Transcriptional divergence and conservation of human and mouse erythropoiesis

by
Novalia Pishesha

B.S. Bioengineering
University of California at Berkeley, 2011

SUBMITTED TO THE DEPARTMENT OF BIOLOGICAL ENGINEERING IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN BIOLOGICAL ENGINEERING
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2014

Signature redacted

Signature of Author: ____________________
Department of Biological Engineering
May 22, 2014

Signature redacted

Certified by: ____________________
Harvey F. Lodish
Professor of Biology and Biological Engineering
Thesis Supervisor

Signature redacted

Accepted by: ____________________
Forest White
Professor of Biological Engineering
Chairman, Biological Engineering Graduate Committee

Transcriptional divergence and conservation of human and mouse erythropoiesis

by

Novalia Pishesha

Submitted to the Department of Biological Engineering on May 23rd, 2014 in Partial Fulfillment of the Requirements for the Degree of Master of Science in Biological Engineering

ABSTRACT

Mouse models have been used extensively for decades and have been instrumental in improving our understanding of mammalian erythropoiesis. Nonetheless, there are several examples of variation between human and mouse erythropoiesis. We performed a comparative global gene expression study using data from morphologically identical stage-matched sorted populations of human and mouse erythroid precursors from early to late erythroblasts. Induction and repression of major transcriptional regulators of erythropoiesis, as well as major erythroid-important proteins, are largely conserved between the species. In contrast, at a global level we identified a significant extent of divergence between the species, both at comparable stages and in the transitions between stages, especially for the 500 most highly expressed genes during development. This suggests that the response of multiple developmentally regulated genes to key erythroid transcriptional regulators represents an important modification that has occurred in the course of erythroid evolution. In developing a systematic framework to understand and study conservation and divergence between human and mouse erythropoiesis, we show how mouse models can fail to mimic specific human diseases and provide predictions for translating findings from mouse models to potential therapies for human disease.

Thesis Supervisor: Harvey F. Lodish
Title: Professor of Biology and Biological Engineering

Acknowledgements

First and foremost, I am immensely indebted to Prof. Harvey F. Lodish for the opportunity and privilege to work in his lab. He has been an amazing mentor, advisor, and friend, providing me with support, guidance, and wisdom in science and, more importantly, in life. He is a constant source of inspiration during my master studies and I have learned so much in this short 1.5 years because he has always supported me to explore beyond the boundary of the laboratory. He, indeed, has taught me to be a better person.

I would also like to thank Prof. Vijay G. Sankaran, who was instrumental in suggesting the topic for this thesis. His guidance on all aspects of its execution, as well as in navigating through graduate study, have been an invaluable gift. His energy and inexhaustible (literally!) enthusiasm are definitely contagious.

Also, I am very grateful to Prathapan Thiru, Jiahai Shi, and Jennifer Eng for their patience in teaching me various techniques as well as in providing me with biological samples used in this work.

I am also indebted to Leif Si-Hun Ludwig and Marko Knoll for their instrumental contributions via numerous discussions on this thesis project, as well as their frequent reminders to work hard.

I also thank all of the members of the Lodish Lab for providing a friendly and supportive environment.

Finally, I would like to thank my family and friends for their constant support and interest in my work and other activities (e.g. to pull me out of the lab every now and then).

Transcriptional divergence and conservation of human and mouse erythropoiesis

Novalia Pishesha[a,b], Prathapan Thiru[a], Jiahai Shi[a], Jennifer C. Eng[a], Vijay G. Sankaran[a,c,d,1], and Harvey F. Lodish[a,b,d,e,1]

[a]Whitehead Institute for Biomedical Research, Cambridge, MA 02142; [b]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02142; [e]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142; [c]Division of Hematology/Oncology, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115; and [d]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA 02142

Contributed by Harvey F. Lodish, February 3, 2014 (sent for review November 3, 2013)

Mouse models have been used extensively for decades and have been instrumental in improving our understanding of mammalian erythropoiesis. Nonetheless, there are several examples of variation between human and mouse erythropoiesis. We performed a comparative global gene expression study using data from morphologically identical stage-matched sorted populations of human and mouse erythroid precursors from early to late erythroblasts. Induction and repression of major transcriptional regulators of erythropoiesis, as well as major erythroid-important proteins, are largely conserved between the species. In contrast, at a global level we identified a significant extent of divergence between the species, both at comparable stages and in the transitions between stages, especially for the 500 most highly expressed genes during development. This suggests that the response of multiple developmentally regulated genes to key erythroid transcriptional regulators represents an important modification that has occurred in the course of erythroid evolution. In developing a systematic framework to understand and study conservation and divergence between human and mouse erythropoiesis, we show how mouse models can fail to mimic specific human diseases and provide predictions for translating findings from mouse models to potential therapies for human disease.

hematopoiesis | comparative genomics | microarray

Model organisms, in particular mouse models, have been extremely valuable for our understanding of hematopoiesis. The study of this process has provided a paradigm for understanding the molecular mechanisms for lineage-specific cellular differentiation in normal and pathologic states (1–3). Hematopoiesis has long been assumed to be largely conserved based on gross comparative studies among mammals, although the extent of this conservation is unclear. Among the various lineages in the hematopoietic system, erythropoiesis has been extensively studied, given the relevance to human disease and the morphologically distinct stages of erythroid terminal maturation. Forward and reverse genetic approaches in mouse and other model systems have helped define many key regulators of erythropoiesis (4). However, there are examples where aspects of mouse erythropoiesis are inconsistent with human erythropoiesis. For example, humans express a fetal hemoglobin not found in the mouse, and their globin gene regulation pattern differs significantly (5–7). Mutations that impair erythropoiesis in humans are often not faithfully recapitulated in mouse models (8–11). Although specific examples of such differences have been documented, no studies have examined the extent to which mouse erythropoiesis recapitulates the global molecular features of human erythropoiesis.

Hematopoietic stem cells give rise to multiple lineage-committed progenitors and precursors, including the erythroid lineage. The earliest committed erythroid progenitor is the burst-forming unit erythroid (BFU-E) (12, 13). BFU-Es undergo limited self-renewal divisions and differentiate to the colony-forming unit erythroid (CFU-E) (14). Morphologically distinguishable stages of maturation compose the process of terminal erythropoiesis that begins at the CFU-E stage and involves three to six terminal divisions (depending on the species and developmental stage); induction of erythroid-important genes, chromatin condensation; and, in mammals, enucleation. In mammals, the stages of terminal erythropoiesis consist of early (proerythroblasts), intermediate (basophilic erythroblasts), and late (polychromatophilic and orthochromatophilic erythroblasts) erythroid precursors that are morphologically distinct (Fig. S1). Several flow cytometric techniques can sort these stages based on surface protein expression, as well as RNA and DNA content (15, 16). Based on surface protein expression, both human and mouse proerythroblasts express modest levels of KIT, high levels of transferrin receptor (CD71), and low levels of glycophorin A (GlyA). Transitioning into the basophilic erythroblast stage, both human and mouse erythroid cells gradually become smaller and begin to express higher levels of GlyA. At the polychromatophilic and orthochromatophilic erythroblast stages, human and mouse erythroid cells have reduced expression of CD71, but retain a high level of GlyA. Polychromatophilic and orthochromatophilic erythroblasts contain less RNA and DNA, compared with the earlier stages, and are also smaller (17).

Recent global comparative transcriptional studies in other hematologic lineages, i.e., immune cells, have shown both conservation and divergence (18, 19). As such, it can be argued that the vast extent of divergence may be due to underlying differences in immune cell function and composition (20, 21). In contrast, human and mouse erythroid cells undergo nearly identical maturation processes that are morphologically and

Significance

Mouse models have been instrumental in advancing our understanding of blood cell production. Although many studies have suggested specific differences between human and mouse red cell production (erythropoiesis), a global study of such similarities and differences has been lacking. By computationally comparing global gene expression data from adult human and mouse erythroid precursors representing the distinct stages of maturation, we showed that, while the overall transcriptional landscape has changed, critical erythroid gene signatures and transcriptional regulators have remained conserved. Importantly, these analyses can serve as a tool to integrate data between human and mouse erythropoiesis research, explain why certain human blood diseases are not faithfully recapitulated in mouse models, and highlight hurdles in translating therapeutic findings from mice to humans.

Author contributions: N.P., V.G.S., and H.F.L. designed research; N.P., J.S., J.C.E., and V.G.S. performed research; N.P., P.T., V.G.S., and H.F.L. analyzed data; and N.P., V.G.S., and H.F.L. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]To whom correspondence may be addressed. E-mail: lodish@wi.mit.edu or sankaran@broadinstitute.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1401598111/-/DCSupplemental.

structurally indistinguishable (15, 16). However, the extent to which this similarity is matched at the molecular level remains to be determined.

Here we report on a systematic comparative study of human and mouse erythroid precursor gene expression during terminal erythropoiesis using multiple publicly available datasets to quantify and delineate the extent of conservation. Comparing gene expression at matched early, intermediate, and late stages of terminal erythropoiesis between the species as well as the conservation and divergence of transcriptional changes occurring between the developmental stages, we find that, although erythropoiesis seems to be morphologically conserved in the ~65 million years of evolution separating humans and mice, its transcriptional landscape has diverged significantly. Our results suggest that these evolutionary changes have been mediated through fine-tuning of gene expression programs that are globally conserved among the species. We also developed a framework to systematically elucidate conserved and divergent aspects of human and mouse erythropoiesis.

Results

Conservation and Divergence in Erythroid Signature Gene Expression. We began by interrogating the global gene expression profiles from early, intermediate, and late erythroid precursors from adult humans and mice (15, 16). We first examined genes whose role in erythropoiesis is well studied, including those encoding cytoskeletal proteins, transmembrane proteins, and heme biosynthetic enzymes. The expression values were mean-centered to depict relative changes across terminal erythropoiesis, independently of magnitude (Fig. 1 and Fig. S2). Red blood cell membrane and cytoskeletal proteins make up the components of the filamentous meshwork of proteins along the cytoplasmic surface of the membrane necessary for mechanical stability, normal morphology, and flexibility (22). Dominant mutations in *ANK1*, *SPTA1*, *SPTB*, *SLC4A1*, *EPB42*, and *EPB41* cause hereditary spherocytosis or elliptocytosis, characterized by poor membrane stability, abnormal morphology, and a high rate of red cell turnover (23, 24). The important function of these proteins implies that their expression must be well conserved among species. Indeed, most of the cytoskeletal proteins show similar patterns of expression in both species. *EPB41*, *TMOD1*, *SPTA1*, *EPB42*, *EPB49*, and *SPTB* increased with maturation in both species, but *ADD3* decreased. However, divergence also exists as shown by several genes that are differentially regulated. Human *ACTB* and *TNFRSF1A* show no significant changes across the stages, but their mouse orthologs decreased in late stage erythroblasts. Expression of human *ANK1* and *ADD1* showed minor changes with differentiation, but their orthologs progressively increased in mice. Finally, during maturation *ADD3* expression increased in humans, but *Add3* decreased in mice (Fig. 1*A*). These findings imply that, although the composition of the red blood cell cytoskeleton is globally conserved in mammals, differences are present in the expression of certain proteins necessary for maintenance of this structural network. These differences may be relevant for our understanding of the human disorders of red blood cell membrane structure, including hereditary spherocytosis and elliptocytosis, and of the mouse models of these diseases (22).

