Excessive fetal growth affects HSC quiescence maintenance through epigenetic programming of EGR1 transcriptional network

Running title: Early epigenetic influence on HSC quiescence maintenance

Alexandre Pelletier^{1,3,*}, Arnaud Carrier^{1,3,*}, Yong Mei Zhao², Mickaël Canouil^{1,3}, Mehdi Derhourhi¹, Emmanuelle Durand¹, Lionel Berberian-Ferrato¹, John Greally⁴, Francine Hughes⁵, 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@cns.fr; (+33) 374008101

Summary

Fetal development is a critical period to shape stem cell identity and functions. Detrimental environments during this period are associated with epigenetics alteration of hematopoietic stem and progenitor cells (HSPC) with unknown functional impacts. We implemented a single-cell resolution integrative analysis combining epigenomics, transcriptomics, and functional data to elucidate the epigenetic influence associated with excessive fetal growth on HSPCs. We showed that hematopoietic stem cells (HSC) from large for gestational age neonates present a coordinated DNA hypermethylation and decrease expression for genes of the EGR1 transcriptional network including *SOCS3*, *KLF2*, and *JUNB* known to sustain stem cell quiescence and pluripotency. Furthermore, these changes were associated with a decreased ability for HSCs to stay undifferentiated and a decreased ability to expand in response to stimulation. Taken together, these results show that fetal overgrowth affects hematopoietic stem cells quiescence maintenance program through an epigenetic programming of the EGR1 related transcriptional network.

Introduction

The hematopoietic stem cells (HSC) are involved in essential processes such as angiogenesis, 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 action of various stressors like aging^{5, 6} can result in alteration of HSC destiny, and ultimately HSC's functions. Although HSCs have been extensively studied, the early mechanisms that control their long-term functions are not well understood, in part due to the extensive heterogeneity in phenotypes and behaviors present in HSCs.⁷

The perinatal period, a time of rapid growth and cell differentiation, represents a critical window for tissue development. Detrimental environmental exposures during this period can lead to long-term consequences as exemplified by the concept of fetal programming of adult chronic diseases.^{8, 9} Extremes of fetal growth, such as large for gestation age (LGA), have been associated with increased susceptibility to obesity, hypertension, and hypertriglyceridemia later in life.¹⁰⁻¹⁴ Even if the number of infants born LGA has increased in the recent years, from less than 1% to 14.9% in developing countries¹⁵, the long-term impact of being born LGA is not completely understood.

Fetal growth impacts the number of circulating CD34+ HSCs^{16, 17} in human. In mice, it has been shown that maternal high-fat diet limits fetal hematopoietic stem and progenitor cells (HSPC) expansion and repopulation ability while inducing myeloid-biased differentiation.¹⁸ Only few molecular studies have addressed the impact of early deleterious environment on the HSPCs development and function in human. In a previous study, we found a global increase of DNA methylation in cord blood derived

CD34+ HSPCs from LGA infants¹⁹, suggesting that epigenetic modifications may play a role in the association between early-life exposures and induce life-long changes within the hematopoietic system. Still, the molecular and functional impacts of these early epigenetic alterations on human HSPCs remained to be elucidated.

To identify signalling pathways under epigenetic influences and characterize the impact of early environment on HSPCs functions, we conducted a combined epigenomic, transcriptomic, and functional analysis on human cord blood derived CD34+ HSPCs from either appropriately grown (CTRL) or LGA neonates in both steady state and in response to extrinsic stimuli (stimulated state). We significantly increased our initial DNA methylation dataset¹⁹ and developed novel analytical approach to improve integration of the epigenomic and transcriptomic data. CD34+ HSPCs are a relatively heterogeneous cell population²⁰ and the interaction between DNA methylation and gene expression depends on cell specific genomic context.^{21, 22} Therefore, we performed single-cell RNA sequencing analysis. Finally, we challenged HSPCs *in vitro* to measure the impact of early epigenetic programming on stem cell capacities to differentiate and proliferate. We found that the epigenetic programming associated with extreme foetal growth affects gene expression across EGR1 associated transcriptional networks impairing HSC maintenance and expansion capacity.

Results

Extreme fetal growth is associated with hypermethylation of key genes regulating HSPCs proliferation and differentiation

In this study, we added 32 new cord-blood derived human CD34+ HSPC samples to the DNA methylation analysis, in addition to our initial DNA methylation dataset ($n=40$).¹⁹ First, we independently retrieved in this new set of data the global DNA hypermethylation initially found in LGA compared to controls (**Supplemental Figure 1**). Thus, we pooled both data and detected a total of 4815 differentially methylated CpGs (DMC) with 4787 CpGs hypermethylated and 28 CpGs hypomethylated in LGA compared to CTRL (p -value <0.001 and $|\text{methylation change}|>25$; **Figure 2A**, **Supplemental Table 1**).

To determine which pathways could be affected by these epigenetic changes, we used gene set enrichment analysis (GSEA). However, a major issue in performing DNA methylation gene set testing is how to assign DMCs to specific genes. Thus, we developed a novel methylation gene-score that better reflects the influence of DNA methylation changes on gene expression than considering only CpG methylation change or significance of this change (see Supplemental Method section, **Supplemental Figure 2**). For this methylation gene-score, we defined gene-CpG association based on both distance and cell-type specific regulatory information for CD34+ HSPCs and summarized this information at gene level. We obtained for each gene ($n=24857$, **Supplemental Table 2**) a unique methylation gene-score recapitulating any change in DNA methylation associated with LGA condition.

