7E**). Indeed, we observed a decrease in number of
315 cells presenting pseudotime associated with the HSC state in LGA samples (*p*-
316 value<0.05) and a shift toward cells presenting elevated pseudotime suggesting that
317 LGA HSCs exit quiescence and differentiate more quickly compared to CTRL HSCs
318 (**Figure 7E**). Altogether, our analysis supports the association between LGA exposure
319 and cell growth signaling targeted by DNA methylation and gene expression changes
320 with alteration of differentiation and proliferation capacities.

321 Discussion

322 Here, we interrogated three major layers of the regulatory landscape in cord-blood
323 derived CD34+ HPSCs, DNA methylation, chromatin conformation, and gene
324 expression. We characterized, in depth and at single-cell resolution, the functional
325 consequences associated with early DNA methylation changes observed in LGA
326 neonates. Through, the integration of multiple datasets and the development of novel
327 analytical approaches, we addressed a very challenging aspect of functional
328 (epi)genomics, the interpretation of DNA methylation changes. Focusing on HSPCs,
329 we believe that we contributed to a better understanding of how early environment
330 shapes the hematopoietic compartment development and long-term function.

331 We demonstrated in LGA neonates, a correlated increase of DNA methylation and
332 change in chromatin accessibility associated with decrease expression of downstream
333 target genes under the influence of key HSC transcription factors EGR1, KLF2, and
334 KLF4. EGR1, KLF2 and KLF4 are zinc-finger transcription factors involved in HSC

335 quiescence and differentiation. EGR1 has a known role in regulating cell growth,
336 development, and stress response in many tissues. In HSPC, EGR1 plays a role in
337 homeostasis of HSCs regulating proliferation[38]. EGR1 promotes quiescence and
338 decreases through differentiation. KLF family is implicated in key stem cell functions.
339 KLF4 is the most well-known factor of this family due to its role in reprogramming
340 somatic cells into induced pluripotent stem cells[47]. KLF4 has been identified as a target
341 for PU.1 transcription factor required for lineage commitment in HSPCs[48]. KLF2 and
342 KLF4 promote self-renewal in embryonic stem cells[34] but no study has looked
343 specifically at KLF2 and KLF2/KLF4 interactions in HSPCs. Our data suggests direct
344 and indirect (shared downstream target) interactions between these 3 transcriptions
345 factors in HSPCs. EGR1, KLF2, and KLF4 represent targets to be further explored in
346 order to challenge causality. Still, our findings lead to a better understanding of how
347 early exposure can affect long-term hematopoietic maintenance in human via
348 epigenetic programming of the EGR1, KLF2, and KLF4 signaling.

349 Interestingly, EGR1, KLF2, and KLF4 are not only involved in the regulation of
350 proliferation and differentiation *per se* but are key factors of the immediate early
351 response involved in stimulation-related cell activation. *EGR1* and *KLF2* expression
352 increase in response to extrinsic stimulation. Elevated *EGR1* and *KLF2* expression
353 promote self-renewal and quiescence in HSC[38,34]. Our transcriptomic data suggests
354 that such activation may be occurring in our samples with the activation of stress-
355 related signaling. The primary scope of our study was not to characterize the
356 environmental exposure that would trigger such response. However, one can
357 speculate that the activation could result from stress due to extended cold exposure
358 and handling time inherent to sample preparation. Still, the decrease activity observed
359 in LGA suggests that LGA HSCs capacities to respond to environmental challenges

360 are diminished. The alteration of HSCs capacity to respond to extrinsic stimulation fits
361 with the concept of early programming in which disease susceptibility relies not only
362 on early impairment of organ development but also on a decreased adaptability to
363 further environmental challenges to trigger disease. Indeed, fine-tuning of HSC
364 quiescence mechanisms is of crucial relevance for correct hematopoiesis. Not
365 responsive dormant HSC would lead to hematopoietic failure due to a lack of
366 differentiated blood cells. While highly responsive HSC would get to exhaustion of the
367 population and lack of long-term maintenance of the hematopoietic system.[46]

368 To validate findings from our integrative approach, we challenged HSPCs *in vitro* and
369 found significant decrease in common myeloid progenitors. These alterations can be
370 associated to difference in HSC population between condition, as the number of HSCs
371 seeded will influence the number of colonies generated, or to decreased differentiation
372 and proliferation capacities of HSC. Our data suggest that both are altered in LGA.
373 Indeed, the cell population analysis at transcriptomic level revealed a decrease in
374 HSCs in LGA neonates. While we also observed epigenomic and transcriptomic
375 alteration of signaling pathways and transcription factors involved in proliferation and
376 differentiation of HSCs. Yet, loss of stemness capacities in HSC is likely to drive the
377 decrease in HSC subpopulation observed in our data as supported by our *in silico*
378 differentiation analysis (pseudotimes). It showed that LGA HSCs appeared to
379 differentiate faster than their control counterpart, thus reducing the pool of HSCs in
380 the HSPCs population. These findings corroborate previous studies on developmental
381 programming of the hematopoietic system.[16,17] A reduction in self-renewal of HSPCs
382 and increase differentiation in both lymphoid and myeloid lineages have been
383 observed in a mouse model of maternal obesity.[18] These effects may drive long-term
384 consequences in human health as illustrated by the study performed by Kotowski et

385 *al.* in which the integrity of the hematopoietic system in neonates was associated with
386 susceptibility to onset of hematopoietic pathologies.[49]

387 Hematopoietic stem cell differentiation and self-renew rely on a synergic interplay
388 between genetically encoded signaling, cell-intrinsic and cell-extrinsic factors as well
389 as epigenetic modifiers.[45] This interplay appears altered in LGA neonates. We here
390 provide a comprehensive model recapitulating the functional influence of the
391 epigenetic early programming on HSPCs fitness to later environmental exposure
392 (**Figure 7F**). We also linked LGA associated epigenetic modifications to gene
393 expression and functional alterations through novel integrative approach. In this
394 regard, we identified targets to be further explored. We also brought a better
395 understanding of how early exposure can affect long-term tissue maintenance via
396 epigenetic programming of EGR1, KLF2, and KLF4 early immediate response
397 signaling associated TFs.

398 ## Methods

399 See the Supplemental Methods for additional information.

400 **Clinical sample collection.** Cord blood samples were obtained from CTRL and LGA
401 neonates. LGA were defined by birth weight and ponderal index values greater than
402 the 90th percentile adjusted for gestational age and sex. Control infants had normal
403 parameters (between 10th and 90th percentiles) for both birth weight and ponderal
404 index. Maternal and infant characteristics are shown in Supplemental Table 8.

405 **Isolation of CD34+ HSPCs.** Mononuclear cells were separated using PrepaCyte-
406 WBC following which CD34+ cells were obtained by positive immunomagnetic bead

407 selection, using the AutoMACS Separator (Miltenyi Biotech). Cells were
408 cryopreserved in 10% dimethyl sulfoxide using controlled rate freezing upon analysis.

409 **Genome-Wide DNA methylation assay.** DNA methylation levels for >1.7M CpGs
410 were obtained using the HELP-tagging assay as previously described.[50]

411 **Single-cell RNA sequencing libraries preparation.** After cell count and viability
412 check, cell suspension was loaded into the Chromium 10x Genomics controller and
413 library was generated using the chromium single-cell v3 chemistry following
414 manufacturer recommendations. Gene expression library was sequenced using
415 100bp paired-end reads on the Illumina NovaSeq 6000 system.

416 **Single-cell ATAC sequencing libraries preparation.** After cell count and viability
417 check, nuclei were isolated from cell suspension and incubated with transposase.
418 Transposed nuclei were then loaded into the Chromium 10x Genomics controller and
419 library was generated using the chromium single-cell ATAC v1.1 chemistry following
420 manufacturer recommendations. Gene expression library was sequenced using
421 150bp paired-end reads on the Illumina NovaSeq 6000 system.

422 **HTO protocol.** After cell counting and viability check and prior to cell suspension
423 loading on the Chromium 10x Genomics controller, cell hashtag (HTO) staining was
424 used following the cell-hashing protocol.[31]

425 **Colony Forming Unit Assay.** To assess clonogenic progenitor frequencies, 3×10^4
426 CD34+ HSPC cells were plated in methylcellulose containing SCF, GM-CSF, IL-3, and
427 EPO (H4434; STEMCELL Technologies). Colonies were scored 14 days later.

428 **Data processing and statistical analysis.** For DNA methylation analysis, low quality
429 CpGs were filtered out based on detection rate and confidence score. 754931 out of

430 1709224 CpGs were conserved for further analysis. Linear regression and statistical
431 modeling using LIMMA R package[51] were used to identify differentially methylated
432 CpGs (DMC). We assessed enrichment for biological pathways performing GSEA
433 using the ClusterProfiler package.[52] We performed transcription factor (TF) motif
434 enrichment analysis using HOMER tool[53] considering a 20bp region around the DMCs.