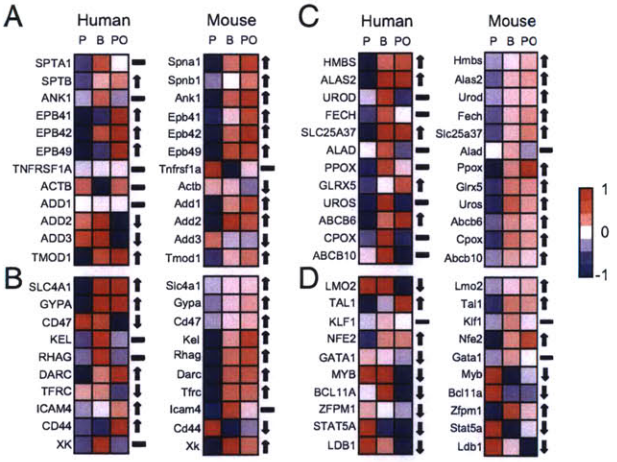


Fig. 1. Conservation and divergence in erythroid signature gene expression. Mean-centered expression of the average expression values across the early, intermediate-, and late-stage erythroblasts (red/blue color scale to the right) of signature erythroid gene groups: (*A*) skeletal proteins, (*B*) transmembrane proteins, (*C*) heme biosynthesis enzymes, and (*D*) erythroid-specific transcription factors in humans (*Left*) and mice (*Right*). P, proerythroblasts; B, basophilic erythroblasts; PO, mixture of polychromatophilic and orthochromatic erythroblasts. ↑, increasing gene expression from P to PO; ↓ decreasing gene expression from P to PO; –, no significant changes from P to PO.

Transmembrane proteins are used to track erythroid cell maturation using flow cytometry, and many compose the clinically relevant blood groups (25, 26). From a limited list (Fig. 1*B*), ~55% are up-regulated with differentiation in both human and mouse, including *GYPA*, *DARC*, and *ICAM4*. However, *CD47*, *RHAG*, *TFRC*, *XK*, and *CD44* showed divergent expression. *CD47*, *RHAG*, *TFRC*, and *XK* decreased during human terminal erythropoiesis, whereas their orthologs increased in mouse terminal erythropoiesis; *CD44* displayed the opposite pattern (Fig. 1*B* and Fig. 2*A*). We confirmed by flow cytometry the differential expression of CD44 protein on human and mouse peripheral red blood cells (Fig. 2*B*).

Genes for erythroid heme biosynthesis also revealed similarities and differences in their expression patterns. Differentially expressed genes include *CPOX*, *ABCB10*, *UROS*, *UROD*, and *FECH*, and conserved genes include *HMBS*, *ALAS2*, *SLC25A37*, *ALAD*, *PPOX*, *GLRX5*, and *ABCB6* (Fig. 1*C*). This research extends earlier experiments suggesting that the kinetics of heme biosynthesis in human and mouse terminal maturation is different (27).

Global Gene Expression Changes During Erythroid Terminal Differentiation. Despite the high conservation of the signature erythroid gene expression patterns, the global gene expression profiles of matched erythroblast stages, i.e., those that compare the expression of all 12,808 orthologous genes, showed a lower degree of conservation. The mean Pearson correlation coefficient between mRNA expression in human and mouse proerythroblasts is 0.66; basophilic erythroblasts, 0.64; and polychromatophilic/orthochromatic erythroblasts, 0.67 (Fig. 3*A*). As essential controls for these interspecies correlations and to confirm that our analyses of these datasets were valid and not skewed by different cell sources, cell purity, and laboratory techniques, we carried out multiple intraspecies and interspecies global gene expression comparisons between these datasets and other publically available datasets.

First, the global gene expression data of the mouse definitive adult datasets in our analysis showed a strong correlation with data from a set of stage-matched primary mouse fetal liver erythroblasts, with mean Pearson correlation coefficients of 0.89, 0.90, and 0.89 for proerythroblasts, basophilic erythroblasts, and polychromatophilic/orthochromatic erythroblasts, respectively (Fig. S3*A*) (28). We also compared the global expression data from adult murine bone marrow to that from the *Gata1*-null mouse cell line G1E-ER4, in which Gata1 activity was reactivated with estradiol treatment (29). After 14 and 21 h of estradiol induction, the G1E-ER4 cells differentiate into proerythroblast-like cells. A comparison of the global gene expression profile of these G1E-ER4 cells to mouse primary proerythroblasts yields a strong mean Pearson correlation coefficient of 0.86 (Fig. S3*B*). At 30 h postinduction, the G1E-ER4 cells resemble basophilic erythroblasts,

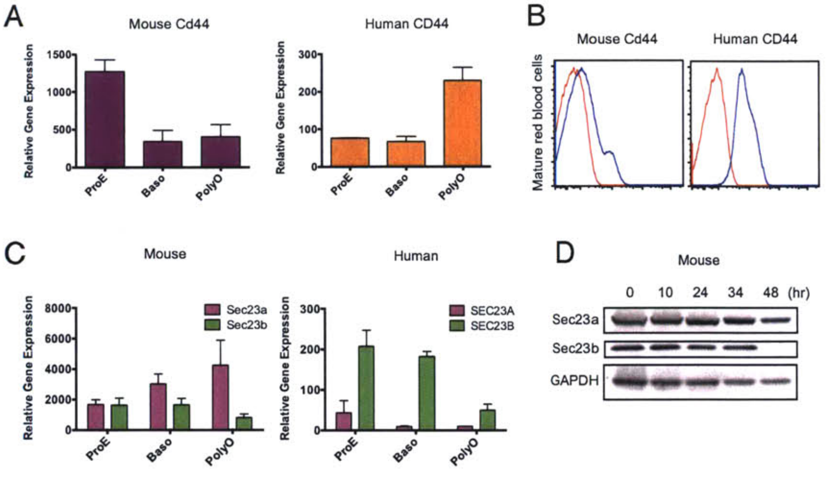


Fig. 2. Differentially regulated genes in human and mouse erythroblasts. (*A*) *CD44* and *Cd44* gene expressions mined from the comparative gene expression framework. (*B*) Flow cytometry of the predicted CD44 expression difference by staining human and mouse peripheral blood (isotype controls are shown in red; CD44 antibody staining is in blue). (*C*) *SEC23B*, *SEC23A*, *Sec23a*, and *Sec23b* gene expression mined from the comparative gene expression framework. (*D*) Western blot of Sec23a and Sec23b at all stages of terminal erythropoiesis in mice fetal liver progenitor cells after 0, 10, 24, 34, and 48 h in culture. ProE, proerythroblasts; Baso, basophilic erythroblasts; PolyO, mixture of polychromatophilic and orthochromatic erythroblasts.

and a comparison of the global gene expression profile of these cells and mouse basophilic erythroblasts yields a mean Pearson correlation coefficient of 0.86 (Fig. S3*B*). For human intraspecies comparisons, we compared the datasets in Fig. 3*A* to those from ex vivo cultured, but unsorted, differentiating erythroid cells from CD34+ hematopoietic stem/progenitor cells (HPSCs) (30). The CD34+ HPSC cells reach a stage resembling proerythroblasts after 5 d and basophilic erythroblasts after 7 d, although quantitation of the purity of these stages was not done in this study. Nonetheless, the mean Pearson correlations with the corresponding proerythroblast and basophilic erythroblast populations are 0.95 and 0.95, respectively (Fig. S3*C*).

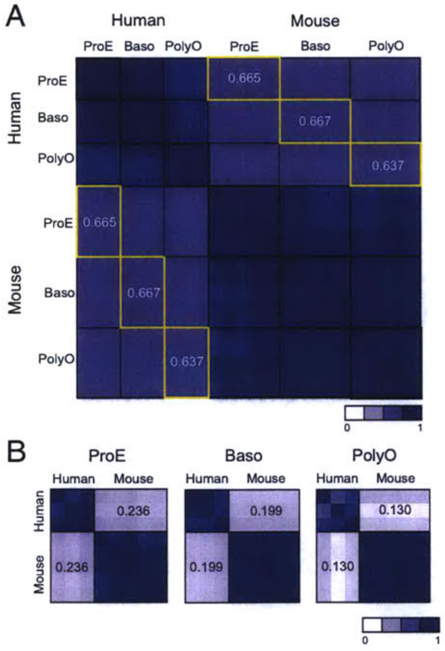


Fig. 3. Gene expression profile comparison between stage-matched populations of terminal erythroid cells in humans and mice. (*A*) Global correlation matrix of the Pearson correlation coefficients (white/blue color scale on the bottom) between each sample of human and mouse erythroid cells at comparable stages (yellow boxes). This matrix uses three human and five mouse datasets for each erythroid terminal differentiation stage, as labeled. The number indicates the mean Pearson correlation between the compared stages (15, 16). (*B*) Matrix of the Pearson correlation coefficients comparing the expression of the top 500 most expressed genes at each stages of human erythroid terminal differentiations with their orthologs in the corresponding mouse erythroblasts. ProE, proerythroblasts; Baso, basophilic erythroblasts; PolyO, mixture of polychromatophilic and orthochromatic erythroblasts.

We also carried out interspecies comparisons for all possible dataset combinations to ensure that the results were not biased by one particular dataset and were consistent across all datasets (Fig. S4). We cross-compared the two human datasets to the three mouse datasets (15, 16, 28–30). All intraspecies comparisons yielded mean Pearson correlations ranging from 0.832 to 0.993, whereas their interspecies equivalent yielded mean Pearson correlations of only 0.637–0.728 (Fig. S4), consistent with our conclusion that there is a substantial divergence in the transcriptional landscape in the two species independent of the datasets used. Although this analysis is confounded by the use of different probesets among the different microarray platforms, the consistency of these findings implies the reliability of such comparisons to delineate species-specific differences.

