# Title:
# Gene-Centric Functional Dissection of Human Genetic Variation Uncovers Regulators of Hematopoiesis

## Authors and Affiliations

Satish K. Nandakumar[1,2,12], Sean K. McFarland[1,2,12], Laura Mateyka[1,2,3,13], Caleb A. Lareau[1,2,4,13], Jacob C. Ulirsch[1,2,4,13], Leif S. Ludwig[1,2], Gaurav Agarwal[1,2,5,6], Jesse M. Engreitz[2,7], Bartlomiej Przychodzen[8], Marie McConkey[9,] Glenn Cowley[2], John G. Doench[2], Jaroslaw P. Maciejewski[8], Benjamin L. Ebert[2,9,10], David E. Root[2], Vijay G. Sankaran[1,2,6,11]

[1] Division of Hematology/Oncology, The Manton Center for Orphan Disease Research, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA

[2] Broad Institute of MIT and Harvard, Cambridge, MA, USA

[3] Biochemistry Center (BZH), Ruprecht-Karls-University Heidelberg, 69120 Heidelberg, Germany

[4] Program in Biological and Medical Sciences, Harvard Medical School, Boston, MA, USA

[5] University of Oxford, Oxford, UK

[6] Harvard Stem Cell Institute, Cambridge, MA, USA

[7] Harvard Society of Fellows, Harvard University, Cambridge, MA USA

[8] Department of Translational Hematology and Oncology Research, Taussig Cancer Institute, Cleveland Clinic, Cleveland, OH, USA

[9] Division of Hematology, Brigham and Women's Hospital, Boston, MA, USA

[10] Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA

[11] Corresponding author

[12] Co-first authors

[13] Co-second authors

Correspondence: sankaran@broadinstitute.org

Keywords: hematopoiesis, genome-wide association studies, human genetics, functional screen, erythropoiesis

## Abstract

Genome-wide association studies (GWAS) have identified thousands of variants associated with human diseases and traits. However, the majority of GWAS-implicated variants are in non-coding regions of the genome and require in depth follow-up to identify target genes and decipher biological mechanisms. Here, rather than focusing on causal variants, we have undertaken a pooled loss-of-function screen in primary hematopoietic cells to interrogate 389 candidate genes contained in 75 loci associated with red blood cell traits. Using this approach, we identify 77 genes at 38 GWAS loci, with most loci harboring 1-2 candidate genes. Importantly, the hit set was strongly enriched for genes validated through orthogonal genetic approaches. Genes identified by this approach are enriched in specific and relevant biological pathways, allowing regulators of human erythropoiesis and modifiers of blood diseases to be defined. More generally, this functional screen provides a paradigm for gene-centric follow up of GWAS for a variety of human diseases and traits.

# Introduction

As genotyping technologies and accompanying analytical capabilities have continued to improve, genome-wide association studies (GWAS) have identified tens of thousands of variants associated with numerous human diseases and traits. Despite these advances, our ability to discern the underlying biological mechanisms for the vast majority of such robust associations has remained limited, with a few exceptions (Claussnitzer et al., 2015; Gupta et al., 2017; Mohanan et al., 2018; Musunuru et al., 2010; Sankaran et al., 2008; Smemo et al., 2014). In general, published successes have required in-depth mechanistic studies of individual loci and implicated genes to decipher biological mechanisms.

Recent innovations in functional and computational genomics have advanced the field and enabled more rapid and higher-throughput identification of putative causal variants. Approaches that have shown the most success include the use of massively parallel reporter assays to examine allelic variation (Tewhey et al., 2016; Ulirsch et al., 2016; Vockley et al., 2015) and perturbation approaches for dissecting the necessity of regulatory elements (Fulco et al., 2016; Simeonov et al., 2017). In addition, genetic fine mapping approaches have improved our ability to identify putative causal variants among larger sets of variants in linkage disequilibrium (Guo et al., 2016; Huang et al., 2017; Lareau et al., 2018). However, even when putative causal variants are identified at a disease or trait-associated locus, they most often localize to non-coding regions of the genome, making it difficult to connect variants to genes that mediate the observed effects in a scalable manner (Claussnitzer et al., 2015; Gupta et al., 2017; Smemo et al., 2014).

In the context of hematopoiesis, GWAS studies have identified thousands of variants associated with various blood cell traits, including hundreds associated with red blood cell traits alone (Astle et al., 2016; van der Harst et al., 2012). Thorough follow-up efforts at individual loci have identified important regulators of hematopoiesis, such as the key regulator of fetal hemoglobin expression, BCL11A (Basak et al., 2015; Liu et al., 2018; Sankaran et al., 2008). However, as in other tissues, the low-throughput with which associated genetic variants can be connected to target genes underlying phenotypes continues to pose a problem for gaining biological insights and clinical actionability in complex traits and diseases.

To accelerate the rate at which genetic variants can be connected to target genes, high-throughput loss-of-function screens involving putative causal genes underlying the genetic associations can be undertaken. This approach is complementary to conventional variant-focused methods and overcomes bottlenecks that can arise during downstream target gene identification. As a proof-of-principle, we connected variants associated with RBC traits to genes regulating erythropoiesis by directly perturbing all candidate genes in primary human hematopoietic stem and progenitor cells (HSPCs) undergoing synchronous differentiation into the erythroid lineage. We demonstrate unique opportunities to rapidly implicate likely causal genes and identify networks of biological actors underlying trait-associated variation. We additionally illustrate the value of such screens to uncover previously unappreciated biological regulators of human hematopoiesis that may serve as key disease modifiers.

# Results

## Design and Execution of an shRNA Screen Using Blood Cell Trait GWAS Hits to Identify Genetic Actors in Erythropoiesis

We applied a gene-centric loss-of-function screening approach to GWAS of RBC traits. We focused on 75 loci associated with RBC traits that were identified by a GWAS performed in up to 135,000 individuals (van der Harst et al., 2012) spanning 6 RBC traits (Figure S1A). Importantly, these 75 loci have been robustly replicated and show large effect sizes in more recently reported association studies performed on larger cohorts and thus represent ideal targets for perturbation studies (Astle et al., 2016; Lareau et al., 2018). We endeavored to select all genes that could potentially underlie these 75 GWAS signals. To do this, each of the 75 sentinel SNPs was first expanded to a linkage disequilibrium (LD) block including all SNPs in high LD ($r^2 > 0.8$, Figure 1A, Figure S1B), then further to the nearest genomic recombination hotspot. Based upon insights from previous expression quantitative trait locus (eQTL) studies (Montgomery and Dermitzakis, 2011; Rossin et al., 2011; Veyrieras et al., 2008), each gene annotated in the genome was expanded to include a wingspan encompassing 110 kb upstream and 40 kb downstream of the transcriptional start and end sites, respectively, to also capture potential functional regulatory elements. This resulted in selection of 389 genes overlapping or in the vicinity of the LD blocks to be tested in the pooled loss-of-function screen. These were distributed at a median of 4 genes per loci (Figure S1C).

Since the majority of common genetic variation underlying RBC traits appears to act in a cell-intrinsic manner within the erythroid lineage, we decided to perturb the candidate genes during the process of human erythropoiesis (Giani et al., 2016; Sankaran et al., 2012; Sankaran et al., 2008; Ulirsch et al., 2016). We chose a pooled short hairpin RNA (shRNA) based loss-of-function approach in primary hematopoietic cells to leverage a number of distinct strengths. First, we have had prior success validating individual genes underlying RBC traits using shRNA-based approaches in primary CD34$^+$ HSPC-derived erythroid cells. Second, shRNA libraries can be much more efficiently packaged into lentiviruses and delivered to primary CD34$^+$ cells compared to alternative CRISPR/Cas9-based guide RNA libraries (Ting et al., 2018). Third, it avoids potential complications like non-uniform loss-of-function or gain-of-function outcomes produced by CRISPR/Cas9 based approaches due to unpredictable DNA repair processes (Mandegar et al., 2016). Furthermore, shRNAs can act rapidly to achieve gene knockdown and thereby avoid compensatory effects that can occur when complete CRISPR knockout is achieved (Rossi et al., 2015), better recapitulating the subtle changes in gene expression that are characteristic of common genetic variation.

Mobilized peripheral blood-derived primary human CD34$^+$ HSPCs from 3 independent healthy donors were infected with a lentiviral-based pooled shRNA library consisting of 2803 hairpins targeting the 389 GWAS-nominated genes, along with 30 control genes (Moffat et al., 2006). Each gene was targeted with 5-7 distinct shRNAs (Figure S1D). The set of control shRNAs encompassed essential control genes, negative controls (e.g. luciferase and other non-expressed genes), and a well-defined set of genes important for erythropoiesis (erythroid controls) (Figure 1B, Table S2). Using lentiviral libraries with defined titers, we achieved an infectivity of 35-50%, which is optimal for obtaining either zero or one stably integrated shRNA per cell, while minimizing the possibility of infection of a single cell by multiple viruses. To achieve sufficient library representation, we infected at least 1000 CD34$^+$ HSPCs per hairpin (7~11 * 10$^6$ cells per experiment). The infected HSPCs were cultured using a three-phase erythroid differentiation method (Giani et al., 2016; Hu et al., 2013) that results in synchronous differentiation and maturation of erythroid progenitors into RBCs. We hypothesized that hairpins targeting potential regulators of erythropoiesis would be depleted or enriched during the three-phase erythroid culture, similar to our prior experience in analyzing specific GWAS-nominated genes (Giani et al., 2016; Sankaran et al., 2012; Ulirsch et al., 2016). To assay these hairpins, we isolated and deep-sequenced genomic DNA from the pool of infected cells at 6 different culture time points that represent distinct stages of erythropoiesis to most broadly assess putative causal genes that may act across the span of differentiation (Figure 1C, S1E).

**Summary Characterization of shRNA Screen Outcomes**

For the vast majority of the ~3000 hairpins included in the library, infection was efficient and consistent. Greater than 95% of hairpins were represented at levels of at least 5 $\log_2$ counts per million (CPM) at day 4, two days post-infection (Figure 2A). Across the two-week time course, a diversity of effects - in terms of both increased and decreased hairpin abundance - were observed. While many hairpins were selected against during the course of erythroid differentiation, as reflected in decreases of those hairpin abundances over time, there were also a number of hairpins that increased in the culture over the time course (Figure 2B, S2A).

The tested set of hairpins targeting genes nominated by the 75 loci showed a variety of activities, forming a broad distribution spanning both decreases and increases in abundance at different time points (Figure 2C). The various controls included in the library behaved as expected. Hairpins targeting genes with known significance to erythropoiesis, such as *GATA1* and *RPS19* (Khajuria et al., 2018; Ludwig et al., 2014), showed markedly decreased abundance across the time course. Likewise, hairpins targeting a set of broadly essential genes (Table S2) were strongly depleted by day 16 when compared to negative control hairpins targeting non-human genes, which showed little if any change (Figure 2C-E). These trends were recapitulated with strong correlation in each of the three donor $CD34^+$ cell backgrounds (Figure S2B).

## Statistical Modeling of Gene Effects and Accounting for Confounders in the shRNA Screen

The resulting longitudinal observations of hairpin abundance at each time point were used to model the importance of each targeted gene during the process of erythropoiesis. A linear mixed model was implemented to account for the longitudinal nature of the time course data (Li et al., 2015) and to handle the confounding off-target and efficiency effects inherent to the shRNA modality (Riba et al., 2017; Tsherniak et al., 2017). Since we wanted our model to be able to detect significant changes in hairpin abundance at any time point throughout the differentiation process, we converted the absolute hairpin abundances at each of the six time points to a $\log_2$ fold change relative to the initial hairpin abundances at the start of the differentiation. Using this metric as our response variable, we specified a fixed effect for each gene to capture the contribution that suppressing it with shRNAs would have on the respective abundances for each of the resulting five time intervals. Given the potential variability that could emerge by using shRNAs, we fit a random effect for each hairpin to minimize the chance of conflating inefficiency or off-target effects with the specific on-target gene effect.

After fitting this model to the data, we selected our hit set using a two-threshold approach in which both the magnitude and statistical confidence of the estimated gene effect size were considered. Specifically, genes were called as hits if they had a fitted slope > 0.1 $\log_2$ fold change per day within the interval while simultaneously possessing a Wald chi-square FDR-adjusted q value < 0.1. This combined approach allowed us to avoid focusing on genes with large, but highly variable or conflicted effects, as well as genes with highly confident but miniscule effects. In total, this approach identified 77 genes at 38 of the 75 targeted loci which, when suppressed, had a significant effect on the slope of shRNA-encoding DNA abundance at any point during the time course. A majority of these hit loci (27 loci) had 1-2 gene targets prioritized (Figure 3A, S3A). These candidate genes were found to be distributed across all 6 of the originally annotated RBC GWAS traits (Figure S3B).

To evaluate the validity of this hit set, we began by assaying for enrichment of erythroid essentiality, as recently quantified for each gene in the K562 erythroid cell line (Wang et al., 2015). A permutation comparing the sum of K562 essentiality scores for the hit genes with those of randomly drawn, identically-sized gene sets from the library of targeted genes revealed that the hit set was indeed enriched with p = 0.0269 (Figure 3B). Likewise, when compared to permuted sets of 77 genes randomly chosen from the genome (Figure S3C), there was even stronger enrichment for erythroid essentiality with p = 0.00021, consistent with the idea that genes in the library likely have stronger essentiality due to their genomic proximity to the GWAS hits. We further explored whether the enrichment could be due to an intrinsic bias inherent to GWAS screening itself by permuting sets of genes from libraries nominated by SNPs associated with low-density lipoprotein levels, high-density lipoprotein levels, and triglyceride levels, finding the hit set to be significantly enriched in all comparisons (Willer et al., 2013) (Figure S3D-F).

We further validated the ability of this approach to discover genetically relevant hits by performing a permutation analysis based upon five “gold standard” genes in the library, which possess known genetic underpinnings via identified causal variants: *CCND3* (Lareau et al., 2018; Sankaran et al., 2012), *SH2B3 (Giani et al., 2016)*, *MYB* (Galarneau et al., 2010; Sankaran et al., 2013; Sankaran et al., 2011), *KIT* (Jing et al., 2008; Lareau et al., 2018), and *RBM38* (Ulirsch et al., 2016). Calculating the rank sums of hairpins ordered by our model’s computed FDR scores for 1,000,000 random combinations of five genes from the library yielded a distribution over which enrichment for the five gold standards was seen with p=0.0249 (Figure

3C). While the vast majority of putative causal variants at the RBC trait-associated loci are in non-coding regions, which can be challenging to use to identify a specific target gene, a subset are in coding regions and thereby nominate a specific gene. As a result, we assayed for the presence of coding variants fine-mapped to the interrogated loci from a recent large GWAS that demonstrated a minimum posterior probability of association of 0.1 among the gene hits and compared this with the overall set of genes interrogated in our library (Lareau et al., 2018). Among the 389 GWAS-nominated genes in our library, 20 (~5%) were found to contain at least one coding variant from this list. Of these, there was a significant enrichment observed among the hits (~9%, p=0.03907 as determined by permutation analysis; Figure 3D).

Having established genetic confidence in our hit set, we next investigated whether the selected genes satisfied enrichment criterion within the erythroid branch of hematopoiesis. RNA expression values for each of the 77 hit genes were examined in datasets spanning human hematopoiesis (Corces et al., 2016), as well as adult and fetal erythropoiesis (Yan et al., 2018) (Figure 3E,F; Figure S3G). In the more holistic hematopoiesis dataset, common myeloid progenitors (CMPs) and megakaryocyte-erythroid progenitors (MEPs) were significantly enriched for hit genes ($p < 0.01$). These progenitor populations are known to contain the progenitors that give rise to erythroid cells. Within a more detailed and separate analysis of human adult erythropoiesis, proerythroblast, early basophilic, and late basophilic erythroblast stages were particularly enriched ($p < 0.001$). The stage at which given genes are implicated to play a role in erythropoiesis from the literature likewise often corresponded with the largest magnitude fold changes across the longitudinal time course measurements, as was the case for earlier genes like *RPL7A*, *RPL23A*, *RPS19*, and *KIT* (Gazda et al., 2012; Jing et al., 2008; Moniz et al., 2012) as well as late genes like *SLC4A1* and *ANK1* (Bennett and Stenbuck, 1979; Peters et al., 1996). Taken together, these results show that this functional gene-centric screen allows for the identification of putative causal genes underlying RBC-trait GWAS hits, which demonstrate clear enrichment in independent genetic and cell biological datasets. We are therefore able to validate the utility of such an approach to identify biologically-relevant genes underlying human genetic variation and holistically identify potential stages by which such target genes may act to impact the process of hematopoiesis.

