Cell Reports

Transcriptional States and Chromatin Accessibility Underlying Human Erythropoiesis

Graphical Abstract

Highlights

- Integrated accessible chromatin and transcriptome analysis of human erythropoiesis
- Inference of differentiation stage-specific transcription factor activities
- Mapping of genetic variants underlying diseases and traits to regulatory regions
- Identification of TMCC2 as a regulator in terminal erythropoiesis

Authors

Leif S. Ludwig, Caleb A. Lareau, Erik L. Bao, ..., Martin J. Aryee, Aviv Regev, Vijay G. Sankaran

Correspondence

aregev@broadinstitute.org (A.R.), sankaran@broadinstitute.org (V.G.S.)

In Brief

Ludwig et al. chart the dynamic transcriptional and chromatin landscapes as hematopoietic stem and progenitor cells differentiate into mature red blood cells. This multi-omic profiling reveals dynamic transcription factor activities and human genetic variation that modulate this process.
Transcriptional States and Chromatin Accessibility Underlying Human Erythropoiesis

Leif S. Ludwig,1,2,14 Caleb A. Lareau,1,2,3,4,15 Erik L. Bao,1,2,5 Satish K. Nandakumar,1,2 Christoph Muus,2,6 Jacob C. Ulirsch,1,2,4 Kaitavjeet Chowdhary,1,2,5 Jason D. Buenrostro,2,7 Narla Mohandas,1 Xiuli An,1,9 Martin J. Aryee,1,3,10,11 Aviv Regev,2,12,13,* and Vijay G. Sankaran1,2,15,*

1Division of Hematology/Oncology, Boston Children’s Hospital, and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA
2Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
3Molecular Pathology Unit, Massachusetts General Hospital, Charlestown, MA 02129, USA
4Program in Biological and Biomedical Sciences, Harvard University, Cambridge, MA 02138, USA
5Harvard-MIT Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA
6Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA
7Society of Fellows, Harvard University, Cambridge, MA 02138, USA
8Laboratory of Membrane Biology, New York Blood Center, New York, NY 10065, USA
9School of Life Science, Zhengzhou University, Zhengzhou, Henan 450001, China
10Department of Pathology, Harvard Medical School, Boston, MA 02115, USA
11Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
12Howard Hughes Medical Institute, Chevy Chase, MD 26309, USA
13Department of Biology and Koch Institute, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
14These authors contributed equally
15Lead Contact
*Correspondence: aregev@broadinstitute.org (A.R.), sankaran@broadinstitute.org (V.G.S.)
https://doi.org/10.1016/j.celrep.2019.05.046

SUMMARY

Human erythropoiesis serves as a paradigm of physiologic cellular differentiation. This process is also of considerable interest for better understanding anemias and identifying new therapies. Here, we apply deep transcriptomic and accessible chromatin profiling to characterize a faithful ex vivo human erythroid differentiation system from hematopoietic stem and progenitor cells. We reveal stage-specific transcriptional states and chromatin accessibility during various stages of erythropoiesis, including 14,260 differentially expressed genes and 63,659 variably accessible chromatin peaks. Our analysis suggests differentiation stage-predominant roles for specific master regulators, including GATA1 and KLF1. We integrate chromatin profiles with common and rare genetic variants associated with erythroid cell traits and diseases, finding that variants regulating different erythroid phenotypes likely act at variable points during differentiation. In addition, we identify a regulator of terminal erythropoiesis, TMCC2, more broadly illustrating the value of this comprehensive analysis to improve our understanding of erythropoiesis in health and disease.

INTRODUCTION

Erythropoiesis describes the process of proliferation and differentiation of hematopoietic stem and progenitor cells (HSPCs) through distinct functionally and morphologically defined stages to produce enucleate reticulocytes. In the circulation, these cells mature further into red blood cells (RBCs), which are the key transporters of oxygen and carbon dioxide for cellular respiration (Nandakumar et al., 2016; Sankaran and Weiss, 2015). In adult humans, approximately 2 million RBCs are produced every second in the bone marrow through the tightly coordinated process of erythropoiesis, making RBCs the most abundant cell type in the human body (Palis, 2014). Congenital as well as acquired defects may lead to various forms of anemia, the lack of sufficient RBCs, causing significant morbidity and mortality (Sankaran and Weiss, 2015).

Genetic and cell biological approaches have provided important insights into normal RBC differentiation and how this process is perturbed in different blood diseases, such as anemias (Arlet et al., 2014; Giani et al., 2016; Khajuria et al., 2018; Ludwig et al., 2014, 2016). Our knowledge of this process has been further enhanced by comprehensive transcriptomic and proteomic profiling of distinct stages of neonatal and adult erythropoiesis (An et al., 2014; Gautier et al., 2016; Li et al., 2014; Yan et al., 2018). Additional efforts have recorded the accessible chromatin landscape (Buenrostro et al., 2018; Corces et al., 2016) and epigenetic marks such as histone modifications (Huang et al., 2016) at select stages and investigated the role of genetic variation and chromatin accessibility in murine cellular models (Behera et al., 2018). However, we lack a detailed characterization of the accessible chromatin landscape and transcription factor dynamics throughout the entire process of human erythroid differentiation, even though such characterization would facilitate a more comprehensive understanding of gene regulatory dynamics and their relation to human genetic variation and disease.
Here, we leverage the assay for transposase accessible chromatin using sequencing (ATAC-seq) (Corces et al., 2016, 2017) to assess eight populations that chart the dynamic open chromatin landscapes of HSPCs undergoing the process of erythroid differentiation. We integrate our data with matched deep transcripтомic profiles, results from genome-wide association studies (GWAS), and mutations resulting in human disease, to facilitate a systems-level understanding of the molecular circuits governing erythropoiesis in human health and disease. We show that our chromatin accessibility profiles yield a detailed description of the regulatory elements and transcription factor dynamics that control gene expression during human erythropoiesis. The intersection with variants identified in GWAS and Mendelian diseases reveal the differential and putative temporal contributions of genetic variants in accessible chromatin in affecting human erythroid traits and pathologies. Overall, our integrated analysis framework provides a comprehensive resource to advance our understanding of gene regulation and will expedite downstream functional and validation studies of regulators, as exemplified by our use of this dataset to identify a role for TMCC2 in terminal human erythroid differentiation.

RESULTS

An Epigenomic and Transcriptional Time Course of Human Erythroid Differentiation

To obtain a comprehensive and integrative picture of the chromatin and transcriptional landscape of human adult erythropoiesis, we differentiated CD34+ HSPCs from healthy donors using an established three-phase erythroid differentiation protocol that allows the efficient production of mature enucleated reticulocytes (Giani et al., 2016; Hu et al., 2013). We then applied flow cytometry-activated cell sorting (FACS) using well-characterized erythroid surface markers CD71, CD235a, CD49d, and Band 3 (encoded by the SLC4A1 gene) across multiple stages of the differentiation process to enrich for cells at different stages of maturation (Hu et al., 2013), validated purity by morphology, and processed each population (P1–P8) using ATAC-seq (FAST-ATAC) and RNA sequencing (RNA-seq) (Smart-seq2) (Figures 1A, 1B, and S1A) (Corces et al., 2016; Picelli et al., 2014). Each population was processed in three or four replicates using cultured cells from two or three healthy adult human donors, resulting in a total of 28 paired RNA-seq and ATAC-seq libraries. Overall, our ATAC-seq libraries were sequenced to an average depth of 26.1 million aligned reads/sample (mean 91.1 million/population), and the RNA-seq libraries were sequenced to an average depth of 21.5 million aligned reads/sample (mean 75.3 million/population), and the libraries were sequenced to an average depth of 29.3 million aligned reads/sample (mean 91.1 million/population).