To establish that these results were not skewed by lowly expressed genes, we carried out a similar comparison with only the top 500 most highly expressed genes from the human data (Dataset S1); these are the datasets that we used in the initial comparison of the mouse and human erythroblast datasets (Fig. 3*B*). Surprisingly, there is a much higher degree of divergence with poor correlation between human and mouse proerythroblasts, basophilic erythroblasts, and polychromatophilic/orthochromatophilic erythroblasts, with mean correlations of 0.236, 0.199, and 0.130, respectively (Fig. 3*B*). These results indicate that there is indeed significant transcriptional divergence between human and mouse erythroblasts, which is most prominent among highly expressed genes. Interspecies comparison for the top 500 most highly expressed genes with all other possible dataset combinations yielded low mean correlations of 0.130–0.329 between the comparable stages (Fig. S5). In contrast, intraspecies comparison yielded mean correlation coefficients between 0.620 and 0.987 (Fig. S5). Once again, the fact that different microarray platforms were used could lead to an underestimation of transcriptional conservation. However, the consistency of the results, as addressed by the various cross-comparisons that we noted above, is in support of the divergence that we observed between human and mouse gene expression. Examining the stark difference in mean correlations between intra- and interspecies comparisons, we can safely deduce that the results are unlikely to be due to technical artifacts. Additionally, we confirmed the consistencies of the gene expression patterns of several erythroid signature genes whose expression pattern differs in humans and mice. We focused on several genes whose expression pattern during terminal erythropoiesis was divergent between humans and mice—*ADD1*, *ADD2*, *ADD3*, *CD47*, *TFRC*, *ITGA4*, *CD44*, *LMO2*, and *ZFPM1* (Fig. 1)—and showed that all of the expression patterns are in agreement across datasets (Fig. S6).

To further eliminate potential probeset-derived technical complications, we examined the changes in gene expression during the differentiation of proerythroblasts to basophilic erythroblasts (B-P), basophilic erythroblasts to polychromatophilic/orthochromatic erythroblasts (PO-B), and proerythroblasts to polychromatophilic/orthochromatic erythroblasts (PO-P) (Fig. 4*A*). We calculated changes in gene expression between stages using the LIMMA algorithm, yielding the $\log_2$ ratio of differential expression between any two stages of differentiation for each orthologous gene (31). We then carried out a global comparison of these gene expression changes to indicate how expression of all 12,808 orthologous genes changes upon differentiation. This method has eliminated the discrepancies due to probeset differences, as the numerical fold changes used in the comparative studies were normalized and calculated within the respective microarray platforms, giving us the "absolute" kinetic trend for each gene. Global comparison of the gene expression changes during terminal differentiation shows significant divergence between humans and mice with a mean correlation of 0.28, 0.22, and 0.32 for B-P, PO-B, and PO-P, respectively (Fig. 4*A*). A more stringent comparison, involving only those genes that significantly change from one stage to another (adjusted *P* values < 0.001) in the human data, indicates only slightly better correlation between the two species (Fig. 4*B*).

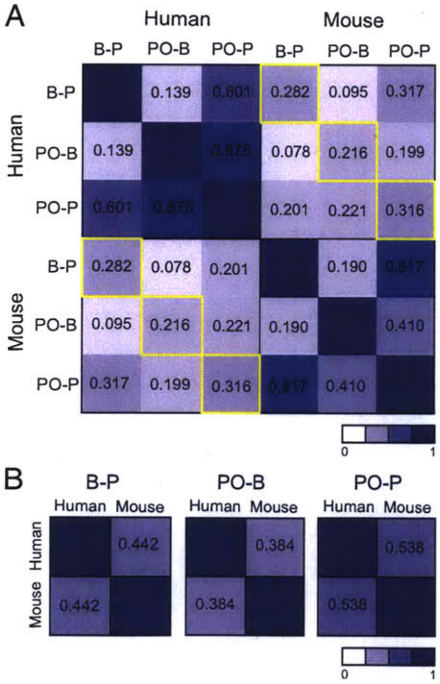


Fig. 4. Global comparison of the gene expression changes during erythroid terminal differentiation in human and mouse. (*A*) Mean Pearson correlation coefficients (white/blue color scale at the bottom of *A* and *B*) comparing the global gene expression changes transitioning from one stage to another during terminal maturation between humans and mice. (*B*) Mean Pearson correlation coefficients comparing the gene expression changes, including only genes that are significantly changed between stages with adjusted *P* values < 0.001 in the human datasets and transitioning from one stage to another during terminal differentiation. Gene changes from proerythroblasts to basophilic erythroblasts (B-P); from basophilic erythroblasts to polychromatophilic/orthochromatic erythroblasts (PO-B); from proerythroblasts to polychromatophilic/orthochromatic erythroblasts (PO-P).

Transcriptional Regulation of Human and Mouse Erythropoiesis. Fig. 1*D* shows that the majority of key transcriptional regulators showed conserved expression patterns during differentiation with few exceptions: *TAL1* and *MYB* expression exhibited altered temporal regulation. Human *TAL1* is up-regulated only during the polychromatophilic and orthochromatic erythroblast stage, whereas mouse *Tal1* is up-regulated earlier at the basophilic erythroblast stage (Fig. 1*D*). These differences suggest that subtle changes can dramatically affect the expression of genes regulated by these transcriptional regulatory factors, particularly because combinatorial interactions of several transcription factors and epigenetic regulators can determine the ultimate transcriptional output at any time during differentiation. Strikingly, we found that, although the transcriptional landscape has diverged significantly, predicted transcriptional regulators at each stage—defined by motifs found in the promoters of target genes—suggest that a similar cohort of transcriptional regulators (with ~70% conservation of regulators at the different stages) are required in both species at the same stages (Table S1). These conserved sites included those potentially bound by the canonical regulators of erythroid differentiation GATA1, NF-E2 (represented by AP-1 sites in data analyzed), and KLF1/EKLF (represented by SP1 sites in data analyzed) (32). This is consistent with the notion that specific master transcriptional regulators of differentiation are critical for the process of erythropoiesis in both species (1, 32, 33).

Contribution of Promoter Region to Transcriptional Divergence. To understand the molecular basis for the interspecies divergence in gene regulation during terminal erythropoiesis, we first analyzed the promoter regions of the 749 genes that were in the top 500 expressed genes at any stage of human erythropoiesis analyzed in Figs. 1–4. For simplicity, we defined promoters as the 1-kb region extending from 900 bases upstream of the transcription start site to 100 bases downstream. We calculated the percentage of sequence identity of each of these promoter regions between humans and mice. In parallel, we also analyzed the expression pattern of these genes during terminal differentiation. For each gene, we calculated a Pearson correlation coefficient comparing the expression patterns generated by plotting the mean values of proerythroblasts, basophilic erythroblasts, and mixtures of polychromatophilic and orthochromatic erythroblasts in human and mouse datasets. Strikingly, when promoter regions are highly conserved, defined as a sequence identity of greater than 65%, the expression patterns of these 15 genes are largely consistent between humans and mice, with an average Pearson coefficient of 0.6 (red box in Fig. S7). When there is only modest or low conservation of promoter sequence identity, the expression patterns are more highly variable (Fig. S7). This supports the notion that modifications to proximal promoters contribute to transcriptional divergence, although specific motifs within the promoters may have a more important contribution to the conservation of expression and are difficult to assess systematically on a global scale.

Contribution of Gene Duplication to Transcriptional Divergence. The comparison presented above focuses on one-to-one orthologs, which make up the majority of the human and mouse genome. Nevertheless, gene duplication, gene loss, and de novo formation play a major role in divergence between species (34). To investigate the contribution of such events in the observed transcriptional divergence, we analyzed a specific clinically relevant example. Both *SEC23B* and *SEC23A* encode proteins that are critical components of the coat protein complex II (COPII) that is necessary for endoplasmic-reticulum-to-Golgi-complex transport of secretory vesicles (35). In humans, recessive mutations of *SEC23B* cause congenital dyserythropoietic anemia (CDA) type II (10, 36). Although the exact basis for the pathophysiology of this disease is not understood, erythroid cells are thought to be particularly sensitive to mutations in SEC23B due to the lack of expression of SEC23A, which can compensate for the loss of *SEC23B* in other tissues (10). In contrast, mutations in *Sec23b* fail to result in erythroid abnormalities in mouse models (11). Using our comparative gene expression framework, we found that *SEC23B* and *Sec23b* expression is comparable during erythroid differentiation in humans and mice. However, *SEC23A* expression is down-regulated in human, whereas *Sec23a* is up-regulated in terminal mouse erythropoiesis, indicating that this paralogous gene may be able to provide compensatory function in the mouse *Sec23b* knockout, suggesting a potential explanation for the lack of hematologic abnormalities in the *Sec23b* knockout mouse model (Fig. 2*C*) (11). To support these findings, we found robust protein expression of Sec23a during all stages of terminal erythropoiesis in mice, whereas Sec23b was greatly reduced at the final stages of differentiation, supporting the concept of compensation in the *Sec23b* knockout animals (Fig. 2*D*) and confirming a recently reported finding on the differential expression of Sec23 orthologs related to CDA type II (37).

Discussion

Although specific differences in human and mouse models of erythropoiesis have been identified, to date there has been no systematic comparison of changes in gene expression. Here we developed a framework for comparing global gene expression changes between the two species and generated a resource allowing users to query expression patterns of numerous orthologous genes in humans and mice during terminal erythropoiesis (Datasets S2 and S3). This framework can be used as a reference map to understand and study conservation and divergence between human and mouse erythropoiesis, as well as to serve as a tool for translating findings between species. To expand this resource, similar compendia can be constructed on other species, as well as on a greater range of erythroid cell subsets.

Applying Pearson correlation analyses to the global gene expression data from morphologically identical stage-matched sorted populations of erythroid precursors that represent the distinct stages of terminal erythroid maturation, we find a significant extent of transcriptional divergence between comparable stages in humans and mice. Significant divergence in transcriptional changes occurring between these stages is also observed in the two species. Despite the technical limitations that we have highlighted, including probeset differences and variation in sample preparation, we consistently found substantial divergence of the transcriptional landscape of erythropoiesis between humans and mice using numerous cross-comparisons. Future cross-species comparative studies can be greatly fine-tuned by using improved purification methods and distinct phenotypic markers to obtain highly purified human and mouse erythroblasts at distinct stages, similar to methods recently reported (12). In addition, the use of RNA sequencing data may help eliminate some issues arising from microarray probeset differences that were faced in this analysis. In particular, the datasets that we used did not allow us to analyze differences in gene expression that may be due to or masked by altered gene splicing programs; studies focused on this aspect of gene regulation will undoubtedly help advance our understanding of erythropoiesis even further.