### Analysis of Interactions Among Members of the Hit Set Identifies Signaling, Structural, and Translation-Related Subnetworks Important to Erythropoiesis

By screening all loci and genes at once, our approach afforded us the immediate value of examining mechanisms underlying the associations in a holistic fashion, unearthing both familiar and more novel core gene network cassettes that play a role in erythropoiesis (Boyle et al., 2017). Using STRING interaction network analyses (version 10.5) (Szklarczyk et al., 2017), we could identify connectivity between the underlying nodes that highlighted a number of interacting biological processes of both known and previously unappreciated importance to erythropoiesis (Figure 4). We observed a number of molecules that play roles in cell signaling or transcriptional regulation. MYB is a master regulator transcription factor that has been implicated in playing a role in fetal hemoglobin regulation and in erythropoiesis more generally (Mucenski et al., 1991; Wang et al., 2018). The *MYB* locus has been associated with numerous red blood cell traits (including mean corpuscular volume, mean corpuscular hemoglobin concentration, and RBC count) (Sankaran et al., 2013; van der Harst et al., 2012). ETO2 (CBFA2T3) is a part of the erythroid transcription factor complex containing TAL1 and is required for expansion of erythroid progenitors (Goardon et al., 2006). Both stem cell factor receptor KIT and erythropoietin receptor (EPOR) mediated signaling are essential for erythropoiesis. Our screen identified KIT as one of the factors underlying common genetic variation. CCND3 fills a critical role in regulating the number of cell divisions during terminal erythropoiesis  and has been validated as a causal gene associated with variation in RBC counts and size (Lareau et al., 2018; Sankaran et al., 2012).

Interacting networks of hits also emerged in other aspects of red blood cell differentiation and function. One of these centered around membrane and structural cytoskeletal proteins. Our method recovered characteristic RBC genes like solute carrier family 4 member 1 (SLC4A1), also known as band 3, (Peters et al., 1996), which serves as a key component of the RBC membrane skeleton. Likewise, it recovered a direct interacting partner for SLC4A1, ankyrin 1 (ANK1), which anchors the cytoskeleton and cell membrane (Bennett and Stenbuck, 1979), as well as N-ethylmaleimide Sensitive Factor, vesicle fusing ATPase (NSF), which facilitates membrane vesicle trafficking within the cell (Glick and Rothman, 1987).

Within the realm of mRNA translation, a number of genes emerged as hits that specifically highlight the role of the ribosome. This is interesting in light of recent work that has begun to illuminate erythroid-specific effects of ribosomal perturbations (Khajuria et al., 2018; Ludwig et al., 2014), although a connection between translation and common genetic variation affecting RBC traits has not been previously appreciated. Both *RPL7A* and *RPL19*, for instance, have been implicated by mutations observed in studies of Diamond-Blackfan anemia (Gazda et al., 2012; Moniz et al., 2012). The common genetic variation affecting these ribosomal protein genes might contribute to the incomplete penetrance and variable expressivity of anemia seen in Diamond-Blackfan anemia patients (Ulirsch et al., 2018). Similar effects have been reported in neurodevelopmental disorders, where common genetic variants may influence phenotypic outcomes in patients (Niemi et al., 2018). Non-ribosomal hits in the mRNA metabolism space were also found with both previously established and unknown ties to erythroid-specific phenotypes. Exosome component 9 (EXOSC9), for instance, has been demonstrated previously to act as part of the exosome complex as a specific gatekeeper of terminal erythroid maturation (McIver et al., 2014). Other unappreciated components, including the tRNA methyltransferase TRMT61A, also were highlighted through this analysis.

### Transferrin receptor 2 is a Negative Regulator of Human Erythropoiesis

We selected several candidate genes identified by our screen for further validation, given their previously unappreciated roles in human hematopoiesis/ erythropoiesis. The first, transferrin receptor 2 (*TFR2*), encodes a protein canonically involved in iron homeostasis that has recently been shown to also regulate EPO receptor signaling (Forejtnikova et al., 2010; Nai et al., 2015). Although TFR2 has been studied in the context of murine erythropoiesis, its role in human erythropoiesis has not been assessed. To validate TFR2 as a regulator of human erythropoiesis, we performed individual knockdown experiments using lentiviral shRNAs in primary human CD34$^+$ HSPCs undergoing erythroid differentiation. Significant knockdown of TFR2 was observed at both the mRNA (Figure 5A) and protein levels (Figure 5B) using two independent shRNAs from among the six targeting TFR2 in the screen. Though two of the six were outliers, the two chosen here for follow-up were part of the consensus group of four showing similar effects. Downregulation of TFR2 increased erythroid differentiation as observed by increased expression of erythroid specific cell surface markers CD235a and CD71 at day 9 (shLUC ~22%; TFR2 sh1 ~42%; TFR2 sh2 ~40%) and day 12 of culture (shLUC ~60%; TF2 sh1 ~80%; TFR2 sh2 ~80%) (Figure 5C, E & S5A). Downregulation of TFR2 also improved the later stages of erythroid differentiation/ maturation, as observed by an increased rate of enucleation and through assessment of cell morphology (Figure 5D & S5B). Previous studies have reported the isolation of TFR2 as a component of the erythropoietin (EPO) receptor complex (Forejtnikova et al., 2010). To test if downregulation of TFR2 can result in increased EPO signaling, we measured EPO-dependent STAT5 phosphorylation after TFR2 knockdown in UT7/EPO cells

(Figure S5C). TFR2 downregulation resulted in significantly higher pSTAT5 phosphorylation in comparison to the control with EPO stimulation from 0.02 U/mL to 200 U/mL (Figure 5F). In addition, the maximal pSTAT5 response could be achieved within a shorter period of EPO stimulation upon TFR2 downregulation (Figure S5D). Given our findings that TFR2 is a negative regulator of EPO signaling, it may be an ideal therapeutic target for conditions characterized by ineffective erythropoiesis like β-thalassemia (Rund and Rachmilewitz, 2005). A recent study has supported this hypothesis, showing that Tfr2 downregulation is beneficial in a mouse model of β-thalassemia (Artuso et al., 2018).

### SF3A2 is a Key Regulator of Human Erythropoiesis and is a Disease Modifier in a Murine Model of Myelodysplastic Syndrome.

Extensive mRNA splicing occurs during the terminal stages of erythropoiesis (Pimentel et al., 2016). However, key regulators of this process remain largely undefined. Our study uncovered splicing factor 3A subunit 2 (SF3A2) in the subnetwork of erythropoiesis signaling and transcription hits (Figure 4). SF3A2 specifically was associated with maximal hairpin drop out at day 12 (FDR = 0.005) – a later time point in erythropoiesis. SF3A2 is a component of the U2SNRP complex whose binding to the branch point is critical for proper mRNA splicing (Gozani et al., 1996; Gozani et al., 1998). Knockdown of SF3A2 in primary human CD34$^+$ HSPCs results in decreased cell numbers during erythroid differentiation starting from day 7 (Figure 6A-C). To measure early effects of SF3A2 and to exclude potential toxicity of puromycin selection, we replaced the puromycin resistance gene with a GFP encoding cDNA in the lentiviral shRNA constructs. We achieved similar infection (30~40% on day 6) at the early time points between controls (shLuc) and shRNAs targeting *SF3A2* (Figure S6A). During erythroid differentiation, we observed a reduction in GFP-expressing cells comparable to the decreased cell numbers seen with the puromycin resistant constructs (Figure S6A). Decreased cell numbers were associated with decreased erythroid differentiation as measured by erythroid surface markers CD71 and CD235a (Figure 6D). We also observed an increase in non-erythroid lineages based on surface marker expression of CD11b (myeloid) and CD41a (megakaryocyte) (Figure S6B).

To identify the molecular mechanisms underlying the reduced differentiation of erythroid cells we sorted stage-matched CD71$^+$/CD235$^+$ cells and performed RNA-Seq analysis. We also ran this analysis in parallel for data from hematopoietic progenitors from patients with myelodysplastic syndrome (MDS), a disorder well-known for significant impairment in terminal erythropoiesis, either with or without somatic mutations in the related splicing factor *SF3B1 (Obeng et al., 2016)*. Cells treated with shRNA to suppress SF3A2 were found to differentially express 6061 genes with an adjusted p value < 0.05 as compared to the shLuc control, whereas only 807 genes were differentially expressed given the same threshold cutoff in the MDS patients with an *SF3B1* mutations compared to those without (Figure 6E). Genes from both the SF3A2 differentially expressed set and the SF3B1 differentially expressed set were significantly enriched for structural constituents of the ribosome ($p < 3.2 \times 10^{-44}$ and $p < 7.5 \times 10^{-24,}$ respectively) among other cellular components and functions (Tables S6-S9). Examining differential splicing in the set of genes not differentially expressed in either condition, both were found to exhibit a similar proportion of altered splicing events, including alternative 3' splice sites, alternative 5' splice sites, mutally exclusive exons, and skipped exons with bayes factor > 10 (Figure 6E, Tables S10-S19).

We therefore wanted to further explore this connection between SF3A2 and its role in common variation in RBC traits with SF3B1 and the role it plays in the pathogenesis of MDS. To this end, we utilized a recently developed faithful mouse model harboring the *Sf3b1*$^{\mathrm{K700E}}$ mutation that displays characteristic features of MDS, including an anemia due to impaired erythropoiesis (Obeng et al., 2016). We tested if downregulation of SF3A2 could worsen the already impaired erythropoiesis seen in these animals. Equal numbers of lineage-negative HSPCs were isolated from bone marrow of wild-type and *Sf3b1*$^{K700E}$ mice and infected with shRNAs targeting SF3A2 and then erythroid differentiation was induced (Figure S6C, D). Consistent with previous reports, we observed that *Sf3b1*$^{K700E}$ cells show reduced erythroid differentiation and cell growth compared to wild-type cells infected with control non-targeting shRNAs (Figure 6F, G, S6E-G). Downregulation of SF3A2 using two independent shRNAs further worsens the defects in both erythroid differentiation and cell growth observed for *Sf3b1*$^{K700E}$ cells (Figure 6F, G, S6E, F, G). This data suggests that modulation of SF3A2 could modify the alterations of erythropoiesis observed in the setting of somatic *SF3B1* MDS-causal mutations. This form of MDS is characterized by significant variation in the degree of anemia found at the time of presentation (Papaemmanuil et al., 2011). We therefore attempted to examine whether such common genetic variation could contribute to such phenotypic variation. We identified a coding SNP, rs25672, that was in LD with the sentinel SNP at this locus, rs2159213 ($r^2$ = 0.737675 in CEU 1000GENOMES phase 3). Prevalence of the alternate "G" allele (which is associated with the prevalence of the "C" effect allele in the van der Harst et al. locus) has a suggestive correlation with an increase in hemoglobin levels (Figure S6H) that was likely insignificant due to the limited number of patients studied here. Unfortunately, larger cohorts in such a relatively rare disorder could not be identified. However, these findings suggest that the subtle variation noted in populations at the *SF3A2* locus may more profoundly cause variation among individuals with an acquired blood disorder, such as MDS, illustrating the value of such a gene-centric study to identify potential disease modifiers.

# Discussion

A major challenge in moving from GWAS-nominated variants to function is to identify potential target genes systematically. While many functional follow up approaches focus on causal variants, we reasoned that a gene-centered approach may be complementary to other emerging methods and represent a scalable approach for gaining broad insights into GWAS. To this end, we designed and executed a GWAS-informed high-throughput loss-of-function screen to identify key players in primary human HSPCs undergoing erythroid differentiation. Such dynamic *in vitro* systems afford a unique window through which to longitudinally screen, enabling unique insight to be gained into inherently non-stationary biological processes like erythropoiesis. The screen identified 77 gene hits at 38 of the original 75 loci used to design the library. Collectively, these hits had strongly amplified essentiality in erythroid cell lines, included a significant proportion of known, genetically-linked "gold standard" erythroid genes, and were enriched for red blood cell trait-associated coding variants orthogonally identified through genetic fine-mapping. From a holistic perspective, the network of interacting gene hits highlighted a number of high-level biological components and pathways important for erythropoiesis, including specific signaling and transcription factors, membrane and structural components, and components involved in mRNA translation.

Functional follow-up on *SF3A2* and *TFR2*, two gene hits identified in the screen, were fruitful in elucidating mechanistic ties between alteration in mRNA splicing and EPO signaling activity, respectively, to observed perturbation of erythroid phenotypes. In addition, our studies suggest that at least SF3A2, and potentially other regulators such as some implicated mRNA translation factors, may be key disease modifiers that alter the impaired erythropoiesis seen in diseases like MDS or Diamond-Blackfan anemia. These outcomes strongly recommend this screening approach as a rapid means to access the genetic mediators underlying human erythroid differentiation, with the potential to much more rapidly derive actionable biology from GWAS studies. Moreover, since shRNA-based loss-of-function screens are readily accessible and offer demonstrated compatibility with primary cell model systems, we believe this approach provides a method that is portable and can be applied across a variety of lines of biological inquiry.

However, it is not a universal solution, and there are certainly a number of considerations that must be kept in mind regarding the extent to which this type of assay can be adopted across other diseases and traits. We acknowledge for it to be useful to a given research question, a suitable system capable of modeling the trait/ disease of interest must first exist, and for many cellular systems this is often challenging. Fortunately, this is a shortcoming that will diminish over time as our understanding of human biology and our ability to faithfully recapitulate *in vivo* microenvironments and processes improves, though this may be a distant prospect for exquisitely complex tissues like the brain or for traits/ diseases that involve a larger number of cell types/ interactions. Likewise, the use of shRNAs as the vehicle for perturbation carries with it unique challenges, chief among them the proclivity of shRNA to exert confounding off-target effects when compared to CRISPR-based methods. While this is true and unavoidable, the inclusion of appropriate controls, both at the experimental level and in modeling off-target contributors to observed phenotypic effects, provide an effective means to address this issue (Tsherniak et al., 2017). We chose to perform our screen in primary hematopoietic cells and thus were partially limited experimentally to the use of shRNA-based suppressive approaches. Finally, evidence has recently been published that the targets of identified non-coding variants are occasionally not within linkage disequilibrium blocks in the genome (Whalen and Pollard, 2018). This does not necessarily conflict with our results, since we identify hits at only 38 of 75 examined loci and provides an intriguing direction for further work that may elucidate how genetic and epigenomic structural blocks in the human genome can provide complementary information.

Our data shows that gene-centric screens are valuable for GWAS follow-up. They are not limited to red cell traits and may be useful for other human traits/ diseases, as has begun to be shown in disease-systems like Type II diabetes (Thomsen et al., 2016). Data from such screens can be integrated with complementary insights gleaned from variant-centric screens. Ultimately this could accelerate our understanding of human hematopoiesis and other biological processes, and aid in the development of applicable therapies.

## Acknowledgments

We thank members of the Sankaran laboratory for valuable comments and suggestions on these studies. This work was supported by the National Institutes of Health grants R01 DK103794 and R33 HL120791, as well as the New York Stem Cell Foundation (to V.G.S.). V.G.S. is a New York Stem Cell Foundation-Robertson Investigator.

## Author Contributions

V.G.S., D.E.R., S.K.N., S.K.M., J.C.U., G.C. and J.G.D. designed the study. S.K.N performed the pooled shRNA screen experiments in primary cells. G.C. and J.G.D prepared pooled shRNA lentiviral libraries and performed downstream deep sequencing of barcodes. D.E.R. oversaw shRNA library preparation and analysis of the screen. S.K.N., L.M., and G.A. performed all the follow-up experiments. S.K.M., C.A.L, J.C.U., and J.M.E. performed all computational analyses. L.S.L performed RNA sequencing on SF3A2 knockdown samples. M.M. and B.L.E. provided samples from *Sf3b1*$^{K700E}$ knock-in mice. B.P. and J.P.M. analyzed and provided MDS patient samples. S.K.N., S.K.M., and V.G.S. wrote the manuscript with input from all authors. V.G.S. supervised all experimental and analytic aspects of this work.

## Competing Interests:

The authors declare no competing interests.

## Figure Legends

**Figure 1: Design and Execution of an shRNA Screen Using Blood Cell Trait GWAS Hits to Identify Genetic Actors in Erythropoiesis** (A) Overview of shRNA library design. 75 loci associated with red blood cell traits (van der Harst et al., 2012) were used as the basis to calculate 75 genomic windows of LD 0.8 or greater from the sentinel SNP. Genes with a start site within 110 kb or end site within 40 kb of the LD-defined genomic windows were chosen as candidates to target in the screen. (B) Compositional makeup of the library, depicted as number of genes and number of hairpins for each of the four included subcategories; GWAS-nominated genes, erythroid genes, essential genes, and negative control genes. (C) Primary CD34$^+$ hematopoietic stem and progenitor cells (HSPCs) isolated from 3 independent donors were cultured for a period of 16 days in erythroid differentiation conditions. At day 2, cells were infected with the shRNA library, and the abundances of each shRNA were measured at days 4, 6, 9, 12, 14, and 16 using deep sequencing.