This methylation gene-score was used to rank genes and perform GSEA. Using the gene ontology (GO) reference database, we found a significant enrichment for signaling regulating fetal development as well as for key stem cell pathways such as Wnt signaling, cell fate specification, and cell fate commitment pathways (adjusted p -value <0.01 , **Figure 2B**). Furthermore, considering as a reference the putative

susceptibility gene for each clinical traits included in the GWAS catalog²³, we found significant enrichment of the differentially methylated genes (adjusted p -value<0.01, **Figure 2C**) for blood related traits (e.g. Red blood cell count, Lymphocyte counts, Monocyte count) and for several metabolic parameters (e.g. Triglyceride levels, Birth weight, Obesity-related traits) indicating that DNA methylation changes target genes influencing hematopoiesis and metabolism.

Finally, to identify regulatory factors likely impacted by these DNA methylation changes, we performed transcription factor (TF) motif analysis using HOMER. Considering proximal regions surrounding each DMCs (± 20 bp), we found significant enrichment for 23 TF motifs (adjusted p -value<0.05, **Figure 2D**). Among them, we found several members of the Kruppel-like factors and specificity protein (KLF/SP) family: KLF14, KLF5, KLF1, KLF6, SP1, SP2, and SP5 that are known to bind to GC-rich regions as well as known to be involved in the regulation of hematopoietic stem cells pluripotency and self-renewal.²⁴⁻²⁷ We also found a significant enrichment for EGR1 motif, another key TF controlling proliferation and activation of HSCs²⁸ suggesting that EGR1 regulatory network may be impacted by early epigenetic programming in LGA neonates.

Epigenetic programming impacts quiescence associated genes in hematopoietic stem cell subpopulation

To get further biological insight into the epigenetic reprogramming observed in LGA HSPCs, we performed single-cell transcriptomic analysis in two conditions steady state and stimulated state to challenge quiescence and proliferation. We first created a hematopoietic reference map (*i.e.* hematomap) by integrating CD34+ cells ($n=18520$) from 7 control neonates (**Figure 3A**). Based on cluster specific gene

expression, we identified 18 distinct clusters representative of major lineages (Long-Term HSC, HSC, Multi-Potent progenitor, Lymphoid, Myeloid, and Erythroid) of the hematopoietic compartment (**Figure 3B, Supplemental Figure 3**). This heatmap was then used as reference to annotate cells through our different conditions to enable lineage specific transcriptomic analysis.

We first analyzed differentially expressed genes (DEGs) at steady state. We integrated single-cell transcriptomic data from 6 LGA ($n=16791$ cells) to the 7 CTRL samples describe above. We only detected few DEGs among the lineages (only one DEGs for Erythroid and Myeloid lineages, but none for the other lineages, adjusted p -value <0.05 and $|\log_2\text{Fold Change (FC)}|>0.5$) indicating no major gene expression changes between CTRL and LGA at steady state (**Supplemental Figure 4**).

We then compared gene expression between steady state cells and stimulated state cells to identify transcriptomic differences between CTRL and LGA upon stimulation of the CD34⁺ HSPCs. First, we validated the cell stimulation considering steady state cells and stimulated state cells from the same CTRL samples ($n=3$ samples representing a total of 6776 steady state cells and 1749 stimulated state cells, **Supplemental Figure 5A, Supplemental Figure 5B**). We identified a consistent response to stimulation with 1518 differentially expressed genes associated to the stimulated state (stimulation signature gene set, adjusted p -value <0.05 and $|\log_2\text{FC}|>0.5$; **Supplemental Table 3**). Using GO Biological Process reference, we demonstrated that the upregulated genes from this signature ($n=1075$) were enriched in key pathways regulating HSPCs differentiation as well as response to stress and extrinsic stimulation (adjusted p -value <0.05 ; **Supplemental Figure 5C, Supplemental Table 4**). Similarly, using KEGG reference, we identified MAPK, FoxO,

TNF, and NF-kappa B signaling pathways known to be involved in the immediate-early response process, further validating our stimulation protocol (adjusted p -value<0.05; **Supplemental Figure 5C, Supplemental Table 4**). Stimulation was further validated looking at cell cycle with increase in cell proportion in S and G2M phases (**Supplemental Figure 5D**). We then performed differential expression analysis across lineages comparing steady state and stimulated state HSPCs and we found a significant enrichment in DEGs in the HSC population compared to the other lineages (p -value< 2.2×10^{-16} , Chi-Squared Test; **Supplemental Figure 5E**). These findings are consistent with a response to stimulation associated to a transition from quiescence to proliferation in HSCs.

Then, we extended our analysis by integrating single-cell transcriptomic data from 6 LGA and 8 CTRL stimulated HSPC samples representing 6861 and 5823 cells, respectively, and conducted a lineage specific analysis. Focusing on the 1518 DEGs previously identified in response to stimulation, we observed a global similar pattern of change in expression in both LGA and CTRL comparing steady state and stimulated state within each group especially in HSCs ($\rho=0.7$ and p -value< 2.2×10^{-16} , Spearman's rank correlation; **Figure 4A; Supplemental Figure 6**). However, when comparing stimulated LGA HSC to stimulated CTRL HSC, we observed in stimulated LGA a shift toward downregulated genes ($n=285$ downregulated genes over 373 DEGs, adjusted p -value<0.05 and $\log_2FC < (-0.5)$, **Figure 4A; Supplemental Table 5**), especially for key genes of the immediate early response promoting quiescence and self-renewal such as *SOCS3*, *EGR1*, *KLF2*, *DUSP2*, *JUNB*, *ID1*, *PLK3*, and *ZFP36* (**Figure 4B**). This shift toward downregulation of expression was also observed at pathway level with MAPK signaling pathway and TNF signaling pathway being now enriched in downregulated genes (adjusted p -value<0.05; **Supplemental Figure 7**).