Time-Dependent Modules of Transcriptional Regulation Define Erythroid Differentiation Programs

To characterize transcriptional modules throughout human erythroid differentiation, we used k-means clustering to identify co-regulated genes. Using the gap statistic to determine an appropriate number of clusters (Figure S2B), we identified seven clusters of differentially expressed genes showing a wide range of expression patterns and dynamics (Figure 2A; Table S3). From the Z score-normalized gene expression values, we observed population-specific expression primarily in presumed MyP (cluster k3) and at the latest stages of erythroid differentiation (OrthoE/Ret; cluster k7). Most other gene expression programs were broadly expressed in early progenitor populations (clusters k1 and k2) and throughout erythroid differentiation (clusters k4–k6).

To determine the underlying cellular processes governed by these patterns of expression variability, we performed Gene Ontology (GO) analyses to identify biological processes that were enriched within these clusters and that may be involved in regulating erythroid differentiation (Figure 2B; Table S4). RNA, non-coding RNA, and DNA processing-related gene sets were most dominant during the earlier stages of terminal erythroid differentiation (cluster k4). Cell cycle-related processes showed maximum activity around the ProE1/2 stage (cluster k5), consistent with the high proliferative capacity of these cells. Heme biosynthetic and oxygen transport processes showed highest activity at the BasoE and PolyE stage, reflecting the peak of hemoglobin synthesis (cluster k6). Regulation of catabolic processes dominate during the OrthoE/Ret stage, as cellular organelles are being recycled before final maturation into erythrocytes (cluster k7) (Figure 2B). Overall, our transcriptional analysis revealed highly consistent findings with those described in prior reports (An et al., 2014; Yan et al., 2016).
Dynamics of the Accessible Chromatin Landscape and Enrichment of Transcriptional Regulatory Programs

Previous genome-wide analyses of accessible chromatin profiles have revealed a comprehensive repertoire of gene regulatory elements, many of which appear to act in a cell type- and tissue-specific manner (Thurman et al., 2012). However, we have a limited understanding of the accessibility dynamics of these elements during cellular differentiation. Our dense profiling of erythroid populations at different stages of differentiation using ATAC-seq enables the most comprehensive description to date of the dynamic accessible chromatin landscape throughout human erythropoiesis (Figure 3A). Quantification of Tn5 insertion density around CTCF motifs as previously described (Buenrostro et al., 2013) verified high-quality libraries throughout our differentiation system (Figure S3A). Using k-means clustering and gap statistic optimization (Figure S3B), we identify seven clusters of differential accessible peaks, showing similar patterns and dynamics at key erythroid gene loci such as ALAS2 (Figure 3B), similar to what we describe for the transcriptomic profiles described above (Figure 2A).

Cluster k1 contained regions of open chromatin showing highest accessibility in HSPCs, which became less accessible during early erythroid differentiation. By contrast, regulatory elements in cluster k2 displayed slower closing dynamics (decreasing accessibility). We also observed profound changes from the ProE1/2 to the OrthoE stage. Some elements were most accessible around the ProE1/2 stage but quickly closed (lost accessibility) in BasoE (cluster k4), whereas cluster k5 contained elements remaining accessible throughout the PolyE stage. Interestingly, we also observed elements showing highest accessibility at the OrthoE stage immediately preceding enucleation. This process has been typically associated with increased nuclear condensation and an anticipated loss of accessibility, also supported by an overall decreasing number of accessibility peaks at late stages of differentiation (Figure 1E) (Zhao et al., 2016). Of note, the late stage-specific peaks were...
enriched in genes displaying the highest levels of induction and expression at the OrthoE stage, as exemplified by the TMCC2 locus, a gene highly induced in terminal erythropoiesis whose role in this process has not been previously studied (Figures 6 and 7).

Our ATAC-seq dataset enables the use of chromVAR to infer transcription factor (TF) variability and dynamics throughout erythroid development (Schep et al., 2017). In brief, chromVAR aggregates accessible regions sharing the same TF motif, then compares the observed accessibility of all peaks containing that TF motif with a background set of peaks normalizing for known technical confounders. Among the most variably accessible sequence motifs determined, we identified important hematopoietic TF motifs such as from SPI1, GATA1, CEBPA, RUNX1, and FOXD4 (Figures 3C and S3C). Notably, we observed profound differences in the accessibility within our sampled populations, including significant loss of inferred activity of erythroid master regulator GATA1 after the BasoE stage (Figures 4A and 4B) (Ludwig et al., 2014; Subramanian et al., 2005). Interestingly, GATA1 activity declined preceding enucleation, suggesting that its expression becomes dispensable at the final stage of erythroid differentiation. Indeed, using western blot analysis, we documented reduced GATA1 protein levels in late-stage erythroid cells in our in vitro culture system (Figures 4C, 4D, S4A, and S4B), consistent with concomitant downregulation of HSP70 protein levels and its protective role in caspase-3-dependent GATA1 cleavage (Gautier et al., 2016; Ribeil et al., 2007) (Figure S4C). In contrast, KLF1 motif accessibility and target gene expression appeared to increase throughout terminal differentiation, consistent with increasing transcript levels and higher chromatin accessibility in the KLF1 locus throughout the late stages compared with the GATA1 locus (Figures S4D–S4G). These findings support the notion that GATA1 and KLF1 play distinct regulatory roles in orchestrating the erythroid maturation program. As such, human KLF1 mutations result in distinct

Figure 2. Time-Dependent Modules of Transcriptional Regulation Define Erythroid Differentiation Programs

(A) Heatmap of all differentially expressed genes, clustered by dynamic expression using k-means clustering across hematopoietic progenitor populations (HSC, MPP, CMP, and MEP) and stages of erythropoiesis (MyP, CFU-E, ProE1, ProE2, BasoE, PolE, OrthoE, and Orth/Ret). Color bar, expression Z score of differentially expressed genes.

(B) Top enriched Gene Ontology biological processes for clusters K4, K5, K6, and K7 of differentially expressed genes. The top five processes are shown for each cluster and their \(-\log_{10}(\text{false discovery rate})\).
erythroid phenotypes (Arnaud et al., 2010; Borg et al., 2010) compared with GATA1-associated pathologies; the latter are characterized by altered erythroid lineage commitment or impaired differentiation of erythroid progenitors (Campbell et al., 2013; Crispino and Horwitz, 2017; Khajuria et al., 2018; Ludwig et al., 2014; Sankaran et al., 2012a).

We next hypothesized that rare variants disrupting GATA1 motifs would also be most accessible at stages of maximal GATA1 transcriptional activity. Investigating the chromatin accessibility of 11 variants known to disrupt GATA1 regulatory elements, including several in human Mendelian erythroid disorders (Campagna et al., 2014; Kaneko et al., 2014; Manco et al., 2000; Wakabayashi et al., 2016), we indeed found that BasoE have maximal chromatin accessibility in regulatory regions encompassing many of these variants (Figure 4E; Table S5). For instance, GATA1 element mutations in uroporphyrinogen III synthase (UROS), a heme biosynthetic enzyme implicated in congenital erythropoietic porphyria (Solis et al., 2001), as well as in the core promoter of the erythroid-specific RH50 antigen (Iwamoto et al., 1998), both fell within chromatin peaks which were maximally accessible in BasoE (Figures 4F and 4G). Altogether, these findings emphasize the role of GATA1 in earlier stages of human terminal erythroid differentiation.