**Figure 2: Summary Characterization of shRNA Screen Outcomes** (A) Kernel density plot showing library representation as $\log_2$ shRNA CPM across all hairpins. (B) shRNA abundance $\log_2$ fold changes from day 4 to day 16. Represented values are the mean of hairpin abundance $\log_2$ fold changes across hairpins for each gene and two standard deviations. (C) Kernel density plots representing the day 4 to day 16 $\log_2$ fold changes of hairpin abundances for each of the subcategories of the library, including GWAS-nominated genes, known erythroid essential genes, essential genes to cell viability, and orthogonal genes serving as negative controls. (D) Violin plot of day 4 and day 16 $\log_2$ CPM for known actors *GATA1* and *RPS19* and negative controls LacZ and luciferase. (E) $\text{Log}_2$ hairpin counts averaged for known actors *GATA1* and *RPS19* as well as negative controls LacZ and luciferase across the course of the experiment. Gray lines depict the universe of all other gene traces in the library for context.

**Figure 3: Statistical Modeling of Gene Effect Accounting for Off-target shRNA Confounders** (A) Bar graph showing the 38 of 75 loci in the screen with at least one corresponding statistically significant (FDR < 0.1, β > 0.1) gene effect causing either a positive or negative $\log_2$ fold change in shRNA abundance. (B) Kernel density plot showing the expected distributions of K562 essentiality scores using permuted gene hit sets from the library. (C) Hairpin rank sums for permuted sets of 5 genes. The red line indicates the enriched rank sums for 5 "gold standard" genes included in the library, *CCND3*, *SH2B3*, *MYB*, *KIT*, and *RBM38*, for each which a genetic basis of action has already been established. (D) Permuted distribution of % inclusion of predicted coding variants among the set of identified hits. (E) Heat map depicting strength of expression (as z scores within each gene) for each of the 77 identified hit genes across hematopoietic lineages (top) and throughout the specific stages of adult erythropoiesis (bottom). Purple boxes highlight the cell types that were enriched for expression of hit genes. (F) Calculated enrichment of the identified hit genes for expression across hematopoietic lineages (top) and throughout the specific stages of adult erythropoiesis (bottom). In both cases, cellular states corresponding to those along the erythropoietic lineage had elevated probability of expressing genes from the hit set as compared to other genes from the library.

**Figure 4: Analysis of interactions among members of the hit set identifies signaling/transcription, membrane, and mRNA translation-related subnetworks important to erythropoiesis.** Comparison of gene hits via the STRING database identified interacting networks related to hematopoietic signaling/transcription, membrane, and mRNA translation-related subnetworks**.** Edges of the network are color-coded according to the evidence supporting the interaction.

**Figure 5: Transferrin receptor 2 is a Negative Regulator of Human Erythropoiesis** (A) Quantitative RT-PCR and (B) Western blot showing the expression of TFR2 in human CD34$^+$ cells five days post-infection with the respective lentiviral shRNAs targeting TFR2 (TFR2 sh1 & sh2) and a control luciferase gene (shLUC). (C) Representative FACS plots of erythroid cell surface markers CD71 (transferrin receptor) and CD235a (Glycophorin A) expression at various time points during erythroid differentiation. Percentages in each quadrant is represented as mean and standard deviation of 3 independent experiments (D) Hoechst staining showing more enucleated cells after TFR2 knockdown at day 21 of erythroid culture. (E) Representative histogram plots showing increased expression of CD235a (Glycophorin A) after TFR2 knockdown (F) Enhanced pSTAT5 response after TFR2 knockdown in UT7/EPO cells.

**Figure 6: SF3A2 is a Key regulator of Human Erythropoiesis and Modulates Erythropoiesis Defects in a Murine Model of MDS** (A) Quantitative RT-PCR and (B) Western blot showing the expression of SF3A2 in human CD34$^+$ cells five days post-infection with the respective lentiviral shRNAs targeting *SF3A2* (sh1 & sh2) and a control luciferase gene (shLUC). (C)  Growth curves showing that downregulation of SF3A2 results in reduced total cell numbers during erythroid differentiation from 3 independent experiments. (D) Representative FACS plots of erythroid cell surface markers CD71 (transferrin receptor) and CD235a (Glycophorin A) expression at various time points during erythroid differentiation. Percentages in each quadrant is represented as mean and standard deviation of 3 independent experiments (E) Altered splicing events identified by RNA-Seq analysis of stage matched erythroid cells (shSF3A2 vs. shLUC). Overlapping changes observed in *SF3B1* mutant BM cells from MDS patients (Obeng et al). (F) Lineage negative bone marrow cells from wildtype (WT) and *Sf3b1*$^{K700E}$ mice were infected with shRNAs targeting murine *Sf3a2* gene co-expressing a reporter GFP gene. Percentage of Ter119$^+$ CD71$^+$ erythroid cells within the GFP compartment

after 48hrs in erythroid differentiation. (G) Total cell numbers of GFP$^+$ erythroid cells after 48hrs in erythroid differentiation.

## Materials and Methods:

### Design of the shRNA Library

Ensembl assembly GRCh37p9 was utilized to expand 75 SNPs previously identified in a RBC trait GWAS to include a genomic region in linkage disequilibrium with $r^2 \geq 0.8$. Each of these regions was then further expanded to the nearest recombination hotspot. All genes in the genome were expanded to include 110 kb upstream and 40 kb downstream of the transcription start and end sites, respectively, to maximize capture of non-coding regulatory interactions based upon previously published observations. Genes with windows calculated in this way found to be overlapping with any of the SNP windows were flagged for inclusion in the screen. In addition, each locus was examined individually, and in cases of gene deserts, unusually proximal recombination hotspots, or other unusual genomic structures, the SNP region was expanded to include additional genes nearby. This resulted in a total of 389 test genes, which were each targeted by 4-7 distinct shRNAs. Also included in the library were shRNAs targeting a set of 8 validated erythroid genes (*GATA1, RPL5, RPS19, EPOR, ALAS2, CDAN1, SEC23B, ZFPM1*). A pooled library of 2803 TRC clones was produced from the sequence-validated TRC shRNA library (Moffat et al., 2006) and included shRNAs targeting control genes and essential genes.

### Pooled shRNA Screening

Mobilized peripheral blood $CD34^+$ cells from three separate donors ($7\sim11 * 10^6$ cells per donor) were differentiated into erythroid cells using a three-stage system that has been previously described (Hu et al., 2013). Cells were cultured using IMDM containing 2% human plasma, 3%

human AB serum, 200 µg/ml human holo-transferrin, 3 IU/mL heparin, and 10 mg/mL insulin (base medium). During days 0 to 7, cells were supplemented with IL-3 (1 ng/mL), SCF (10 ng/ml), and EPO (3 IU/ml). On day 2 of this culture, cells were transduced with the pooled lentiviral shRNA library prepared by Broad Institute Genetic Perturbation Platform (1ml of virus per 0.75 * $10^6$ cells) by spinfection at 2000 rpm for 90 minutes with 6 µg/ml polybrene. During days 7 to 13, cells were supplemented with SCF and EPO only. After day 13, cells were supplemented with EPO alone and the holo-transferrin concentration was increased to 1 mg/ml. A minimum of 10 * $10^6$ cells was re-plated at each time point to ensure appropriate library representation and prevent bottlenecks among the infected cells. Cell pellets were made from 20~80 * $10^6$ cells at days 4, 6, 9, 12, 14, and 16. At the conclusion of the pooled screen, genomic DNA (gDNA) was extracted from the cell pellets using NucleoSpin Blood XL-Maxi kit (Clonetech) according to kit specifications. The shRNA-containing region was PCR amplified from the purified gDNA and barcoded using the following conditions: 0.5 µl P5 primer mix (100µM), 10 µl P7 primer mix (5µM), 8 µl dNTP mix, 1x ExTaq buffer, 1.5 µl of ExTaq DNA polymerase (Takara), and up to 10 µg genomic DNA in a total reaction volume of 100 µl. A total of 40~87.5 µg gDNA was used as template from each conditions. Thermal cycler PCR conditions consisted of heating samples to 95 °C for 5 min; 28 cycles of 95 °C for 30 s, 53 °C for 30 s, and 72 °C for 20 s; and 72 °C for 10 min. Equal amounts of samples were then mixed and purified using AMPure XP for PCR purification (Beckman Coulter). Samples were sequenced using a custom sequencing primer using standard Illumina conditions by the Broad Institute Genetic Perturbation Platform. Sequencing reads were deconvolved and hairpin counts were quantified for subsequent analysis by counting against the barcode reference using PoolQ (https://portals.broadinstitute.org/gpp/public/dir/download?dirpath=software&filename=poolq-2.2.0-manual.pdf).

P5 primer

**AATGATACGGCGACCACCGAGATCT**ACACTCTTTCCCTACACGACGCTCTTCCGATCT[s]**TCTTGTGGAAAGG*A*C*G*A**

A mix of P5 primers with stagger regions [s] of different length was used to maintain sequence diversity across the flow-cell.

P7 primer

**CAAGCAGAAGACGGCATACGAGAT**NNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**TCTACTATTCTTTCCCCTGCA*C*T*G*T**

Independently barcoded P7 primers was used for each condition.

NNNNNNNN – barcode region

Analysis of the shRNA Screen

Three separate donor primary CD34$^+$ cells populations were run as replicates in the shRNA screen. A pseudocount of 1 was added to all shRNA-encoding DNA count totals and these counts were subsequently normalized to counts per million (CPM) and log$_2$ transformed. A linear mixed model was constructed to fit fixed effects for each gene (*g*) using the log$_2$ fold change from initial hairpin counts as the response variable (*y*). A random effect was included to capture variations in efficacy and off-target effects for each shRNA (*h*) used to target a given gene. The resulting model, $y \sim g + (0 + h|g)$, was fit in R-3.4 using the lme4 package. Genes hits were called from the set of genes with β coefficient effect size > 0.1 and the Wald chi-square test adjusted q value < 0.1. Enrichment of erythroid essential genes within the hit set was calculated by running 1 million permutations against the distribution of K562 essentiality for all genes included in the library, panels of genes nominated by sets of significant GWAS-associated lipid trait SNPs (Willer et al., 2013), and against all genes in the genome (Ensembl GRCh37p9).

Enrichment for identification of the included 5 "gold standard" genes and for red blood cell trait-associated coding variants were each accomplished using identical permutation schemes. Expression of the hit genes in various cell states/ stages of differentiation were derived from the cited datasets and permuted across all unique stages to determine stage-specific enrichment. The interaction network surrounding the 77 hits identified in the screen was generated in the latest version of STRING (10.5) and filtered for the purposes of display to only those nodes with at least one edge to another node among the hits.

### RNA-Seq

Stage matched CD71$^+$/ CD235a$^+$ cells derived from CD34$^+$ HSPCs infected with SF3A2 sh3, sh4 and shLUC were FACS sorted at day 8 of erythroid differentiation. RNA was isolated using a RNAqueous Micro kit (Invitrogen) according to the manufacturer’s instructions. DNase digestion was performed before RNA was quantified using a Qubit RNA HS Assay kit (Invitrogen). 1-10 ng of RNA were used as input to a modified SMART-seq2 (Picelli et al., 2014) protocol and after reverse transcription, 8-9 cycles of PCR were used to amplify transcriptome library. Quality of whole transcriptome libraries was validated using 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). All libraries were sequenced using Nextseq High Output Cartridge kits and a Nextseq 500 sequencer (Illumina). Libraries were sequenced using 2x38bp paired end reads.

### RNA-seq Differential Expression Analysis

For differential expression analysis, paired end sequencing reads from our SF3A2 shRNA knockdown experiments and obtained from the SF3B1 mutant datasets (Obeng et al., 2016) were quantified using Salmon version 0.11.1 (Patro et al., 2017) with default parameters and an index constructed from Gencode annotations version 28. Differential expression of quantified counts were calculated using DESeq2 (Love et al., 2014) in R-3.4. Enrichment for functions and components of the cell among the differentially expressed gene sets were quantified using GOrilla (Eden et al., 2007; Eden et al., 2009).

RNA-seq Differential Splicing Analysis

Paired end sequencing reads from our SF3A2 shRNA knockdown experiments and obtained from the cited SF3B1 mutant datasets were aligned using STAR version 2.5.2 in two-pass mode. Differential splicing was quantified using MISO version 0.5.4 in Python 2.7 using the instructions and annotation files provided with the package (Katz et al., 2010).

Analysis of Hemoglobin Levels for MDS Patients with or without SF3A2 Mutations

Genotyped MDS patient hemoglobin level measurements were obtained from the laboratory of J. Maciejewski. 1000GENOMES phase 3 data was used to find a SNP encoded in whole-exome sequencing data (rs25672) in high LD ($r^2$ = 0.737675) with the SF3A2-associated sentinel SNP (rs2159213). An ordinary least squares linear regression was used to fit the patient hemoglobin levels to the number of SF3A2 minor alleles present in each patient (log likelihood ratio test p = 0.140).

Phosphorylated STAT5 assessment with Intracellular Flow Cytometry

UT-7/EPO cells were cultured in DMEM medium supplemented with 10% Fetal Bovine Serum and 2 U/mL EPO. 5 days post-infection with TFR2 shRNAs, UT-7/EPO cells were cytokine starved overnight. On the next day, cells were treated with EPO in a dose dependent manner ((0 U/mL, 0.002 U/mL, 0.02 U/mL, 0.2 U/mL, 2 U/mL, 20 U/mL and 200 U/mL) and incubated 37 °C for 30 min. Alternatively the cells were treated with 2U/ml EPO in a time dependent manner (15, 30, 60, 120,180 min). Treated cells were gently mixed with pre-warmed Fixation Buffer (BD Bioscience) at 37°C for 10 min to fix cells. To permeabilize cells for intracellular staining, cells were resuspended in pre-chilled Perm Buffer III (BD Bioscience) for 30 min at 4°C. After three washes with 3% FBS in PBS, samples were stained either with Alexa Fluor-647 Mouse Anti-phospho-STAT5 (pY694; 1:20 dilution) for 1 hr in the dark at room temperature. A BD Accuri C6 Cytometer (BD Bioscience) was used to acquire mean fluorescent intensity (MFI) of phospho-STAT5-Alexa Fluor 647. The MFI of phospho-STAT5-Alexa Fluor 647 of gated single cells was calculated using FlowJo (version 10.0.8r1). Unstimulated UT7/EPO cells were used as a negative control.

### May-Grünwald-Giemsa staining

Approximately 50,000 – 200,000 cells were harvested, washed once at 300 x g for 5 min, resuspended in 200 µL FACS buffer and spun onto poly-L-lysine coated glass slides (Sigma Aldrich) with a Shandon 4 (Thermo Fisher) cytocentrifuge at 300 rpm for 4 min. Visibly dry slides were stained with May-Grünwald solution for 5 min, rinsed 4 times for each 30 s in $H_2O$, transferred to Giemsa solution for 15 min and washed as described above. Slides were dried overnight and mounted with coverslip. All images were taken with AxioVision software (Zeiss) at 100 x magnification.

### Mouse erythroid differentiation culture

Bone marrow cells were isolated from SF3B1$^{K700E\ +/-}$ mice and littermate controls were lineage depleted using Lineage Cell Depletion Kit, mouse (Miltenyi Biotech) according to manufacturer's protocols. Lineage negative cells were immediately transduced with lentiviral shRNAs targeting SF3A2 or controls (MOI -90) by spinfection at 2000rpm for 90 min. The cells were cultured in erythroid maintenance medium (StemSpan-SFEM; StemCell Technologies) supplemented with 100 ng/mL recombinant mouse stem cell factor (SCF) (R&D Systems), 40 ng/mL recombinant mouse IGF1 (R&D Systems), 100 nM dexamethasone (Sigma), and 2 U/mL erythropoietin (Amgen) and cultured at 37°C for 36 hours. Following this the cells were cultured for another 48 hours in erythroid differentiation medium (Iscove modified Dulbecco's medium containing 15% (vol/vol) FBS (Stemcell), 1% detoxified BSA (Stemcell), 500 µg/mL holo-transferrin (Sigma-Aldrich), 0.5 U/mL Epoetin (Epo; Amgen), 10 µg/mL recombinant human insulin (Sigma-Aldrich), and 2 mM L-glutamine (Invitrogen)) at 37 °C.