435 For single-cell RNAseq (scRNA-seq) analysis, data were preprocessed using the
436 CellRanger count pipeline. Data filtering, normalization and integration as well as
437 cluster identifications were performed using Seurat (v4) pipeline. Pseudo-bulk
438 differential expression analysis between LGA and CTRL cells within each
439 hematopoietic lineage was performed using DESeq2 R package.[54] Over
440 representation test was performed on differentially expressed genes (DEGs) using
441 enrichGO and enrichKEGG of the ClusterProfiler Package. The SCENIC workflow[55]
442 was used to identify co-regulated modules associated to a TF (regulons) and to
443 generate cell-specific activity score for each regulon. Differentiation trajectory analysis
444 and pseudotime attribution were conducted with Monocle.[32]

445 For single-cell ATAC-seq, data were preprocessed using the CellRanger ATAC
446 pipeline. Data filtering, normalization and integration as well as clustering were
447 performed using Signac pipeline. Cell type identification was based on scRNA-seq
448 annotation using a label transfer approach. Peaks calling at lineage level was
449 performed using MACS2 tool. Peaks specific to each lineage or differentially
450 accessible between LGA and Control were identified using FindMarkers function with
451 Logistic Regression (LR) models including cellular sequencing depth as latent
452 variable. TF motif enrichment on lineage or group specific peaks was performed using
453 FindMotifs function. All peaks enrichment analysis was performed using

454 hypergeometric tests. For final Gene Regulatory Network (GRN) construction, TF
455 target interactions inferred with SCENIC was filtered out based on presence of
456 corresponding TF motif in the peak associated to the target. Supplemental Table 9
457 contains information on number of cells per sample.

458 **Gene-methylation score**

459 To compute the gene-methylation score, 2 steps were needed: 1) to generate a CpG
460 score that reflects the association between CpG and gene, and 2) to concatenate
461 CpG-scores at gene level.

462 1) <u>CpG-score</u>

463 $$\text{CpGScore} = (-\log10(p_{cpg}) \times \text{meth.change}) \times \text{LinkWeight} \times \text{RegWeight}$$

464 Where p_{cpg} is the nominal *p*-value of the differential methylation analysis, and
465 meth.change is the difference between the percentage of methylation in LGA and the
466 percentage of methylation in CTRL. LinkWeight represents the confidence in CpG-
467 gene association and RegWeight represents the estimate regulatory influence of the
468 considered CpG based on CD34+ specific genomic annotation defined using CD34+
469 specific histone marks as previously described[4] and EnsRegScore refers to regulatory
470 regions defined based on the Ensembl Regulatory build hg19 genome annotation.[5]

471 <u>2) To concatenate CpG-Scores at gene level: gene-methylation score</u>

472 To summarize the CpG methylation change at the gene level, we aggregated the CpG-
473 Scores into a methylation gene-score by taking care of i) alleviate the arbitrary number
474 of CpGs per gene and ii) interpret differently CpG influences located on the promoter
475 of them in others genomic region.

476 The gene-methylation score is defined as:

477 $$\text{Gene-methylation score} = (\sum CpG\ Score \times Weight_{n_cpg})_{\text{promoter}} + (\sum |CpG\ Score| \times Weight_{n_cpg})_{\text{other_regions}}$$

478 Where the $Weight_{ncpg}$ was optimized to alleviate the influence of the number of CpGs
479 linked to a gene and defined as:

480 $$Weight_{nCpG} = \sqrt[3,8]{\frac{1}{\sum \frac{1}{|CpGScore+1|}}}$$

481

482 The code to perform the analyses in this manuscript is available at
483 https://github.com/umr1283/LGA_HSPC_PAPER.git.

484 ## Declarations

485 **Data sharing statement.** The DNA methylation and gene expression data will be
486 made available upon request to A.P., P.F., or F.D.

487 **Ethics approval.** This study was approved by the Institutional Review Board of the
488 Montefiore Medical Center and the Committee on Clinical Investigation at the Albert
489 Einstein College of Medicine and is in accordance with Health Insurance Portability
490 and Accountability Act regulations. Written informed consent was obtained from all
491 subjects before participation.

492 **Acknowledgments.** The authors thank the UMR 8199 LIGAN-PM Genomics platform
493 (Lille, France) which belongs to the 'Federation de Recherche' 3508 Labex EGID
494 (European Genomics Institute for Diabetes; ANR-10-LABX-46) and was supported by
495 the ANR Equipex 2010 session (ANR-10-EQPX-07-01; 'LIGAN-PM'). The LIGAN-PM
496 Genomics platform (Lille, France) is also supported by the FEDER and the Region
497 Nord-Pas-de-Calais-Picardie. This project is cofunded in the frame of CPER CTRL
498 program by the European Union - European Regional Development Fund (ERDF),
499 Hauts de France Region (contract n°17003781), Métropole Européenne de Lille

(contract n°2016_ESR_05), and French State (contract n°2017-R3-CTRL-Phase 1). The present work was also supported by the National Center for Precision Diabetic Medicine – PreciDIAB, which is jointly supported by the French National Agency for Research (ANR-18-IBHU-0001), by the European Union (FEDER), by the Hauts-de-France Regional Council and by the European Metropolis of Lille (MEL) and by the European Research Council (ERC Reg-Seq – 715575).

Support for this project was provided by the Roadmap Epigenomics Program, R01 HD063791 (Einstein/Greally). Support was also provided by Einstein's Center for Epigenomics, including the Epigenomics Shared Facility and Computational Epigenomics Group.

Author contributions. AP, AC, YMZ, ED, LBF, and FD were responsible for conducting research and analyzing data. MD and MC provided feedback on the data analysis. AP, AC, FD, AB, and PF contributed to writing the manuscript. FD, FH, and JG were responsible for designing the study.

Competing Interest. The authors declare no competing financial interests in relation to the work described.

References

1. Doulatov S, Notta F, Laurenti E, Dick JE. Hematopoiesis: a human perspective. *Cell Stem Cell*. 2012; **10**(2): p. 120-36.
2. Eaves CJ. Hematopoietic stem cells: concepts, definitions, and the new reality. *Blood*. 2015; **125**(17): p. 2605-13.
3. Balistreri CR, Garagnani P, Madonna R, Vaiserman A, Melino G. Developmental programming of adult haematopoiesis system. *Ageing Res Rev*. 2019; **54**: p. 100918.
4. Pardali E, Dimmeler S, Zeiher AM, Rieger MA. Clonal hematopoiesis, aging, and cardiovascular diseases. *Exp Hematol*. 2020; **83**: p. 95-104.
5. Lee CC, Fletcher MD, Tarantal AF. Effect of age on the frequency, cell cycle, and lineage maturation of rhesus monkey (Macaca mulatta) CD34+ and hematopoietic progenitor cells. *Pediatr Res*. 2005; **58**(2): p. 315-22.

528 6. Chandel NS, Jasper H, Ho TT, Passegue E. Metabolic regulation of stem cell function in
529 tissue homeostasis and organismal ageing. *Nat Cell Biol*. 2016; **18**(8): p. 823-32.
530 7. Guenechea G, Gan OI, Dorrell C, Dick JE. Distinct classes of human stem cells that differ
531 in proliferative and self-renewal potential. *Nat Immunol*. 2001; **2**(1): p. 75-82.
532 8. Barker DJ, Osmond C. Infant mortality, childhood nutrition, and ischaemic heart
533 disease in England and Wales. *Lancet*. 1986; **1**(8489): p. 1077-81.
534 9. Barker DJ, Osmond C, Kajantie E, Eriksson JG. Growth and chronic disease: findings in
535 the Helsinki Birth Cohort. *Ann Hum Biol*. 2009; **36**(5): p. 445-58.
536 10. Cnattingius S, Villamor E, Lagerros YT, Wikstrom AK, Granath F. High birth weight and
537 obesity--a vicious circle across generations. *Int J Obes (Lond)*. 2012; **36**(10): p. 1320-4.
538 11. Wang Y, Gao E, Wu J, et al. Fetal macrosomia and adolescence obesity: results from a
539 longitudinal cohort study. *Int J Obes (Lond)*. 2009; **33**(8): p. 923-8.
540 12. Eriksson J, Forsen T, Tuomilehto J, Osmond C, Barker D. Size at birth, childhood growth
541 and obesity in adult life. *Int J Obes Relat Metab Disord*. 2001; **25**(5): p. 735-40.
542 13. Harder T, Rodekamp E, Schellong K, Dudenhausen JW, Plagemann A. Birth weight and
543 subsequent risk of type 2 diabetes: a meta-analysis. *Am J Epidemiol*. 2007; **165**(8): p.
544 849-57.
545 14. Pocobelli G, Dublin S, Enquobahrie DA, Mueller BA. Birth Weight and Birth Weight for
546 Gestational Age in Relation to Risk of Hospitalization with Primary Hypertension in
547 Children and Young Adults. *Matern Child Health J*. 2016; **20**(7): p. 1415-23.
548 15. Koyanagi A, Zhang J, Dagvadorj A, et al. Macrosomia in 23 developing countries: an
549 analysis of a multicountry, facility-based, cross-sectional survey. *Lancet*. 2013;
550 **381**(9865): p. 476-83.
551 16. Al-Sweedan SA, Musalam L, Obeidat B. Factors predicting the hematopoietic stem
552 cells content of the umbilical cord blood. *Transfus Apher Sci*. 2013; **48**(2): p. 247-52.
553 17. Chandra T, Afreen S, Kumar A, Singh U, Gupta A. Does umbilical cord blood-derived
554 CD34+ cell concentration depend on the weight and sex of a full-term infant? *J Pediatr*
555 *Hematol Oncol*. 2012; **34**(3): p. 184-7.
556 18. Kamimae-Lanning AN, Krasnow SM, Goloviznina NA, et al. Maternal high-fat diet and
557 obesity compromise fetal hematopoiesis. *Mol Metab*. 2015; **4**(1): p. 25-38.
558 19. Delahaye F, Wijetunga NA, Heo HJ, et al. Sexual dimorphism in epigenomic responses
559 of stem cells to extreme fetal growth. *Nat Commun*. 2014; **5**: p. 5187.
560 20. Stella CC, Cazzola M, De Fabritiis P, et al. CD34-positive cells: biology and clinical
561 relevance. *Haematologica*. 1995; **80**(4): p. 367-87.
562 21. Thurman RE, Rynes E, Humbert R, et al. The accessible chromatin landscape of the
563 human genome. *Nature*. 2012; **489**(7414): p. 75-82.
564 22. Delahaye F, Do C, Kong Y, et al. Genetic variants influence on the placenta regulatory
565 landscape. *PLoS Genet*. 2018; **14**(11): p. e1007785.
566 23. Dong D, Tian Y, Zheng SC, Teschendorff AE. ebGSEA: an improved Gene Set Enrichment
567 Analysis method for Epigenome-Wide-Association Studies. *Bioinformatics*. 2019;
568 **35**(18): p. 3514-3516.
569 24. Ren X, Kuan PF. methylGSA: a Bioconductor package and Shiny app for DNA
570 methylation data length bias adjustment in gene set testing. *Bioinformatics*. 2019;
571 **35**(11): p. 1958-1959.
572 25. Wijetunga NA, Delahaye F, Zhao YM, et al. The meta-epigenomic structure of purified
573 human stem cell populations is defined at cis-regulatory sequences. *Nat Commun*.
574 2014; **5**: p. 5195.