Our results suggest that, although the expression of major red cell proteins and the morphological process of erythropoiesis have remained largely conserved among mammals, a significant divergence of transcriptional regulation has occurred. However, our finding of conservation of transcriptional regulators and their target genes suggests that a key modification occurring in the course of erythroid evolution was in the responses of multiple target genes to their cognate transcriptional regulators. In particular, and focusing specifically on the top 500 most abundantly expressed genes in human Dataset S1, the interspecies divergence of gene expression patterns was even more prominent. Whereas expression of individual genes can vary significantly between humans and mice, our analysis indicates that, for each stage-matched population, ~30% of the top 500 genes in human erythroid cells have their orthologous equivalent in the top 500 genes expressed in stage-matched mouse erythroid cells (Dataset S1). Additionally, several metabolic pathways, including iron metabolism and heme biosynthesis, are well known to be shared in mouse and human erythroid cells at the same developmental stage, and this is largely confirmed by our Gene Ontology analysis (Dataset S4). Therefore, the altered expression patterns of many of the genes that comprise these pathways are likely to accommodate the substantial physiological differences between the two species.

Of the top 500 genes expressed in humans and mice at each stage of terminal erythropoiesis, most demonstrate modest amounts of promoter conservation—on the order of 40–60% sequence identity—and there is little correlation between the expression patterns of these genes in humans and mice (Fig. S7). In contrast, high sequence conservation of promoters, as indicated by greater than 65% sequence identity, is associated with similar gene expression patterns in the two species. Clearly, mutations in proximal promoters likely contribute to the interspecies divergence in gene expression, but extensive analyses of binding of multiple transcription factors and chromatin-modifying enzymes to these promoters will be essential to determine precisely how promoter mutations might drive evolutionary transcriptional divergence. Additionally, numerous species-specific properties of distal enhancers likely also contribute significantly to interspecies transcriptional divergence (38). In addition, many noncoding RNAs are expressed in terminally differentiating murine erythroid cells, and at least 12 are essential for red cell formation; additional work is essential to determine whether evolutionary changes in noncoding RNAs also contribute to the divergence in interspecies gene expression patterns that we have uncovered (39, 40).

Given both the similarities and the differences in gene expression in mammalian erythropoiesis, it is difficult to predict a priori whether a particular perturbation in mice may recapitulate what is seen in humans. For example, in mammals there exist two paralogs, *SEC23A* and *SEC23B*, with presumably identical functions in COPII vesicular transport. Transcriptional divergence in these genes might be attributed to gene duplication leading to two paralogous genes with compensatory function, and therefore their expression can diverge significantly between humans and mice. However, *CD44*, which is present as a single ortholog in humans and mice, also demonstrates transcriptional divergence despite transcriptional conservation of other membrane proteins expressed on human and mouse erythroid cells. These differences necessitate a systematic framework to understand and study conservation and divergence between human and mouse erythropoiesis. This becomes evident in the case of recessive *SEC23B* mutations, which in humans results in CDA II. However, the failure of mouse models to recapitulate this disease is most likely attributable to the observed variation in expression of Sec23a and Sec23b in comparison with their human counterparts. Indeed, analyses of published ChIP-seq data show binding of GATA1 to the proximal promoter region of the Sec23A gene in mouse erythroleukemia cells but not in human K562 cells (41, 42). This is only one of several such diseases that fail to be faithfully recapitulated by mouse models (2). Thus, developing a compendium of differences derived from global gene expression patterns can help to explain why mouse models fail to mimic specific human diseases and whether studying certain human blood diseases in transgenic mouse models will be appropriate.

This framework can also serve as a predictive tool and a resource to design studies for drug candidates in humans. For example, CD44 is a critical factor in chronic lymphocytic leukemia disease progression, and the use of antibodies specifically targeting CD44 leads to complete clearance of engrafted leukemic cells in xenograft mouse models (43, 44). However, examining the *CD44* expression profile from erythroid gene expression datasets indicates that, unlike in mice where *Cd44* expression decreases upon maturation, human red cells highly express *CD44*, an observation that we substantiated using flow cytometric analysis of mature red blood cells (Fig. 2*B*). Therefore, caution needs to be applied to translate such models between species: CD44 antibodies might be readily absorbed by mature human red blood cells in the circulation, decreasing the effectiveness of such treatments, which could not be identified in preclinical xenograft models (43, 44).

This study emphasizes both the value and the limitations of mouse models. Clearly, many aspects of erythropoiesis are conserved, although gene expression has varied considerably in the course of mammalian evolution. By using such a systematic gene expression framework, logical predictions on the effect of perturbations in human and mouse erythropoiesis can be made and better mouse models of human disease can be rationally designed. In addition, this framework will assist in our understanding of how data on human erythropoiesis derived from primary cell culture, pluripotent stem cells, xenograft models, and human genetics can be merged with the significant amount of information derived from studies of mouse erythropoiesis (2, 3, 45,).

Materials and Methods

Microarray data from humans and mice were obtained as discussed in *SI Materials and Methods*. The data were preprocessed with the GC-RMA algorithm using custom probeset definitions for National Center for Biotechnology Information Entrez Genes derived from the method of Dai et al. (46). Details of the microarray analysis, ortholog mapping, differential gene expression analysis, and experimental approaches (flow cytometry, cell culture, and Western blotting) are provided in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank Tom DiCesare for assistance with illustrations and Marko Knoll and Leif Si-Hun Ludwig for discussion and critical proofreading of the manuscript. This work was supported by National Institutes of Health (NIH) Grants P01 HL32262 (to H.F.L.) and T32 HL007574-30 (to V.G.S.). N.P. was supported by grants from the Schlumberger Foundation Faculty for the Future, the American Association of University Women International Fellowship, and the Philanthropic Educational Organization International Peace Scholarship Fund.

1. Orkin SH, Zon LI (2008) Hematopoiesis: An evolving paradigm for stem cell biology. *Cell* 132(4):631–644.
2. Sankaran VG, Orkin SH (2013) Genome-wide association studies of hematologic phenotypes: A window into human hematopoiesis. *Curr Opin Genet Dev* 23(3): 339–344.
3. Doulatov S, Notta F, Laurenti E, Dick JE (2012) Hematopoiesis: A human perspective. *Cell Stem Cell* 10(2):120–136.
4. Orkin SH, Zon LI (1997) Genetics of erythropoiesis: Induced mutations in mice and zebrafish. *Annu Rev Genet* 31:33–60.
5. Sankaran VG, et al. (2009) Developmental and species-divergent globin switching are driven by BCL11A. *Nature* 460(7259):1093–1097.
6. Sankaran VG, et al. (2011) A functional element necessary for fetal hemoglobin silencing. *N Engl J Med* 365(9):807–814.
7. McGrath KE, et al. (2011) A transient definitive erythroid lineage with unique regulation of the β-globin locus in the mammalian embryo. *Blood* 117(17):4600–4608.
8. Li Z, et al. (2005) Developmental stage-selective effect of somatically mutated leukemogenic transcription factor GATA1. *Nat Genet* 37(6):613–619.
9. Sankaran VG, et al. (2012) Exome sequencing identifies GATA1 mutations resulting in Diamond-Blackfan anemia. *J Clin Invest* 122(7):2439–2443.
10. Schwarz K, et al. (2009) Mutations affecting the secretory COPII coat component SEC23B cause congenital dyserythropoietic anemia type II. *Nat Genet* 41(8):936–940.
11. Tao J, et al. (2012) SEC23B is required for the maintenance of murine professional secretory tissues. *Proc Natl Acad Sci USA* 109(29):E2001–E2009.
12. Hu J, et al. (2013) Isolation and functional characterization of human erythroblasts at distinct stages: Implications for understanding of normal and disordered erythropoiesis in vivo. *Blood* 121(16):3246–3253.
13. Flygare J, Rayon Estrada V, Shin C, Gupta S, Lodish HF (2011) HIF1alpha synergizes with glucocorticoids to promote BFU-E progenitor self-renewal. *Blood* 117(12): 3435–3444.
14. Zhang L, et al. (2013) ZFP36L2 is required for self-renewal of early burst-forming unit erythroid progenitors. *Nature* 499(7456):92–96.
15. Merryweather-Clarke AT, et al. (2011) Global gene expression analysis of human erythroid progenitors. *Blood* 117(13):e96–e108.
16. Kingsley PD, et al. (2013) Ontogeny of erythroid gene expression. *Blood* 121(6): e5–e13.
17. Sankaran VG, et al. (2012) Cyclin D3 coordinates the cell cycle during differentiation to regulate erythrocyte size and number. *Genes Dev* 26(18):2075–2087.
18. Seok J, et al.; Inflammation and Host Response to Injury, Large Scale Collaborative Research Program (2013) Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc Natl Acad Sci USA* 110(9):3507–3512.
19. Shay T, et al.; ImmGen Consortium (2013) Conservation and divergence in the transcriptional programs of the human and mouse immune systems. *Proc Natl Acad Sci USA* 110(8):2946–2951.
20. Mestas J, Hughes CCW (2004) Of mice and not men: Differences between mouse and human immunology. *J Immunol* 172(5):2731–2738.
21. Davis MM (2008) A prescription for human immunology. *Immunity* 29(6):835–838.
22. Mohandas N, Gallagher PG (2008) Red cell membrane: Past, present, and future. *Blood* 112(10):3939–3948.
23. Conboy JG, et al. (1991) Hereditary elliptocytosis due to both qualitative and quantitative defects in membrane skeletal protein 4.1. *Blood* 78(9):2438–2443.
24. Delaunay J (2002) Molecular basis of red cell membrane disorders. *Acta Haematol* 108(4):210–218.
25. Chen K, et al. (2009) Resolving the distinct stages in erythroid differentiation based on dynamic changes in membrane protein expression during erythropoiesis. *Proc Natl Acad Sci USA* 106(41):17413–17418.
26. Zhang J, Socolovsky M, Gross AW, Lodish HF (2003) Role of Ras signaling in erythroid differentiation of mouse fetal liver cells: Functional analysis by a flow cytometry-based novel culture system. *Blood* 102(12):3938–3946.
27. Ajioka RS, Phillips JD, Kushner JP (2006) Biosynthesis of heme in mammals. *Biochim Biophys Acta* 1763(7):723–736.
28. Hattangadi SM, Burke KA, Lodish HF (2010) Homeodomain-interacting protein kinase 2 plays an important role in normal terminal erythroid differentiation. *Blood* 115(23): 4853–4861.
29. Cheng Y, et al. (2009) Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Res* 19(12):2172–2184.
30. Xu J, et al. (2012) Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Dev Cell* 23(4): 796–811.
31. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3:e3.
32. Cantor AB, Orkin SH (2002) Transcriptional regulation of erythropoiesis: An affair involving multiple partners. *Oncogene* 21(21):3368–3376.
33. Wu W, et al. (2011) Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. *Genome Res* 21(10):1659–1671.
34. Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449(7158):54–61.
35. Zanetti G, Pahuja KB, Studer S, Shim S, Schekman R (2012) COPII and the regulation of protein sorting in mammals. *Nat Cell Biol* 14(1):20–28.
36. Khoriaty R, Vasievich MP, Ginsburg D (2012) The COPII pathway and hematologic disease. *Blood* 120(1):31–38.
37. Satchwell TJ, et al. (2013) Characteristic phenotypes associated with congenital dyserythropoietic anemia (type II) manifest at different stages of erythropoiesis. *Haematologica* 98(11):1788–1796.
38. Bulger M, Groudine M (2011) Functional and mechanistic diversity of distal transcription enhancers. *Cell* 144(3):327–339.
39. Hu W, Alvarez-Dominguez JR, Lodish HF (2012) Regulation of mammalian cell differentiation by long non-coding RNAs. *EMBO Rep* 13(11):971–983.
40. Alvarez-Dominguez JR, et al. (2014) Global discovery of erythroid long noncoding RNAs reveals novel regulators of red cell maturation. *Blood* 123(4):570–581.
41. Karolchik D, et al. (2013) (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* 42(1):D764–D770.
42. Rosenbloom KR, et al. (2013) ENCODE data in the UCSC Genome Browser: Year 5 update. *Nucleic Acids Res* 41(Database issue):D56–D63.
43. Fedorchenko O, et al. (2013) CD44 regulates the apoptotic response and promotes disease development in chronic lymphocytic leukemia. *Blood* 121(20):4126–4136.
44. Zhang S, et al. (2013) Targeting chronic lymphocytic leukemia cells with a humanized monoclonal antibody specific for CD44. *Proc Natl Acad Sci USA* 110(15):6127–6132.
45. Bouhassira EE (2012) Concise review: Production of cultured red blood cells from stem cells. *Stem Cells Transl Med* 1(12):927–933.
46. Dai M, et al. (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* 33(20):e175.