Flow cytometry analyses and antibodies

All flow cytometry data was acquired using either using LSR II SORP or LSR Fortessa flow cytometers (BD Biosciences). All staining was carried out in FACS buffer (2% FBS in PBS) for 30 minutes on ice unless otherwise described. The following antibodies were used anti-human CD235a-APC (eBioscience, Clone HIR2), anti-human CD71-FITC (eBioscience, Clone OKT9), anti-human CD71-PEcy7 (eBioscience, Clone OKT9), ant-human CD49d-PE (Miltenyi, Clone MZ18-24A9), anti-human CD41a-PE (eBioscience, Clone HIP8), anti-human CD11b-PE (eBioscience, Clone ICRF44), anti-mouse Ter119-APC (eBioscience, Clone TER119), anti-mouse CD71-PE (eBioscience, Clone R17217) and Alexa Fluor-647 anti-phospho STAT5 (pY694) (BD Bioscience Cat#: 612599). Hoechst 33342 (Life Technologies, H1399) was used to visualize nuclei.

shRNA sequences

The following lentiviral shRNA constructs were generated in Polymerase III based shRNA backbone pLKO.1-puro (Sigma Aldrich).

shLUC

5’-CCGGCGCTGAGTACTTCGAAATGTCCTCGAGGACATTTCGAAGTACTCAGCGTTTTTG-3’

TFR2 sh1

5'-CCGGGCCAGATCACTACGTTGTCATCTCGAGATGACAACGTAGTGATCTGGCTTTTTG-3

TFR2 sh2

5'-CCGGCAACAACATCTTCGGCTGCATCTCGAGATGCAGCCGAAGATGTTGTTGTTTTTG-3'

SF3A2 sh1 (human)

5'-CCGGCTACGAGACCATTGCCTTCAACTCGAGTTGAAGGCAATGGTCTCGTAGTTTTT-3

SF3A2 sh2 (human)

5'-CCGGCCTGGGCTCCTATGAATGCAACTCGAGTTGCATTCATAGGAGCCCAGGTTTTT-3'

SF3A2 sh3 (human)

5'-CCGGCAAAGTGACCAAGCAGAGAGACTCGAGTCTCTCTGCTTGGTCACTTTGTTTTT-3

SF3A2 sh4 (human)

5'-CCGGACATCAACAAGGACCCGTACTCTCGAGAGTACGGGTCCTTGTTGATGTTTTTT-3'

The following lentiviral shRNA constructs were generated in Polymerase II based mir30 shRNA backbone developed in the lab SFFV-Venus-mir30 shRNA backbone.

shNT(non-targeting)

5’_TGCTGTTGACAGTGAGCGATCTCGCTTGGGCGAGAGTAAGTAGTGAAGCCACAGATGTA
CTTACTCTCGCCCAAGCGAGAGTGCCTACTGCCTCGGA_3’

*Sf3a2* sh1 (mouse)

5’_TGCTGTTGACAGTGAGCGCGGAGGTGAAGAAGTTTGTGAATAGTGAAGCCACAGATGTA
TTCACAAACTTCTTCACCTCCATGCCTACTGCCTCGGA_3’

*Sf3a2* sh2 (mouse)

5’_TGCTGTTGACAGTGAGCGACCACCGTTTCATGTCTGCTTATAGTGAAGCCACAGATGTATAAGCAGACATGAAACGGTGGCTGCCTACTGCCTCGGA_3’

*Sf3a2* sh3 (mouse)

5’_TGCTGTTGACAGTGAGCGATCCTGCCTTGAGCCTATTAAATAGTGAAGCCACAGATGTATTTAATAGGCTCAAGGCAGGACTGCCTACTGCCTCGGA_3’

*Sf3a2* sh4 (mouse)

5’_TGCTGTTGACAGTGAGCGACCACTGGAACAGAGAAACCAATAGTGAAGCCACAGATGTATTGGTTTCTCTGTTCCAGTGGGTGCCTACTGCCTCGGA_3’

*Sf3a2* sh5 (mouse)

5’_TGCTGTTGACAGTGAGCGATGGAGGTGAAGAAGTTTGTGATAGTGAAGCCACAGATGTATCACAAACTTCTTCACCTCCACTGCCTACTGCCTCGGA_3’

qPCR primers

*TFR2* Fwd: 5'-ATCCTTCCCTCTTCCCTCCC-3'

*TFR2* Rev: 5'-CCATCCAGCCACATGGTTCT-3

*SF3A2* Fwd: 5'-CCTGAGAAGGTCAAGGTGGA-3'

*SF3A2* Rev: 5'-CTCCGAGTCTCTCTGCTTGG-3'

Western Blot antibodies

Anti-GAPDH (Santa Cruz Biotechnology, sc-32233); anti-TFR2 (Santa Cruz Biotechnology, sc-sc-32271); anti-SF3A2 (Santa Cruz Biotechnology, sc-390444)

# Supplemental Information

## Supplementary Figure Legends

**Supplementary Figure S1: Summary Characteristics of Designed shRNA Library** (A) Counts of loci from among the original 75 annotated with linkage to each of the six RBC traits, hemoglobin (Hb), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), mean corpuscular volume (MCV), packed cell volume (PCV), and red blood cell count (RBC). Some loci were associated with multiple traits. (B) Kernel density plot showing the $\log_{10}$ sizes in bp of the LD-defined genomic windows used to find overlapping genes. (C) Histogram showing distribution of number of genes selected using the LD window method at each locus. A median of 4 genes were present at each. (D) Histogram showing distribution of number of independent hairpins included in the library to target each of the candidate's genes. (E)  Representative FACS plots of erythroid cell surface markers CD71 (transferrin receptor) and CD235a (Glycophorin A) expression at various time points during erythroid differentiation in uninfected (Mock) or CD34$^+$ cells infected with the shRNA library (Pool). Percentages in each quadrant is represented as mean and standard deviation of 3 experiments from independent donors.

**Supplementary Figure S2: Additional Metrics of Library Performance** (A) shRNA abundance $\log_2$ fold changes from day 4 to each of the other time points. Represented values are the mean of hairpin abundance $\log_2$ fold changes across hairpins for each gene and two standard deviations. (B) Scatter plots showing agreement of replicate observations across independent CD34$^+$ donor populations.

**Supplementary Figure S3: Additional Characterization of Modeling Outcomes** (A) Histogram showing the number of gene hits identified at each of the 40 loci with at least one significant gene effect detected. (B) Bar graph showing the number of gene hits identified for each of the 6 red blood cell traits used in the original GWAS to identify the studied loci. (C) Permuted enrichment of essentiality among the set of hit genes vs. randomly chosen sets of genes from the human genome. (D) Permuted enrichment of essentiality among the set of hit genes vs. genes implicated by a separate GWAS for LDL cholesterol levels. (E) Permuted enrichment of essentiality among the set of hit genes vs. genes implicated by a separate GWAS for HDL cholesterol levels. (F) Permuted enrichment of essentiality among the set of hit genes vs. genes implicated by a separate GWAS for blood triglyceride levels. (G) Heat map depicting strength of expression (as z scores within each gene) for each of the 77 identified hit genes throughout the specific stages of fetal erythropoiesis. Purple boxes highlight the cell types that were enriched for expression of hit genes.

**Supplementary Figure S5: Transferrin Receptor 2 is a Negative Regulator of Human Erythropoiesis** (A) Representative FACS plots of alternate erythroid cell surface markers CD49d (α4 integrin) and CD235a (Glycophorin A) expression at various time points during erythroid differentiation. (B) May-Grunwald Giemsa staining showing more differentiated erythroid cells after TFR2 knockdown at day 18 of erythroid culture. (C) Western blot showing downregulation of TFR2 in UT7/EPO cells. (D) Time dependent absolute value of MFI of STAT5 in UT7/ Epo cells after TFR2 knockdown.

**Supplementary Figure S6: SF3A2 is Required for Human Erythropoiesis and Modulates Erythropoiesis Defects in a Murine Model of MDS** (A) shRNAs targeting *SF3A2* co-expressing a reporter GFP gene was infected into CD34$^+$ cells and cultured in erythroid conditions. GFP expression at various time points from three independent experiments show that downregulation of SF3A2 results in reduced cell numbers. (B) Representative FACS plots of erythroid (CD235a) and non-erythroid cell surface markers (CD11b / CD41a) and at various time points showing an increase in non-erythroid lineages upon SF3A2 downregulation. Cells were gated on the GFP positive population. (C) Knockdown efficiency of shRNAs targeting *SF3A2* in murine erythroleukemia (MEL) cells by western blot. (D) Total cell numbers of GFP+ cells at the start of murine erythroid differentiation. (E) Percentage of Ter119$^+$ CD71$^+$ erythroid cells within GFP compartment and (F) Total cell numbers of GFP$^+$ erythroid cells after 24hrs in erythroid differentiation. (G) Growth curves of GFP$^+$ erythroid cells during erythroid culture. (H) Putative but insignificant interaction between *SF3A2* variant alleles (rs25672) and hemoglobin levels in MDS patients with *SF3B1* mutations.

## Supplemental Tables

**Table S1:** Table containing annotations and information for the 75 SNPs used to seed the shRNA library.

**Table S2:** Table containing annotations and information for all hairpins, as well as shRNA counts for each time point and replicate.

**Table S3:** Table containing the R model output for each gene.

**Table S4:** Table containing the DESeq2 output for differentially expressed genes in cells undergoing SF3A2 knockdown or control shRNA treatment.

**Table S5:** Table containing the DESeq2 output for differentially expressed genes in MDS patients with and without mutations in *SF3B1*.

**Tables S6, S7:** Tables containing the GO component (S6) and function (S7) enrichments calculated using GOrilla for cells undergoing SF3A2 knockdown or control shRNA treatment.

**Tables S8, S9:** Tables containing the GO component (S8) and function (S9) enrichments calculated using GOrilla for MDS patient samples with and without mutations in *SF3B1*.

**Tables S10-S14:** Tables containing the differential splicing analysis for cells undergoing SF3A2 knockdown or control shRNA treatment. Categories of splice mutations presented in each table are alternative 3' splice sites, alternative 5' splice sites, mutually exclusive exons, retrained introns, and skipped exons, respectively.

**Tables S15-S19:** Tables containing the differential splicing analysis for MDS patient patient samples with and without mutations in *SF3B1*. Categories of splice mutations presented in each table are alternative 3' splice sites, alternative 5' splice sites, mutually exclusive exons, retrained introns, and skipped exons, respectively.

# References

Artuso, I., Lidonnici, M.R., Altamura, S., Mandelli, G., Pettinato, M., Muckenthaler, M.U., Silvestri, L., Ferrari, G., Camaschella, C., and Nai, A. (2018). Transferrin Receptor 2 is a potential novel therapeutic target for beta-thalassemia: evidence from a murine model. Blood.

Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A.*, et al.* (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell *167*, 1415-1429.e1419.

Basak, A., Hancarova, M., Ulirsch, J.C., Balci, T.B., Trkova, M., Pelisek, M., Vlckova, M., Muzikova, K., Cermak, J., Trka, J.*, et al.* (2015). BCL11A deletions result in fetal hemoglobin persistence and neurodevelopmental alterations. The Journal of clinical investigation *125*, 2363-2368.

Bennett, V., and Stenbuck, P.J. (1979). Identification and partial purification of ankyrin, the high affinity membrane attachment site for human erythrocyte spectrin. J Biol Chem *254*, 2533-2541.

Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell *169*, 1177-1186.

Claussnitzer, M., Dankel, S.N., Kim, K.H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V.*, et al.* (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. The New England journal of medicine *373*, 895-907.

Corces, M.R., Buenrostro, J.D., Wu, B., Greenside, P.G., Chan, S.M., Koenig, J.L., Snyder, M.P., Pritchard, J.K., Kundaje, A., Greenleaf, W.J.*, et al.* (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet *48*, 1193-1203.

Eden, E., Lipson, D., Yogev, S., and Yakhini, Z. (2007). Discovering motifs in ranked lists of DNA sequences. PLoS Comput Biol *3*, e39.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics *10*, 48.

Forejtnikova, H., Vieillevoye, M., Zermati, Y., Lambert, M., Pellegrino, R.M., Guihard, S., Gaudry, M., Camaschella, C., Lacombe, C., Roetto, A.*, et al.* (2010). Transferrin receptor 2 is a component of the erythropoietin receptor complex and is required for efficient erythropoiesis. Blood *116*, 5357-5367.

Fulco, C.P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S.R., Perez, E.M., Kane, M., Cleary, B., Lander, E.S., and Engreitz, J.M. (2016). Systematic mapping of functional enhancer-promoter connections with CRISPR interference. Science *354*, 769-773.

Galarneau, G., Palmer, C.D., Sankaran, V.G., Orkin, S.H., Hirschhorn, J.N., and Lettre, G. (2010). Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. Nat Genet *42*, 1049-1051.

Gazda, H.T., Preti, M., Sheen, M.R., O'Donohue, M.F., Vlachos, A., Davies, S.M., Kattamis, A., Doherty, L., Landowski, M., Buros, C., *et al.* (2012). Frameshift mutation in p53 regulator RPL26 is associated with multiple physical abnormalities and a specific pre-rRNA processing defect in Diamond-Blackfan anemia. Hum Mutat *33*, 1037-1044.

Giani, F.C., Fiorini, C., Wakabayashi, A., Ludwig, L.S., Salem, R.M., Jobaliya, C.D., Regan, S.N., Ulirsch, J.C., Liang, G., Steinberg-Shemer, O., *et al.* (2016). Targeted Application of Human Genetic Variation Can Improve Red Blood Cell Production from Stem Cells. Cell Stem Cell *18*, 73-78.

Glick, B.S., and Rothman, J.E. (1987). Possible role for fatty acyl-coenzyme A in intracellular protein transport. Nature *326*, 309-312.

Goardon, N., Lambert, J.A., Rodriguez, P., Nissaire, P., Herblot, S., Thibault, P., Dumenil, D., Strouboulis, J., Romeo, P.H., and Hoang, T. (2006). ETO2 coordinates cellular proliferation and differentiation during erythropoiesis. Embo j *25*, 357-366.

Gozani, O., Feld, R., and Reed, R. (1996). Evidence that sequence-independent binding of highly conserved U2 snRNP proteins upstream of the branch site is required for assembly of spliceosomal complex A. Genes Dev *10*, 233-243.

Gozani, O., Potashkin, J., and Reed, R. (1998). A potential role for U2AF-SAP 155 interactions in recruiting U2 snRNP to the branch site. Molecular and cellular biology *18*, 4752-4760.

Guo, M.H., Nandakumar, S.K., Ulirsch, J.C., Zekavat, S.M., Buenrostro, J.D., Natarajan, P., Salem, R.M., Chiarle, R., Mitt, M., and Kals, M. (2016). Comprehensive population-based genome sequencing provides insight into hematopoietic regulatory mechanisms. Proceedings of the National Academy of Sciences, 201619052.

Gupta, R.M., Hadaya, J., Trehan, A., Zekavat, S.M., Roselli, C., Klarin, D., Emdin, C.A., Hilvering, C.R.E., Bianchi, V., Mueller, C., *et al.* (2017). A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. Cell *170*, 522-533.e515.

Hu, J., Liu, J., Xue, F., Halverson, G., Reid, M., Guo, A., Chen, L., Raza, A., Galili, N., Jaffray, 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*, 3246-3253.

Huang, H., Fang, M., Jostins, L., Umicevic Mirkov, M., Boucher, G., Anderson, C.A., Andersen, V., Cleynen, I., Cortes, A., Crins, F., *et al.* (2017). Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature *547*, 173-178.

Jing, H., Vakoc, C.R., Ying, L., Mandat, S., Wang, H., Zheng, X., and Blobel, G.A. (2008). Exchange of GATA factors mediates transitions in looped chromatin organization at a developmentally regulated gene locus. Mol Cell *29*, 232-242.

Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods *7*, 1009-1015.

Khajuria, R.K., Munschauer, M., Ulirsch, J.C., Fiorini, C., Ludwig, L.S., McFarland, S.K., Abdulhay, N.J., Specht, H., Keshishian, H., Mani, D.R., *et al.* (2018). Ribosome Levels Selectively Regulate Translation and Lineage Commitment in Human Hematopoiesis. Cell *173*, 90-103.e119.

Lareau, C.A., Ulirsch, J.C., Bao, E.L., Ludwig, L.S., Guo, M.H., Benner, C., Satpathy, A.T., Salem, R., Hirschhorn, J.N., Finucane, H.K., *et al.* (2018). Interrogation of human hematopoiesis at single-cell and single-variant resolution.

Li, W., Koster, J., Xu, H., Chen, C.H., Xiao, T., Liu, J.S., Brown, M., and Liu, X.S. (2015). Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. Genome biology *16*, 281.

