nature genetics

Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci

Matthew C Canver[1], Samuel Lessard[2], Luca Pinello[3], Yuxuan Wu[1], Yann Ilboudo[2], Emily N Stern[1], Austen J Needleman[1], Frédéric Galactéros[4], Carlo Brugnara[5], Abdullah Kutlar[6], Colin McKenzie[7], Marvin Reid[7], Diane D Chen[1], Partha Pratim Das[1], Mitchel A Cole[1], Jing Zeng[1], Ryo Kurita[8], Yukio Nakamura[9,10], Guo-Cheng Yuan[11], Guillaume Lettre[2], Daniel E Bauer[1,13] & Stuart H Orkin[1,12,13]

Cas9-mediated, high-throughput, saturating *in situ* mutagenesis permits fine-mapping of function across genomic segments. Disease- and trait-associated variants identified in genome-wide association studies largely cluster at regulatory loci. Here we demonstrate the use of multiple designer nucleases and variant-aware library design to interrogate trait-associated regulatory DNA at high resolution. We developed a computational tool for the creation of saturating-mutagenesis libraries with single or multiple nucleases with incorporation of variants. We applied this methodology to the *HBS1L-MYB* intergenic region, which is associated with red-blood-cell traits, including fetal hemoglobin levels. This approach identified putative regulatory elements that control *MYB* expression. Analysis of genomic copy number highlighted potential false-positive regions, thus emphasizing the importance of off-target analysis in the design of saturating-mutagenesis experiments. Together, these data establish a widely applicable high-throughput and high-resolution methodology to identify minimal functional sequences within large disease- and trait-associated regions.

Genome-wide association studies (GWAS) are a powerful approach for the identification of disease- and trait-associated variants. More than 90% of GWAS variants lie within noncoding DNA[1]. However, linkage disequilibrium (LD) often obscures the causal variant and hence the biological mechanisms producing the trait association. Reliable methods to identify the underlying functional sequences remain elusive. Clustered regularly interspaced short palindromic repeats (CRISPR)-based genome-editing systems have emerged as highly efficient tools to study regulatory DNA. Targeted deletion provides a valuable tool for loss of function[2,3]. However, targeted deletion has limited throughput, efficiency, and resolution[4]. Alternatively, the homology-directed repair (HDR) pathway can be exploited after cleavage by a designer nuclease to insert putative causal variants into endogenous DNA sequence by using a customized extra-chromosomal template. However, HDR used to insert variants has low throughput and is limited by efficiency. Furthermore, individual trait-associated variants may underestimate the effect of the underlying haplotype and consequently may underestimate the biological importance of the given genetic element[2,3,5].

Saturating a region with insertions/deletions (indels) by using every available protospacer-adjacent motif (PAM)-restricted single guide RNA (sgRNA) is a powerful strategy to identify minimal functional sequences within regulatory DNA[3]. Saturating mutagenesis relies on pooled screening to take advantage of the typical indel spectrum after nonhomologous end joining (NHEJ) repair of 1 to 10 bp[3,4,6–9]. The ability to saturate a region with indels is a function of PAM availability. Moreover, genomic variants that attenuate sgRNA activity may decrease resolution through false negatives. We hypothesized that combining multiple nucleases with unique PAM sequences would enhance mutagenesis resolution and that incorporating variants into sgRNA library design would minimize false negatives associated with libraries based on the reference genome. To test this hypothesis, we applied this methodology to the *HBS1L-MYB* intergenic region.

[1]Division of Hematology/Oncology, Boston Children's Hospital; Department of Pediatric Oncology, Dana-Farber Cancer Institute; Harvard Stem Cell Institute; and Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, USA. [2]Montreal Heart Institute, Université de Montréal, Montréal, Québec, Canada. [3]Department of Molecular Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. [4]Red Cell Genetic Disease Unit, Hôpital Henri-Mondor, Assistance Publique–Hôpitaux de Paris (AP-HP), UPeC, IMRB U955 Equipe no. 2, Créteil, France. [5]Department of Laboratory Medicine, Boston Children's Hospital, Boston, Massachusetts, USA. [6]Department of Medicine, Sickle Cell Center, Augusta University, Augusta, Georgia, USA. [7]The Caribbean Institute for Health Research, University of the West Indies, Mona, Kingston, Jamaica. [8]Department of Research and Development, Central Blood Institute, Japanese Red Cross Society, Tokyo, Japan. [9]Cell Engineering Division, RIKEN BioResource Center, Tsukuba, Japan. [10]Faculty of Medicine, University of Tsukuba, Tsukuba, Japan. [11]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [12]Howard Hughes Medical Institute, Boston, Massachusetts, USA. [13]These authors jointly supervised this work. Correspondence should be addressed to S.H.O. (Stuart_Orkin@dfci.harvard.edu) or D.E.B. (Daniel.Bauer@childrens.harvard.edu).