Taken together, these observations suggest that the LGA condition is associated with an alteration of the stem cell response to stimulation at molecular level especially in HSC lineage.

We then analyzed the correlation between DNA methylation and gene expression changes using the methylation gene-score to compare DNA methylation and expression at similar gene-level resolution. For expression data, we considered gene expression fold change in stimulated state (stimulated LGA vs stimulated CTRL) as no differences were found in steady state. DEGs also found in the stimulation signature have a higher methylation gene-score (p -value<0.001) compared to other DEGs or non-DEG genes (**Figure 4C**). More specifically, the downregulated genes involved in the regulation of proliferation and differentiation of HSPCs such as *SOCS3*, *PLK3*, *EGR1*, *KLF2*, *ZFP36*, *DUSP2*, and *JUNB* were hypermethylated according to their methylation gene-score (**Figure 4D**) indicating that early epigenetic programming observed in LGA was able to modulate expression of target genes regulating stem cell activation, proliferation, and differentiation processes.

Epigenetic programming affects key regulons essential to HSCs quiescence and differentiation

A challenge when modeling epigenetic influences on gene expression is the identification of upstream regulators. Transcription factors (TF) have been often suggested as such, with key role in modulating gene expression as well as demonstrated susceptibility to DNA methylation.²⁹ TFs are also essential in regulating HSCs function and differentiation with well-established lineage specific TFs.³⁰ We therefore performed coregulatory network analysis to identify co-regulated genes, *i.e.* regulons, and performed analysis of cis-regulatory motif to assign TF to each regulon

(SCENIC). We observed a lineage specific pattern of regulons associated with key hematopoietic TFs such as SPI1, GATA1, GATA2, GATA3, MEIS1, TAL1, TCF3, EGR1, CEBPB, HOXB4, and STAT1/3 (**Supplemental Figure 8, Supplemental Table 6**).

To identify master regulators associated with the changes in gene expression observed in stimulated LGA HSCs, we used the AUCell algorithm proposed by SCENIC to score the activity of the entire regulon in each cell. We observed six regulons with a significant decrease in activity in the stimulated LGA HSC population (adjusted p -value <0.001 and $|\text{activity score difference}|>0.04$, **Supplemental Table 7**). These regulons were associated to ARID5A, EGR1, KLF2, KLF4, FOSB, and JUN consistent with the hypothesis that regulatory mechanisms underlying proliferation and differentiation of HSCs were affected (**Figure 5A**).

We then assessed if the correlation observed between DNA methylation and expression was preserved at the regulon level. Thus, we performed GSEA using our list of regulons as reference, to first identify regulons enriched for differentially methylated genes. We ranked genes by their methylation gene-score comparing LGA vs CTRL samples. Among the enriched regulons (adjusted p -value <0.01 and normalized enrichment score >1.6 , $n=33$), we found the regulons associated with ARID5A, EGR1, FOSB, JUN, KLF2, and KLF4. We also found enrichment for key HSPC specific regulons such as ATF3 and ATF4 regulating HSC stress response;^{31, 32} CEBPA and SPI1 promoting myeloid differentiation;³³ and HOX family (HOXA9, HOXA10, HOXB4) promoting HSC expansion.³⁴⁻³⁶ Interestingly, we identified master functional modules (*i.e.*, network of co-regulated regulons) considering the top enriched regulons for hypermethylated genes ($n=33$; **Figure 5B**). The principal module

included ARID5A, EGR1, FOSB, JUN, JUNB, KLF2, KLF4, and KLF10 highlighting how these key regulons are co-regulated and altered in LGA. We then identified another module based on HOXA9, HOXA10, and HOXB4. The last module including FOXJ3, FOXO1, FOXP1, IRF8 and XBP1 is more involved in proliferation and lineage specific differentiation process.

Furthermore, considering only the regulons enriched for differentially methylated genes ($n=33$), we looked for enrichment on DEGs. We observed that our top affected regulons ARID5A, JUN, FOSB, KLF4, KLF2 and EGR1 were enriched in DMCs and in downregulated genes (adjusted p -value <0.01 , **Figure 5C**). Looking at a gene regulatory network constructed based on the top affected regulons (**Figure 5D**), we confirmed that *SOCS3*, *ID1*, *DUSP2* and *ZFP36*, genes previously identified as downregulated in stimulated LGA (**Figure 4B**), were targets of these affected regulons. Interestingly, SPI1 and XBP1 regulons associated with HSPC expansion were enriched in DMCs and in upregulated genes (adjusted p -value <0.01 , **Figure 5C**). Together, these results confirm at regulon level, the correlation between change in expression and change in DNA methylation observed at gene level. They also demonstrate an alteration of signaling pathways involved in the balance between stem cell quiescence and differentiation, notably through programming of the early immediate response regulatory network, in LGA.