Liu, N., Hargreaves, V.V., Zhu, Q., Kurland, J.V., Hong, J., Kim, W., Sher, F., Macias-Trevino, C., Rogers, J.M., Kurita, R., *et al.* (2018). Direct Promoter Repression by BCL11A Controls the Fetal to Adult Hemoglobin Switch. Cell *173*, 430-442.e417.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550.

Ludwig, L.S., Gazda, H.T., Eng, J.C., Eichhorn, S.W., Thiru, P., Ghazvinian, R., George, T.I., Gotlib, J.R., Beggs, A.H., Sieff, C.A., *et al.* (2014). Altered translation of GATA1 in Diamond-Blackfan anemia. Nature medicine *20*, 748-753.

Mandegar, M.A., Huebsch, N., Frolov, E.B., Shin, E., Truong, A., Olvera, M.P., Chan, A.H., Miyaoka, Y., Holmes, K., Spencer, C.I., *et al.* (2016). CRISPR Interference Efficiently Induces Specific and Reversible Gene Silencing in Human iPSCs. Cell Stem Cell *18*, 541-553.

McIver, S.C., Kang, Y.A., DeVilbiss, A.W., O'Driscoll, C.A., Ouellette, J.N., Pope, N.J., Camprecios, G., Chang, C.J., Yang, D., Bouhassira, E.E., *et al.* (2014). The exosome complex establishes a barricade to erythroid maturation. Blood *124*, 2285-2297.

Moffat, J., Grueneberg, D.A., Yang, X., Kim, S.Y., Kloepfer, A.M., Hinkle, G., Piqani, B., Eisenhaure, T.M., Luo, B., Grenier, J.K., *et al.* (2006). A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. Cell *124*, 1283-1298.

Mohanan, V., Nakata, T., Desch, A.N., Levesque, C., Boroughs, A., Guzman, G., Cao, Z., Creasey, E., Yao, J., Boucher, G., *et al.* (2018). C1orf106 is a colitis risk gene that regulates stability of epithelial adherens junctions. Science *359*, 1161-1166.

Moniz, H., Gastou, M., Leblanc, T., Hurtaud, C., Cretien, A., Lecluse, Y., Raslova, H., Larghero, J., Croisille, L., Faubladier, M., *et al.* (2012). Primary hematopoietic cells from DBA patients with mutations in RPL11 and RPS19 genes exhibit distinct erythroid phenotype in vitro. Cell Death Dis *3*, e356.

Montgomery, S.B., and Dermitzakis, E.T. (2011). From expression QTLs to personalized transcriptomics. Nature reviews Genetics *12*, 277-282.

Mucenski, M.L., McLain, K., Kier, A.B., Swerdlow, S.H., Schreiner, C.M., Miller, T.A., Pietryga, D.W., Scott, W.J., Jr., and Potter, S.S. (1991). A functional c-myb gene is required for normal murine fetal hepatic hematopoiesis. Cell *65*, 677-689.

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K.V., Li, X., Li, H., Kuperwasser, N., Ruda, V.M.*, et al.* (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature *466*, 714-719.

Nai, A., Lidonnici, M.R., Rausa, M., Mandelli, G., Pagani, A., Silvestri, L., Ferrari, G., and Camaschella, C. (2015). The second transferrin receptor regulates red blood cell production in mice. Blood *125*, 1170-1179.

Niemi, M.E.K., Martin, H.C., Rice, D.L., Gallone, G., Gordon, S., Kelemen, M., McAloney, K., McRae, J., Radford, E.J., Yu, S.*, et al.* (2018). Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. Nature *562*, 268-271.

Obeng, E.A., Chappell, R.J., Seiler, M., Chen, M.C., Campagna, D.R., Schmidt, P.J., Schneider, R.K., Lord, A.M., Wang, L., Gambe, R.G.*, et al.* (2016). Physiologic Expression of Sf3b1(K700E) Causes Impaired Erythropoiesis, Aberrant Splicing, and Sensitivity to Therapeutic Spliceosome Modulation. Cancer cell *30*, 404-417.

Papaemmanuil, E., Cazzola, M., Boultwood, J., Malcovati, L., Vyas, P., Bowen, D., Pellagatti, A., Wainscoat, J.S., Hellstrom-Lindberg, E., Gambacorti-Passerini, C.*, et al.* (2011). Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. The New England journal of medicine *365*, 1384-1395.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods *14*, 417-419.

Peters, L.L., Shivdasani, R.A., Liu, S.C., Hanspal, M., John, K.M., Gonzalez, J.M., Brugnara, C., Gwynn, B., Mohandas, N., Alper, S.L.*, et al.* (1996). Anion exchanger 1 (band 3) is required to prevent erythrocyte membrane surface loss but not to form the membrane skeleton. Cell *86*, 917-927.

Pimentel, H., Parra, M., Gee, S.L., Mohandas, N., Pachter, L., and Conboy, J.G. (2016). A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. Nucleic Acids Res *44*, 838-851.

Riba, A., Emmenlauer, M., Chen, A., Sigoillot, F., Cong, F., Dehio, C., Jenkins, J., and Zavolan, M. (2017). Explicit Modeling of siRNA-Dependent On- and Off-Target Repression Improves the Interpretation of Screening Results. Cell systems *4*, 182-193.e184.

Rossi, A., Kontarakis, Z., Gerri, C., Nolte, H., Holper, S., Kruger, M., and Stainier, D.Y. (2015). Genetic compensation induced by deleterious mutations but not gene knockdowns. Nature *524*, 230-233.

Rossin, E.J., Lage, K., Raychaudhuri, S., Xavier, R.J., Tatar, D., Benita, Y., Cotsapas, C., and Daly, M.J. (2011). Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS genetics *7*, e1001273.

Rund, D., and Rachmilewitz, E. (2005). Beta-thalassemia. N Engl J Med *353*, 1135-1146.

Sankaran, V.G., Joshi, M., Agrawal, A., Schmitz-Abe, K., Towne, M.C., Marinakis, N., Markianos, K., Berry, G.T., and Agrawal, P.B. (2013). Rare complete loss of function provides insight into a pleiotropic genome-wide association study locus. Blood *122*, 3845-3847.

Sankaran, V.G., Ludwig, L.S., Sicinska, E., Xu, J., Bauer, D.E., Eng, J.C., Patterson, H.C., Metcalf, R.A., Natkunam, Y., Orkin, S.H.*, et al.* (2012). Cyclin D3 coordinates the cell cycle during differentiation to regulate erythrocyte size and number. Genes Dev *26*, 2075-2087.

Sankaran, V.G., Menne, T.F., Scepanovic, D., Vergilio, J.A., Ji, P., Kim, J., Thiru, P., Orkin, S.H., Lander, E.S., and Lodish, H.F. (2011). MicroRNA-15a and -16-1 act via MYB to elevate fetal hemoglobin expression in human trisomy 13. Proceedings of the National Academy of Sciences of the United States of America *108*, 1519-1524.

Sankaran, V.G., Menne, T.F., Xu, J., Akie, T.E., Lettre, G., Van Handel, B., Mikkola, H.K., Hirschhorn, J.N., Cantor, A.B., and Orkin, S.H. (2008). Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. Science *322*, 1839-1842.

Simeonov, D.R., Gowen, B.G., Boontanrart, M., Roth, T.L., Gagnon, J.D., Mumbach, M.R., Satpathy, A.T., Lee, Y., Bray, N.L., Chan, A.Y.*, et al.* (2017). Discovery of stimulation-responsive immune enhancers with CRISPR activation. Nature *549*, 111-115.

Smemo, S., Tena, J.J., Kim, K.H., Gamazon, E.R., Sakabe, N.J., Gomez-Marin, C., Aneas, I., Credidio, F.L., Sobreira, D.R., Wasserman, N.F.*, et al.* (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature *507*, 371-375.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P.*, et al.* (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res *45*, D362-d368.

Tewhey, R., Kotliar, D., Park, D.S., Liu, B., Winnicki, S., Reilly, S.K., Andersen, K.G., Mikkelsen, T.S., Lander, E.S., and Schaffner, S.F. (2016). Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. Cell *165*, 1519-1529.

Thomsen, S.K., Ceroni, A., van de Bunt, M., Burrows, C., Barrett, A., Scharfmann, R., Ebner, D., McCarthy, M.I., and Gloyn, A.L. (2016). Systematic Functional Characterization of Candidate Causal Genes for Type 2 Diabetes Risk Variants. Diabetes *65*, 3805-3811.

Ting, P.Y., Parker, A.E., Lee, J.S., Trussell, C., Sharif, O., Luna, F., Federe, G., Barnes, S.W., Walker, J.R., Vance, J.*, et al.* (2018). Guide Swap enables genome-scale pooled CRISPR-Cas9 screening in human primary cells. Nat Methods *15*, 941-946.

Tsherniak, A., Vazquez, F., Montgomery, P.G., Weir, B.A., Kryukov, G., Cowley, G.S., Gill, S., Harrington, W.F., Pantel, S., Krill-Burger, J.M.*, et al.* (2017). Defining a Cancer Dependency Map. Cell *170*, 564-576.e516.

Ulirsch, J.C., Nandakumar, S.K., Wang, L., Giani, F.C., Zhang, X., Rogov, P., Melnikov, A., McDonel, P., Do, R., Mikkelsen, T.S.*, et al.* (2016). Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. Cell *165*, 1530-1545.

Ulirsch, J.C., Verboon, J.M., Kazerounian, S., Guo, M.H., Yuan, D., Ludwig, L.S., Handsaker, R.E., Abdulhay, N.J., Fiorini, C., Genovese, G.*, et al.* (2018). The Genetic Landscape of Diamond-Blackfan Anemia.

van der Harst, P., Zhang, W., Mateo Leach, I., Rendon, A., Verweij, N., Sehmi, J., Paul, D.S., Elling, U., Allayee, H., Li, X.*, et al.* (2012). Seventy-five genetic loci influencing the human red blood cell. Nature *492*, 369-375.

Veyrieras, J.B., Kudaravalli, S., Kim, S.Y., Dermitzakis, E.T., Gilad, Y., Stephens, M., and Pritchard, J.K. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS genetics *4*, e1000214.

Vockley, C.M., Guo, C., Majoros, W.H., Nodzenski, M., Scholtens, D.M., Hayes, M.G., Lowe, W.L., Jr., and Reddy, T.E. (2015). Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. Genome research *25*, 1206-1214.

Wang, T., Birsoy, K., Hughes, N.W., Krupczak, K.M., Post, Y., Wei, J.J., Lander, E.S., and Sabatini, D.M. (2015). Identification and characterization of essential genes in the human genome. Science *350*, 1096-1101.

Wang, X., Angelis, N., and Thein, S.L. (2018). MYB - A regulatory factor in hematopoiesis. Gene *665*, 6-17.

Whalen, S., and Pollard, K.S. (2018). Most regulatory interactions are not in linkage disequilibrium (Cold Spring Harbor Laboratory).

Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S.*, et al.* (2013). Discovery and refinement of loci associated with lipid levels. Nat Genet *45*, 1274-1283.

Yan, H., Hale, J., Jaffray, J., Li, J., Wang, Y., Huang, Y., An, X., Hillyer, C., Wang, N., Kinet, S.*, et al.* (2018). Developmental differences between neonatal and adult human erythropoiesis. Am J Hematol *93*, 494-503.

Figure 1

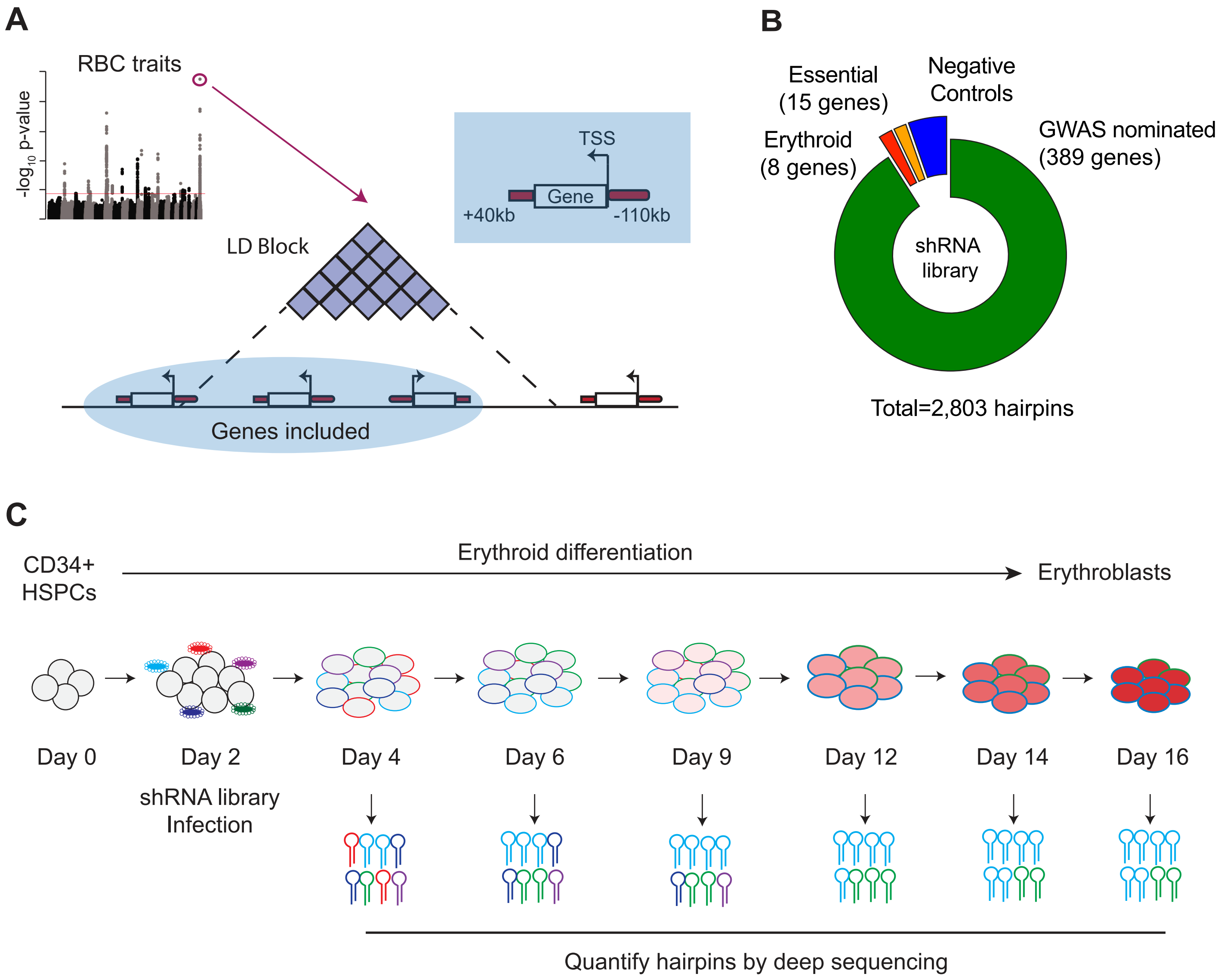

A
RBC traits
$-\log_{10}$ p-value
LD Block
TSS
Gene
+40kb
-110kb
Genes included
B
Essential
(15 genes)
Negative
Controls
Erythroid
(8 genes)
GWAS nominated
(389 genes)
shRNA
library
Total=2,803 hairpins
C
CD34+
HSPCs
Erythroid differentiation
Erythroblasts
Day 0
Day 2
Day 4
Day 6
Day 9
Day 12
Day 14
Day 16
shRNA library
Infection
Quantify hairpins by deep sequencing