Received 31 October 2016; accepted 25 January 2017; published online 20 February 2017; doi:10.1038/ng.3793

RESULTS

The *HBS1L-MYB* intergenic region is associated with erythroid traits

GWAS, quantitative-trait-loci studies, and other human genetic studies of fetal hemoglobin (HbF) levels (or the related trait F-cell number) have highlighted the *HBS1L-MYB* interval[10–17]. The *HBS1L-MYB* interval has also been associated with erythroid traits including levels of hemoglobin, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, mean corpuscular volume, packed-cell volume, and red-blood-cell count[18–23]. These associations have been suggested to reflect changes in the expression of *MYB*, owing to distant variants localizing kilobases away, approximately equidistant to the *HBS1L* gene[15]. Genotyping in multiple cohorts of individuals with sickle-cell disease (SCD) (n = 2,222) was conducted to refine the genetic association with HbF levels (**Fig. 1a** and **Supplementary Table 1**).

This HbF meta-analysis identified single-nucleotide polymorphisms (SNPs) with clustering similar to that in a previously published meta-analysis of variants associated with erythroid traits[22] (**Fig. 1b** and **Supplementary Table 2**). Owing to extensive LD and limited sample size, conditional analysis of HbF-associated SNPs could not confidently pinpoint a specific set of causal variants (**Supplementary Table 3**). Recent studies using lineage-restricted expression patterns, clustering of erythroid transcription factor–binding sites affecting *MYB* expression, and chromatin conformation capture have suggested that HbF-associated variants modulate *MYB* expression by altering GATA1- or GATA1–TAL1-binding motifs within regulatory elements 71 and 84 kb upstream of the *MYB* transcription start site (TSS)[15]. However, our meta-analysis, which, to our knowledge is the largest performed to date for HbF levels in SCD patients, was unable to discriminate between the previously reported causal variant (rs66650371) and other markers in strong LD.

The *HBS1L-MYB* region is composed of 98 DNase I–hypersensitive sites (DHSs), as identified from erythroid precursors[2] (**Fig. 1**). The trait-associated SNPs from both meta-analyses are concentrated in an 83-kb intergenic super-enhancer (**Fig. 1** and **Supplementary Fig. 1**). To interrogate the *HBS1L-MYB* locus in a comprehensive fashion, we subjected each of the 98 DHSs to saturating mutagenesis.

Distribution of PAM sequences in the genome and outline of the *DNA Striker* algorithm

Maximizing the degree of saturating mutagenesis depends on minimizing the genomic distance between potential adjacent cleavages. To functionally fine-map the *HBS1L-MYB* intergenic region, we reasoned that use of multiple highly saturating nucleases in combination might increase resolution. We further hypothesized that designing a variant-aware saturating-mutagenesis library might limit false negatives resulting from diminished sgRNA activity due to variants present in the cells used for study, a consideration highlighted by the region's trait association with common genetic variants. To design a variant-aware saturating-mutagenesis library by using multiple nucleases, we created the *DNA Striker* computational tool (**Fig. 2** and **Supplementary Fig. 2**). It facilitates design of saturating-mutagenesis libraries, using single or multiple designer nucleases, and alternative sgRNAs based on haplotype structure, whole-genome sequencing (WGS), or a custom list of variants. The algorithm is summarized in **Figure 2** (details in Online Methods).

Saturating-mutagenesis-library design

CRISPR-associated nucleases with unique PAM-recognition sequences have been reported for genome editing[6,7,24–29]. The frequency of each PAM varies throughout the genome (**Fig. 3a**, **Supplementary Fig. 3**, and **Supplementary Table 4**). Given the sequence dependence of PAM availability, feature-specific variation in cleavage density for each nuclease was observed in DHSs, enhancers, and repressed regions as well as genes (**Supplementary Figs. 4–7** and **Supplementary Table 5**).