**Human Genetic Variation in the Context of Chromatin Dynamics**

Motivated by our finding of differential GATA1 activity in rare disease-associated erythroid variants, we next investigated whether differential chromatin dynamics could extend more broadly to common variants associated with RBC traits. GWAS have identified thousands of common genetic variants that fall within accessible regions of chromatin in diverse tissues and cell types and have been associated with various human phenotypes, including blood cell traits (Astle et al., 2016). However, it remains an open question as to how and at what stage of differentiation or cell state these variants primarily affect hematopoietic development and traits. Given that the majority of nominated genetic variants fall into gene-regulatory regions and are thought to alter gene expression, we reasoned that our comprehensive open chromatin profiles could provide insights into the precise stages of erythroid development during which such regulation may occur. First, we intersected hematopoietic cell type-specific regions of open chromatin with fine-mapped variants affecting six commonly measured erythroid cell traits nominated by a recent GWAS (Ulirsch et al., 2019). Among the 114,905 individuals comprising this GWAS cohort, 4,327 (3.76%) had hemoglobin concentrations meeting the

---

**Figure 3. Dynamics of the Open Chromatin Landscape and Enrichment of Transcriptional Regulators**

(A) Heatmap of all differentially accessible peaks, clustered using k-means clustering across hematopoietic progenitor populations (HSC, MPP, CMP, and MEP) and stages of erythropoiesis (MyP, CFU-E, ProE1, ProE2, BasoE, PolyE, OrthoE, and Orth/Ret). Color bar, accessibility Z score of differentially peaks identified by ATAC-seq.

(B) Distribution of accessible peaks in the ALAS2 locus across sampled populations.

(C) Rank order plot of transcription factor binding sites with greatest variability in chromatin accessibility across sampled cell populations.

(D) Heatmap showing temporal changes in chromatin accessibility for the top 20 TFs with greatest accessibility variability between populations profiled. Color bar, chromVAR accessibility deviation Z score.
criteria for clinical anemia as defined by the World Health Organization (2011), suggesting that our analysis may be relevant for better understanding the genetics of both normal variation, as well as disease risk. Overall, 347 of 1,789 (19.4%) fine-mapped GWAS variants with posterior probability (PP) > 0.10 for causal association fell within a chromatin-accessible region in one or more of the characterized populations. Upon k-means clustering of these variants by chromatin dynamics (Figures 5A and S5A; Table S6), we observed that a majority showed maximum accessibility at specific stages of erythroid differentiation rather than being uniformly accessible, suggesting potential windows of active gene regulation.

We next sought to leverage this comprehensive data to more finely assess the stages of erythropoiesis during which RBC trait-associated variants may act using g-chromVAR, a recently developed high-resolution cell type enrichment tool (Ulirsch et al., 2019). Interestingly, cells at the earlier stages of differentiation (CFU-E and ProE1/2) were significantly enriched for variants regulating hematocrit (HCT) and hemoglobin (HGB) levels (Figure 5B). In contrast, the BasoE and PolyE stages were significantly enriched for variants affecting mean corpuscular volume (MCV), mean corpuscular HGB (MCH), MCH concentration (MCHC), and RBC count (Figure 5B). To dissect this duality, we looked at high-confidence regulatory ATAC-seq peaks containing fine-mapped variants with a combined PP of 0.10 or higher. Although fewer absolute variants were associated with HGB and/or HCT relative to MCV, MCH, and/or RBC count, the density of HGB and/or HCT variants in chromatin-accessible regions was shifted toward earlier stages of erythropoiesis, whereas a more balanced distribution was observed for MCV, MCH, MCHC, and/or RBC count variants (Figures 5C, S5B, and S5C). We note that HCT is calculated as the product of RBC count and MCV, two traits that are negatively correlated, such that when a variant alters the RBC count and MCV in opposing directions, the net effect on HCT is neutralized. Indeed, 41% of fine-mapped variants associated with RBC count were not associated with HGB or HCT (Figure S5D), consistent with previous studies showing higher genetic correlations between HGB and HCT compared with those with a RBC count (Ulirsch et al., 2019). Thus, the enrichment for HGB and/or HCT variants acting in earlier stages of erythropoiesis is likely driven in part by HGB and/or HCT-associated variants affecting the RBC count or MCV alone or in consistent directions, but not in opposite directions.
Our previous work has shown that variants rs112233623 and rs9349205 in CCND3 (Figure 5D), associated with MCV and RBC count, act by modulating the expression levels of cyclin D3, thereby affecting the number of cell divisions during terminal erythropoiesis (Sankaran et al., 2012b; Ulirsch et al., 2019). Although this had pronounced effects on the RBC count and MCV (in opposite directions), HGB and HCT were less affected in Ccnd3-knockout mice, providing an example of how common genetic variants modulate different erythroid traits at distinct stages of erythropoiesis. Here, we show that both rs112233623 and rs9349205 fall within a single chromatin peak that opens up in the BasoE, PolyE, and OrthoE stages (Figure 5D), consistent with the expected differentiation stage of these variants’ phenotypic effects. Similarly, in the KLF1 locus, we identified a region...
Figure 6. Erythroid-Specific Isoform Expression and Regulation of TMCC2
(A) Line plot showing 14,640 differentially expressed genes across indicated populations, with TMCC2 highlighted.
(B) TMCC2 protein copy number across human erythroid populations.
(C) TMCC2 locus with corresponding RNA-seq (top) and ATAC-seq data (bottom) across indicated populations. Track plot depicting two isoforms of TMCC2 is shown in the middle. Compared with MyP-PolyE stages, certain regions lose chromatin accessibility (blue shade), whereas others gain accessibility in late erythropoiesis (OrthoE, pink shade). The shorter TMCC2 isoform is the dominant isoform expressed in blood cells.
(D) Schematic showing the KRAB domain fused to dCas9 used to target the enhancer (pink shade) or promoter (gray shade) non-coding DNA regulatory regions in the TMCC2 locus. ATAC-seq tracks of HUDEP-2 cells of the TMCC2 locus are shown in green.
(E) Guide RNAs targeting the enhancer and promoter result in reduction of TMCC2 expression as determined using RT-qPCR in HUDEP-2 cells. Error bars represent ±1 SD between replicates.
with maximum chromatin accessibility at the later stages of erythropoiesis (BasoE to OrthoE encompassing rs80326512 (Figure S5E), a fine-mapped (\( P = 0.24 \)) variant associated with MCV. When considered together with the expression of \( KLF1 \) and its target gene set both peaking during terminal stages (Figures S4D and S4E), these findings suggest that rs80326512 acts primarily during the late stages of erythropoiesis. As a third example, a fine-mapped MCV variant, rs117747069 (\( P = 1.00 \)), falls in a region that is both (1) maximally accessible during late stages of erythropoiesis and (2) correlated with alpha-globin expression (Figure S9F) (Iotchikova et al., 2016).

Going the other direction, genetic fine-mapping identified two putative causal variants (rs28357093, \( P = 0.33 \); rs833070, \( P = 0.28 \)) in the vascular endothelial growth factor A (VEGFA) locus associated with HGB and HCT, both falling in regions with higher chromatin accessibility at earlier stages of erythropoiesis (CFU-E, ProE1/2) (Figure 5E). VEGFA has been shown to regulate erythropoietin signaling in response to hypoxia (Rehn et al., 2014; Tam et al., 2006), suggesting that VEGFA may preferentially act during early erythropoiesis to modulate expansion of total erythroid cell mass. Other variants in early-accessible regions included HGB-associated rs1175550 (\( P = 0.36 \)), a known modifier of \( SMIM1 \) that encodes the Vel blood group antigen (Ulirsch et al., 2016), and rs218264 (\( P = 0.31 \)) upstream of \( KIT \), an essential receptor tyrosine kinase for hematopoietic progenitor cells (Figures S5G and S5H) (Edling and Hallberg, 2007). Our integrative findings suggest that these variants and associated genes may play a more dominant role in the regulation of HGB and HCT during early rather than late stages of erythroid differentiation.