Epigenetic programming is associated with a HSC shift toward more differentiated cells

HSPCs represent a population of cells from progenitors to progressively restricted cells of the erythroid, myeloid, or lymphoid lineages. To follow cell distribution through these levels of differentiation and assess the influence of the LGA environment we

used Monocle.³⁷ Monocle generates a pseudotime, *i.e.* a measure that reflects how far an individual cell is in the differentiation process. Collecting these pseudotimes across our different cell populations, we observed a positive correlation between pseudotime and lineage differentiation as expected ($r=0.99$, Pearson correlation, **Figure 6A, Supplemental Figure 9A**). We then compared the distribution of the pseudotime between LGA and CTRL using long-term HSCs as roots. At steady state, no major differences were observed (**Supplemental Figure 9B**). However, upon stimulation, we observed at population level an increase in pseudotime in stimulated LGA (p -value <0.001), suggesting that in response to stimulation LGA cells fail to maintain their pool of progenitor cells. Indeed, we observed a decrease in number of cells presenting pseudotime associated with the HSC state in stimulated LGA (p -value <0.05) and a shift toward cells presenting elevated pseudotime suggesting that stimulated LGA HSCs leave quiescence but failed to properly self-renew compared to stimulated CTRL HSCs (**Figure 6B**).

Epigenetic programming is associated with alteration of the hematopoietic compartment function and subpopulation distribution

To further assess the influence of LGA exposure on stem compartment homeostasis, we monitored cell populations distribution across conditions at molecular resolution using our single-cell expression dataset. At steady state, no significant changes in lineage distribution were observed (**Supplemental Figure 10**). After stimulation, we observed a decrease in HSC cells (p -value $=0.015$) and a trend toward increased MPP cells (p -value $=0.13$, **Figure 6C**). Such a shift in cell populations is likely to reflect the altered balance between quiescence and differentiation in LGA HSCs.

Finally, we tested *in vitro* differentiation and proliferation capacities of HSPCs using colony forming unit (CFU) assays on 4 CTRL and 4 LGA samples. After 14 days of expansion, colonies were classified in 3 categories: common myeloid progenitors (CFU_GEMM), erythroid progenitors (BFU-E), and granulocyte-macrophage progenitors (CFU_GM) based on the morphology of each colony. We observed a significant decrease in the number of common myeloid progenitors in LGA samples (p -value<0.05; **Figure 6D**) as well as striking differences in shape and size of more differentiated colonies (**Figure 6E**) validating that HSPCs from LGA were less able to expand and confirming the functional impact of the early epigenetic programming.

Discussion

Hematopoietic stem cells differentiation and self-renew rely on a synergic interplay between genetically encoded signaling as well as cell-intrinsic and cell-extrinsic factors. The integration of these inputs toward a coherent lineage specific differentiation program is thought to rely on epigenetic modifiers.³⁸ In our previous study¹⁹, we considered bulk DNA methylation and we suggested an alteration of this program in response to extreme fetal growth. Here, our integrative analysis combining single-cell expression, DNA methylation, and *in vitro* data allowed us to build, at cell subpopulation resolution, a model around EGR1 associated networks in which early epigenetic modifications are linked to defective HSC self-renewal and differentiation capacities in neonates exposed to excessive fetal growth (**Figure 7**). This model, based on human samples, demonstrated how early epigenetic programming influences HSC response to further extrinsic stimulation with consequences toward quiescence maintenance. It presents a possible mechanism to explain the association found between LGA and increased susceptibility to adult chronic diseases. Indeed,

fine-tuning of HSC quiescence mechanisms is of crucial relevance for correct hematopoiesis. While, not responsive dormant HSCs would lead to hematopoietic failure due to a lack of differentiated blood cells. Highly responsive HSCs, as identified in LGA, would get to exhaustion of the population and lack of long-term maintenance of the hematopoietic system.³⁹

We demonstrated correlated increase of DNA methylation and decrease of gene expression associated with extreme fetal growth, targeting interconnected regulons under the influence of key transcription factors known to regulate HSC self-renew and differentiation ARID5A, EGR1, FOSB, JUN, JUNB, KLF2, and KLF4. ARID5A, a quiescence associated transcription factor⁴⁰ interacts with STAT3 and IL6^{41, 42}, two key regulators of HSC self-renewal and proliferation, respectively.^{43, 44} EGR1, FOSB, and JUN are member of the immediate-early response transcription factor family with essential roles in stress response and in differentiation⁴⁵. EGR1 promotes quiescence in HSCs.²⁸ JUNB regulates myelopoiesis⁴⁶, HSC proliferation and differentiation.⁴⁷ A recent study in human has demonstrated that downregulation of *JUN* promotes HSC expansion.⁴⁸ FOS members are thought to be gatekeepers to HSC mitotic entry⁴⁹ with *FOSB* being downregulated in highly proliferative HSCs.⁵⁰ Finally, KLF family is implicated in key stem cell functions. KLF4 is the most well-known factor of this family due to its role in reprogramming somatic cells into induced pluripotent stem cells suggesting a function in preservation of stemness.⁵¹ Thus, our approach supports a model in which the influence of early exposure to deleterious environment (LGA) on HSPCs is mediated through an epigenetically programmed down regulation of key signaling pathways involved in HSC quiescence maintenance.

Interestingly, ARID5A, EGR1, and KLFs are known to interact with epigenetic regulators. ARID5A is a subunit of the SWI/SNF family of chromatin-remodeling complexes implicated in the control of cell differentiation and lineage specification. EGR1 forms a complex with the DNA methyltransferase 3 responsible for targeted *de novo* methylation.⁵² While it was reported that TET2 binds to KLF4 through protein-to-protein interaction to drive locus-specific demethylation during reprogramming of B cells into iPSCs.⁵³ These observations further support the interaction between epigenetic modifications and regulatory activities of the transcription factors founds in our study.