Supporting Information

Pishesha et al. 10.1073/pnas.1401598111

SI Materials and Methods

Datasets and Analysis of Microarray Data. Human microarray datasets were downloaded from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) GSE22552, whereas mouse microarray datasets were downloaded from www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1035/ (1, 2). We specifically selected only adult bone marrow-derived definitive erythroblast datasets from the mouse microarray corpus to better match the differentiating adult erythroblast populations in the human microarray datasets. The human dataset is composed of biological triplicates of FACS-sorted human adult primary differentiating erythroblasts, whereas the mouse dataset is composed of five biological replicates of FACS-sorted mouse adult primary bone marrow erythroblasts at each of the three stages of terminal differentiation (1, 2). These datasets were preprocessed and normalized using the GC Robust Multichip Array algorithm from the affy package in Bioconductor using custom probeset definitions for NCBI Entrez Genes derived from the method of Dai et al. (3), yielding $\log_2$-transformed intensity values. We used these $\log_2$-transformed values, which represent relative gene expression with a normal distribution of values, for subsequent analyses.

Differentially Expressed Erythroid Signature Genes. Gene lists of cytoskeletal proteins, transmembrane proteins, heme biosynthesis enzymes, and erythroid-specific transcription factors were manually curated from available literature (4–6). The GC-RMA normalized data were mean-centered for display on heat maps.

Global Gene Expression Correlation. Using the human and mouse 1–1 ortholog list from Ensembl, we mapped 12,808 orthologs from our processed microarray data, and this gene subset was then used in our analysis of global comparison studies as well as subsequent interspecies comparisons. Pearson correlation coefficients were calculated in R between datasets.

Control Comparisons. Mouse intraspecies global gene expression comparisons were conducted using microarray datasets downloaded from the NCBI GEO GSE20391 and GSE18042, whereas another human microarray dataset was downloaded from NCBI GEO GSE36984 (7–9). We also conducted cross-species gene expression comparison between all combinations of different datasets from both species (1, 2, 7–9). Matched stages from these different datasets were identified appropriately for analyses.

Global Gene Expression Changes Comparisons. Differential expression was assayed using a moderated *t* test, as implemented by the LIMMA package in Bioconductor, corrected for the false discovery rate (10). The resulting fold changes between stages for each orthologous gene were used to calculate the Pearson correlation coefficients between datasets.

Transcription-Factor-Binding Site Analysis. Transcription-factor-binding site analysis was performed using GeneGo MetaCore software. The gene list was submitted to the GeneGo MetaCore database for integrated transcriptional binding site analyses composed of the 800 most expressed genes at each stage of terminal differentiation in human and mouse.

Pathway Analysis. Pathway analysis of the top 500 most expressed genes at each stage of terminal differentiation in humans and mice was performed using GeneGo MetaCore software. The gene list was submitted to the GeneGo MetaCore database for integrated pathway analyses.

Promoter Region Analysis. The top 500 genes in human proerythroblasts, basophilic erythroblasst, and mixture of polychromatophilic and orthochromatic erythroblasts were combined, resulting in 749 genes. Promoters were defined as 900 bases upstream from the transcription start site (TSS) and 100 bases downstream from the TSS for a total of 1 kb. Interspecies global alignment was done using the Needle program, which calculates the percentage of sequence identity. When there were multiple isoforms, the longest one was used.

Gene Expression Pattern Correlation. The 749 genes used in the Promoter Region Analysis were also used in our analysis of gene expression patterns during terminal differentiation. For each gene, we calculated a Pearson correlation coefficient by comparing the expression patterns generated by plotting the mean values of proerythroblasts, basophilic erythroblasts, and mixture of polychromatophilic and orthochromatic erythroblasts in human and mouse datasets.

Western Blot. Murine fetal liver progenitor cells (2.5×10^6) were harvested at indicated time points in culture and lysed in RIPA buffer followed by a 10-min incubation at 60 °C. Proteins were separated by SDS gel electrophoresis using the NuPAGE Bis-Tris gel system and MOPS running buffer under reducing conditions. Proteins were transferred onto a nitrocellulose membrane using the NuPAGE transfer buffer (Invitrogen). Membranes were blocked with TBST (50 mM Tris, 150 mM NaCl, 0.05% Tween 20, pH 7.6)/5% (wt/vol) BSA for 1 h and probed with Sec23A goat monoclonal antibody (SAB2501341, Sigma-Aldrich), Sec23B chicken polyclonal antibody (ab37739, Abcam), or GAPDH mouse monoclonal antibody (sc-32233, Santa Cruz Biotechnology) all at a 1:1,000 dilution in TBST/5% BSA overnight at 4 °C. Membranes were washed three times, incubated with donkey anti-goat (100188-432, VWR International), goat anti-chicken (PA1-28658, Thermo Scientific), or goat anti-mouse (100187-616, VWR International) peroxidase-coupled secondary antibodies at a 1:10,000 dilution in TBST/5% BSA for 1 h at room temperature, washed three times, and developed with Western Lightning Plus-ECL substrate (Perkin-Elmer). Proteins were visualized by exposure to scientific imaging film (Kodak).

Flow Cytometry Analysis. For FACS, peripheral human or C57BL/6J mouse (The Jackson Laboratory) blood was washed in PBS/5% BSA, incubated in blocking buffer containing human serum type AB (S40110, Atlanta Biologicals), and stained with 1:400 PE (Phycoerythrin)-conjugated anti-human/mouse CD44 (12–0441-82e, Bioscience). For control, rat IgG2b κ-isotype control PE (eBioscience, 12–4031-81) was used. FACS analysis was carried out using the BD Bioscience LSR II flow cytometer, and the resulting data were analyzed using FlowJo 8.6.9 (TreeStar).

Study Approval. All discarded human samples were obtained through protocols approved by the institutional review board at Boston Children's Hospital. All animal procedures were performed under protocols approved by the animal use committee at the Massachusetts Institute of Technology.

1. Merryweather-Clarke AT, et al. (2011) Global gene expression analysis of human erythroid progenitors. *Blood* 117(13):e96–e108.
2. Kingsley PD, et al. (2013) Ontogeny of erythroid gene expression. *Blood* 121(6): e5–e13.
3. Dai M, et al. (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* 33(20):e175.
4. Chen K, et al. (2009) Resolving the distinct stages in erythroid differentiation based on dynamic changes in membrane protein expression during erythropoiesis. *Proc Natl Acad Sci USA* 106(41):17413–17418.
5. Nilsson R, et al. (2009) Discovery of genes essential for heme biosynthesis through large-scale gene expression analysis. *Cell Metab* 10(2):119–130.
6. Hattangadi SM, Wong P, Zhang L, Flygare J, Lodish HF (2011) From stem cell to red cell: Regulation of erythropoiesis at multiple levels by multiple proteins, RNAs, and chromatin modifications. *Blood* 118(24):6258–6268.
7. Hattangadi SM, Burke KA, Lodish HF (2010) Homeodomain-interacting protein kinase 2 plays an important role in normal terminal erythroid differentiation. *Blood* 115(23): 4853–4861.
8. Cheng Y, et al. (2009) Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Res* 19(12):2172–2184.
9. Xu J, et al. (2012) Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Dev Cell* 23(4): 796–811.
10. Cantor AB, Orkin SH (2002) Transcriptional regulation of erythropoiesis: An affair involving multiple partners. *Oncogene* 21(21):3368–3376.

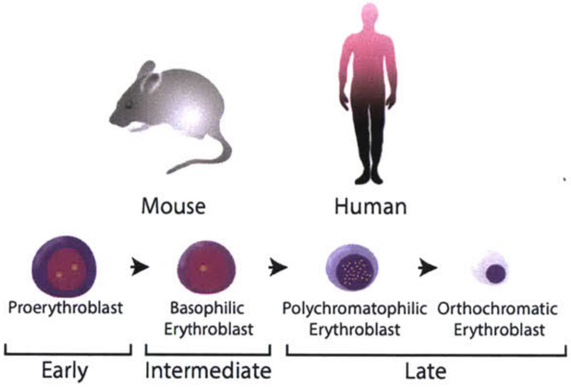


Fig. S1. Simplified scheme depicting terminal erythropoiesis. Both human and mouse erythroid terminal differentiations consists of three distinct stages: early (pro), intermediate (basophilic), and late (polychromatophilic and orthochromatic) erythroblasts.