575 26. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published
576 genome-wide association studies, targeted arrays and summary statistics 2019.
577 *Nucleic Acids Res*. 2019; **47**(D1): p. D1005-D1012.
578 27. Cabezas-Wallscheid N, Klimmeck D, Hansson J, et al. Identification of regulatory
579 networks in HSCs and their immediate progeny via integrated proteome,
580 transcriptome, and DNA methylome analysis. *Cell Stem Cell*. 2014; **15**(4): p. 507-522.
581 28. Karamitros D, Stoilova B, Aboukhalil Z, et al. Single-cell analysis reveals the continuum
582 of human lympho-myeloid progenitor cells. *Nat Immunol*. 2018; **19**(1): p. 85-97.
583 29. Zhang YH, Hu Y, Zhang Y, Hu LD, Kong X. Distinguishing three subtypes of
584 hematopoietic cells based on gene expression profiles using a support vector machine.
585 *Biochim Biophys Acta Mol Basis Dis*. 2018; **1864**(6 Pt B): p. 2255-2265.
586 30. Zheng S, Papalexi E, Butler A, Stephenson W, Satija R. Molecular transitions in early
587 progenitors during human cord blood hematopoiesis. *Mol Syst Biol*. 2018; **14**(3): p.
588 e8041.
589 31. Stoeckius M, Zheng S, Houck-Loomis B, et al. Cell Hashing with barcoded antibodies
590 enables multiplexing and doublet detection for single cell genomics. *Genome Biol*.
591 2018; **19**(1): p. 224.
592 32. Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C. Single-cell mRNA quantification and
593 differential analysis with Census. *Nat Methods*. 2017; **14**(3): p. 309-315.
594 33. Zhu H, Wang G, Qian J. Transcription factors as readers and effectors of DNA
595 methylation. *Nat Rev Genet*. 2016; **17**(9): p. 551-65.
596 34. Jiang J, Chan YS, Loh YH, et al. A core Klf circuitry regulates self-renewal of embryonic
597 stem cells. *Nat Cell Biol*. 2008; **10**(3): p. 353-60.
598 35. Zhao T, Liu C, Chen L. Roles of Klf5 Acetylation in the Self-Renewal and the
599 Differentiation of Mouse Embryonic Stem Cells. *PLoS One*. 2015; **10**(9): p. e0138168.
600 36. Lecuyer E, Herblot S, Saint-Denis M, et al. The SCL complex regulates c-kit expression
601 in hematopoietic cells through functional interaction with Sp1. *Blood*. 2002; **100**(7): p.
602 2430-40.
603 37. Gilmour J, O'Connor L, Middleton CP, et al. Robust hematopoietic specification
604 requires the ubiquitous Sp1 and Sp3 transcription factors. *Epigenetics Chromatin*.
605 2019; **12**(1): p. 33.
606 38. Min IM, Pietramaggiori G, Kim FS, Passegue E, Stevenson KE, Wagers AJ. The
607 transcription factor EGR1 controls both the proliferation and localization of
608 hematopoietic stem cells. *Cell Stem Cell*. 2008; **2**(4): p. 380-91.
609 39. Zhu J, Emerson SG. Hematopoietic cytokines, transcription factors and lineage
610 commitment. *Oncogene*. 2002; **21**(21): p. 3295-313.
611 40. Venezia TA, Merchant AA, Ramos CA, et al. Molecular signatures of proliferation and
612 quiescence in hematopoietic stem cells. *PLoS Biol*. 2004; **2**(10): p. e301.
613 41. Tenen DG, Hromas R, Licht JD, Zhang DE. Transcription factors, normal myeloid
614 development, and leukemia. *Blood*. 1997; **90**(2): p. 489-519.
615 42. Antonchuk J, Sauvageau G, Humphries RK. HOXB4-induced expansion of adult
616 hematopoietic stem cells ex vivo. *Cell*. 2002; **109**(1): p. 39-45.
617 43. Sun Y, Zhou B, Mao F, et al. HOXA9 Reprograms the Enhancer Landscape to Promote
618 Leukemogenesis. *Cancer Cell*. 2018; **34**(4): p. 643-658 e5.
619 44. Magnusson M, Brun AC, Miyake N, et al. HOXA10 is a critical regulator for
620 hematopoietic stem cells and erythroid/megakaryocyte development. *Blood*. 2007;
621 **109**(9): p. 3687-96.

622 45. Yu VWC, Yusuf RZ, Oki T, et al. Epigenetic Memory Underlies Cell-Autonomous
623 Heterogeneous Behavior of Hematopoietic Stem Cells. *Cell*. 2017; **168**(5): p. 944-945.
624 46. Wilson A, Laurenti E, Trumpp A. Balancing dormant and self-renewing hematopoietic
625 stem cells. *Curr Opin Genet Dev*. 2009; **19**(5): p. 461-8.
626 47. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic
627 and adult fibroblast cultures by defined factors. *Cell*. 2006; **126**(4): p. 663-76.
628 48. Feinberg MW, Wara AK, Cao Z, et al. The Kruppel-like factor KLF4 is a critical regulator
629 of monocyte differentiation. *EMBO J*. 2007; **26**(18): p. 4138-48.
630 49. Kotowski M, Safranow K, Kawa MP, et al. Circulating hematopoietic stem cell count is
631 a valuable predictor of prematurity complications in preterm newborns. *BMC Pediatr*.
632 2012; **12**: p. 148.
633 50. Suzuki M, Jing Q, Lia D, Pascual M, McLellan A, Greally JM. Optimized design and data
634 analysis of tag-based cytosine methylation assays. *Genome Biol*. 2010; **11**(4): p. R36.
635 51. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for
636 RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015; **43**(7): p. e47.
637 52. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological
638 themes among gene clusters. *OMICS*. 2012; **16**(5): p. 284-7.
639 53. Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining
640 transcription factors prime cis-regulatory elements required for macrophage and B
641 cell identities. *Mol Cell*. 2010; **38**(4): p. 576-89.
642 54. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for
643 RNA-seq data with DESeq2. *Genome Biol*. 2014; **15**(12): p. 550.
644 55. Aibar S, Gonzalez-Blas CB, Moerman T, et al. SCENIC: single-cell regulatory network
645 inference and clustering. *Nat Methods*. 2017; **14**(11): p. 1083-1086.
646

647 ## Figure Legend

648 **Figure 1: LGA is associated with DNA hypermethylation targeting key stem cell**
649 **signaling pathways.** (A) Volcano plot of DNA methylation score differences for LGA
650 compared to CTRL in cohort 1, cohort 2, and cohort 1 +2. Differentially methylated loci
651 with *p*-value <0.001 and |methylation difference| > 25 % are shown in red. (B)
652 Summary of calculation for the gene-methylation score. (C) Validation of the gene-
653 methylation score. Gene-methylation score distribution. Bar plot of the association
654 between gene-methylation score and genomic or methylation related features. Bar
655 plots representing the significance of the difference of gene-methylation score of
656 DEGs compared to non-DEGs considering different metrics. eQTR, region with
657 expression quantitative traits loci; DMC, differentially methylated CpGs. (D) Network

658 representation of GO Biological Process enriched in hypermethylated genes.
659 Significantly enriched GO terms were identify using GSEA based on the gene-
660 methylation score. Edges represent interactions (gene overlap) between pathways.
661
662 **Figure 2: Lineage specific transcriptomic analysis.** (A) Hematomap, UMAP
663 representation of distinct HSPC lineages. (B) Dot plot representing key markers used
664 to annotate cell populations. LT-HSC, long-term hematopoietic stem cell; HSC,
665 hematopoietic stem cell; MPP, multipotent progenitor; LMPP, lymphoid-primed
666 multipotent progenitors; Erythro-Mas, erythroid and mast precursor; Mk/Er,
667 megakaryocyte and erythrocyte; DC, dendritic cell. (C) MA plots representing gene
668 expression analysis in HSCs comparing LGA vs CTRL. Differentially expressed genes
669 with adjusted *p*-value <0.05 and $|\log_2 FC|$ >0.5 are shown in red. (D) Network
670 representation of significantly enriched pathways identified through GO GSEA
671 analysis comparing LGA vs CTRL. Non redundant pathways annotations have been
672 used. Edges represent interactions between pathways.
673
674 **Figure 3: Association between changes in DNA methylation and in gene**
675 **expression.** (A) Dot plot representing the correlation between DNA methylation and
676 gene expression changes. Differentially expressed genes with adjusted *p*-value <0.05
677 and $|\log_2 FC|$ >0.5 are shown in red. (B) Boxplots representing gene-methylation score
678 distribution associated to non DEG, up-regulated and down-regulated genes
679 (Wilcoxon test). (C) Network representation of significantly enriched pathways
680 identified through GO GSEA analysis based on DEG identified comparing LGA vs
681 CTRL. Nodes are color coded based on enrichment for differentially methylated genes
682 using the gene-methylation score. Edges represent interactions between pathways.