We then challenged differentiation and proliferation of HSPCs *in vitro* and demonstrated that the LGA associated epigenetic programming was linked to an alteration of HSC preservation and expansion. Even if we did not directly measure HSC quiescence, the study of the cell populations at molecular level reveals a decrease in HSC population associated with an increase of multipotent progenitor population suggesting that upon stimulation, HSCs are prone to leave quiescence toward differentiation failing to properly maintain the initial pool of HSCs. Alterations at progenitor level were further confirm using lineage specific cytokine exposure (CFU assays). The decreased in common myeloid progenitors (CMPs) colonies observed in LGA may be associated with a decrease in the initial HSC cells proportion or with an alteration of their capacity to respond to stimulation leading to decreased proliferation as CMPs originate from HSCs. These findings corroborate previous studies on developmental programming of the hematopoietic system.^{16, 17} A reduction in self-renewal of HSPCs and increase differentiation in both lymphoid and myeloid lineages have been observed in a mouse model of maternal obesity.¹⁸ These effects may drive long-term consequences in human health as illustrated by the study performed by

Kotowski et *al.* in which the integrity of the hematopoietic system in neonates was associated with susceptibility to onset of hematopoietic pathologies, that can be further associated to age-related diseases.⁵⁴

Interestingly, the LGA associated epigenetic influence at the transcriptomic and functional levels was exacerbated upon stimulation of the HSC compartment. Indeed, in the absence of environmental stimuli (steady state), no major difference was detected at transcriptomic level as HSCs linger in a quiescent state. However, under extrinsic stimulation (stimulated state), HSCs adaptative response was altered leading to a decrease in early progenitors' pool likely to impact long-term system function. This observation fits with the concept of developmental programming which relies not only on early impairment of organ development but most importantly on a decreased adaptability to further environmental challenges that will enhance disease susceptibility and highlights that an interaction between epigenetic programming and further environmental factors is needed to trigger disease.

In summary, we provided a comprehensive model recapitulating the epigenetic early programming of the hematopoietic system and its influence on the hematopoietic stem cell fitness in response to stimulation. This study by identifying epigenetic modifications associated with gene expression and functional alterations contribute to elucidate how early exposure can affect long-term tissue maintenance and possibly increase risk to chronic diseases.

Acknowledgments

The authors thank the UMR 8199 LIGAN-PM Genomics platform (Lille, France) which belongs to the 'Federation de Recherche' 3508 Labex EGID (European Genomics

Institute for Diabetes; ANR-10-LABX-46) and was supported by the ANR Equipex 2010 session (ANR-10-EQPX-07-01; 'LIGAN-PM'). The LIGAN-PM Genomics platform (Lille, France) is also supported by the FEDER and the Region Nord-Pas-de-Calais-Picardie. This project is cofunded in the frame of CPER CTRL program by the European Union - European Regional Development Fund (ERDF), 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.

Figure Legend

Figure 1: Overview of the experimental design.

Figure 2: LGA is associated with DNA hypermethylation targeting key stem cell signaling pathways. (A) Volcano plot of DNA methylation score differences for LGA compared to CTRL based on 754931 loci throughout the genome. Differentially methylated loci with p -value < 0.001 and $|\text{methylation difference}| > 25$ are shown in red. (B) Network representation of GO pathways. Significantly enriched pathways were identified using GSEA based on the methylation gene-score. Edges represent interactions between pathways. (C) Gene-set enrichment analysis for susceptibility genes associated to pathophysiological traits based on GWAS catalog. Dot plot representing enrichment for selected relevant traits. Gene ratio corresponds to the ratio between the total number of genes associated to the trait (using GWAS catalog) and the number of differentially methylated genes. (D) Dot plot representing enrichment for TF binding motifs using HOMER considering a $\pm 20\text{bp}$ region around DMCs. Dots are color coded based on significance of the enrichment and y-axis represent the number of regions with binding motif among DMCs.

Figure 3: HSPCs lineage annotation using single-cell RNAseq. (A) Hematomap, UMAP representation of distinct HSPC lineages. (B) Dot plot representing key markers

used to annotate cell populations. LT-HSC, long-term hematopoietic stem cell; HSC, hematopoietic stem cell; MPP, multipotent progenitor; LMPP, lymphoid-primed multipotent progenitors; Erythro-Mas, erythroid and mast precursor; Mk/Er, megakaryocyte and erythrocyte; DC, dendritic cell.

Figure 4: Response to stimulation is altered in LGA. (A) Heatmap representing change in expression for stimulation associated DEG ($n=1518$) in HSC population. (B) Volcano plot representing DEG comparing LGA vs CTRL in stimulated state. Differentially expressed genes with adjusted p -value <0.05 and $|\log_2FC| >0.5$ are shown in red if also part of the stimulation associated DEGs, otherwise in pink. (C) Boxplots representing methylation gene-score distribution associated to non DEG, DEG and stimulation associated DEG (p -value <0.001 , Wilcoxon test). (D) Scatter plot representing the correlation between change in expression and change in methylation. x-axis represents the $\log_2FC$ of expression in stimulated HSCs and y-axis represents the percentile of methylation gene-score comparing LGA vs CTRL. Differentially expressed gene with adjusted p -value <0.05 and $|\log_2FC| >0.5$ are shown in red if also part of the stimulation associated DEGs, otherwise in pink.