Finally, certain loci may contain multiple regulatory elements active at distinct stages of erythropoiesis, as exemplified by the \( HBS1L-MYB \) intergenic region. One HCT-associated (\( P = 0.18 \)) variant from our data, rs66650371, falls within a region with maximal chromatin accessibility at the CFU-E, ProE1/2 stages (Figure 5F) and has been previously implicated as a functional variant significantly associated with fetal HGB and \( MYB \) expression (Stadhouders et al., 2014). In contrast, another fine-mapped \( HBS1L-MYB \) intergenic variant associated with several erythroid traits (HCT, \( P = 0.27 \); HGB, \( P = 0.15 \); RBC count, \( P = 0.10 \)), rs9402685, falls within a region with increased chromatin accessibility in BasoE to OrthoE populations. Strikingly, both rs66650371 and rs9402685 localize within functional enhancer elements (–84 and –83 kb upstream of the \( MYB \) transcription start site) previously shown to block erythroid differentiation and reduce \( MYB \) expression upon their deletion (Canver et al., 2017). Although complex linkage disequilibrium (LD) structure prevents us from ruling out the possibility of other causal variants in these two loci, the present analysis expands upon our previous work (Guo et al., 2017) and adds insights into the complex regulation at the \( HBS1L-MYB \) region by suggesting that the –84 element likely acts at early stages of erythropoiesis, whereas the –83 element may act at later stages of erythropoiesis. Altogether, our stage-specific matched profiles of ATAC-seq and RNA-seq are beginning to provide a comprehensive resource to delineate specific differentiation stages at which various erythroid phenotypes and genes are most highly regulated, allowing functional follow-up of nominated variants and genes at high temporal and stage-specific resolution.

**An Erythroid-Specific Isoform of TMCC2 Is an Essential Regulator of Terminal Human Erythropoiesis**

Our dataset provides a rich resource to expedite functional studies of regulators of erythropoiesis. Highlighting one such example, we identified \( TMCC2 \) as one of the most strongly induced genes during the terminal stages of erythroid differentiation (Figure 6A). \( TMCC2 \) is highly expressed in the human brain and has been implicated in neurodegeneration (Hopkins, 2013), but a role in hematopoietic development has not been previously described (Figure S6A). Of note, \( TMCC2 \) showed high expression in whole blood, appeared to be selectively expressed in erythroblasts (Figures S6A and S6B) with maximum protein expression levels at the OrthoE stage (Figure 6B) (Gautier et al., 2016), and was strongly induced in differentiating G1E-ER cells, suggesting a cell type-specific role in erythropoiesis (Figure S6C). Surprisingly, we observed decreasing chromatin accessibility near the promoter of the reference transcript (Figure 6C, bottom, blue shade), but an increase in accessible chromatin at an alternative promoter and nearby putative enhancer in late stages, supporting an isoform (ENST00000329800.7) of \( TMCC2 \) that is blood specific (Figure S6D). Indeed, coverage tracks of the RNA-seq data uniquely supported the expression of the shorter isoform, indicating that this transcript may have a specific role in erythropoiesis (Figure 6C, top). To confirm the functional roles of these regulatory elements, we used a CRISPR interference (CRISPRi) strategy, taking advantage of the local heterochromatin formation and functional silencing mediated by a Kruppel-associated box (KRAK) domain fused to a catalytically dead Cas9 (dCas9) (Figures 6D and S6E) (Gilbert et al., 2014). Consistent with a role in \( TMCC2 \) gene regulation, we observed significant downregulation of transcript levels as measured by qRT-PCR upon targeting the promoter or enhancer of \( TMCC2 \) with two independent guide RNAs in HUDEP-2 cells (Kurita et al., 2013), which was more pronounced when targeting the promoter (Figure 6E). Importantly, the enhancer showed great specificity for \( TMCC2 \) regulation as determined using RNA-seq analysis (Figures 6F–6H).

To interrogate a potential functional role and the relevance of \( TMCC2 \) in RBC differentiation, we introduced two independent short hairpin RNAs (shRNAs sh1 and sh2, both targeting the 3’ end of the gene) by lentiviral infection to knockdown its expression in human erythroid progenitors. Both shRNAs resulted in pronounced reduction of \( TMCC2 \) mRNA levels compared with a non-targeting control (shLuc) as measured by

---

(F and G) Bar plots showing RNA-seq expression counts per million (CPM) for \( TMCC2 \) (F) and other expressed genes (G) in the locus when treated with guide RNAs (gRNAs) targeting the enhancer and a non-targeting guide. Error bars represent \( \pm 1 \) SD between replicates. (G) Pools replicates between gRNAs.

(H) Volcano plot of all genes considered for differential expression between the pooled enhancer gRNAs and a non-targeting control.
apoptotic annexin V+ cells (C) and impaired erythroid differentiation (D) as

TMCC2

sent

3, 5, and 8 post-infection at the three indicated time points. Error bars repre-

resent ± 1 SD between replicates. (A) Short hairpin RNAs (sh1 and sh2) knockdown TMCC2 expression as determined using RT-qPCR in in vitro cultured primary hematopoietic cells. Error bars represent ± 1 SD between replicates. (B) Erythroid cells with shRNA knockdown of TMCC2 demonstrate reduced proliferative capacity compared with a non-targeting control (shluc). Day (x axis) represents the day following the start of the culture, representing days 3, 5, and 8 post-infection at the three indicated time points. Error bars represent ± 1 SD between replicates. (C and D) TMCC2-knockdown cells further show increased frequency of apoptotic annexin V+ cells (C) and impaired erythroid differentiation (D) as shown by altered morphology and an increased frequency of disrupted cells (arrows, 63× magnification, MayGrunwald staining). In (C), error bars represent ± 1 SD between replicates.

qRT-PCR (Figure 7A). Knockdown of TMCC2 resulted in altered cell proliferation (Figure 7B) and reduced levels of cellular viability as indicated by flow cytometry with altered forward and side scatter properties (Figure S7A). We also noted increased cell apoptosis as measured by annexin V staining and use of DNA binding dyes (Figures 7C, S7A, and S7B). We did not observe significant alterations in the expression of erythroid markers CD71 and CD235a in knockdown compared with control cells (shluc) (Figure S7B). However, terminal erythroid maturation appeared significantly impaired as demonstrated by altered cellular morphology and an increased frequency of “disrupted” cells on Cytospin images of TMCC2-knockdown cells (Figure 7D, arrows).

As such, the observed decrease in proliferation (Figure 7B) may also be in part a consequence of increased cell lysis in these late stages of erythroid differentiation. Finally, from population-based sequencing studies of healthy individuals (ExAC) (Lek et al., 2016), we observed fewer than expected loss of function (observed = 3, expected = 12) and missense (observed = 216, expected = 330) variants in TMCC2, suggesting that mutations in this gene are poorly tolerated and selected against (constraint metric \( z = 3.07 \)). Taken together, these results identify an erythroid-specific isoform of TMCC2 as an essential regulator of human terminal erythroid differentiation.

**DISCUSSION**

Here, we paired RNA-seq and ATAC-seq to provide deep transcriptomic profiles and describe the dynamic accessible chromatin landscape across distinct stages of adult human erythroid differentiation. Our work provides a rich resource for functional studies and complements previously published transcriptomic and proteomic datasets (An et al., 2014; Gautier et al., 2016; Li et al., 2014; Yan et al., 2018). Although prior studies have performed epigenetic analyses at specific stages (Buenrostro et al., 2018; Corces et al., 2016; Huang et al., 2016), our comprehensive analysis across the entire process of human erythroid differentiation provides substantial insights. We comprehensively characterized differential regions of accessible chromatin across erythroid development and identified regulatory elements with maximum accessibility specific to PolyE and OrthoE, stages associated with nuclear condensation and overall decreasing accessibility (Figures 1E and 6). Furthermore, we inferred the TF dynamics and their expression programs that govern the molecular circuits of RBC differentiation (Figures 3 and 4), inferred their maximum activity, and identified TF motifs most accessible in polychromatic and OrthoEs, although their exact roles and transcriptional programs remain to be functionally validated. Although the chromatin accessibility profiles identified in this work enable robust inference of accessibility peaks and global TF dynamics, these data have limited capacity to identify site-specific TF binding (via TF-footprinting) or nucleosome positioning. Thus, additional assays applied in our in vitro system, such as ChIP-seq of specific factors, may further reveal insights into dynamic chromatin states during erythropoiesis.