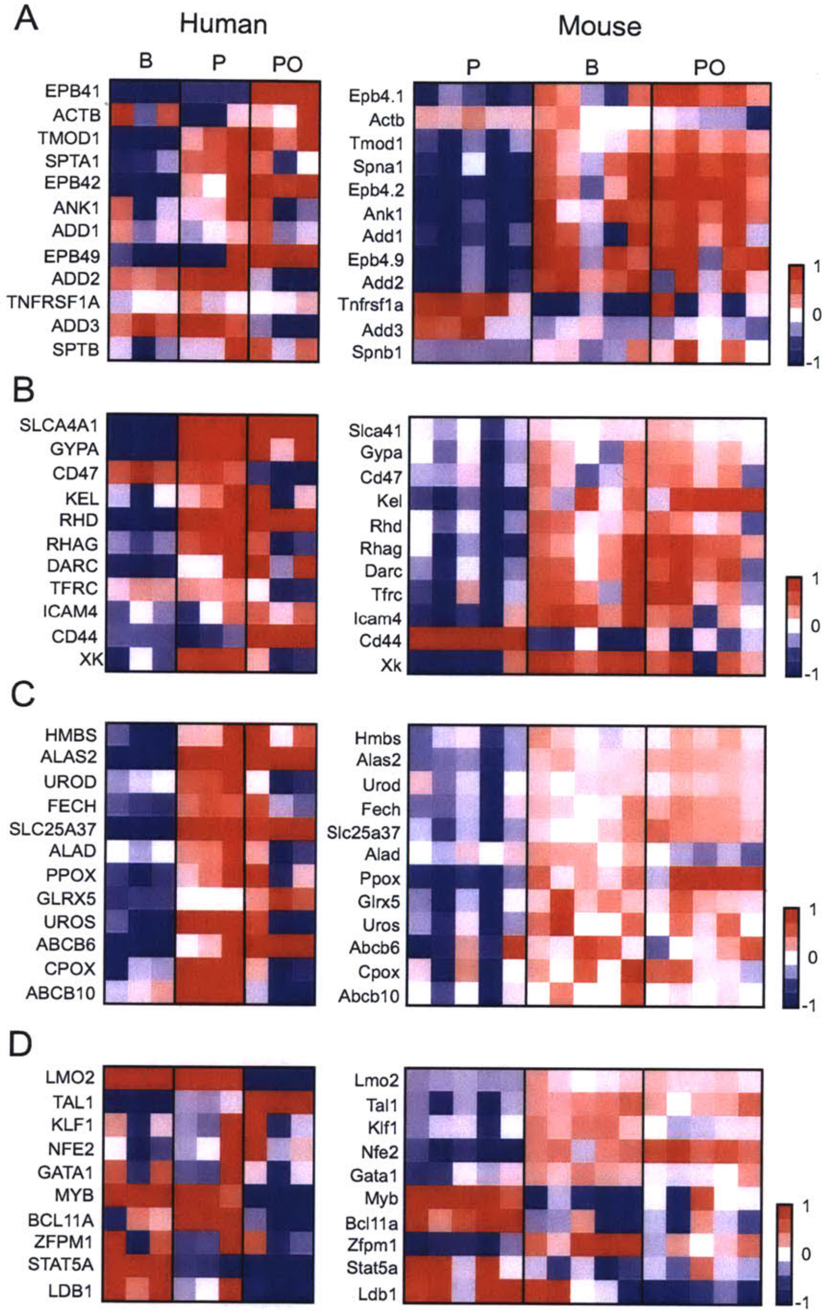


Fig. S2. Mean-centered expression of individual expression values across all samples of the early, intermediate, and late stage erythroblasts (red/white/blue color scale to the right) of several signature erythroid gene groups: (*A*) skeletal proteins, (*B*) transmembrane proteins, (*C*) heme biosynthesis enzymes, and (*D*) erythroid-specific transcription factors in humans (*Left*) and mice (*Right*). Red, white, and blue color indicates higher expression than the mean, values close to the mean, and expression below the mean, respectively. The scale −1 to +1 signifies how far the average expression values are from mean-centered values. P, proerythroblasts; B, basophilic erythroblasts; PO, mixture of polychromatophilic and orthochromatic erythroblasts.

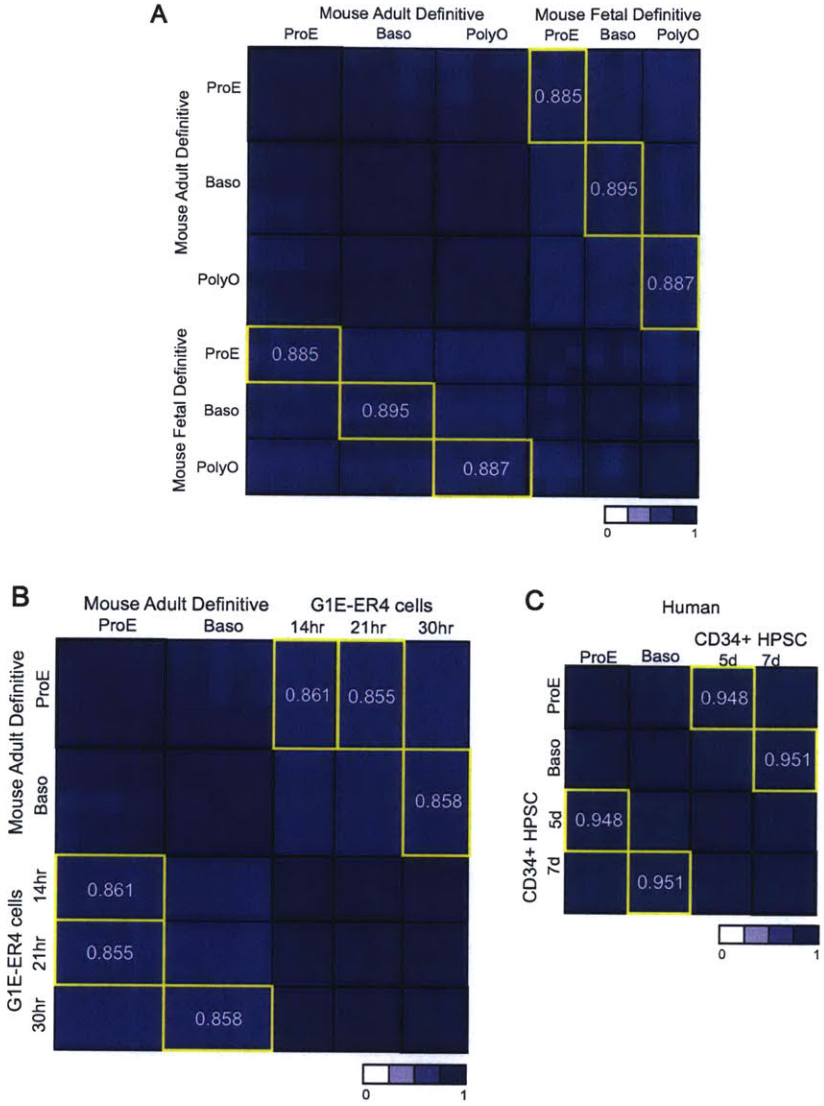


Fig. S3. Intraspecies transcriptional profiles comparison between stage-matched populations of terminal erythroid cells. (*A*) Global correlation matrix of the Pearson correlation coefficients (white/blue color scale at the bottom) between each sample of mouse erythroid cells at comparable stages (yellow boxes). The number indicates the mean Pearson correlation between the compared stages. A total of five adult mouse definitive datasets and five fetal mouse definitive datasets for each erythroid terminal differentiation stage are laid out in rows and columns (1, 2). (*B*) Global correlation matrix of the Pearson correlation coefficients between each sample of mouse erythroid cells at comparable stages. The comparable stages are the mouse adult definitive proerythroblasts with G1E-ER4 cells cultured in estradiol infused G1E medium for 14 and 21 h (14 h and 21 h) as well as the mouse adult definitive basophilic erythroblasts with G1E-ER4 cells cultured in estradiol infused G1E medium for 30 h (30hr) (1, 3) (*C*) Global correlation matrix of the Pearson correlation coefficients between each sample of human erythroid cells at comparable stages. The ex vivo-cultured CD34+ HPSC cells resemble the proerythroblast stage at day 5 in culture (5d) whereas basophilic erythroblasts emerges after 7 d in culture (7d). CD34+ HPSC, CD34+ hematopoietic stem/progenitor cells (4, 5).

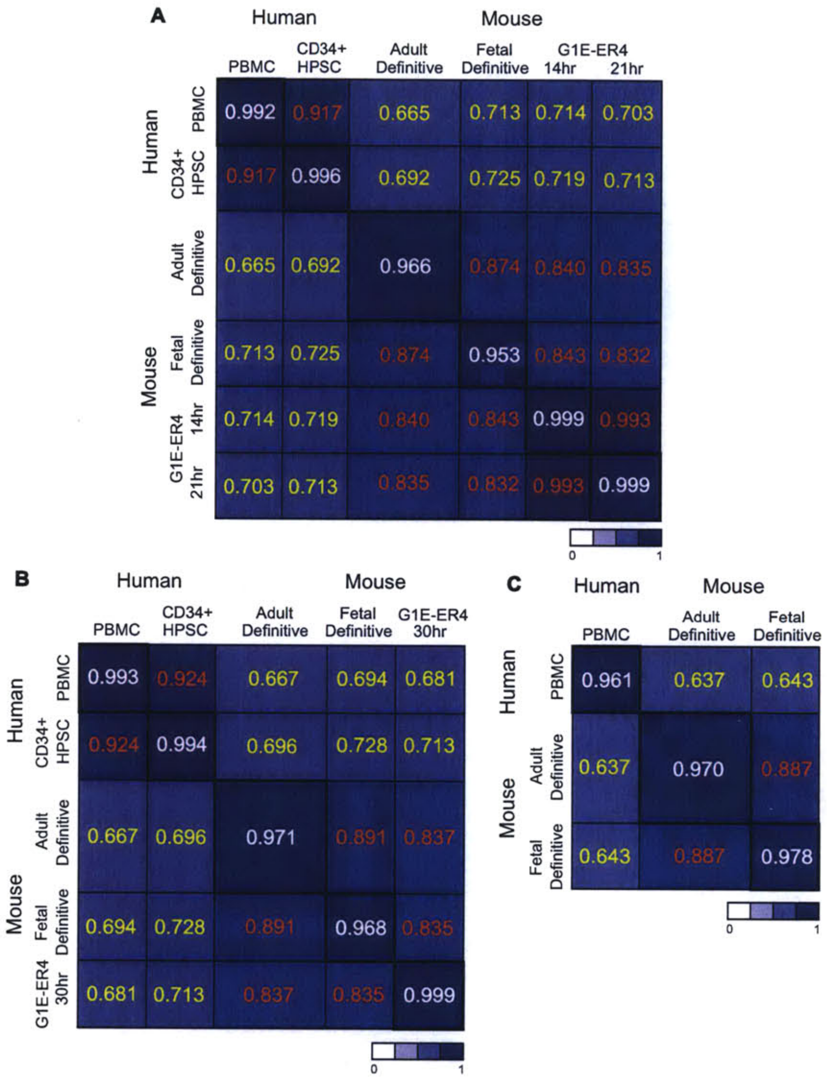
1. Kingsley PD, et al. (2013) Ontogeny of erythroid gene expression. *Blood* 121(6):e5–e13.