We reasoned that combining multiple species of Cas9 nucleases with unique PAM sequences would enhance the resolution of saturating mutagenesis. To evaluate this approach, we used the regions of each DHS summit (peak of DNase I sensitivity) ±200 bp within the *HBS1L-MYB* intergenic region for saturating mutagenesis. NGG- and NGA-PAM-restricted sgRNAs were chosen because these PAM sequences resulted in the lowest mean and median gap distance between adjacent genomic cleavages in DHSs (**Supplementary Fig. 4** and **Supplementary Table 5**).

To demonstrate the feasibility of using these nucleases, and to evaluate the specificity and efficiency of *Streptococcus pyogenes* Cas9 (SpCas9; NGG PAM) and *S. pyogenes* VQR-variant Cas9 (SpCas9-VQR; NGA PAM)[28], we used Cas9 reporter constructs that delivered *GFP* as well as either an NGG-restricted or NGA-restricted sgRNA targeting *GFP*. Cells stably expressing SpCas9, SpCas9-VQR, or no Cas9 were transduced with the reporter construct at low multiplicity and selected for 14 d. The analysis demonstrated that the SpCas9 and SpCas9-VQR Cas9 proteins were both specific and efficient nucleases, because SpCas9 led to decreased GFP with only the NGG-restricted sgRNA, and SpCas9-VQR led to decreased GFP with only the NGA-restricted sgRNA (**Fig. 3b**).

Therefore, we used *DNA Striker* to design a high-resolution saturating-mutagenesis library consisting of all 20-mer sequences upstream of an NGG or NGA PAM sequence on the top or bottom strand within the *HBS1L-MYB*-region DHSs, as well as controls including *BCL11A* exon 2, the core of the +58 DHS within the *BCL11A* enhancer[3], *HBS1L* exon 4, and *MYB* exon 5 (**Fig. 3c** and **Supplementary Tables 6** and **7**). The median and 90th-percentile gap distance between adjacent genomic cleavages with SpCas9 was 5 bp and 22.5 bp, respectively and was 6 bp and 18 bp for SpCas9-VQR (**Fig. 3d**). The combination of both SpCas9 and SpCas9-VQR nucleases led to a decrease in the median and 90th-percentile gap between adjacent genomic cleavages, to 3 bp and 11 bp, respectively (**Fig. 3d**). Furthermore, use of both nucleases decreased the maximum gap size from 115 bp for SpCas9 and 82 bp for SpCas9-VQR to a maximum of 41 bp for the combination (**Supplementary Fig. 8**). Therefore, the inclusion of sgRNAs restricted by two distinct nucleases resulted in higher resolution by decreasing the 50th and 90th percentiles of distances between adjacent genomic cleavages as well as decreasing the maximum gap between adjacent cleavages. The use of multiple nucleases allows for minimization of the distance of double-strand breaks (DSBs) to SNPs and motifs of interest, thereby enhancing functional interrogation of regions of interest (**Supplementary Fig. 9**).

To construct a variant-informed library, phased variants within these regions were taken from the 1000 Genomes Project database from all populations and incorporated into sgRNA design with *DNA Striker* to identify potential altered sgRNAs and novel sgRNAs resulting from variant-induced PAM creation (**Figs. 2** and **3c**). Haplotype-associated sgRNAs were included in the library if they were present at a frequency ≥1% (NGG, 176/1,350 haplotype-associated sgRNAs; NGA, 186/1,551 haplotype-associated sgRNAs) (**Figs. 2** and **3c**, and **Supplementary Fig. 10a**,**b**). Both NGG- and NGA-restricted sgRNA libraries were synthesized and successfully batch-cloned into lentiviral constructs (**Supplementary Fig. 10c**,**d**).

Cutting-frequency determination (CFD) has previously been used to assess the activity of imperfect-match sgRNAs[30]. We used CFD