Integration of ATAC-seq profiles with summary statistics from GWAS and Mendelian erythroid disorders further enable the nomination of genetic variants acting in gene regulatory elements and imply a potential window of activity (Figures 4 and 5). Our analyses suggest that regulatory elements and the variants harbored within may show maximum activity at different stages of erythroid differentiation, thereby affecting distinct erythroid traits or causing different blood disease phenotypes. Variants in regulatory regions associated with HGB and HCT appeared to be primarily accessible and regulate gene expression at the early stages of RBC development. In contrast, RBC count and MCV were primarily affected by regulatory elements containing variants most accessible in late stage erythroblasts, as showcased by our previous work characterizing the role of cyclin D3 in terminal mouse and human erythropoiesis (Sankaran et al., 2012b). We nominate multiple additional gene regulatory elements containing common variants associated with distinct erythroid traits and suggest stage-specific activities, suggesting that these elements may be functionally important in regulating human erythroid development.
that our data may aid in the identification of non-coding variants underlying GWAS results.

Because ~90% of variants identified in GWAS (including hematopoietic traits) are in the non-coding genome (Ulirsch et al., 2019), cell type-specific epigenomic profiling provides a powerful means for more precise annotations of functionally relevant variants underlying human phenotypes. As such, we anticipate that our resource will inform functional validation studies, which may further include the use of chromatin conformation capture, massively parallel reporter assays, and genome editing approaches via CRISPR and related methods (Canver et al., 2017; Ulirsch et al., 2016; Wakabayashi et al., 2016). As an example, our data identified a blood-specific isoform of TMCC2 that is highly induced at the PolyE/OrthoE stage (Figure 6). Knockdown of TMCC2 resulted in pronounced proliferation and differentiation defects, enabling us to identify this gene as an essential regulator of terminal human erythropoiesis (Figure 7). In summary, our data provide a comprehensive resource on the expression dynamics, TF elements and their transcriptional programs, as well as their relation to genetic variation to inform functional studies of human erythroid differentiation.

**STAR METHODS**

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **CONTACT FOR REAGENT AND RESOURCE SHARING**
- **EXPERIMENTAL MODEL AND SUBJECT DETAILS**
  - Human erythroid cell culture
  - HDUEP-2 cell culture
  - 293T cell culture
- **METHOD DETAILS**
  - Flow cytometry analysis and apoptosis
  - Flow cytometry activated cell sorting
  - May-Grünewald-Giemsa staining
  - FAST-ATAC-seq
  - RNA-seq
  - Sequencing
  - Western blotting
  - Lentiviral shRNA vectors and infection
  - Construction of the lentiviral CRISPRi constructs
  - CRISPRi experiments in HDUEP-2 cells
  - Quantitative RT-PCR
- **QUANTIFICATION AND STATISTICAL ANALYSIS**
  - Data processing and read alignment
  - Principal component analyses
  - Differential gene expression and chromatin occupancy
  - Gene set enrichments
  - Downstream ATAC-seq analyses
  - Integration with Genome-Wide Association Studies
- **DATA AND SOFTWARE AVAILABILITY**

**SUPPLEMENTAL INFORMATION**

Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2019.05.046.

**ACKNOWLEDGMENTS**

We thank members of the Sankaran, Regev, and Aryee laboratories for valuable comments and the Whitehead Institute and Broad Institute Flow Cytometry facilities for assistance with cell sorting and flow cytometric analysis. We thank Ryo Kurita and Yukio Nakamura for providing HDUEP-2 cells. E.L.B. received support from the Howard Hughes Medical Institute Medical Research Fellows program. K.C. received support from NIH training grant T32GM07753. J.D.B. received support from the Broad Institute Fellows program and the Allen Institute Distinguished Investigator Award. This research was supported by NIH grant P01 DK02094 (N.M.), the Howard Hughes Medical Institute (A.R.), the Klaman Cell Observatory (A.R.), NIH grants R01 DK103794 and R33 HL120791 (V.G.S.), and the New York Stem Cell Foundation (NYSCF; V.G.S.). V.G.S. is a NYSCF-Robertson Investigator.

**AUTHOR CONTRIBUTIONS**


**DECLARATION OF INTERESTS**

A.R. is a founder and equity holder in Celsius Therapeutics and an SAB member of the ThermoFisher Scientific and Syros Pharmaceuticals.

Received: January 8, 2019
Revised: March 3, 2019
Accepted: May 14, 2019
Published: June 11, 2019

**REFERENCES**


### STAR★METHODS

#### KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Antibodies</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FITC anti-human CD71, clone OKT9</td>
<td>eBioscience</td>
<td>Cat#: 14-0719-82; RRID: AB_467338</td>
</tr>
<tr>
<td>APC anti-human CD235a, clone HIR2</td>
<td>eBioscience</td>
<td>Cat#: 17-9987-42; RRID: AB_2043823</td>
</tr>
<tr>
<td>APC anti-human CD49d, clone 9F10</td>
<td>Biolegend</td>
<td>Cat#: 304308; RRID: AB_2130041</td>
</tr>
<tr>
<td>FITC anti-human BAND3</td>
<td>Laboratory of Narla Mohandas; Hu et al., 2013</td>
<td>N/A</td>
</tr>
<tr>
<td>Goat polyclonal anti-GATA1 (M-20)</td>
<td>Santa Cruz Biotechnology</td>
<td>Cat#: sc-1234; RRID: AB_2263157</td>
</tr>
<tr>
<td>Mouse monoclonal anti-ACTB (AC-15)</td>
<td>Sigma-Aldrich</td>
<td>Cat#: A1978; RRID: AB_476692</td>
</tr>
<tr>
<td>Anti-goat peroxidase-coupled antibody</td>
<td>Jackson ImmunoResearch</td>
<td>Cat#: 715-035-147; RRID: AB_2313587</td>
</tr>
<tr>
<td>Anti-mouse peroxidase-coupled antibody</td>
<td>Jackson ImmunoResearch</td>
<td>Cat#: 715-035-150; RRID: AB_2340770</td>
</tr>
<tr>
<td><strong>Biological Samples</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human CD34+ hematopoietic stem and progenitor cells, adult</td>
<td>Fred Hutchinson Cancer Research Center</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Chemicals, Peptides, and Recombinant Proteins</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PBS</td>
<td>GIBCO</td>
<td>Cat#: 10010-023</td>
</tr>
<tr>
<td>Dulbecco’s Modified Eagle Medium-High Glucose (DMEM)</td>
<td>GIBCO</td>
<td>Cat#: 11965-118</td>
</tr>
<tr>
<td>Iscove’s Modified Dulbecco’s Medium (IMDM)</td>
<td>GIBCO</td>
<td>Cat#: 12440-053</td>
</tr>
<tr>
<td>Opti-MEM</td>
<td>GIBCO</td>
<td>Cat#: 31985-062</td>
</tr>
<tr>
<td>StemSpan SFEM II medium</td>
<td>STEMCELL Technologies</td>
<td>Cat#: 09655</td>
</tr>
<tr>
<td>Human AB serum</td>
<td>Atlanta Biologicals</td>
<td>Cat#: S40110</td>
</tr>
<tr>
<td>Human AB plasma</td>
<td>SeraCare</td>
<td>Cat#: 1810-0001</td>
</tr>
<tr>
<td>Recombinant Erythropoietin (EPO)</td>
<td>Amgen</td>
<td>Cat#: NDC 55513-267-10</td>
</tr>
<tr>
<td>Interleukin 3 (IL-3)</td>
<td>Peprotech</td>
<td>Cat#: 200-03</td>
</tr>
<tr>
<td>Stem cell factor (SCF)</td>
<td>Peprotech</td>
<td>Cat#: 300-07</td>
</tr>
<tr>
<td>Human Holo-Transferrin</td>
<td>Sigma-Aldrich</td>
<td>Cat#: T0665</td>
</tr>
<tr>
<td>Dexamethasone Sodium Phosphate</td>
<td>Mylan</td>
<td>Cat#: NDC67457-420-10</td>
</tr>
<tr>
<td>Heparin</td>
<td>Hospira</td>
<td>Cat#: NDC 00409-2720-01</td>
</tr>
<tr>
<td>Doxycycline</td>
<td>Sigma-Aldrich</td>
<td>Cat#: D3072</td>
</tr>
<tr>
<td>Fetal Bovine Serum (FBS)</td>
<td>Atlanta Biologicals</td>
<td>Cat#: S11150</td>
</tr>
<tr>
<td>Penicillin-Streptomycin</td>
<td>GIBCO</td>
<td>Cat#: 15140-122</td>
</tr>
<tr>
<td>FuGENE 6 Transfection Reagent</td>
<td>Promega</td>
<td>Cat#: E2691</td>
</tr>
<tr>
<td>Polybrene Infection/Transfection reagent</td>
<td>Millipore</td>
<td>Cat#: TR-1003-G</td>
</tr>
<tr>
<td>SYTOX Blue Dead Cell Stain</td>
<td>Thermo Fisher</td>
<td>Cat#: S34857</td>
</tr>
<tr>
<td>SYTOX Orange Dead Cell Stain</td>
<td>Thermo Fisher</td>
<td>Cat#: S34861</td>
</tr>
<tr>
<td>Buffer RLT</td>
<td>QIAGEN</td>
<td>Cat#: 79216</td>
</tr>
<tr>
<td>2-mercaptoethanol</td>
<td>Sigma</td>
<td>Cat#: M6250</td>
</tr>
<tr>
<td>Recombinant Ribonuclease Inhibitor (40U/ul)</td>
<td>Clontech</td>
<td>Cat#: 2313B</td>
</tr>
<tr>
<td>DigiKinin</td>
<td>Promega</td>
<td>Cat#: G9441</td>
</tr>
<tr>
<td>Trehalose Solution, 1M, Sterile</td>
<td>Life Sciences Advanced Technologies</td>
<td>Cat#: TS1M-100</td>
</tr>
<tr>
<td>dNTP mix (10mM)</td>
<td>Thermo Fisher</td>
<td>Cat#: R0193</td>
</tr>
<tr>
<td>Magnesium Chloride</td>
<td>Sigma-Aldrich</td>
<td>Cat#: M1028-10X1ML</td>
</tr>
<tr>
<td>Buffer EB</td>
<td>QIAGEN</td>
<td>Cat#: 19086</td>
</tr>
<tr>
<td>TE Buffer</td>
<td>Thermo Fisher</td>
<td>Cat#: 12090015</td>
</tr>
</tbody>
</table>