683

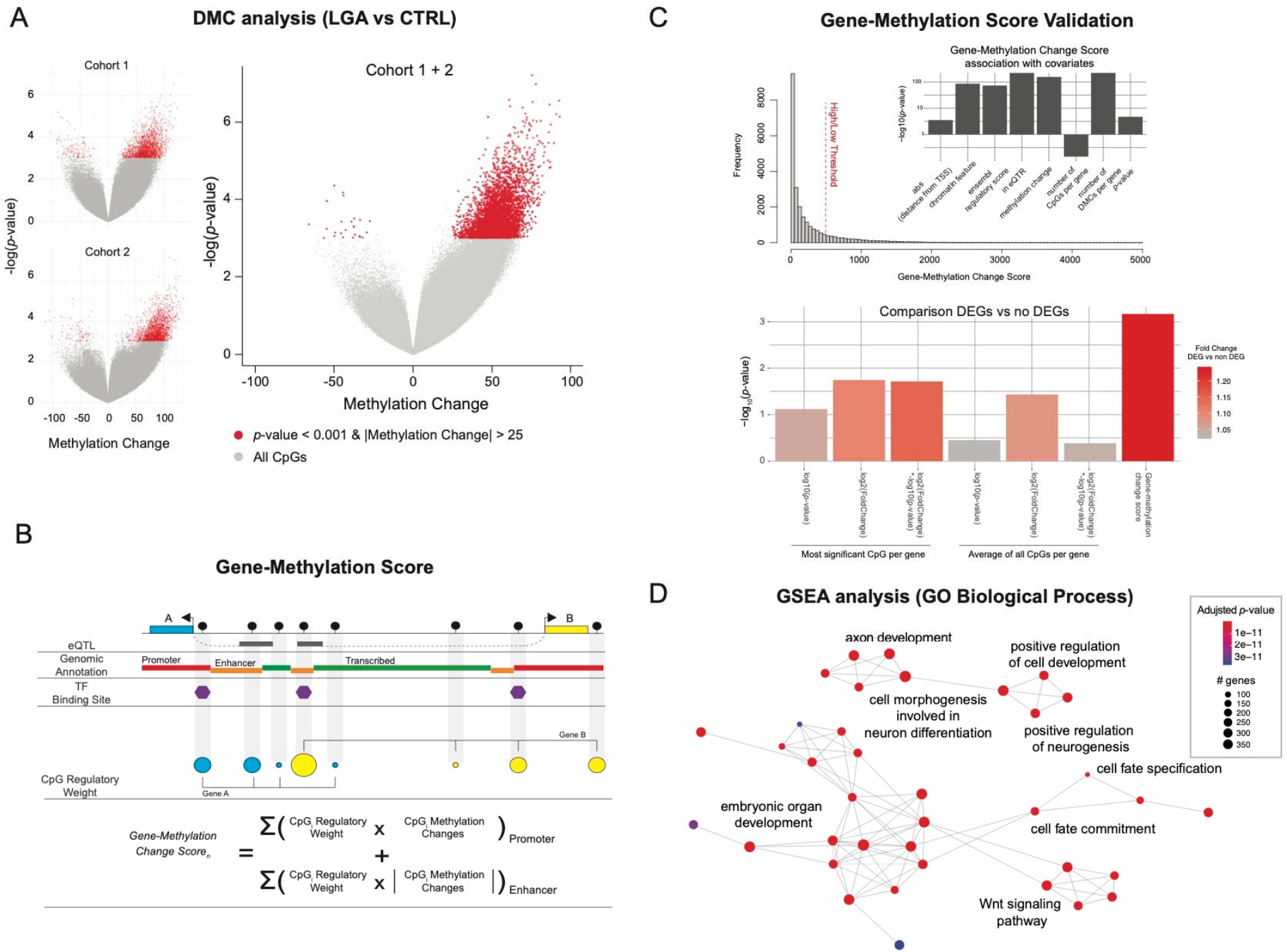
684 **Figure 4: Chromatin accessibility analysis.** (A) UMAP representing HSPCs lineage
685 based on chromatin accessibility. Annotations is based on the Hematomap using
686 transfer label approach. (B) Dot plot representing enrichment for DMC within lineage
687 specific peaks. (C) Bar plots representing enrichment for peaks containing CpG or
688 peaks containing DMC associated to DEGs, up-regulated and down-regulated genes
689 (* *p*-value<0.05; *** *p*-value<0.001). (D) Dot plot representing enrichment for
690 transcription factor motif within Down peaks identified comparing chromatin
691 accessibility between LGA and CTRL. Dots are color coded based on percentage of
692 peaks with motif and y-axis represent the significance of the enrichment. (E) Bar plots
693 representing enrichment analysis considering accessible, down and up peaks.
694 Enrichment is performed using peaks in expressed genes (expressed), peaks with
695 CpGs (CpGs), peaks in DEG (DEGs), peaks with DMC (DMCs) and peaks in DEG
696 with DMC (DMCs_DEGs) as reference gene sets.