Figure 2

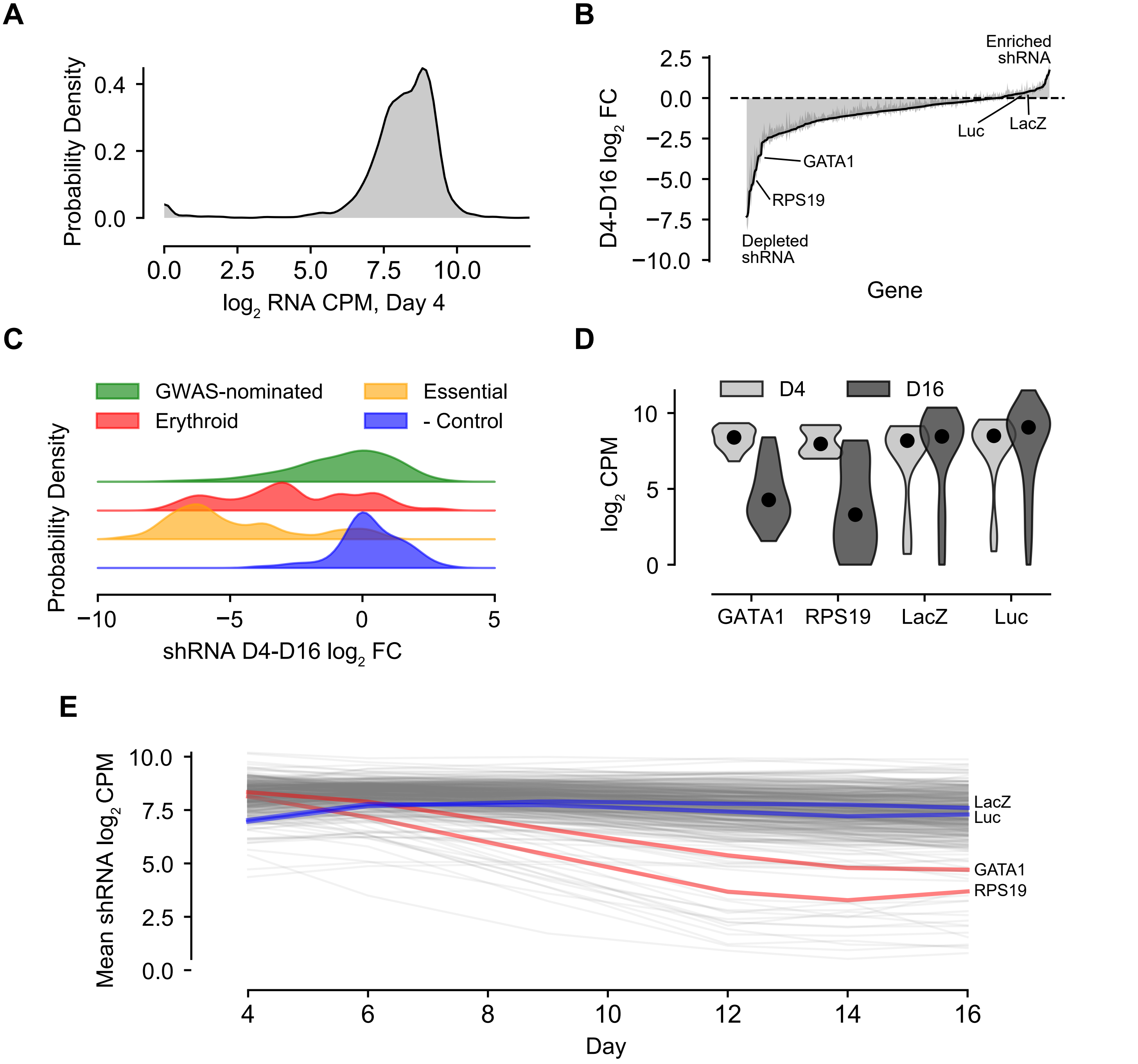

A
Probability Density
0.4
0.2
0.0
0.0
2.5
5.0
7.5
10.0
$\log_2$ RNA CPM, Day 4
B
D4-D16 $\log_2$ FC
2.5
0.0
−2.5
−5.0
−7.5
−10.0
Enriched shRNA
Luc
LacZ
GATA1
RPS19
Depleted shRNA
Gene
C
GWAS-nominated
Essential
Erythroid
- Control
Probability Density
−10
−5
0
5
shRNA D4-D16 $\log_2$ FC
D
D4
D16
10
5
0
$\log_2$ CPM
GATA1
RPS19
LacZ
Luc
E
Mean shRNA $\log_2$ CPM
10.0
7.5
5.0
2.5
0.0
LacZ
Luc
GATA1
RPS19
4
6
8
10
12
14
16
Day

Figure 3

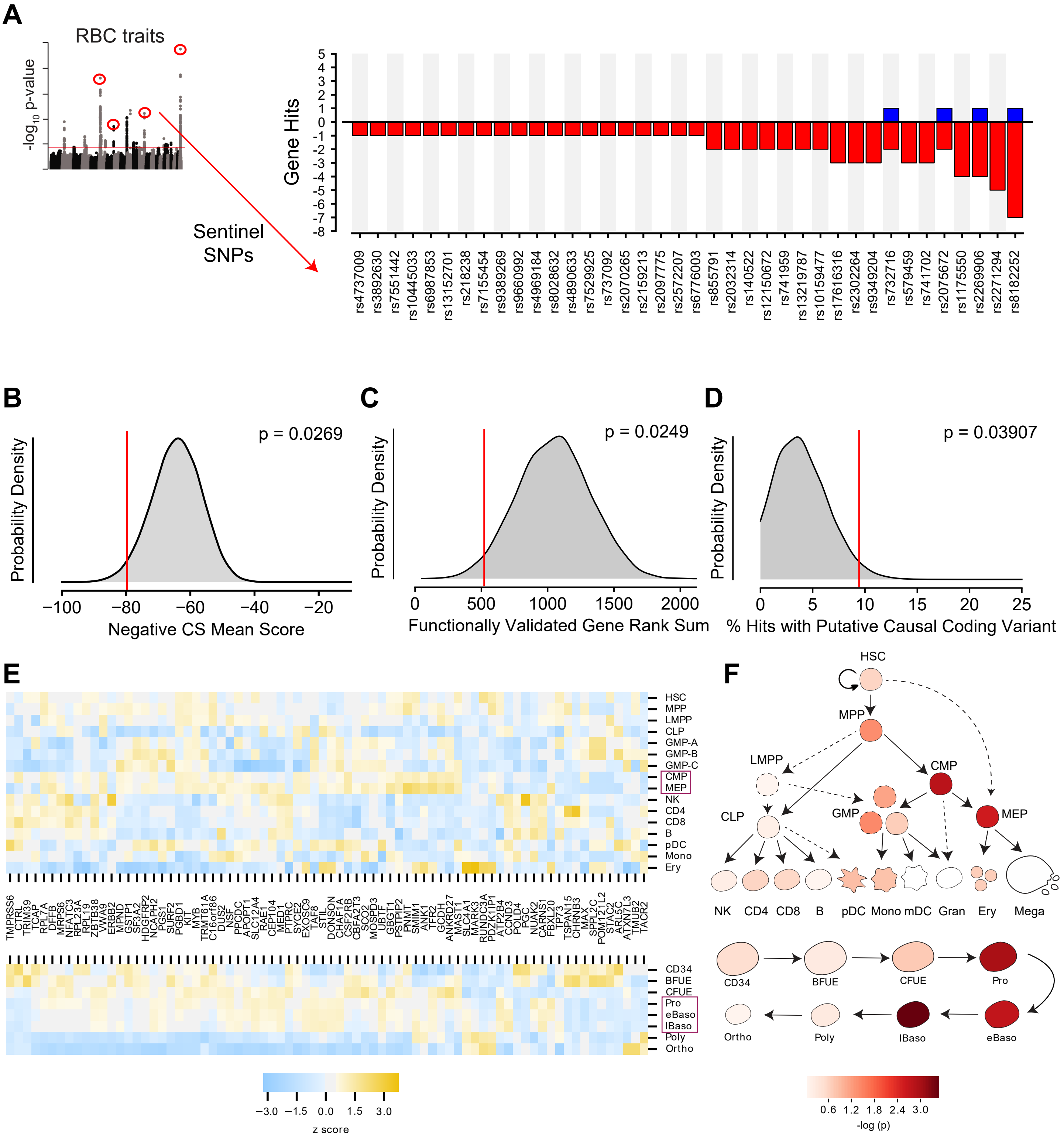

A
RBC traits
$-\log_{10}$ p-value
Sentinel SNPs
Gene Hits
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-7
-8
rs4737009
rs3892630
rs7551442
rs10445033
rs6987853
rs13152701
rs218238
rs7155454
rs9389269
rs9660992
rs4969184
rs8028632
rs4890633
rs7529925
rs737092
rs2070265
rs2159213
rs2097775
rs2572207
rs6776003
rs855791
rs2032314
rs140522
rs12150672
rs741959
rs13219787
rs10159477
rs17616316
rs2302264
rs9349204
rs732716
rs579459
rs741702
rs2075672
rs1175550
rs2269906
rs2271294
rs8182252
B
Probability Density
p = 0.0269
−100
−80
−60
−40
−20
Negative CS Mean Score
C
Probability Density
p = 0.0249
0
500
1000
1500
2000
Functionally Validated Gene Rank Sum
D
Probability Density
p = 0.03907
0
5
10
15
20
25
% Hits with Putative Causal Coding Variant
E
HSC
MPP
LMPP
CLP
GMP-A
GMP-B
GMP-C
CMP
MEP
NK
CD4
CD8
B
pDC
Mono
Ery
TMPRSS6
CTRL
TRIM39
TCAP
RPL7A
DFFB
MRPS6
NFATC3
RPL23A
RPL19
ZBTB38
VWA9
ERBB2
MPND
GSTP1
SF3A2
HDGFRP2
NCAPH2
PGS1
SURF2
PGBD1
KIT
MYB
TRMT61A
C16orf86
DUS2
NSF
PPCDC
APOPT1
SLC12A4
RAE1
CEP104
MED1
PTPRC
SYCE2
EXOSC9
TAF8
STIL
DONSON
CHAF1A
CSF2RB
CBFA2T3
SCO2
MOSPD3
UBTF
GBGT1
PSTPIP2
PNMT
SMIM1
ANK1
TFR2
GCDH
ANKRD27
MAST1
SLC4A1
MARK3
RUNDC3A
PDZK1IP1
ATP2B4
CCND3
POLD4
PGC
NUAK2
CARNS1
FBXL20
TP73
TSPAN15
CHRNB3
MAX
SPPL2C
POM121L2
STAC2
ARL5C
ATXN7L3
TMUB2
TACR2
CD34
BFUE
CFUE
Pro
eBaso
lBaso
Poly
Ortho
−3.0
−1.5
0.0
1.5
3.0
z score
F
HSC
MPP
LMPP
CMP
CLP
GMP
MEP
NK
CD4
CD8
B
pDC
Mono
mDC
Gran
Ery
Mega
CD34
BFUE
CFUE
Pro
Ortho
Poly
lBaso
eBaso
0.6
1.2
1.8
2.4
3.0
-log (p)

Figure 4

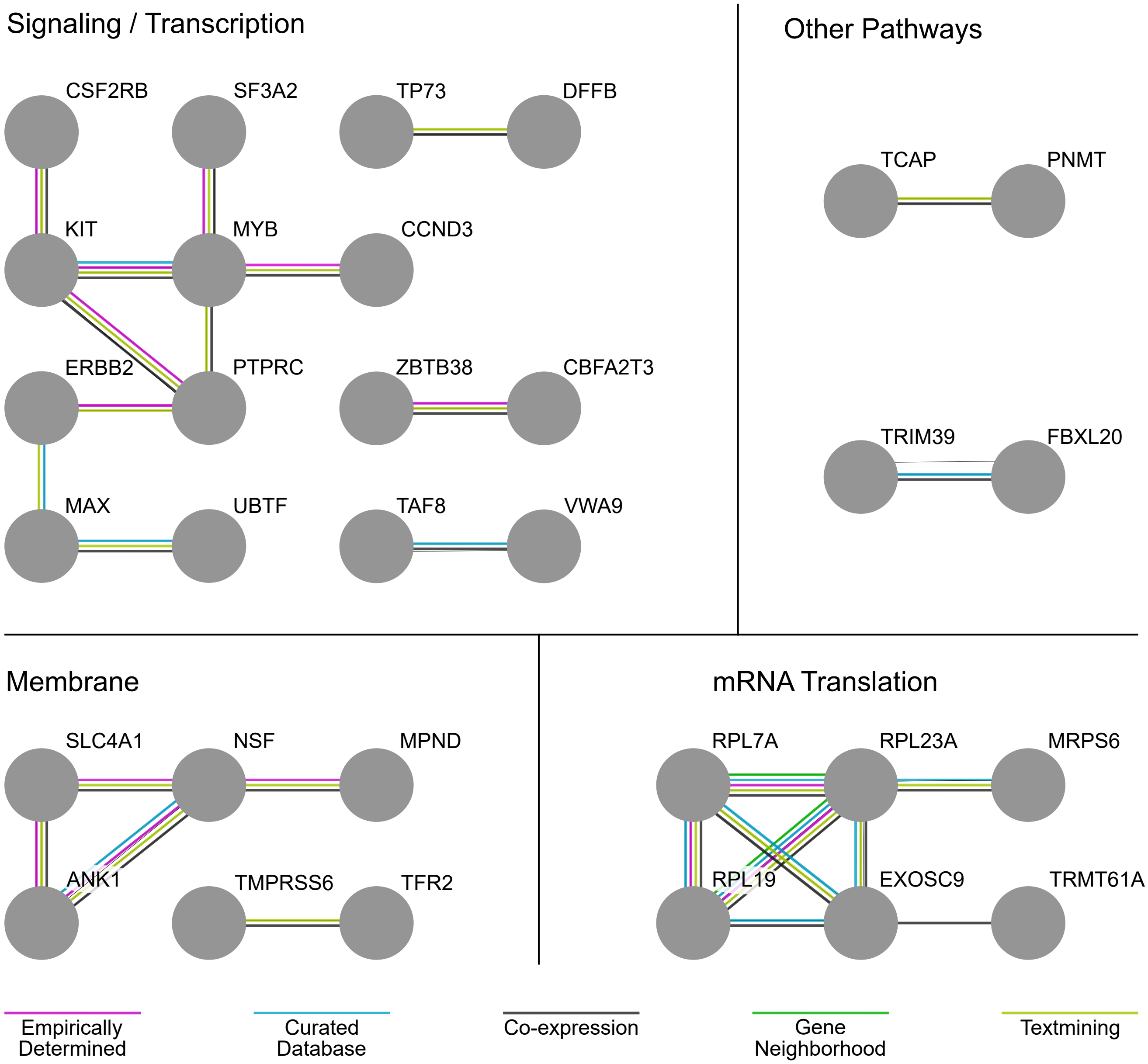

Signaling / Transcription
CSF2RB
SF3A2
TP73
DFFB
KIT
MYB
CCND3
ERBB2
PTPRC
ZBTB38
CBFA2T3
MAX
UBTF
TAF8
VWA9
Other Pathways
TCAP
PNMT
TRIM39
FBXL20
Membrane
SLC4A1
NSF
MPND
ANK1
TMPRSS6
TFR2
mRNA Translation
RPL7A
RPL23A
MRPS6
RPL19
EXOSC9
TRMT61A
Empirically Determined
Curated Database
Co-expression
Gene Neighborhood
Textmining

Figure 5

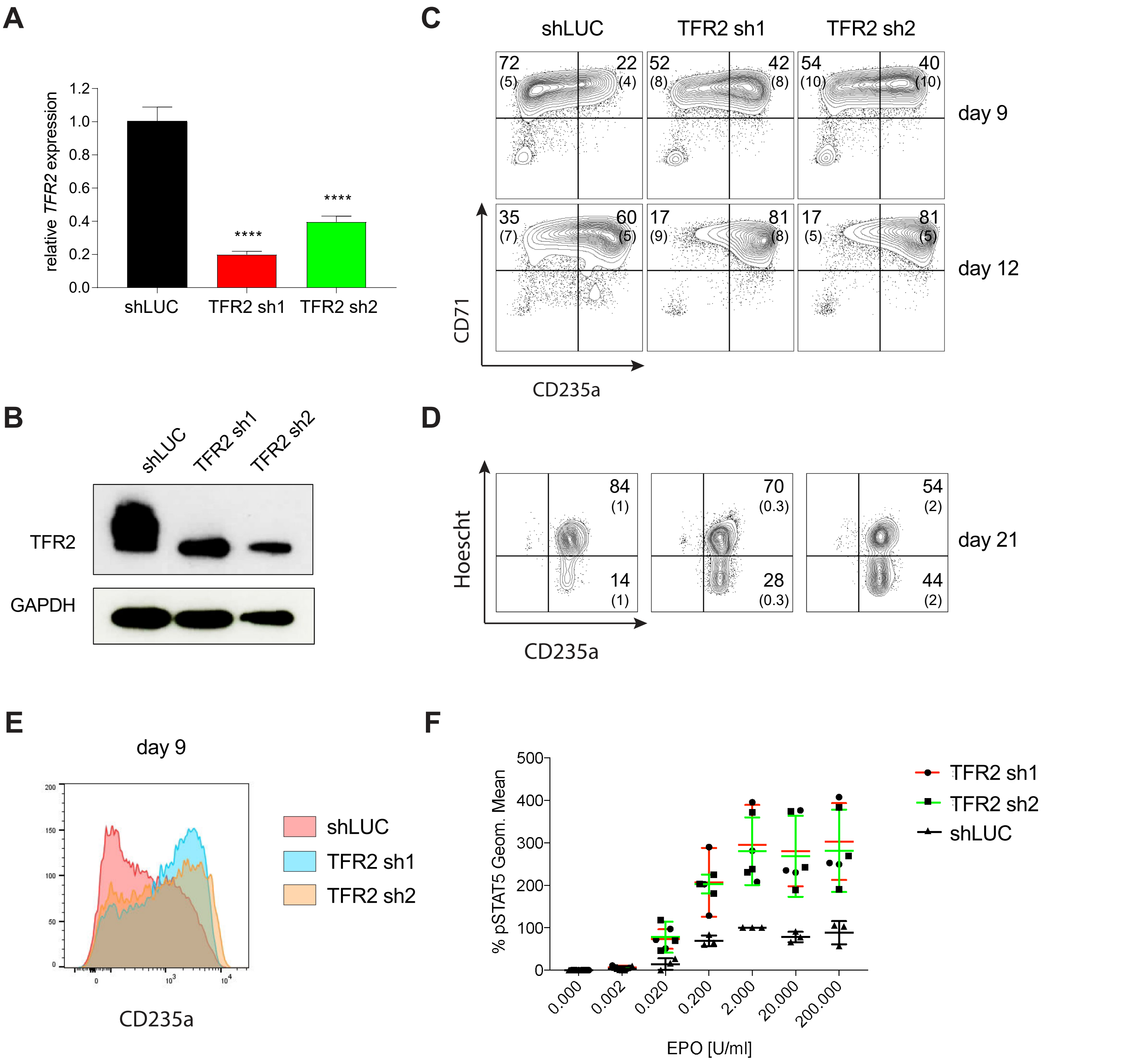

A
relative TFR2 expression
1.2
1.0
0.8
0.6
0.4
0.2
0.0
****
****
shLUC
TFR2 sh1
TFR2 sh2
C
shLUC
TFR2 sh1
TFR2 sh2
72 (5)
22 (4)
52 (8)
42 (8)
54 (10)
40 (10)
day 9
35 (7)
60 (5)
17 (9)
81 (8)
17 (5)
81 (5)
day 12
CD71
CD235a
B
shLUC
TFR2 sh1
TFR2 sh2
TFR2
GAPDH
D
Hoescht
84 (1)
14 (1)
70 (0.3)
28 (0.3)
54 (2)
44 (2)
day 21
CD235a
E
day 9
200
150
100
50
0
0
10³
10⁴
CD235a
shLUC
TFR2 sh1
TFR2 sh2
F
% pSTAT5 Geom. Mean
500
400
300
200
100
0
0.000
0.002
0.020
0.200
2.000
20.000
200.000
EPO [U/ml]
TFR2 sh1
TFR2 sh2
shLUC