(Continued on next page)
## Critical Commercial Assays

<table>
<thead>
<tr>
<th>REAGENT Description</th>
<th>SOURCE Information</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>BV421-Annexin V</td>
<td>BD Biosciences</td>
<td>Cat#: 563973</td>
</tr>
<tr>
<td>RNase-Free DNase Set</td>
<td>QIAGEN</td>
<td>Cat#: 79254</td>
</tr>
<tr>
<td>NEBNext High-Fidelity 2X PCR Master Mix</td>
<td>New England Biolabs</td>
<td>Cat#: M0541L</td>
</tr>
<tr>
<td>Maxima H-minus RT (200 u/µL)</td>
<td>Thermo Fisher</td>
<td>Cat#: EP0752</td>
</tr>
<tr>
<td>KAPA HiFi HotStart PCR ReadyMix</td>
<td>Kapa Biosystems</td>
<td>Cat#: KK2602</td>
</tr>
<tr>
<td>Agencourt AMPure XP</td>
<td>Beckman-Coulter</td>
<td>Cat#: A63881</td>
</tr>
<tr>
<td>Agencourt RNA Clean XP</td>
<td>Beckman-Coulter</td>
<td>Cat#: A63987</td>
</tr>
<tr>
<td>Qubit dsDNA HS Assay Kit</td>
<td>Thermo Fisher</td>
<td>Cat#: Q32854</td>
</tr>
<tr>
<td>Qubit RNA HS Assay Kit</td>
<td>Thermo Fisher</td>
<td>Cat#: Q32852</td>
</tr>
<tr>
<td>Bioanalyzer High Sensitivity DNA Analysis</td>
<td>Agilent</td>
<td>Cat#: 5067-4626</td>
</tr>
<tr>
<td>E-Gel EX Gel, 2%</td>
<td>Thermo Fisher</td>
<td>Cat#: G402002</td>
</tr>
<tr>
<td>Tn5 enzyme from Nextera DNA Library Preparation Kit</td>
<td>Illumina</td>
<td>Cat#: FC-121-1031</td>
</tr>
<tr>
<td>Nextera XT DNA Library Preparation Kit</td>
<td>Illumina</td>
<td>Cat#: FC-131-1096</td>
</tr>
<tr>
<td>NextSeq 500/550 High Output Kit v2.5 (75 Cycles)</td>
<td>Illumina</td>
<td>Cat#: 20024906</td>
</tr>
</tbody>
</table>

## Deposited Data

<table>
<thead>
<tr>
<th>Deposited Data Description</th>
<th>Source Information</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw and processed data</td>
<td>This paper</td>
<td>GEO: GSE115684</td>
</tr>
<tr>
<td>Code and processed data</td>
<td>This paper</td>
<td><a href="https://github.com/sankaranlab/erythroid-profiling">https://github.com/sankaranlab/erythroid-profiling</a></td>
</tr>
<tr>
<td>Human reference genome UCSC build 19, hg19</td>
<td>University of California Santa Cruz</td>
<td></td>
</tr>
<tr>
<td>Human erythroblast RNA-seq expression profiles</td>
<td>Yan et al., 2018</td>
<td>GEO: GSE107218</td>
</tr>
<tr>
<td>Human hematopoietic progenitor ATAC-seq datasets</td>
<td>Corces et al., 2016</td>
<td>GEO: GSE74310</td>
</tr>
<tr>
<td>GTEx RNA-seq expression profiles</td>
<td>dbGAP</td>
<td><a href="https://gtex%E9%97%A8%E6%88%B7%E7%BD%91%E7%AB%99/home/">https://gtex门户网站/home/</a></td>
</tr>
</tbody>
</table>

## Experimental Models: Cell Lines

<table>
<thead>
<tr>
<th>Experimental Models: Cell Lines Description</th>
<th>Source Information</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>293T cells</td>
<td>ATCC</td>
<td>Cat#: CRL-3216</td>
</tr>
<tr>
<td>HUDEP-2</td>
<td>Kurita et al., 2013</td>
<td>N/A</td>
</tr>
</tbody>
</table>

(Continued on next page)
CONTACT FOR REAGENT AND RESOURCE SHARING

Requests for further information or reagents should be directed to and will be fulfilled by the Lead Contact, Vijay G. Sankaran (sankaran@broadinstitute.org).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human erythroid cell culture

Human adult CD34+ hematopoietic stem and progenitor cells were obtained from the Fred Hutchinson Hematopoietic Cell Processing and Repository (Seattle, USA). No human subjects were involved in this study. The CD34+ samples were deidentified and approval for use of these samples was provided by the Institutional Review Board and Biosafety Committees at Boston Children’s Hospital. Donor characteristics: ID R003221 (Female, 65 years old), ID R003498 (Female, 46 years old), ID R003172 (Male, 54 years old). These cells were differentiated into mature erythroid cells utilizing a three-phase culture protocol (Giani et al., 2016; Hu et al., 2013). In phase 1 (day 0 – 7), cells were cultured at a density of 10^5 - 10^6 cells per milliliter (ml) in IMDM supplemented with 2% human AB plasma, 3% human AB serum, 1% penicillin/streptomycin, 3 IU/ml heparin, 10 μg/ml insulin, 200 μg/ml holo-transferin, 1 IU erythropoietin (Epo), 10 ng/ml stem cell factor (SCF) and 1 ng/ml IL-3. In phase 2 (day 7 – 12), IL-3 was omitted from the medium.
In phase 3 (day 12 – 18), cells were cultured at a density of $10^6$ cells per milliliter, with both IL-3 and SCF omitted from the medium and the holo-transferrin concentration was increased to 1 mg/ml. Cells were cultured at 37°C and 5% CO₂.