Figure 5: Epigenetic programming of HSC specific regulons altered in LGA neonates. Regulons and TF target information were obtained through the SCENIC workflow. (A) Boxplots representing regulon activity score in stimulated CTRL and stimulated LGA HSC lineage. Barplot representing the change in regulon activity and significance comparing stimulated LGA vs stimulated CTRL. Only significantly affected regulons are represented (adjusted p -value <0.001 and $|\text{activity score difference}| >0.04$). (B) Network analysis representing regulons interactions considering regulons

significantly enriched for differentially methylated genes (adjusted p -value<0.01, GSEA). Edges represent overlap between regulons (a minimum of 20% overlap is required to connect regulons). (C) Dot plot representing enrichment for DEGs in regulons enriched for differentially methylated genes. (D) Network representing interactions between target genes and transcription factors considering our top affected regulons ARID5A, EGR1, KLF2, KLF4, FOSB and JUN. Each dot represents a gene within the network, triangle represents transcription factor, arrow represents interaction between transcription factor and target genes, shapes are color coded to reflect the change in methylation gene-score, and DEGs are labeled in red. Size of the shape represents the number of interactions. Only genes with 2 or more interactions are represented.

Figure 6: LGA is associated with a HSC shift toward more differentiated cells and decrease expansion capacity.

(A) UMAPs representing pseudotimes across lineages. (B) Box plots representing the cumulative percentage of cells per pseudotime in stimulated CTRL and stimulated LGA. Boxplots in the vignette represent overall pseudotime distribution in stimulated CTRL and stimulated LGA. Density plots correspond to cell populations distribution across pseudotimes. (C) Dot plot representing change in cell distribution comparing stimulated LGA vs stimulated CTRL. Dot sizes represent for each lineage the percent distribution of total distribution. Boxplots representing the cell distribution across hematopoietic main lineages in stimulated CTRL and stimulated LGA. (D) Bar plot representing colonies distribution after CFU assays. (E) Representative capture of colonies morphological differences found in CTRL and LGA. LT-HSC, long-term hematopoietic stem cell; HSC, hematopoietic stem cell; MPP, multipotent progenitor; LMPP, lymphoid-primed

476 multipotent progenitors; Erythro-Mas, erythroid and mast precursor; Mk/Er,
477 megakaryocyte and erythrocyte; DC, dendritic cell; CFU-GEMM, common myeloid
478 progenitors; BFU-E, erythroid progenitors; CFU-GM, granulocyte-macrophage
479 progenitors.

480
481 **Figure 7: Model recapitulating the epigenomic, transcriptomic, and functional**
482 **influence of extreme fetal growth on HSCs.** HSC, hematopoietic stem cell; MPP,
483 multipotent progenitor.

484 485 **Methods**

486 See the Supplemental Methods for additional information. Overview of the study
487 design is presented in Figure 1.

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

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

Isolation of CD34+ HSPCs. Mononuclear cells were separated using PrepaCyte-WBC following which CD34+ cells were obtained by positive immunomagnetic bead selection, using the AutoMACS Separator (Miltenyi Biotech). Cells were cryopreserved in 10% dimethyl sulfoxide using controlled rate freezing upon analysis.

Genome-Wide DNA methylation assay. DNA methylation levels for >1.7M CpGs were obtained using the HELP-tagging assay as previously described.⁵⁵

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

HSPCs extrinsic stimulation. After cell counting and viability check and prior to cell suspension loading on the Chromium 10x Genomics controller, cell hashtag (HTO) staining was used following the cell-hashing protocol.⁵⁶ This protocol resulting in extended cell manipulation was used to mimic response to extrinsic stimulation. Response to stimulation was validated at transcriptomic level.

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

Data processing and statistical analysis. For DNA methylation analysis, low quality CpGs were filtered out based on detection rate and confidence score. 754931 out of 1709224 CpGs were conserved for further analysis. Linear regression and statistical modeling using LIMMA R package⁵⁷ were used to identify differentially methylated

CpGs (DMC). To link gene to CpGs and performed gene set enrichment analysis (GSEA), we generated a methylation gene-score; first by scoring association between individual CpG and unique gene, based on evidence for association between gene and CpG (distance, eQTL) and on belonging in regulatory regions; and then, by concatenating CpG scores across each gene. We assessed enrichment for biological pathways performing GSEA using the ClusterProfiler package.⁵⁸ We performed transcription factor (TF) motif enrichment analysis using HOMER tool⁵⁹ considering a 20bp region around the DMCs.

For single-cell RNAseq (sc-RNAseq) analysis, data were preprocessed using the CellRanger count pipeline. Data filtering, normalization and integration as well as cluster identifications were performed using Seurat (v4) pipeline. Pseudo-bulk differential expression analysis between LGA and CTRL cells within each hematopoietic lineage was performed using DESeq2 R package.⁶⁰ Over representation test was performed on differentially expressed genes (DEGs) using enrichGO and enrichKEGG of the ClusterProfiler Package. The SCENIC workflow⁶¹ was used to identify co-regulated modules associated to a TF (regulons) and to generate cell-specific activity score for each regulon. Differentiation trajectory analysis and pseudotime attribution were conducted with Monocle.³⁷ The code to perform the analyses in this manuscript is available at https://github.com/umr1283/LGA_HSPC_PAPER.git.