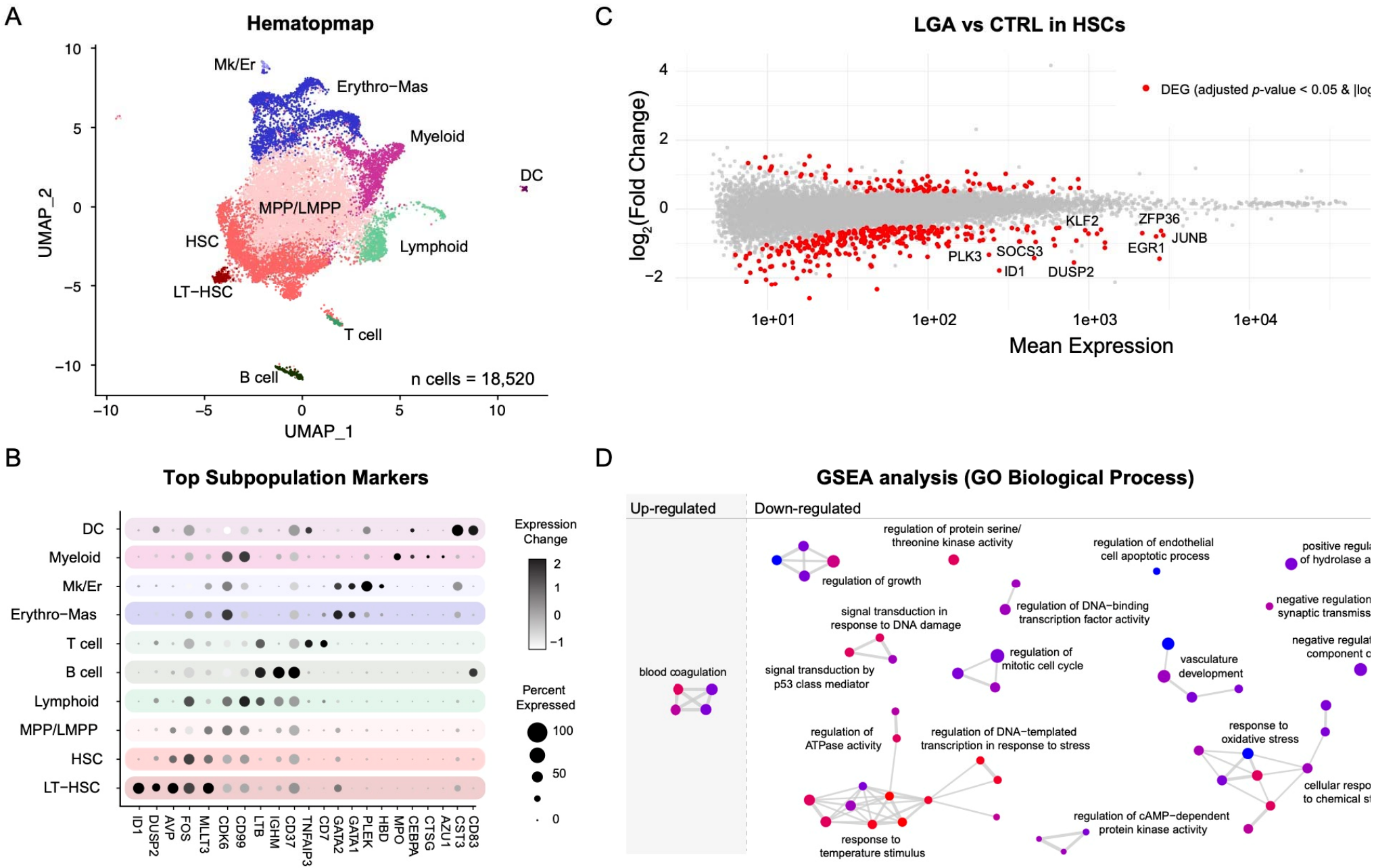
697

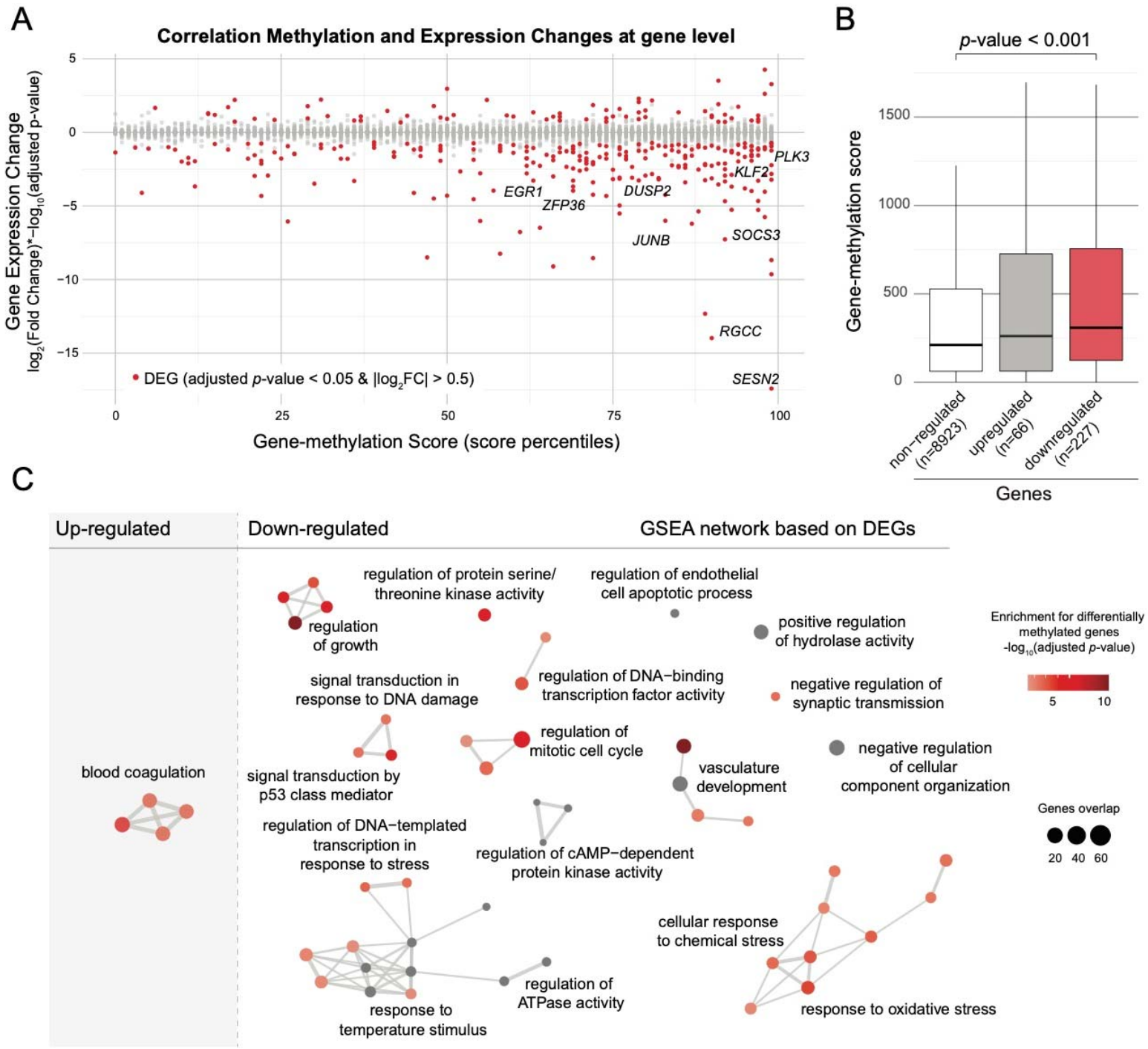
698 **Figure 5: Epigenetic programming of HSC specific regulons altered in LGA**
699 **neonates.** Regulons and TF target information were obtained through the SCENIC
700 workflow. (A) Boxplots representing regulon activity score in CTRL and LGA HSC
701 lineage. Barplot representing the change in regulon activity and significance
702 comparing LGA vs CTRL. Only significantly affected regulons are represented
703 (adjusted *p*-value <0.001 and |activity score fold change| >10%). (B) Heatmap
704 representing association between altered regulons and selected gene sets annotation.
705 (C) Volcano plot representing enrichment in change in expression and DNA
706 methylation in regulons. Regulons enriched considering both expression and
707 methylation (adjusted p-value<0.01 and NES>1.6) are in red. (D) Dot plot representing

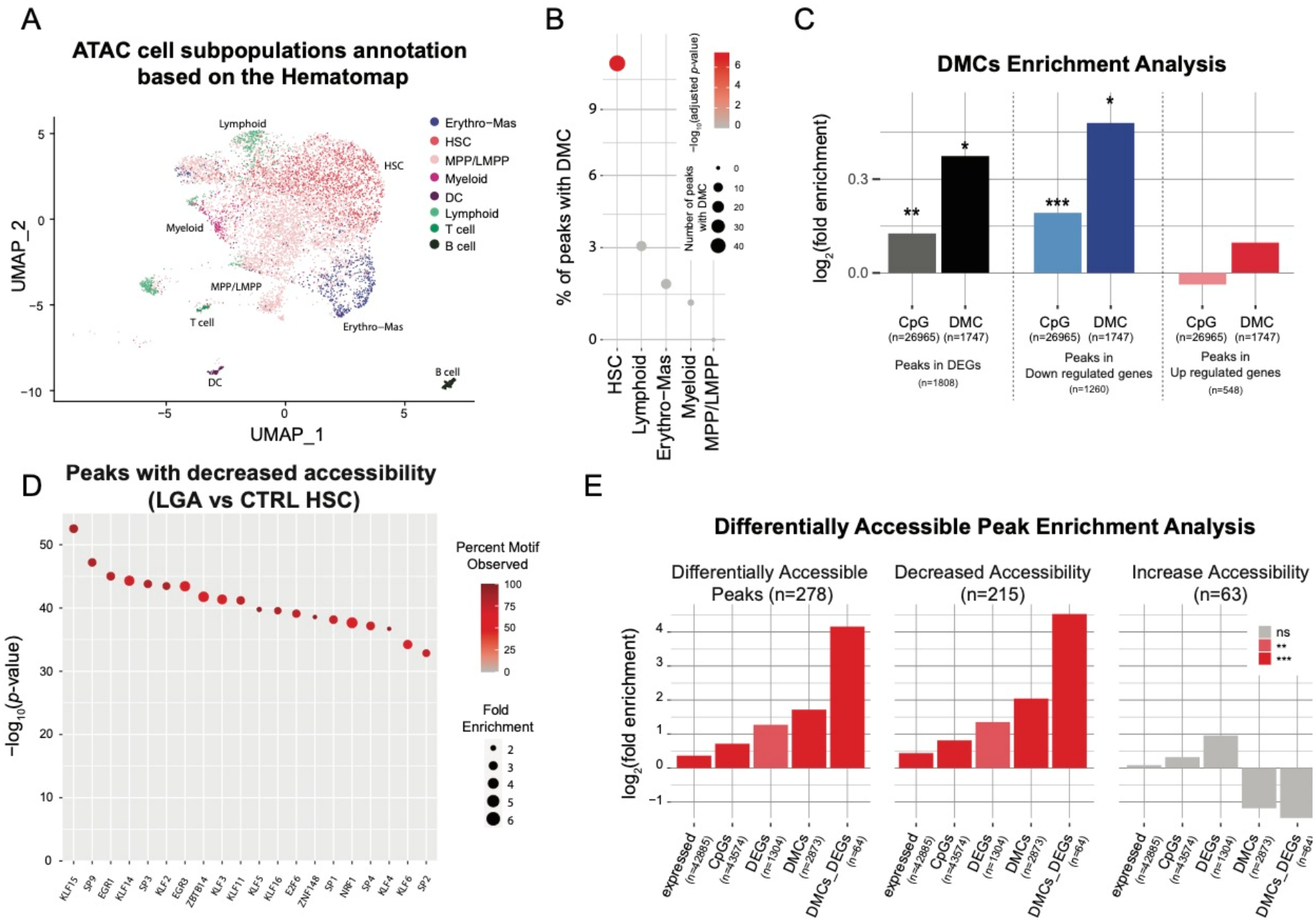
708 enrichment for TF binding motifs using HOMER considering a ±20bp region around
709 DMCs. Dots are color coded based on significance of the enrichment and y-axis
710 represent the number of regions with binding motif among DMCs. (E) Dot plot
711 representing enrichment for TF binding motifs using HOMER considering peaks with
712 DMCs. Dots are color coded based on the fold-enrichment and y-axis represent the
713 significance of the enrichment.
714
715 **Figure 6: Network recapitulating interaction between the epigenomic and**
716 **transcriptomic alterations in LGA.** (A) Network representing interactions between
717 target genes and transcription factors considering our top affected regulons ARID5A,
718 EGR1, KLF2, KLF4, FOSB and JUN. Each dot represents a gene within the network,
719 triangle represents transcription factor, arrow represents interaction between
720 transcription factor and target genes, shapes are color coded to reflect the change in
721 gene-methylation score, and DEGs are labeled in red. Size of the shape represents
722 the number of interactions. Only genes with 2 or more interactions are represented.
723 (B) Tracks representing DNA methylation and chromatin accessibility for selected
724 representative regions. Histogram representing change in DNA methylation at CpG
725 level comparing LGA vs CTRL. Violin plot representing gene expression for selected
726 genes (C) Network based on integration of DNA methylation, gene expression and
727 chromatin accessibility representing transcription factors and downstream target
728 interactions within EGR1, KLF2 and KLF4 regulons. Only genes associated to peaks
729 with TF motifs of interest are annotated. Donuts represent different level of
730 interactions.
731