2. Hattangadi SM, Burke KA, Lodish HF (2010) Homeodomain-interacting protein kinase 2 plays an important role in normal terminal erythroid differentiation. *Blood* 115(23):4853–4861.

3. Cheng Y, et al. (2009) Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Res* 19(12): 2172–2184.

4. Merryweather-Clarke AT, et al. (2011) Global gene expression analysis of human erythroid progenitors. *Blood* 117(13):e96–e108.

5. Xu J, et al. (2012) Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Dev Cell* 23(4):796–811.



A

		Human		Mouse			
		PBMC	CD34+ HPSC	Adult Definitive	Fetal Definitive	G1E-ER4 14hr	G1E-ER4 21hr
Human	PBMC	0.992	0.917	0.665	0.713	0.714	0.703
	CD34+ HPSC	0.917	0.996	0.692	0.725	0.719	0.713
Mouse	Adult Definitive	0.665	0.692	0.966	0.874	0.840	0.835
	Fetal Definitive	0.713	0.725	0.874	0.953	0.843	0.832
	G1E-ER4 14hr	0.714	0.719	0.840	0.843	0.999	0.993
	G1E-ER4 21hr	0.703	0.713	0.835	0.832	0.993	0.999

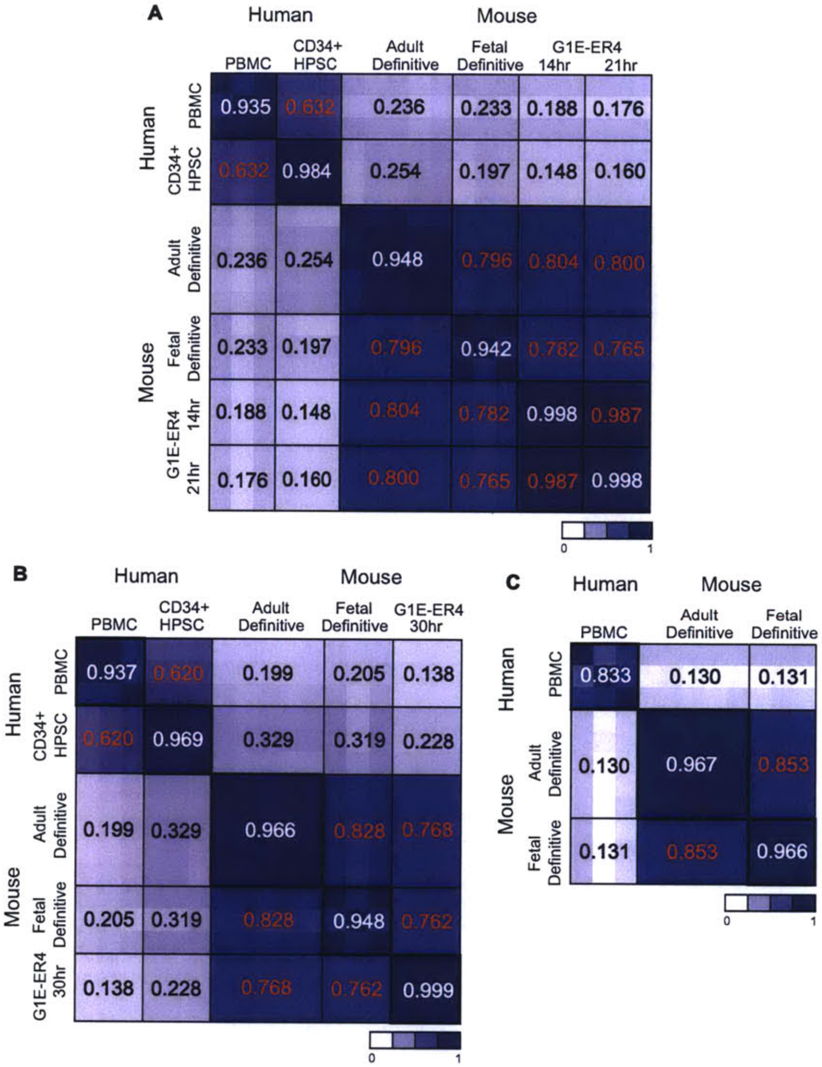
B

		Human		Mouse		
		PBMC	CD34+ HPSC	Adult Definitive	Fetal Definitive	G1E-ER4 30hr
Human	PBMC	0.993	0.924	0.667	0.694	0.681
	CD34+ HPSC	0.924	0.994	0.696	0.728	0.713
Mouse	Adult Definitive	0.667	0.696	0.971	0.891	0.837
	Fetal Definitive	0.694	0.728	0.891	0.968	0.835
	G1E-ER4 30hr	0.681	0.713	0.837	0.835	0.999

C

		Human	Mouse	
		PBMC	Adult Definitive	Fetal Definitive
Human	PBMC	0.961	0.637	0.643
Mouse	Adult Definitive	0.637	0.970	0.887
	Fetal Definitive	0.643	0.887	0.978

Fig. S4. Global transcriptional profiles comparison between stage-matched populations of terminal erythroid cells. Global correlation matrix of the Pearson correlation coefficients (white/blue color scale at the bottom) between each sample of human and mouse erythroid cells at (*A*) proerythroblast, (*B*) basophilic erythroblasts, and (*C*) mixture of polychromatophilic and orthochromatic erythroblasts. The white number indicates the mean Pearson correlation among biological replicates of a specific stage. The yellow number indicates the mean Pearson correlation of interspecies comparison at a specific stage of terminal erythroid differentiation, and the red number indicates the mean Pearson correlation of intraspecies comparison at a specific stage of terminal erythroid differentiation. We compared two human data sources: the in vitro-cultured peripheral blood mononuclear cells (PBMC; GSE22552), and the ex vivo-cultured CD34+ hematopoietic stem/progenitor cells (CD34+ HPSC; GSE36984). The mouse data sources were the following: adult definitive (sorted bone marrow cells; http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1035/), fetal definitive (sorted fetal liver cells; GSE20391), and G1E-ER4 cells cultured in estradiol-infused G1E medium. The G1-ER4 cells resemble the adult mouse definitive proerythroblasts with after 14 and 21 h (14hr and 21hr) of culture and resemble basophilic erythroblasts (GSE18042) after 30 h (30hr).



A

		Human	Human	Mouse	Mouse	Mouse	Mouse
		PBMC	CD34+ HPSC	Adult Definitive	Fetal Definitive	G1E-ER4 14hr	G1E-ER4 21hr
Human	PBMC	0.935	0.632	0.236	0.233	0.188	0.176
Human	CD34+ HPSC	0.632	0.984	0.254	0.197	0.148	0.160
Mouse	Adult Definitive	0.236	0.254	0.948	0.796	0.804	0.800
Mouse	Fetal Definitive	0.233	0.197	0.796	0.942	0.782	0.765
Mouse	G1E-ER4 14hr	0.188	0.148	0.804	0.782	0.998	0.987
Mouse	G1E-ER4 21hr	0.176	0.160	0.800	0.765	0.987	0.998

B

		Human	Human	Mouse	Mouse	Mouse
		PBMC	CD34+ HPSC	Adult Definitive	Fetal Definitive	G1E-ER4 30hr
Human	PBMC	0.937	0.620	0.199	0.205	0.138
Human	CD34+ HPSC	0.620	0.969	0.329	0.319	0.228
Mouse	Adult Definitive	0.199	0.329	0.966	0.828	0.768
Mouse	Fetal Definitive	0.205	0.319	0.828	0.948	0.762
Mouse	G1E-ER4 30hr	0.138	0.228	0.768	0.762	0.999

C

		Human	Mouse	Mouse
		PBMC	Adult Definitive	Fetal Definitive
Human	PBMC	0.833	0.130	0.131
Mouse	Adult Definitive	0.130	0.967	0.853
Mouse	Fetal Definitive	0.131	0.853	0.966

Fig. S5. Matrix of the Pearson correlation coefficients (white/blue color scale at the bottom) comparing the expression of the top 500 most expressed genes at stage-matched populations of terminal erythroid cells: (*A*) proerythroblast, (*B*) basophilic erythroblasts, and (*C*) mixture of polychromatophilic and orthochromatic erythroblasts. The white number indicates the mean Pearson correlation among biological replicates of a specific stage. The yellow number indicates the mean Pearson correlation of interspecies comparison at a specific stage of terminal erythroid differentiation, and the red number indicates the mean Pearson correlation of intraspecies comparison at a specific stage of terminal erythroid differentiation. The two human data sources were the in vitro-cultured peripheral blood mononuclear cells (PBMC; GSE22552) and the ex vivo cultured CD34+ hematopoietic stem/progenitor cells (CD34+ HPSC; GSE36984). The mouse data sources were adult definitive (sorted bone marrow cells, www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1035/), fetal definitive (sorted fetal liver cells; GSE20391), and G1E-ER4 cells cultured in estradiol-infused G1E medium. The G1-ER4 cells resemble the mouse adult definitive proerythroblasts after 14 and 21 h (14hr and 21hr) of culture and resemble basophilic erythroblasts after 30 h (30hr) (GSE18042).