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

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.
6. Chandel NS, Jasper H, Ho TT, Passegue E. Metabolic regulation of stem cell function in tissue homeostasis and organismal ageing. *Nat Cell Biol*. 2016; **18**(8): p. 823-32.
7. Guenechea G, Gan OI, Dorrell C, Dick JE. Distinct classes of human stem cells that differ in proliferative and self-renewal potential. *Nat Immunol*. 2001; **2**(1): p. 75-82.
8. Barker DJ, Osmond C. Infant mortality, childhood nutrition, and ischaemic heart disease in England and Wales. *Lancet*. 1986; **1**(8489): p. 1077-81.
9. Barker DJ, Osmond C, Kajantie E, Eriksson JG. Growth and chronic disease: findings in the Helsinki Birth Cohort. *Ann Hum Biol*. 2009; **36**(5): p. 445-58.
10. Cnattingius S, Villamor E, Lagerros YT, Wikstrom AK, Granath F. High birth weight and obesity--a vicious circle across generations. *Int J Obes (Lond)*. 2012; **36**(10): p. 1320-4.
11. Wang Y, Gao E, Wu J, et al. Fetal macrosomia and adolescence obesity: results from a longitudinal cohort study. *Int J Obes (Lond)*. 2009; **33**(8): p. 923-8.
12. Eriksson J, Forsen T, Tuomilehto J, Osmond C, Barker D. Size at birth, childhood growth and obesity in adult life. *Int J Obes Relat Metab Disord*. 2001; **25**(5): p. 735-40.
13. Harder T, Rodekamp E, Schellong K, Dudenhausen JW, Plagemann A. Birth weight and subsequent risk of type 2 diabetes: a meta-analysis. *Am J Epidemiol*. 2007; **165**(8): p. 849-57.
14. Pocobelli G, Dublin S, Enquobahrie DA, Mueller BA. Birth Weight and Birth Weight for Gestational Age in Relation to Risk of Hospitalization with Primary Hypertension in Children and Young Adults. *Matern Child Health J*. 2016; **20**(7): p. 1415-23.
15. Koyanagi A, Zhang J, Dagvadorj A, et al. Macrosomia in 23 developing countries: an analysis of a multicountry, facility-based, cross-sectional survey. *Lancet*. 2013; **381**(9865): p. 476-83.
16. Al-Sweedan SA, Musalam L, Obeidat B. Factors predicting the hematopoietic stem cells content of the umbilical cord blood. *Transfus Apher Sci*. 2013; **48**(2): p. 247-52.
17. Chandra T, Afreen S, Kumar A, Singh U, Gupta A. Does umbilical cord blood-derived CD34+ cell concentration depend on the weight and sex of a full-term infant? *J Pediatr Hematol Oncol*. 2012; **34**(3): p. 184-7.
18. Kamimae-Lanning AN, Krasnow SM, Goloviznina NA, et al. Maternal high-fat diet and obesity compromise fetal hematopoiesis. *Mol Metab*. 2015; **4**(1): p. 25-38.
19. Delahaye F, Wijetunga NA, Heo HJ, et al. Sexual dimorphism in epigenomic responses of stem cells to extreme fetal growth. *Nat Commun*. 2014; **5**: p. 5187.
20. Stella CC, Cazzola M, De Fabritiis P, et al. CD34-positive cells: biology and clinical relevance. *Haematologica*. 1995; **80**(4): p. 367-87.

21. Thurman RE, Rynes E, Humbert R, et al. The accessible chromatin landscape of the human genome. *Nature*. 2012; **489**(7414): p. 75-82.
22. Delahaye F, Do C, Kong Y, et al. Genetic variants influence on the placenta regulatory landscape. *PLoS Genet*. 2018; **14**(11): p. e1007785.
23. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019; **47**(D1): p. D1005-D1012.
24. Jiang J, Chan YS, Loh YH, et al. A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nat Cell Biol*. 2008; **10**(3): p. 353-60.
25. Zhao T, Liu C, Chen L. Roles of Klf5 Acetylation in the Self-Renewal and the Differentiation of Mouse Embryonic Stem Cells. *PLoS One*. 2015; **10**(9): p. e0138168.
26. Lecuyer E, Herblot S, Saint-Denis M, et al. The SCL complex regulates c-kit expression in hematopoietic cells through functional interaction with Sp1. *Blood*. 2002; **100**(7): p. 2430-40.
27. Gilmour J, O'Connor L, Middleton CP, et al. Robust hematopoietic specification requires the ubiquitous Sp1 and Sp3 transcription factors. *Epigenetics Chromatin*. 2019; **12**(1): p. 33.
28. Min IM, Pietramaggiori G, Kim FS, Passegue E, Stevenson KE, Wagers AJ. The transcription factor EGR1 controls both the proliferation and localization of hematopoietic stem cells. *Cell Stem Cell*. 2008; **2**(4): p. 380-91.
29. Zhu H, Wang G, Qian J. Transcription factors as readers and effectors of DNA methylation. *Nat Rev Genet*. 2016; **17**(9): p. 551-65.
30. Zhu J, Emerson SG. Hematopoietic cytokines, transcription factors and lineage commitment. *Oncogene*. 2002; **21**(21): p. 3295-313.
31. van Galen P, Mbong N, Kreso A, et al. Integrated Stress Response Activity Marks Stem Cells in Normal Hematopoiesis and Leukemia. *Cell Rep*. 2018; **25**(5): p. 1109-1117 e5.
32. Liu Y, Chen Y, Deng X, Zhou J. ATF3 Prevents Stress-Induced Hematopoietic Stem Cell Exhaustion. *Front Cell Dev Biol*. 2020; **8**: p. 585771.
33. Tenen DG, Hromas R, Licht JD, Zhang DE. Transcription factors, normal myeloid development, and leukemia. *Blood*. 1997; **90**(2): p. 489-519.
34. Antonchuk J, Sauvageau G, Humphries RK. HOXB4-induced expansion of adult hematopoietic stem cells ex vivo. *Cell*. 2002; **109**(1): p. 39-45.
35. Sun Y, Zhou B, Mao F, et al. HOXA9 Reprograms the Enhancer Landscape to Promote Leukemogenesis. *Cancer Cell*. 2018; **34**(4): p. 643-658 e5.
36. Magnusson M, Brun AC, Miyake N, et al. HOXA10 is a critical regulator for hematopoietic stem cells and erythroid/megakaryocyte development. *Blood*. 2007; **109**(9): p. 3687-96.
37. Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. *Nat Methods*. 2017; **14**(3): p. 309-315.
38. Yu VWC, Yusuf RZ, Oki T, et al. Epigenetic Memory Underlies Cell-Autonomous Heterogeneous Behavior of Hematopoietic Stem Cells. *Cell*. 2017; **168**(5): p. 944-945.
39. Wilson A, Laurenti E, Trumpp A. Balancing dormant and self-renewing hematopoietic stem cells. *Curr Opin Genet Dev*. 2009; **19**(5): p. 461-8.
40. Anwar T, Sen B, Aggarwal S, et al. Differentially regulated gene expression in quiescence versus senescence and identification of ARID5A as a quiescence associated marker. *J Cell Physiol*. 2018; **233**(5): p. 3695-3712.