**HUDEP-2 cell culture**

HUDEP-2 cells were maintained in StemSpan SFEM II medium (Stem Cell Technologies) supplemented with hSCF (50 ng/ml), erythropoietin (3 IU/ml), dexamethasone (10⁻⁶ M) and doxycycline (1 μg/ml). For erythroid differentiation cells were cultured in differentiation media (IMDM, 2% human AB plasma, 3% human AB serum, 1 mg/mL Holo-human transferrin, 10 μg/mL insulin and 3 IU/mL erythropoietin. Cells were cultured at 37°C and 5% CO₂.

**293T cell culture**

293T cells (ATCC) were maintained in DMEM (GIBCO) supplemented with 10% FBS, 1% penicillin/streptomycin and 2 mM L-glutamine. Cells were cultured at 37°C and 5% CO₂.

**METHOD DETAILS**

**Flow cytometry analysis and apoptosis**

For flow cytometry analysis, *in vitro* cultured erythroid cells were washed in FACS buffer (1% FBS in PBS) before antibody staining. The following antibodies at indicated dilutions were used in this study: 1:50 APC-conjugated CD235a (Glycophorin A, clone HIR2, 50-153-69, eBioscience), 1:50 FITC-conjugated CD71 (OKT9, 14-0719-82, eBioscience), 1:50 APC-conjugated CD49d (9F10, Biolegend, 304308) and 1:200 FITC-conjugated BAND3 (An et al., 2014; Hu et al., 2013). For live/ dead cell discrimination Sytox Blue was used according to the manufacturer’s instructions (Thermo Fisher, S34857). For apoptosis analysis, BV421-Annexin V (BD Bioscience, 563973) was used in combination with Sytox Orange according to the manufacturer’s instructions (Thermo Fisher, S34861). FACS analysis was conducted on a BD Bioscience Fortessa flow cytometer at the Whitehead Institute Flow Cytometry core. Data were analyzed using FlowJo software v10.4.2.

**Flow cytometry activated cell sorting**

*In vitro* cultured erythroid cells were washed in FACS buffer and stained as described above using antibodies against CD235a, CD71, CD49d and BAND3. Sytox Blue (Thermo Fisher, S34857) was used for live/ dead cell discrimination. Cells were sorted using a Sony SH800 sorter with a 100 μm chip at the Broad Institute Flow Cytometry facility.

**May-Grünwald-Giemsa staining**

50,000 – 100,000 cells were harvested, washed once at 300 x g for 5 min, resuspended in 200 μL of FACS Buffer, and spun onto poly-L-lysine coated microscope slides with a Shandon 4 (Thermo Scientific) cytocentrifuge at 300 rpm for 4 min. When visibly dry slides were transferred into May-Grünwald solution (Sigma-Aldrich) for 5 min, rinsed 4 times for 30 s in water, and transferred to Giemsa solution (Sigma-Aldrich) for 15 min. Slides were washed as described above, dry mounted with coverslips, and examined. All images shown were taken using a Metafer slide scanning platform and software (Metasystems) at 63X magnification.

**FAST-ATAC-seq**

For ATAC-seq library preparations 10,000-15,000 cells were washed in PBS, pelleted by centrifugation and lysed and tagmented in 1x TD buffer, 2.5 μl Tn5 (Illumina), 0.01% Digitonin (Promega, G9441), 0.3x PBS in a 50 μl reaction volume as described (Corces et al., 2016). Samples were incubated at 37°C for 30 min at 300 rpm. Tagmented DNA was purified using the MinElute PCR kit (QIAGEN). The complete eluate underwent PCR, as follows. After initial extension, 5 cycles of pre-amplification using indexed primers (Buenrostro et al., 2015) and NEBNext High-Fidelity 2X PCR Master Mix (NEB) were conducted, before the number of additional cycles was assessed by quantitative PCR using SYBR Green. Typically, 7-9 additional cycles were run. The final library was purified using a MinElute PCR kit (QIAGEN) and quantified using a Qubit dsDNA HS Assay kit (Invitrogen) and a High Sensitivity DNA chip run on a Bioanalyzer 2100 system (Agilent).

**RNA-seq**

Cells were lysed in RLT lysis buffer (QIAGEN) supplemented with beta-mercaptoethanol and RNA was isolated using a RNeasy Micro kit (QIAGEN) according to the manufacturer’s instructions. An on-column DNase digestion was performed before RNA was quantified using a Qubit RNA HS Assay kit (Invitrogen) and a High Sensitivity DNA chip run on a Bioanalyzer 2100 system (Agilent), followed by library preparation using the Nextera XT kit (Illumina) and custom index primers according to the manufacturer’s instructions. Final libraries were quantified using a Qubit dsDNA HS Assay kit (Invitrogen) and a High Sensitivity DNA chip run on a Bioanalyzer 2100 system (Agilent).
Sequencing
All libraries were sequenced using Nextseq High Output Cartridge kits and a Nextseq 500 sequencer (Illumina). Libraries were consistently sequenced paired end (2x38bp).

Western blotting
Cells were washed twice in PBS, resuspended in RIPA lysis buffer (Santa Cruz Biotechnology) supplemented with 1x Complete Protease Inhibitor Cocktail (Roche) and incubated for 30 min on ice. After centrifugation at 14,000 rpm for 10 min at 4°C to remove cellular debris, the remaining supernatant was transferred to a new tube, supplemented with 4x Laemmli sample buffer (Bio-Rad) and beta-mercaptoethanol and incubated for 10 min at 95°C. Equal amounts of proteins were separated by gel electrophoresis using the Mini-PROTEAN gel system and Tris/Glycine/SDS running buffer (Bio-Rad). Subsequently, proteins were transferred onto a PVDF membrane using Tris/Glycine transfer buffer supplemented with methanol (Bio-Rad). Membranes were blocked with 3% BSA-TBST for 1 h and probed with GATA1 goat polyclonal antibody (M-20, sc-1234, Santa Cruz Biotechnology) at a 1:500 dilution or β-actin mouse monoclonal (AC-15, Sigma) at a 1:5,000 dilution in 3% BSA-TBST for 1 h at room temperature or overnight at 4°C. Membranes were washed four times with TBST, incubated with donkey anti-mouse or anti-goat peroxidase-coupled secondary antibodies (715-035-150 and 705-035-147, Jackson ImmunoResearch) at a 1:10,000 dilution in 3% BSA-TBST for 1 h at room temperature, washed three times with TBST and incubated for 1 min with Western Lightning Plus-ECL substrate (PerkinElmer). Proteins were visualized by exposure to scientific imaging film (Kodak).

Lentiviral shRNA vectors and infection
The shRNA constructs targeting human TMCC2 (sh1 and sh2, SHCLNG-NM_014858) were obtained from the Mission shRNA collection (Sigma-Aldrich). The sequences of the shRNAs used in this study are

- sh1: CCGGGCAAGTGTTCGAGAAGAAGAACTCGAGTTCTTCTTCTCGAACACTTGCTTTTTTG
- sh2: CCGGCCTGACTGAGCTTCATCAGAACTCGAGTTCTGATGAAGCTCAGTCAGGTTTTTTG

As controls, the lentiviral vectors pLKO-GFP and shluc were used. For production of lentiviruses, 293T cells were transfected with the appropriate viral packaging and genomic vectors (pVSV-G and pDelta8.9) using FuGene 6 reagent (Promega) according to the manufacturer’s protocol. The medium was changed the day after transfection to phase I medium (described above). After 24–30 h, this medium was collected and filtered using an 0.22 μm filter immediately before infection of primary hematopoietic cells. The cells were mixed with viral supernatant supplemented with cytokines in the presence of 8 μg/ml polybrene (Millipore) in a 6-well plate at a density of 250,000–500,000 cells per well. The cells were spun at 2,000 rpm for 90 min at 22°C and left in viral supernatant overnight. The medium was replaced the morning after infection. Selection of infected cells was started 24 h after infection with 1 μg/ml puromycin for up to 48 h. Infection efficiency of pLKO-GFP–infected cells was assessed by measuring the frequency of GFP+ cells by flow cytometry 48-72 h post infection. Typically, the frequency of GFP+ cells was between 40%–70%.