Figure 6

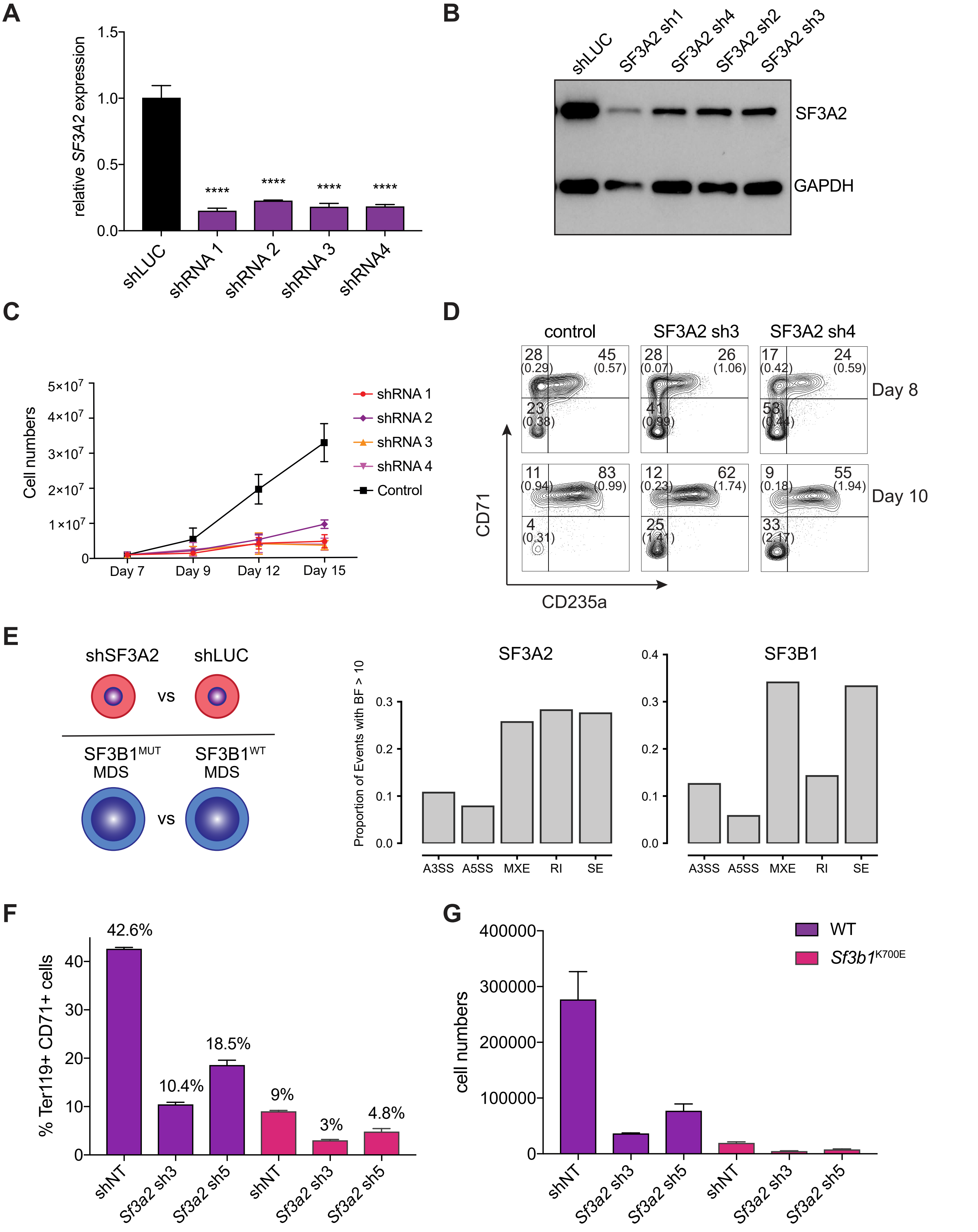

A
relative SF3A2 expression
1.5
1.0
0.5
0.0
****
****
****
****
shLUC
shRNA 1
shRNA 2
shRNA 3
shRNA4
B
shLUC
SF3A2 sh1
SF3A2 sh4
SF3A2 sh2
SF3A2 sh3
SF3A2
GAPDH
C
Cell numbers
5×10^7
4×10^7
3×10^7
2×10^7
1×10^7
0
Day 7
Day 9
Day 12
Day 15
shRNA 1
shRNA 2
shRNA 3
shRNA 4
Control
D
control
SF3A2 sh3
SF3A2 sh4
28 (0.29)
45 (0.57)
23 (0.38)
28 (0.07)
26 (1.06)
41 (0.39)
17 (0.42)
24 (0.59)
58 (0.44)
Day 8
11 (0.94)
83 (0.99)
4 (0.31)
12 (0.23)
62 (1.74)
25 (1.41)
9 (0.18)
55 (1.94)
33 (2.17)
Day 10
CD71
CD235a
E
shSF3A2
vs
shLUC
SF3B1^MUT MDS
vs
SF3B1^WT MDS
Proportion of Events with BF > 10
SF3A2
0.3
0.2
0.1
0.0
A3SS
A5SS
MXE
RI
SE
SF3B1
0.4
0.3
0.2
0.1
0.0
A3SS
A5SS
MXE
RI
SE
F
% Ter119+ CD71+ cells
40
30
20
10
0
42.6%
10.4%
18.5%
9%
3%
4.8%
shNT
Sf3a2 sh3
Sf3a2 sh5
shNT
Sf3a2 sh3
Sf3a2 sh5
G
cell numbers
400000
300000
200000
100000
0
WT
Sf3b1^K700E
shNT
Sf3a2 sh3
Sf3a2 sh5
shNT
Sf3a2 sh3
Sf3a2 sh5

Figure S1

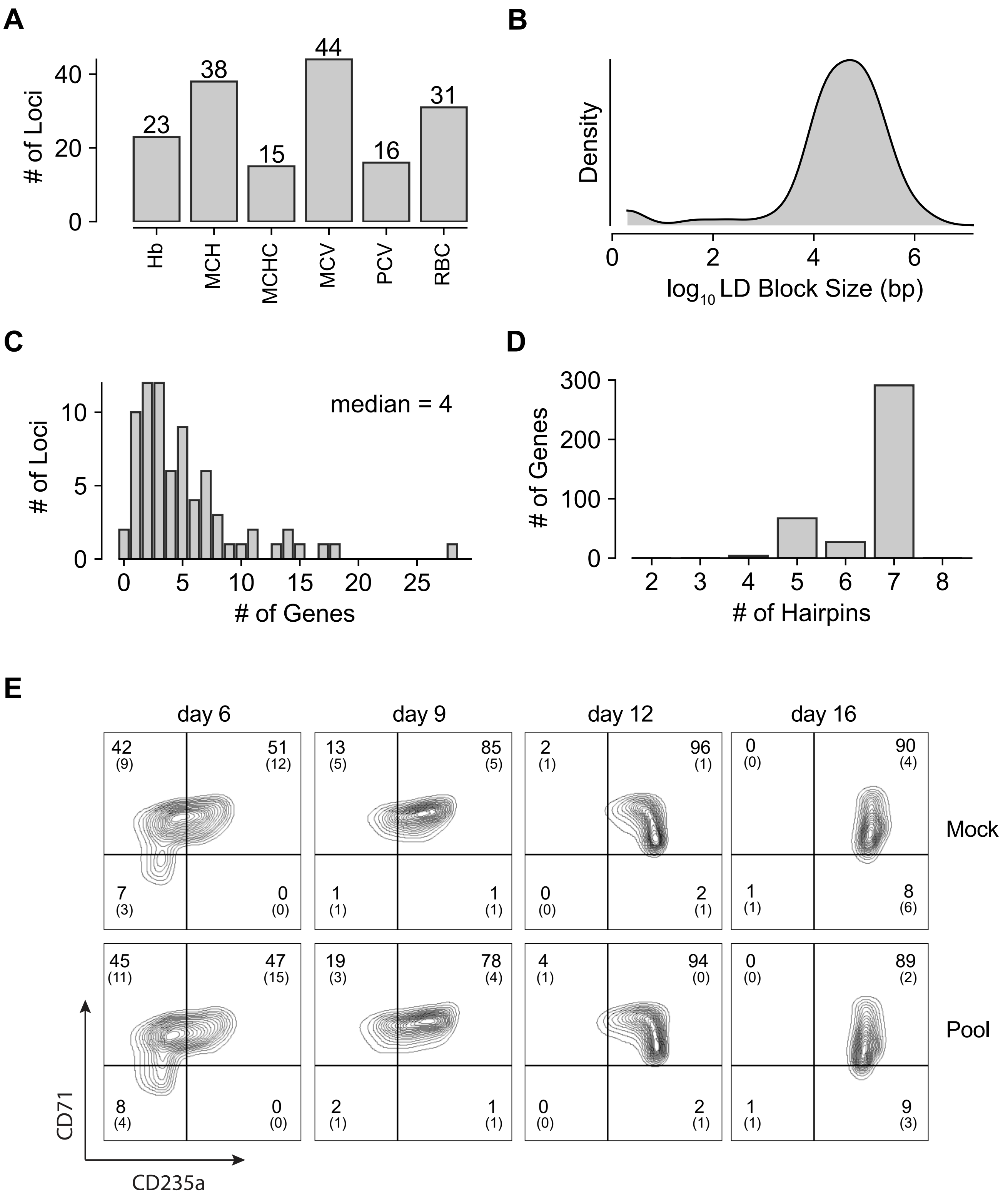

A
# of Loci
40
20
0
23
38
15
44
16
31
Hb
MCH
MCHC
MCV
PCV
RBC
B
Density
0
2
4
6
log10 LD Block Size (bp)
C
# of Loci
10
5
0
median = 4
0
5
10
15
20
25
# of Genes
D
# of Genes
300
200
100
0
2
3
4
5
6
7
8
# of Hairpins
E
day 6
day 9
day 12
day 16
42 (9)
51 (12)
7 (3)
0 (0)
13 (5)
85 (5)
1 (1)
1 (1)
2 (1)
96 (1)
0 (0)
2 (1)
0 (0)
90 (4)
1 (1)
8 (6)
Mock
45 (11)
47 (15)
8 (4)
0 (0)
19 (3)
78 (4)
2 (1)
1 (1)
4 (1)
94 (0)
0 (0)
2 (1)
0 (0)
89 (2)
1 (1)
9 (3)
Pool
CD71
CD235a

Figure S2

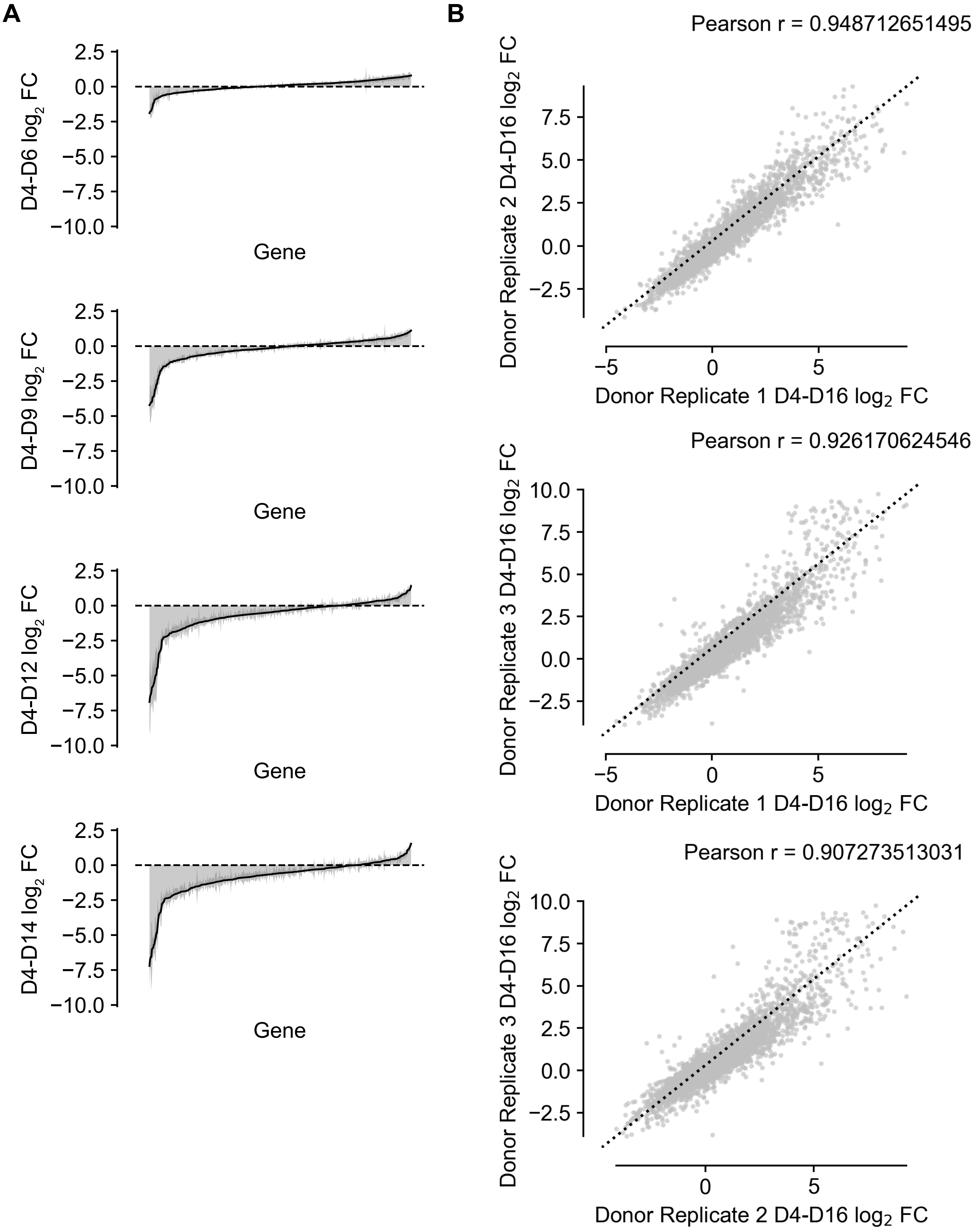

A
D4-D6 $\log_2$ FC
2.5
0.0
−2.5
−5.0
−7.5
−10.0
Gene
D4-D9 $\log_2$ FC
2.5
0.0
−2.5
−5.0
−7.5
−10.0
Gene
D4-D12 $\log_2$ FC
2.5
0.0
−2.5
−5.0
−7.5
−10.0
Gene
D4-D14 $\log_2$ FC
2.5
0.0
−2.5
−5.0
−7.5
−10.0
Gene
B
Pearson r = 0.948712651495
Donor Replicate 2 D4-D16 $\log_2$ FC
7.5
5.0
2.5
0.0
−2.5
−5
0
5
Donor Replicate 1 D4-D16 $\log_2$ FC
Pearson r = 0.926170624546
Donor Replicate 3 D4-D16 $\log_2$ FC
10.0
7.5
5.0
2.5
0.0
−2.5
−5
0
5
Donor Replicate 1 D4-D16 $\log_2$ FC
Pearson r = 0.907273513031
Donor Replicate 3 D4-D16 $\log_2$ FC
10.0
7.5
5.0
2.5
0.0
−2.5
0
5
Donor Replicate 2 D4-D16 $\log_2$ FC