41. Masuda K, Ripley B, Nishimura R, et al. Arid5a controls IL-6 mRNA stability, which contributes to elevation of IL-6 level in vivo. *Proc Natl Acad Sci U S A*. 2013; **110**(23): p. 9409-14.
42. Nyati KK, Masuda K, Zaman MM, et al. TLR4-induced NF-kappaB and MAPK signaling regulate the IL-6 mRNA stabilizing protein Arid5a. *Nucleic Acids Res*. 2017; **45**(5): p. 2687-2703.
43. Goodsell DS. The molecular perspective: estrogen sulfotransferase. *Oncologist*. 2006; **11**(4): p. 418-9.
44. Tie R, Li H, Cai S, et al. Interleukin-6 signaling regulates hematopoietic stem cell emergence. *Exp Mol Med*. 2019; **51**(10): p. 1-12.
45. Bahrami S, Drablos F. Gene regulation in the immediate-early response process. *Adv Biol Regul*. 2016; **62**: p. 37-49.
46. Passegue E, Wagner EF, Weissman IL. JunB deficiency leads to a myeloproliferative disorder arising from hematopoietic stem cells. *Cell*. 2004; **119**(3): p. 431-43.
47. Santaguida M, Schepers K, King B, et al. JunB protects against myeloid malignancies by limiting hematopoietic stem cell proliferation and differentiation without affecting self-renewal. *Cancer Cell*. 2009; **15**(4): p. 341-52.
48. Xiao X, Lai W, Xie H, et al. Targeting JNK pathway promotes human hematopoietic stem cell expansion. *Cell Discov*. 2019; **5**: p. 2.
49. Okada S, Fukuda T, Inada K, Tokuhisa T. Prolonged expression of c-fos suppresses cell cycle entry of dormant hematopoietic stem cells. *Blood*. 1999; **93**(3): p. 816-25.
50. McKinney-Freeman S, Cahan P, Li H, et al. The transcriptional landscape of hematopoietic stem cell ontogeny. *Cell Stem Cell*. 2012; **11**(5): p. 701-14.
51. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*. 2006; **126**(4): p. 663-76.
52. Cartron PF, Blanquart C, Hervouet E, Gregoire M, Vallette FM. HDAC1-mSin3a-NCOR1, Dnmt3b-HDAC1-Egr1 and Dnmt1-PCNA-UHRF1-G9a regulate the NY-ESO1 gene expression. *Mol Oncol*. 2013; **7**(3): p. 452-63.
53. Sardina JL, Collombet S, Tian TV, et al. Transcription Factors Drive Tet2-Mediated Enhancer Demethylation to Reprogram Cell Fate. *Cell Stem Cell*. 2018; **23**(5): p. 727-741 e9.
54. Kotowski M, Safranow K, Kawa MP, et al. Circulating hematopoietic stem cell count is a valuable predictor of prematurity complications in preterm newborns. *BMC Pediatr*. 2012; **12**: p. 148.
55. Suzuki M, Jing Q, Lia D, Pascual M, McLellan A, Grealley JM. Optimized design and data analysis of tag-based cytosine methylation assays. *Genome Biol*. 2010; **11**(4): p. R36.
56. Stoeckius M, Zheng S, Houck-Loomis B, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol*. 2018; **19**(1): p. 224.
57. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015; **43**(7): p. e47.
58. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012; **16**(5): p. 284-7.
59. Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*. 2010; **38**(4): p. 576-89.

60. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014; **15**(12): p. 550.
61. Aibar S, Gonzalez-Blas CB, Moerman T, et al. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods.* 2017; **14**(11): p. 1083-1086.

CTRL
neonate

LGA
neonate

CD34+ cells isolation from
cord blood samples

Steady state

Stimulated state

HTO staining

10X Genomics
scRNA-seq

in vitro analysis

14 days

Scoring

Colony Forming Unit Assay

Integrative analysis

Differential methylation analysis

Clustering

Differential expression analysis

Subpopulations distribution

Network analysis

A

C

B

D

A

B

A

Stimulation signature gene set across group in HSC
($n = 1,518$)

B

C

D