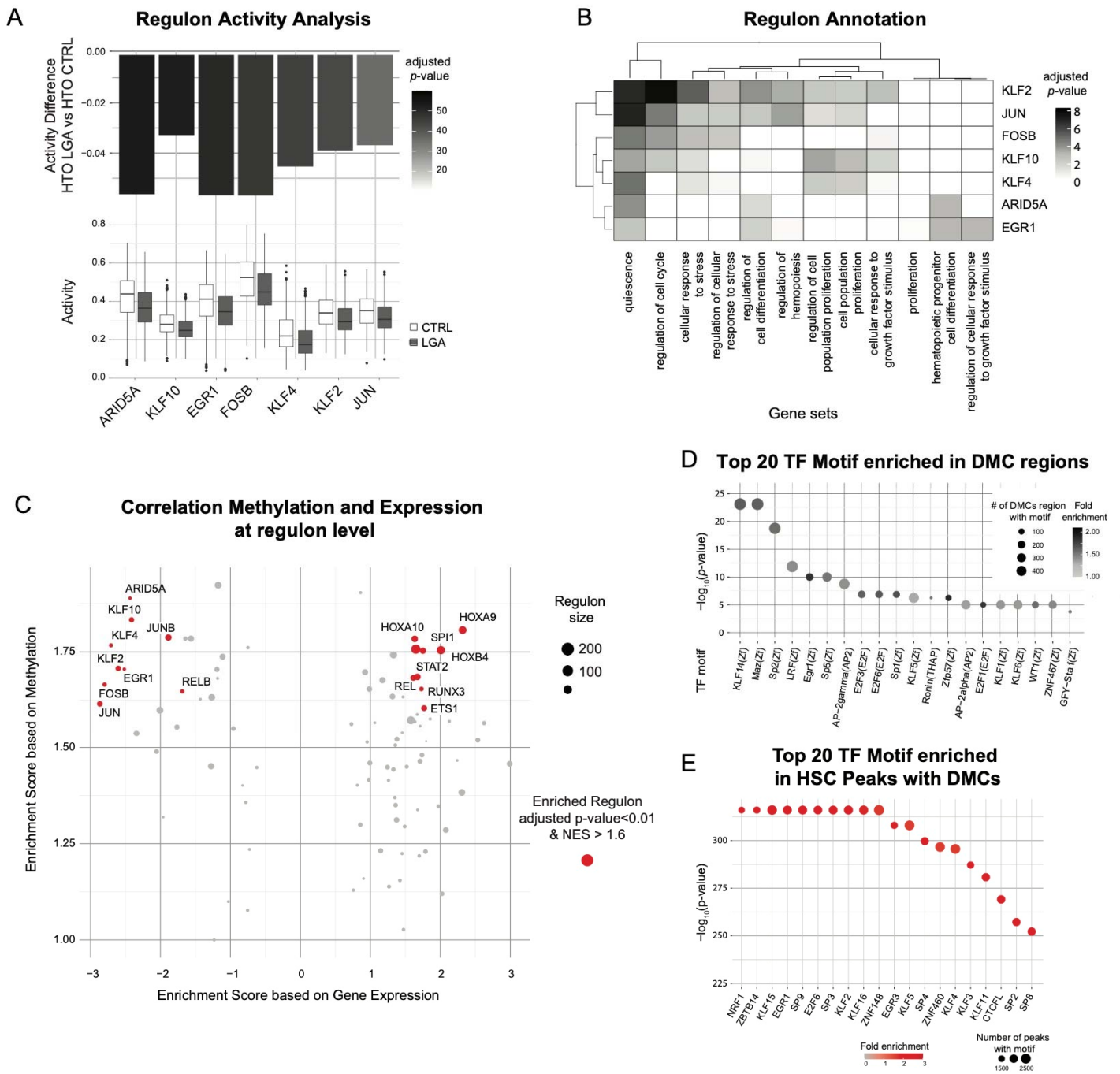
732 **Figure 7: LGA is associated with decreased expansion capacity and a HSC shift**
733 **toward more differentiated cells.** (A) Bar plot representing colonies distribution after
734 CFU assays. (B) Representative capture of colonies morphological differences found
735 in CTRL and LGA. (C) Boxplots representing the cell distribution across hematopoietic
736 main lineages in CTRL and LGA. (D) UMAPs representing pseudotimes across
737 lineages. (E) Box plots representing the cumulative percentage of cells per
738 pseudotime in CTRL and LGA. Boxplots in the vignette represent overall pseudotime
739 distribution in CTRL and LGA. Density plots correspond to cell populations distribution
740 across pseudotimes. (F) Model recapitulating the influence of LGA on the
741 hematopoietic compartment. (LT-HSC, long-term hematopoietic stem cell; HSC,
742 hematopoietic stem cell; MPP, multipotent progenitor; LMPP, lymphoid-primed
743 multipotent progenitors; Erythro-Mas, erythroid and mast precursor; Mk/Er,
744 megakaryocyte and erythrocyte; DC, dendritic cell; CFU-GEMM, common myeloid
745 progenitors; BFU-E, erythroid progenitors; CFU-GM, granulocyte-macrophage
746 progenitors.
747
748
749
750