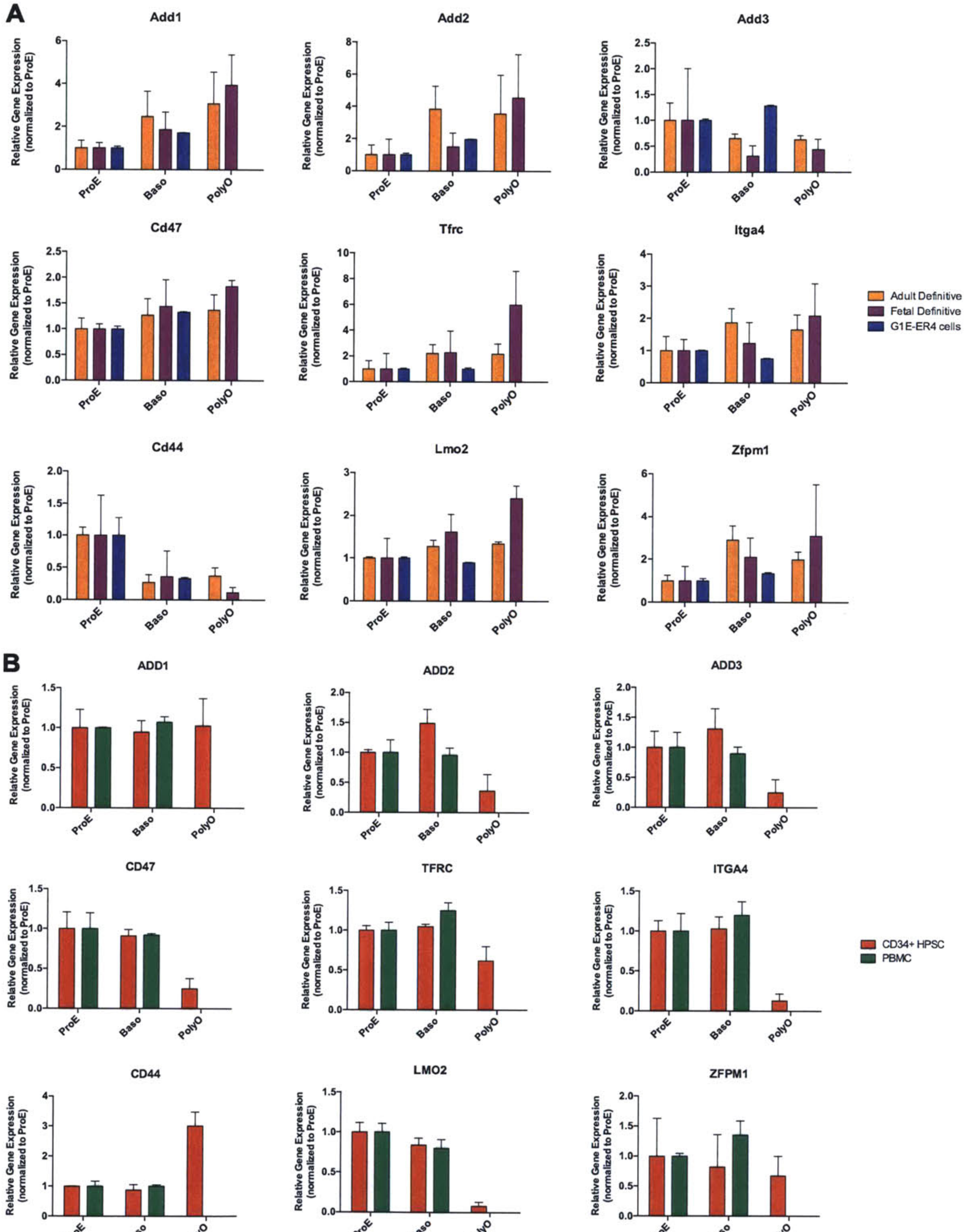


Fig. S6. Differentially regulated genes between human and mouse erythroblasts. (*A*) Mouse and (*B*) human gene expressions mined from the comparative gene expression framework. ProE, proerythroblasts; Baso, basophilic erythroblasts; PolyO, mixture of polychromatophilic and orthochromatic erythroblasts. The two human data sources were the in vitro-cultured peripheral blood mononuclear cells (PBMC; GSE22552) and the ex vivo-cultured CD34+ hematopoietic stem/progenitor cells (CD34+ HPSC; GSE36984). The mouse data sources were adult definitive (sorted bone marrow cells; http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1035/), fetal definitive (sorted fetal liver cells; GSE20391), and G1E-ER4 cells cultured in estradiol-infused G1E medium. The G1-ER4 cells resemble the mouse adult definitive proerythroblasts after 14 and 21 h (14hr and 21hr) of culture and resemble basophilic erythroblasts (GSE18042) after 30 h (30hr).

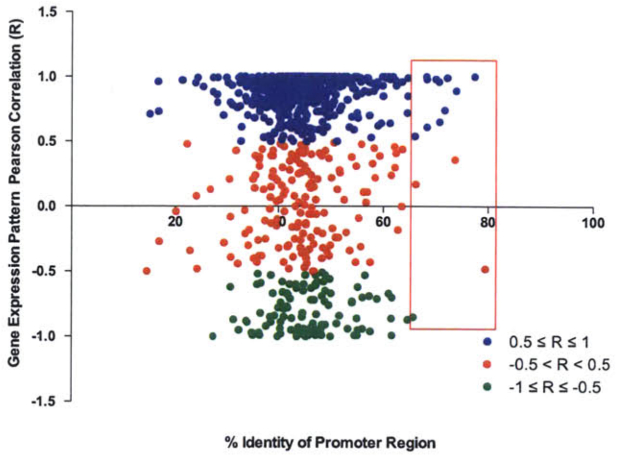


Fig. S7. Association between promoter region sequence conservation and gene expression pattern during erythroid terminal differentiation. The top 500 genes in human proerythroblasts, basophilic erythroblasts, and a mixture of polychromatophilic and orthochromatic erythroblasts were combined, resulting in 749 genes. Promoters were defined as 900 bases upstream from the TSS and 100 bases downstream from the TSS for a total of 1 kb. Interspecies global alignment of the promoter region is represented in percentage of sequence identity. Pearson correlation coefficients (R) were calculated by comparing the expression patterns generated by plotting the mean values of proerythroblasts, basophilic erythroblasts, and the mixture of polychromatophilic and orthochromatic erythroblasts in human and mouse datasets. The closer the R values are to 1 indicates a more similar expression pattern of gene expression in differentiating human erythroid cells and its ortholog in differentiating mouse erythroid cells, whereas −1 indicates the opposite expression patterns between species.

Table S1. Transcription-factor-binding site enrichment

Mouse ProE	Human ProE	Mouse Baso	Human Baso	Mouse PolyO	Human PolyO
SP1	SP1	SP1	SP1	SP1	SP1
HNF4-α	HNF4-α	HNF4-α	HNF4-α	HNF4-α	HNF4-α
c-Myc	c-Myc	c-Myc	c-Myc	c-Myc	c-Myc
p53	p53	p53	p53	p53	p53
CREB1	CREB1	CREB1	CREB1	CREB1	CREB1
E2F1	E2F1	E2F1	E2F1	E2F1	E2F1
HIF1-A	HIF1-A	HIF1-A	HIF1-A	HIF1-A	HIF1-A
YY1	YY1	YY1	YY1	YY1	YY1
Androgen receptor	Androgen receptor	Androgen receptor	Androgen receptor	Androgen receptor	Androgen receptor
GATA-1	GATA-1	GATA-1	GATA-1	GATA-1	GATA-1
GCR-α	GCR-α	GCR-α	GCR-α	GCR-α	GCR-α
TBP	TBP	TBP	TBP	TBP	TBP
NF-Y	NF-Y	NF-Y	NF-Y	NF-Y	NF-Y
ESR1 (nuclear)	ESR1 (nuclear)	ESR1 (nuclear)	ESR1 (nuclear)	ESR1 (nuclear)	ESR1 (nuclear)
AP-1	AP-1	AP-1	AP-1	C/EBPβ	C/EBPβ
Elk-1	Elk-1	Elk-1	Elk-1	NF-κB	NF-κB
E2F4	E2F4	EGR1	EGR1	NRF2	NRF2
EGR-1	NRF-1	C/EBPβ	C/EBPβ	Elk-1	RelA
C/EBPβ	NRF-2	E2F4	AP-2A	ETS1	AP-1
ETS1	AP-2A	ETS1	NRF2	EGR1	Oct-3/4
NF-κB	Oct-3/4	NF-κB	Oct-3/4	c-Jun	USF1

Red highlights the predicted transcription-factor-binding sites shared by the most expressed genes at all stages of mouse and human erythroid terminal differentiations. Blue with yellow background, green with gray background, and brown with blue background indicate transcription-factor-binding sites shared by the most highly expressed genes in stage-matched human and mouse proerythroblasts, basophilic erythroblasts, and polychromatophilic and orthochromatic erythroblasts, respectively. Black denotes factors expressed at a specific stage in one species but not in another. Baso, basophilic erythroblasts; PolyO, mixture of polychromatophilic and orthochromatic erythroblasts; ProE, proerythroblasts.

Dataset S1. Top 500 genes in mouse and human proerythroblast (ProE), basophilic erythroblast (Baso), and a mixture of polychromatophilic and orthochromatic erythroblasts (PolyO). Each list is organized from the genes that have the highest gene expression level to the lowest. For each stage-matched population, 152, 158, and 154 genes of the top 500 genes in ProE, Baso, and PolyO, respectively, have their orthologous equivalent in the top 500 genes expressed in stage-matched mouse erythroid cells

Dataset S1

Dataset S2. The log2-transformed intensity values for the 12,808 orthologus genes for each sample. ProE, proerythroblast; Baso, basophilic erythroblast; PolyO, mixture of polychromatophilic and orthochromatic erythroblasts

Dataset S2

Dataset S3. The mean log2-transformed intensity values for the 12,808 orthologus genes. ProE, proerythroblast; Baso, basophilic erythroblast; PolyO, mixture of polychromatophilic and orthochromatic erythroblasts

Dataset S3

Dataset S4. Enrichment analysis reports (generated by GeneGo Software) on the top 500 genes in mouse and human proerythroblasts (ProE), basophilic erythroblasts (Baso), and a mixture of polychromatophilic and orthochromatic erythroblasts (PolyO). Each list is organized from the highest number of input genes to the lowest. There are 20, 29, and 20 pathways (of 50) constituted by these top 500 genes at the same developmental stage that are shared in mouse and human ProE, Baso, and PolyO, respectively

Dataset S4