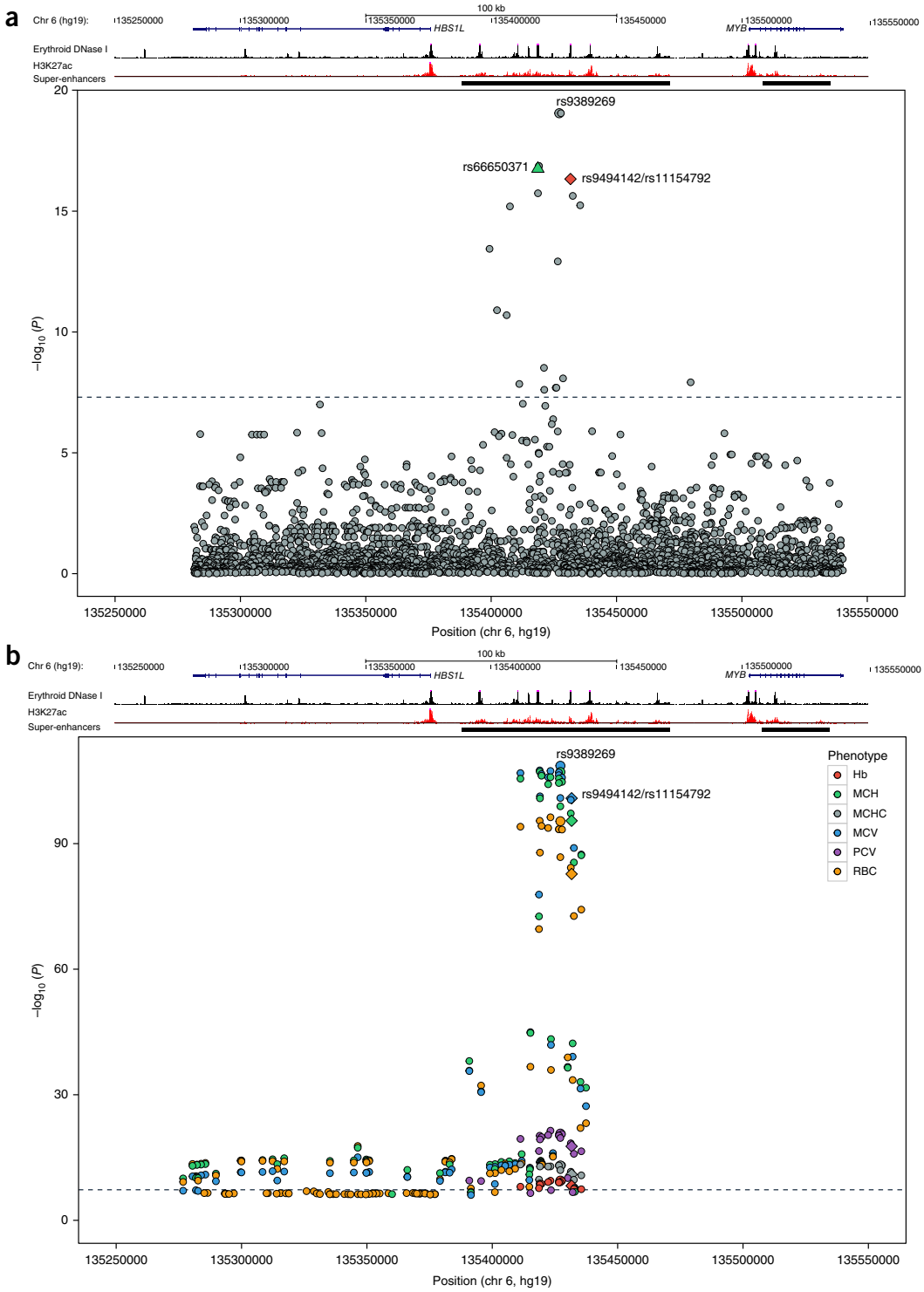


Figure 1 Trait associations of the *HBS1L-MYB* intergenic region. (**a**) Meta-analysis of HbF-associated SNPs from SCD cohorts (n = 2,222). rs66650371 (green triangle) and rs9494142/rs11154792 (red diamond) have previously been implicated as possible functional SNPs affecting *MYB* expression[15]. The larger dot (gray) corresponds to the top HbF-associated SNP, rs9389269. The super-enhancer region is indicated by a black horizontal bar. Genome-wide significance is indicated by a horizontal dotted line ($P < 5 \times 10^{-8}$; P values were calculated with linear regression, as described in the Online Methods). Schematic of the *HBS1L-MYB* interval region (hg19) with erythroid DNase I hypersensitivity and acetylated histone H3 K27 (H3K27ac) is shown above the meta-analysis. (**b**) Previously published meta-analyses of SNPs associated with erythroid traits including hemoglobin (Hb, red), mean corpuscular hemoglobin (MCH, green), mean corpuscular hemoglobin concentration (MCHC, gray), mean corpuscular volume (MCV, blue), packed-cell volume (PCV, purple), and red-blood-cell count (RBC, orange)[22]. Only SNPs with $P < 10^{-6}$ are displayed. The super-enhancer region is indicated by a black horizontal bar. Genome-wide significance is indicated by a horizontal dotted line ($P < 5 \times 10^{-8}$). The larger dots correspond to the top HbF-associated SNP, rs9389269. The diamonds correspond to rs9494142/rs11154792. Schematic of the *HBS1L-MYB* interval region (hg19) with erythroid DNase I hypersensitivity and H3K27ac is shown above the meta-analysis.