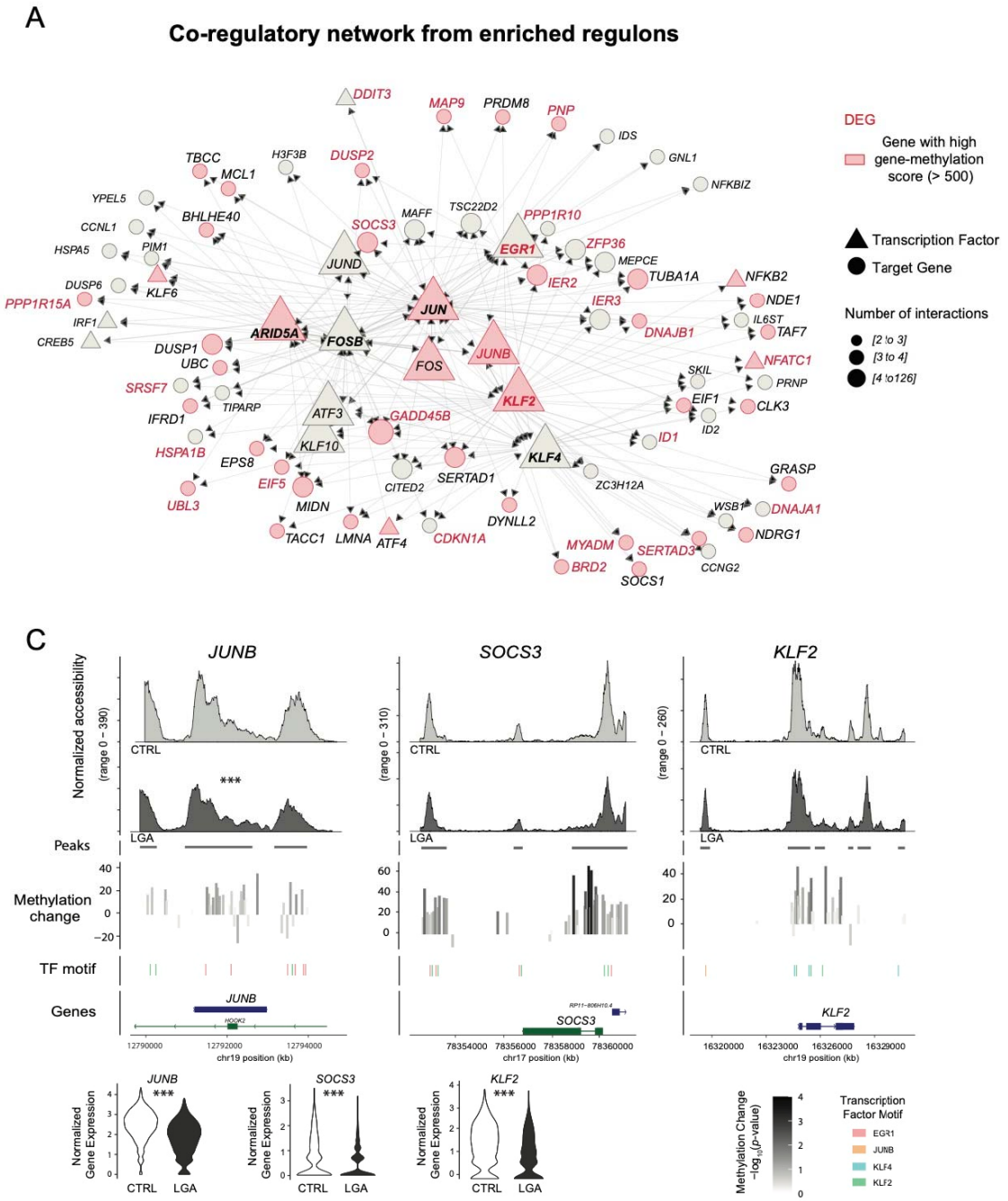

A
DMC analysis (LGA vs CTRL)
Cohort 1
Cohort 2
Cohort 1 + 2
-log(p-value)
Methylation Change
p-value < 0.001 & |Methylation Change| > 25
All CpGs
B
Gene-Methylation Score
eQTL
Genomic Annotation
Promoter
Enhancer
Transcribed
TF Binding Site
CpG Regulatory Weight
Gene A
Gene B
Gene-Methylation Change Score
CpG Regulatory Weight
CpG Methylation Changes
Promoter
Enhancer
C
Gene-Methylation Score Validation
Gene-Methylation Change Score association with covariates
Frequency
High/Low Threshold
Gene-Methylation Change Score
abs (distance from TSS)
chromatin feature
ensembl regulatory score
in eQTR
methylation change
number of CpGs per gene
number of DMCs per gene
p-value
Comparison DEGs vs no DEGs
Fold Change DEG vs non DEG
log10(p-value)
log2(FoldChange)
log2(FoldChange) * -log10(p-value)
Gene-methylation change score
Most significant CpG per gene
Average of all CpGs per gene
D
GSEA analysis (GO Biological Process)
Adjusted p-value
genes
axon development
positive regulation of cell development
cell morphogenesis involved in neuron differentiation
positive regulation of neurogenesis
cell fate specification
embryonic organ development
cell fate commitment
Wnt signaling pathway

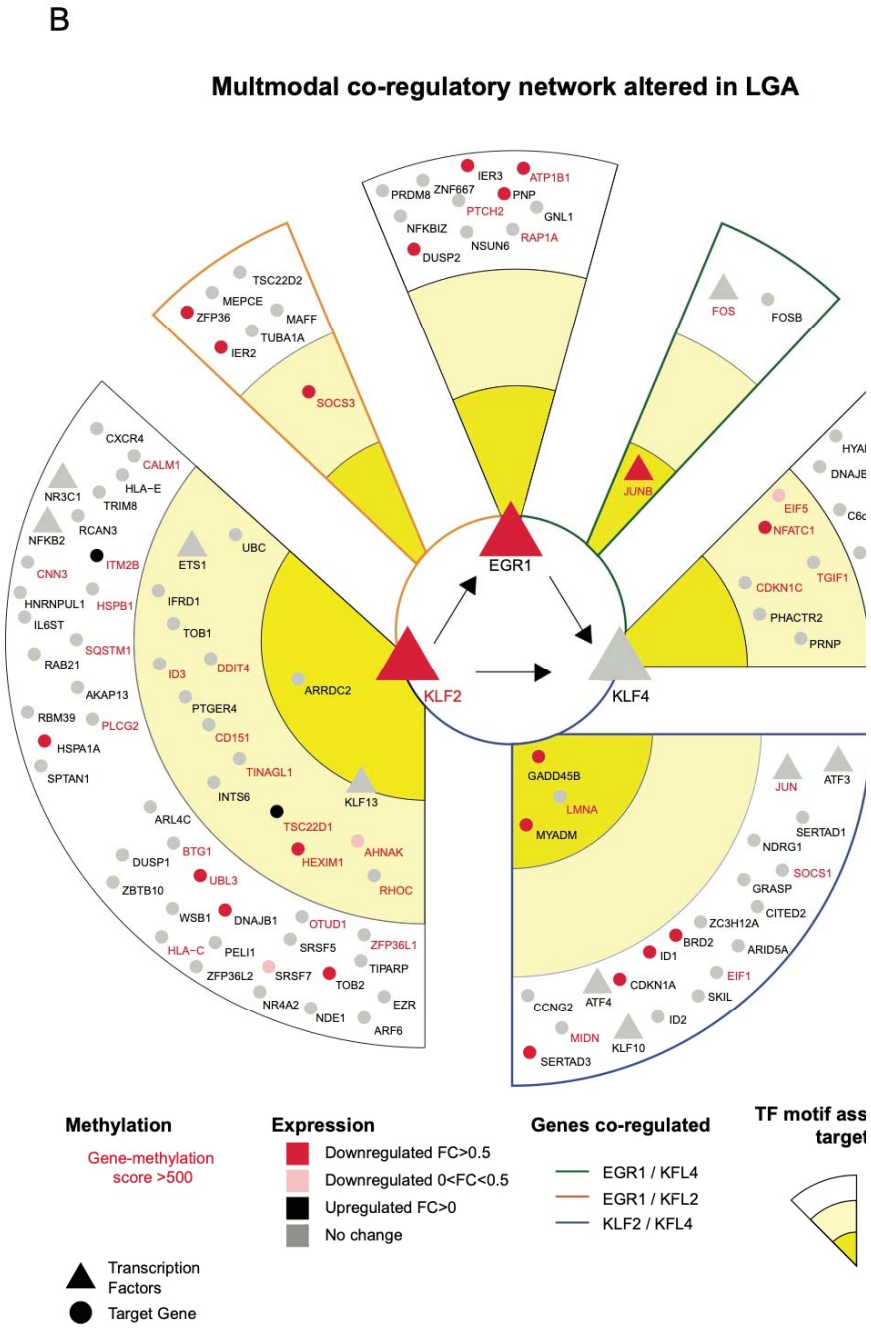

A
Hematopmap
Mk/Er
Erythro−Mas
Myeloid
DC
MPP/LMPP
HSC
Lymphoid
LT−HSC
T cell
B cell
n cells = 18,520
UMAP_2
UMAP_1
C
LGA vs CTRL in HSCs
DEG (adjusted p-value < 0.05 & |log
log2(Fold Change)
Mean Expression
KLF2
ZFP36
JUNB
EGR1
PLK3
SOCS3
ID1
DUSP2
B
Top Subpopulation Markers
DC
Myeloid
Mk/Er
Erythro−Mas
T cell
B cell
Lymphoid
MPP/LMPP
HSC
LT−HSC
ID1
DUSP2
AVP
FOS
MLLT3
CDK6
CD99
LTB
IGHM
CD37
TNFAIP3
CD7
GATA2
GATA1
PLEK
HBD
MPO
CEBPA
CTSG
AZU1
CST3
CD83
Expression Change
Percent Expressed
D
GSEA analysis (GO Biological Process)
Up-regulated
Down-regulated
blood coagulation
regulation of protein serine/ threonine kinase activity
regulation of endothelial cell apoptotic process
regulation of growth
regulation of DNA−binding transcription factor activity
signal transduction in response to DNA damage
signal transduction by p53 class mediator
regulation of mitotic cell cycle
vasculature development
regulation of ATPase activity
regulation of DNA−templated transcription in response to stress
response to oxidative stress
response to temperature stimulus
regulation of cAMP−dependent protein kinase activity