Construction of the lentiviral CRISPRi constructs
We utilized the LentiCRISPRv2 backbone for constructing the CRISPRi constructs (Sanjana et al., 2014). The sgRNA scaffold in this construct was replaced by sgRNA-(F+E)-combined optimized scaffold from the sgOpti plasmid (Addgene # 85681) (Fulco et al., 2016). The puromycin resistance gene was replaced with the GFP gene using BamHI and SacII sites. To achieve better expression in hematopoietic cells, the EF1α promoter was replaced by the MSCV promoter using Gibson assembly. The SpCas9 was replaced with the KRAB-dCas9 from the pH-R-SFFV-KRAB-dCas9-P2A-mCherry construct using BamHI sites (Addgene: #60954) (Gilbert et al., 2014). A scheme of the vector construct is shown in Figure S6E. The sequences of the guide RNAs used in this study are:

- Control / non-targeting guide: ATCGCGAGGACCCTCCGCC
- TMCC2 promoter gRNA1: CCTGGCAAGCATATTACAT
- TMCC2 promoter gRNA2: GTATAGTTTCCATGAGCCCA
- TMCC2 enhancer gRNA1: GTCGTGCTGCAGGTGAAGTG
- TMCC2 enhancer gRNA2: CCATCCCTACAGTAAACAG

CRISPRi experiments in HUDEP-2 cells
The cells were infected with all-in-one CRISPRi lentiviral constructs containing KRABdcas9 driven by the MSCV promoter linked to GFP (by a self-cleaving 2A peptide sequence from porcine teschovirus) and an U6 promoter driving respective guide RNAs targeting the TMCC2 gene locus. HUDEP-2 cells were infected as described above and GFP+ cells were FACs sorted 48 h post-infection and expanded for an additional 6 days. HUDEP-2 cells were then differentiated in erythroid differentiation media as described above for 72 h and RNA was extracted using a RNeasy Mini kit (QIAGEN) for gene expression analysis. RNA-seq libraries were constructed for three replicates of the non-targeting gRNA and two replicates for each of the on-target enhancer gRNAs. Differential gene expression analysis was performed as stated below for the FACS-sorted primary cells.
Quantitative RT-PCR
Isolation of RNA was performed using the RNeasy Mini Kit (QIAGEN). An on-column DNase (QIAGEN) digestion was performed according to the manufacturer’s instructions. RNA was quantified by a NanoDrop spectrophotometer (Thermo Scientific). Reverse transcription was carried out using the iScript cDNA synthesis kit (Bio-Rad). Real-time PCR was performed using the CFX384 Real-Time PCR system and iQ SYBR Green Supermix (Bio-Rad). Quantification was performed using the ΔΔCT method. Normalization was performed using ACTB mRNA as a standard. The primers used for quantitative RT-PCR are:

ACTB forward: 5’-AGAAAAATCTGGCACCACACC-3’
ACTB reverse: 5’-GGGGTGTTGAAAGGTCTCAA-3’
TMCC2 forward: 5’-GCAGCGATGATGAGTGCTC-3’
TMCC2 reverse: 5’-GTGCATTGGACTTAGGGCTCC-3’
TMCC2 forward 2: 5’-CTACATGACCAGTGCTGC-3’
TMCC2 reverse 2: 5’-CTCCTGCTTCAGGTTCGTCA-3’

QUANTIFICATION AND STATISTICAL ANALYSIS

Data processing and read alignment
For each sequencing library generated in this study, libraries were sequenced on an Illumina NextSeq 500 and demultiplexed using the bc2fastq program. For each library, raw .fastq reads were aligned using either Bowtie2 version 2.3.3 (ATAC-seq) (Langmead and Salzberg, 2012) or STAR version 2.5.1b (RNA-seq) (Dobin et al., 2013) to the hg19 reference genome. Chromatin accessibility peaks were called using MACS2 using custom parameters for ATAC-seq (–nomodel–nolambda–keep-dup all–call-summits) (Zhang et al., 2008). Gene expression counts were summarized using the–quantMode functionality in STAR. Downstream processing of sequencing data was performed using Samtools (Li et al., 2009). Per population ATAC and RNA-seq values were visualized as heatmaps were generated using Complexheatmap (Gu et al., 2016).

Principal component analyses
To establish a uniform feature set for the ATAC-seq data, we used all 1 bp summits from the MACS2 peak calls per population and expanded them a uniform 250 bp, similar to a strategy implemented previously (Corces et al., 2016). Summits overlapping the same window were iteratively centered at the summit with the strongest signal (measured by -log10 FDR) until all summits were accounted. Gene x sample and Peak x sample matrices were log2 counts-per-million normalized and then transformed to Z scores before the first two principle components were computed via the implicitly restarted Lanczos bidiagonalization algorithm (irlba).

Differential gene expression and chromatin occupancy
To test for differential gene expression from our RNA-seq data and differential chromatin accessibility in individual loci, we used the DEseq2 method (Love et al., 2014). Statistically significant genes varying between two populations were identified at an independent hypothesis weighting (IHW) value of 0.01 (Ignatiadis et al., 2016). A total of 56 comparisons between populations were performed (28 for ATAC-seq; 28 for RNA-seq).

Gene set enrichments
Gene set enrichments of 7 clusters of differentially expressed genes were performed using the Functional Mapping and Annotation of Genome-Wide Associations (FUMA) platform, using all protein-coding genes as background model and requiring a minimum overlap of 2 genes and an FDR-adjusted p < 0.01 for each gene-set. Only Gene Ontology (GO) biological processes were considered.

Downstream ATAC-seq analyses
All motif analyses were performed using the motifmatchr package as part of the chromVAR suite of tools (Schep et al., 2017). For insertion footprinting analyses, reads were offset by +4/-5 base pairs as originally described (Buenrostro et al., 2013). Transcription factor activity scoring was performed using combined replicates for each population for both the data generated in this study (Figures 3C and 3D) and augmented with populations previously described (Corces et al., 2016) (Figures S3C and S3D). We note that as chromVAR builds a background peak set based on the population intensities, deviation z-scores are not absolute but relative to the populations considered.

Integration with Genome-Wide Association Studies
We obtained GWAS summary statistics and fine-mapping results of 6 erythroid traits (HGB, HCT, MCV, RBC count, MCH, and MCHC), which were performed on ~115,000 individuals of European ancestry from the UK Biobank, as described extensively elsewhere (Ullrich et al., 2019). For g-chromVAR enrichment analysis, the set of all fine-mapped variants with PP > 0.001 for each trait were used as input.

Next, to nominate high-confidence fine-mapped variants located in strong chromatin peaks, we took the consensus peak set for all stages of erythropoiesis, performed row and column quantile normalization on the counts matrix, and kept only peaks that had a
maximum count in the top 80% for at least one of the eight sampled cell populations. We then overlapped fine-mapped variants with posterior probability (PP) > 0.10 with this subset of strong peaks.

DATA AND SOFTWARE AVAILABILITY

The accession number for the raw sequencing data reported in this paper is GEO: GSE115684 and is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE115684. A UCSC Track Hub for the ATAC-seq data generated for these eight populations is available here: https://s3.amazonaws.com/atacerythropoiesis/hub.txt. A git repository containing code for analyses described here and processed data files is available at www.github.com/sankranlab/erythroid-profiling.