Figure S3

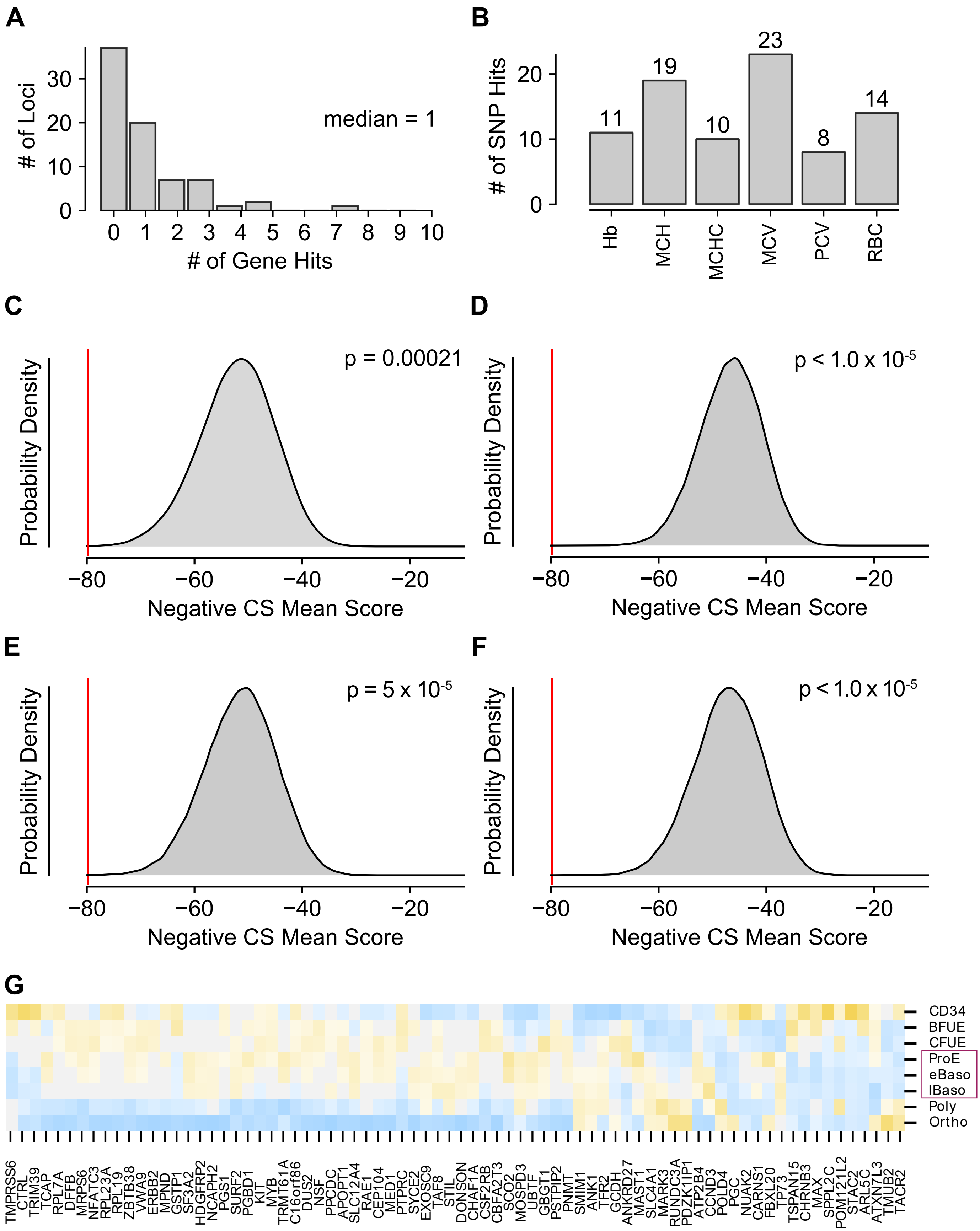

A
# of Loci
0
10
20
30
0 1 2 3 4 5 6 7 8 9 10
# of Gene Hits
median = 1
B
# of SNP Hits
0
10
20
11
19
10
23
8
14
Hb
MCH
MCHC
MCV
PCV
RBC
C
Probability Density
p = 0.00021
−80 −60 −40 −20
Negative CS Mean Score
D
Probability Density
$p < 1.0 \times 10^{-5}$
−80 −60 −40 −20
Negative CS Mean Score
E
Probability Density
$p = 5 \times 10^{-5}$
−80 −60 −40 −20
Negative CS Mean Score
F
Probability Density
$p < 1.0 \times 10^{-5}$
−80 −60 −40 −20
Negative CS Mean Score
G
CD34
BFUE
CFUE
ProE
eBaso
lBaso
Poly
Ortho
TMPRSS6
CTRL
TRIM39
TCAP
RPL7A
DFFB
MRPS6
NFATC3
RPL23A
RPL19
ZBTB38
VWA9
ERBB2
MPND
GSTP1
SF3A2
HDGFRP2
NCAPH2
PGS1
SURF2
PGBD1
KIT
MYB
TRMT61A
C16orf86
DUS2
NSF
PPCDC
APOPT1
SLC12A4
RAE1
CEP104
MED1
PTPRC
SYCE2
EXOSC9
TAF8
STIL
DONSON
CHAF1A
CSF2RB
CBFA2T3
SCO2
MOSPD3
UBTF
GBGT1
PSTPIP2
PNMT
SMIM1
ANK1
TFR2
GCDH
ANKRD27
MAST1
SLC4A1
MARK3
RUNDC3A
PDZK1IP1
ATP2B4
CCND3
POLD4
PGC
NUAK2
CARNS1
FBXL20
TP73
TSPAN15
CHRNB3
MAX
SPPL2C
POM121L2
STAC2
ARL5C
ATXN7L3
TMUB2
TACR2

Figure S5

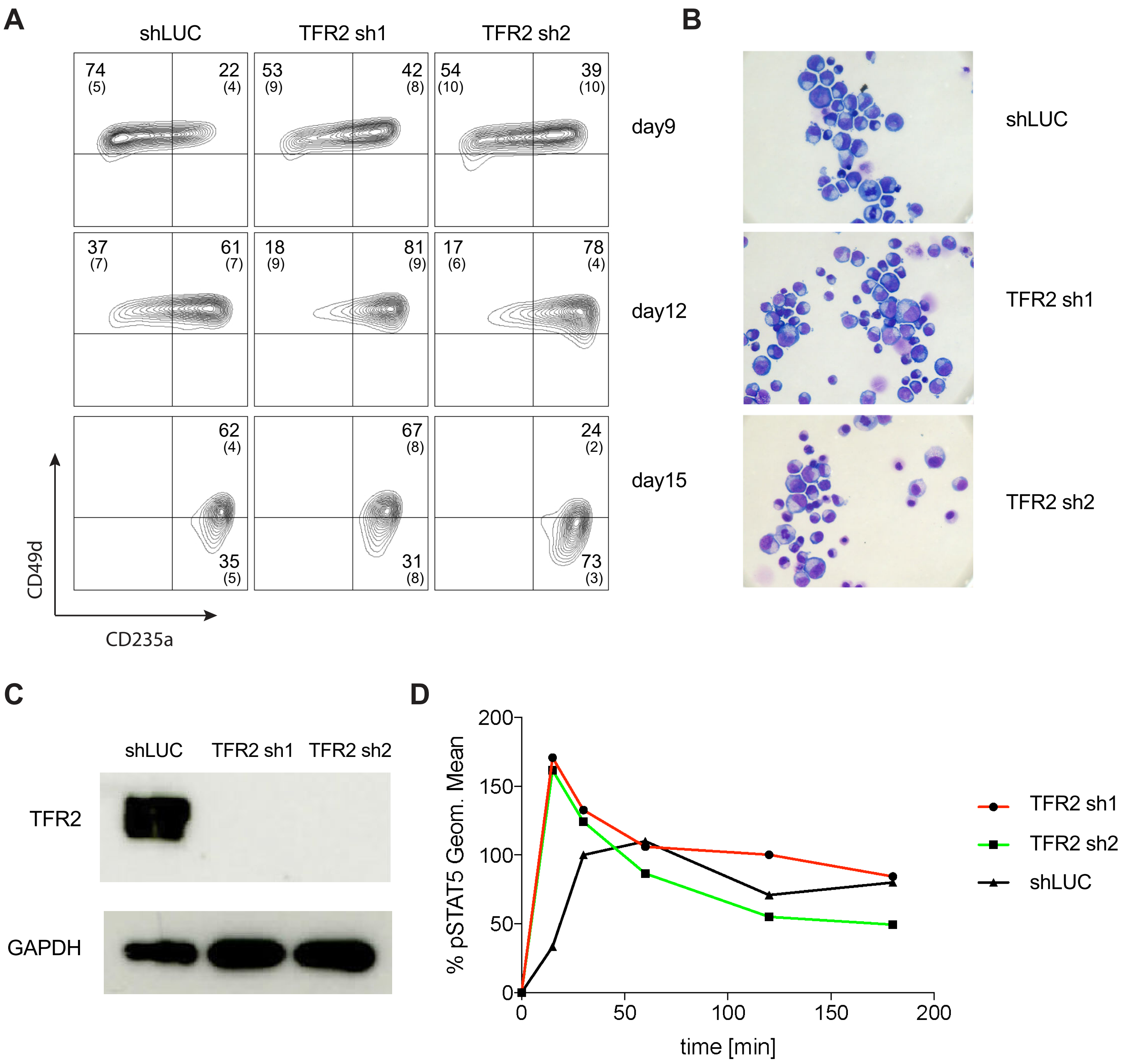
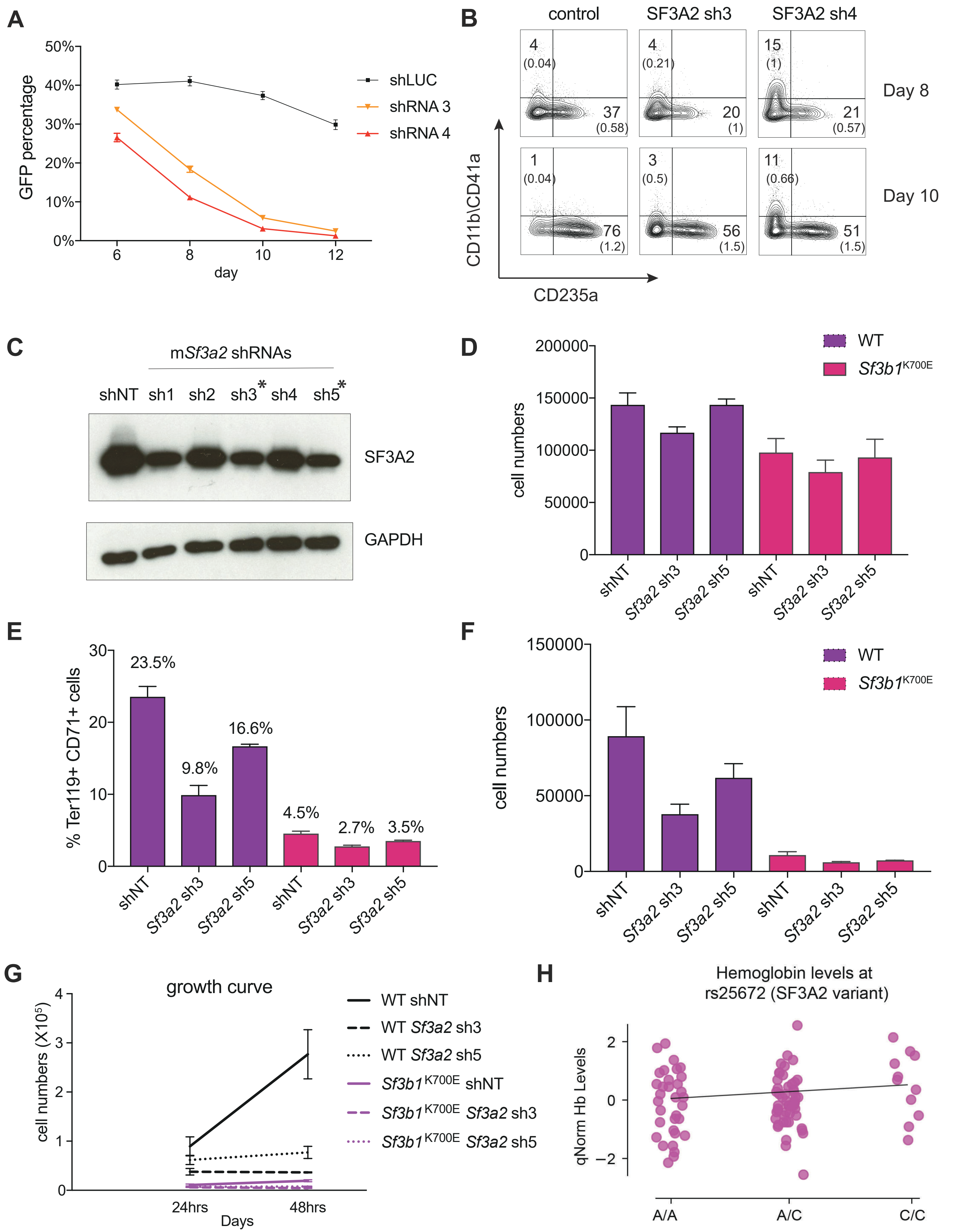

A
shLUC
TFR2 sh1
TFR2 sh2
74 (5)
22 (4)
53 (9)
42 (8)
54 (10)
39 (10)
day9
37 (7)
61 (7)
18 (9)
81 (9)
17 (6)
78 (4)
day12
62 (4)
35 (5)
67 (8)
31 (8)
24 (2)
73 (3)
day15
CD49d
CD235a
B
shLUC
TFR2 sh1
TFR2 sh2
C
shLUC
TFR2 sh1
TFR2 sh2
TFR2
GAPDH
D
% pSTAT5 Geom. Mean
200
150
100
50
0
0
50
100
150
200
time [min]
TFR2 sh1
TFR2 sh2
shLUC

Figure S6


A
GFP percentage
50%
40%
30%
20%
10%
0%
6
8
10
12
day
shLUC
shRNA 3
shRNA 4
B
control
SF3A2 sh3
SF3A2 sh4
4 (0.04)
37 (0.58)
4 (0.21)
20 (1)
15 (1)
21 (0.57)
Day 8
1 (0.04)
76 (1.2)
3 (0.5)
56 (1.5)
11 (0.66)
51 (1.5)
Day 10
CD11b\CD41a
CD235a
C
mSf3a2 shRNAs
shNT
sh1
sh2
sh3*
sh4
sh5*
SF3A2
GAPDH
D
200000
150000
100000
50000
0
cell numbers
WT
Sf3b1K700E
shNT
Sf3a2 sh3
Sf3a2 sh5
shNT
Sf3a2 sh3
Sf3a2 sh5
E
% Ter119+ CD71+ cells
30
20
10
0
23.5%
9.8%
16.6%
4.5%
2.7%
3.5%
shNT
Sf3a2 sh3
Sf3a2 sh5
shNT
Sf3a2 sh3
Sf3a2 sh5
F
150000
100000
50000
0
cell numbers
WT
Sf3b1K700E
shNT
Sf3a2 sh3
Sf3a2 sh5
shNT
Sf3a2 sh3
Sf3a2 sh5
G
growth curve
cell numbers (X10⁵)
4
3
2
1
0
24hrs
48hrs
Days
WT shNT
WT Sf3a2 sh3
WT Sf3a2 sh5
Sf3b1K700E shNT
Sf3b1K700E Sf3a2 sh3
Sf3b1K700E Sf3a2 sh5
H
Hemoglobin levels at rs25672 (SF3A2 variant)
qNorm Hb Levels
2
0
−2
A/A
A/C
C/C