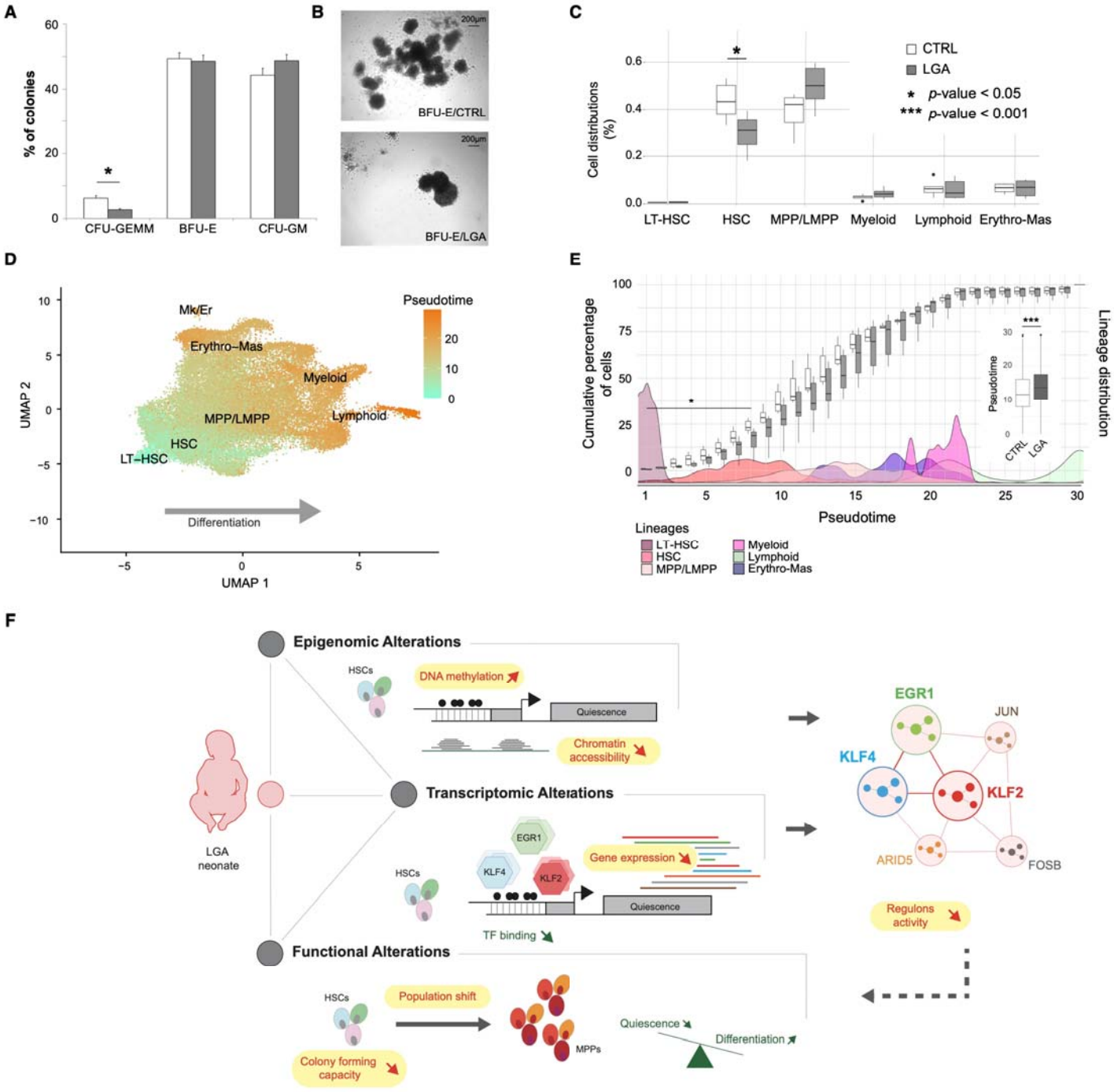

A
Correlation Methylation and Expression Changes at gene level
Gene Expression Change
log2(Fold Change)*−log10(adjusted p-value)
5
0
−5
−10
−15
PLK3
KLF2
EGR1
ZFP36
DUSP2
JUNB
SOCS3
RGCC
SESN2
DEG (adjusted p-value < 0.05 & |log2FC| > 0.5)
0
25
50
75
100
Gene-methylation Score (score percentiles)
B
p-value < 0.001
Gene-methylation score
1500
1000
500
0
non−regulated (n=8923)
upregulated (n=66)
downregulated (n=227)
Genes
C
Up-regulated
Down-regulated
GSEA network based on DEGs
blood coagulation
regulation of protein serine/ threonine kinase activity
regulation of endothelial cell apoptotic process
regulation of growth
positive regulation of hydrolase activity
signal transduction in response to DNA damage
regulation of DNA−binding transcription factor activity
negative regulation of synaptic transmission
regulation of mitotic cell cycle
signal transduction by p53 class mediator
vasculature development
negative regulation of cellular component organization
regulation of DNA−templated transcription in response to stress
regulation of cAMP−dependent protein kinase activity
cellular response to chemical stress
regulation of ATPase activity
response to temperature stimulus
response to oxidative stress
Enrichment for differentially methylated genes
-log10(adjusted p-value)
5
10
Genes overlap
20
40
60


A
ATAC cell subpopulations annotation based on the Hematomap
Lymphoid
HSC
Myeloid
MPP/LMPP
T cell
Erythro−Mas
DC
B cell
Erythro−Mas
HSC
MPP/LMPP
Myeloid
DC
Lymphoid
T cell
B cell
UMAP_2
UMAP_1
B
% of peaks with DMC
−log10(adjusted p-value)
Number of peaks with DMC
HSC
Lymphoid
Erythro−Mas
Myeloid
MPP/LMPP
C
DMCs Enrichment Analysis
log2(fold enrichment)
CpG (n=26965)
DMC (n=1747)
Peaks in DEGs (n=1808)
Peaks in Down regulated genes (n=1260)
Peaks in Up regulated genes (n=548)
D
Peaks with decreased accessibility (LGA vs CTRL HSC)
−log10(p-value)
Percent Motif Observed
Fold Enrichment
KLF15
SP9
EGR1
KLF14
SP3
KLF2
EGR3
ZBTB14
KLF3
KLF11
KLF5
KLF16
E2F6
ZNF148
SP1
NR1
SP4
KLF4
KLF6
SP2
E
Differentially Accessible Peak Enrichment Analysis
Differentially Accessible Peaks (n=278)
Decreased Accessibility (n=215)
Increase Accessibility (n=63)
log2(fold enrichment)
ns
**
***
expressed (n=42885)
CpGs (n=43574)
DEGs (n=1304)
DMCs (n=2873)
DMCs_DEGs (n=64)


A
Regulon Activity Analysis
Activity Difference
HTO LGA vs HTO CTRL
adjusted
p-value
Activity
CTRL
LGA
ARID5A
KLF10
EGR1
FOSB
KLF4
KLF2
JUN
B
Regulon Annotation
adjusted
p-value
KLF2
JUN
FOSB
KLF10
KLF4
ARID5A
EGR1
quiescence
regulation of cell cycle
cellular response to stress
regulation of cellular response to stress
regulation of cell differentiation
regulation of hemopoiesis
regulation of cell population proliferation
cell population proliferation
cellular response to growth factor stimulus
proliferation
hematopoietic progenitor cell differentiation
regulation of cellular response to growth factor stimulus
Gene sets
C
Correlation Methylation and Expression at regulon level
Enrichment Score based on Methylation
Enrichment Score based on Gene Expression
ARID5A
KLF10
JUNB
KLF4
KLF2
EGR1
RELB
FOSB
JUN
HOXA10
HOXA9
SPI1
HOXB4
STAT2
REL
RUNX3
ETS1
Regulon size
200
100
Enriched Regulon
adjusted p-value<0.01
& NES > 1.6
D
Top 20 TF Motif enriched in DMC regions
−log10(p-value)
TF motif
# of DMCs region with motif
Fold enrichment
KLF14(Zf)
Max(Zf)
Sp2(Zf)
LRF(Zf)
Egr1(Zf)
Sp5(Zf)
AP-2gamma(AP2)
E2F3(E2F)
E2F6(E2F)
Sp1(Zf)
KLF5(Zf)
Ronin(THAP)
Zfp57(Zf)
AP-2alpha(AP2)
E2F1(E2F)
KLF1(Zf)
KLF6(Zf)
WT1(Zf)
ZNF467(Zf)
GFY-Sta f(Zf)
E
Top 20 TF Motif enriched in HSC Peaks with DMCs
−log10(p-value)
NRF1
ZBTB14
KLF15
EGR1
SP9
E2F6
SP3
KLF2
KLF16
ZNF148
EGR3
KLF5
SP4
ZNF460
KLF4
KLF3
KLF11
CTCFL
SP2
SP8
Fold enrichment
Number of peaks with motif


A
Co-regulatory network from enriched regulons
DEG
Gene with high gene-methylation score (> 500)
Transcription Factor
Target Gene
Number of interactions
[2 to 3]
[3 to 4]
[4 to126]
JUND
JUN
FOSB
ARID5A
FOS
JUNB
KLF2
KLF4
EGR1
ATF3
KLF10
SOCS3
C
JUNB
SOCS3
KLF2
Normalized accessibility
(range 0 – 390)
(range 0 – 310)
(range 0 – 260)
CTRL
LGA
***
Peaks
Methylation change
TF motif
Genes
chr19 position (kb)
chr17 position (kb)
Normalized Gene Expression
Methylation Change −log10(p-value)
Transcription Factor Motif
EGR1
JUNB
KLF4
KLF2


B
Multmodal co-regulatory network altered in LGA
EGR1
KLF2
KLF4
Methylation
Gene-methylation score >500
Expression
Downregulated FC>0.5
Downregulated 0<FC<0.5
Upregulated FC>0
No change
Genes co-regulated
EGR1 / KFL4
EGR1 / KFL2
KLF2 / KFL4
TF motif ass target
Transcription Factors
Target Gene


A
% of colonies
CFU-GEMM
BFU-E
CFU-GM
B
BFU-E/CTRL
BFU-E/LGA
200µm
C
Cell distributions (%)
CTRL
LGA
* p-value < 0.05
*** p-value < 0.001
LT-HSC
HSC
MPP/LMPP
Myeloid
Lymphoid
Erythro-Mas
D
Pseudotime
Mk/Er
Erythro–Mas
Myeloid
MPP/LMPP
Lymphoid
HSC
LT–HSC
Differentiation
UMAP 1
UMAP 2
E
Cumulative percentage of cells
Pseudotime
Lineage distribution
Lineages
LT-HSC
HSC
MPP/LMPP
Myeloid
Lymphoid
Erythro-Mas
F
Epigenomic Alterations
HSCs
DNA methylation
Quiescence
Chromatin accessibility
LGA neonate
Transcriptomic Alterations
EGR1
KLF4
KLF2
Gene expression
TF binding
Functional Alterations
Population shift
MPPs
Quiescence
Differentiation
Colony forming capacity
JUN
ARID5
FOSB
Regulons activity