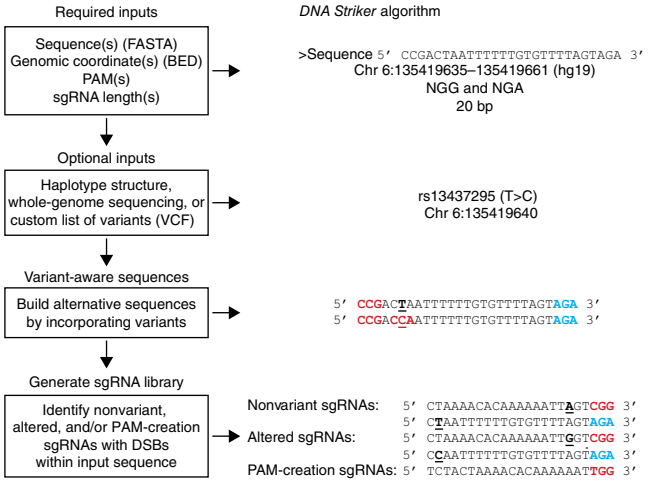


Figure 2 *DNA Striker* algorithm. Description of the *DNA Striker* algorithm for sgRNA design, which allows for creation of variant-aware saturating-mutagenesis libraries from haplotype structure, WGS, or custom lists of variants. *DNA Striker* can output libraries by using any combination of PAM sequences. NGG and NGA library design is shown as a representative example. In this example, NGG PAMs are shown in red, NGA PAMs are shown in blue, and the positions of variants are underlined.

analysis to assess the predicted activity of the haplotype-associated (nonreference) sgRNAs in the presence of the reference genome and the predicted activity of the nonvariant (reference) sgRNAs in the presence of the haplotype-derived variants (nonreference genome). This analysis demonstrated a decrease in CFD for reference sgRNAs in the presence of a nonreference genome, thus suggesting the utility of variant-aware library design (**Fig. 3e**). Furthermore, CFD analysis suggested that the majority of nonreference sgRNAs had diminished activity against the reference genome (**Fig. 3e**).

Functional saturating-mutagenesis screens with SpCas9 and SpCas9-VQR

For the *HBS1L-MYB* saturating-mutagenesis experiments, we used the immortalized human erythroid cell line HUDEP-2, which has previously been used to examine erythroid maturation and HbF regulation[3,31]. Briefly, HUDEP-2 cells stably expressing SpCas9 or SpCas9-VQR were transduced at low multiplicity with the NGG-restricted or NGA-restricted sgRNA library, respectively. Cells were expanded, differentiated, sorted on the basis of high and low HbF expression, and deep-sequenced to enumerate sgRNAs present in the HbF-high and HbF-low pools[3]. Three independent experiments were performed for both libraries. Unexpectedly, sgRNAs targeting *HBS1L* exon 4 and *MYB* exon 5 did not show significant HbF enrichment, although the positive-control sgRNAs targeting *BCL11A* exon 2 and *BCL11A* DHS +58 showed enrichment in the HbF-high pool, as expected (**Fig. 4a**). Interestingly, sgRNAs targeting *MYB* showed a tendency to 'drop out' (decrease in abundance) in the screen, a result consistent with MYB's known essential role in erythropoiesis[32] (**Fig. 4b**). *BCL11A* +58 DHS–targeted sgRNAs were not underrepresented, whereas *BCL11A* exon 2 sgRNAs showed modest dropout, in agreement with previous findings[3] (**Fig. 4b**). In addition, sgRNAs targeting *HBS1L* coding sequences did not drop out, thus suggesting that this gene does not contribute to the fitness of the HUDEP-2 cells. Mann–Whitney testing showed no significant differences in dropout between SpCas9 and SpCas9-VQR species ($P > 0.05$).

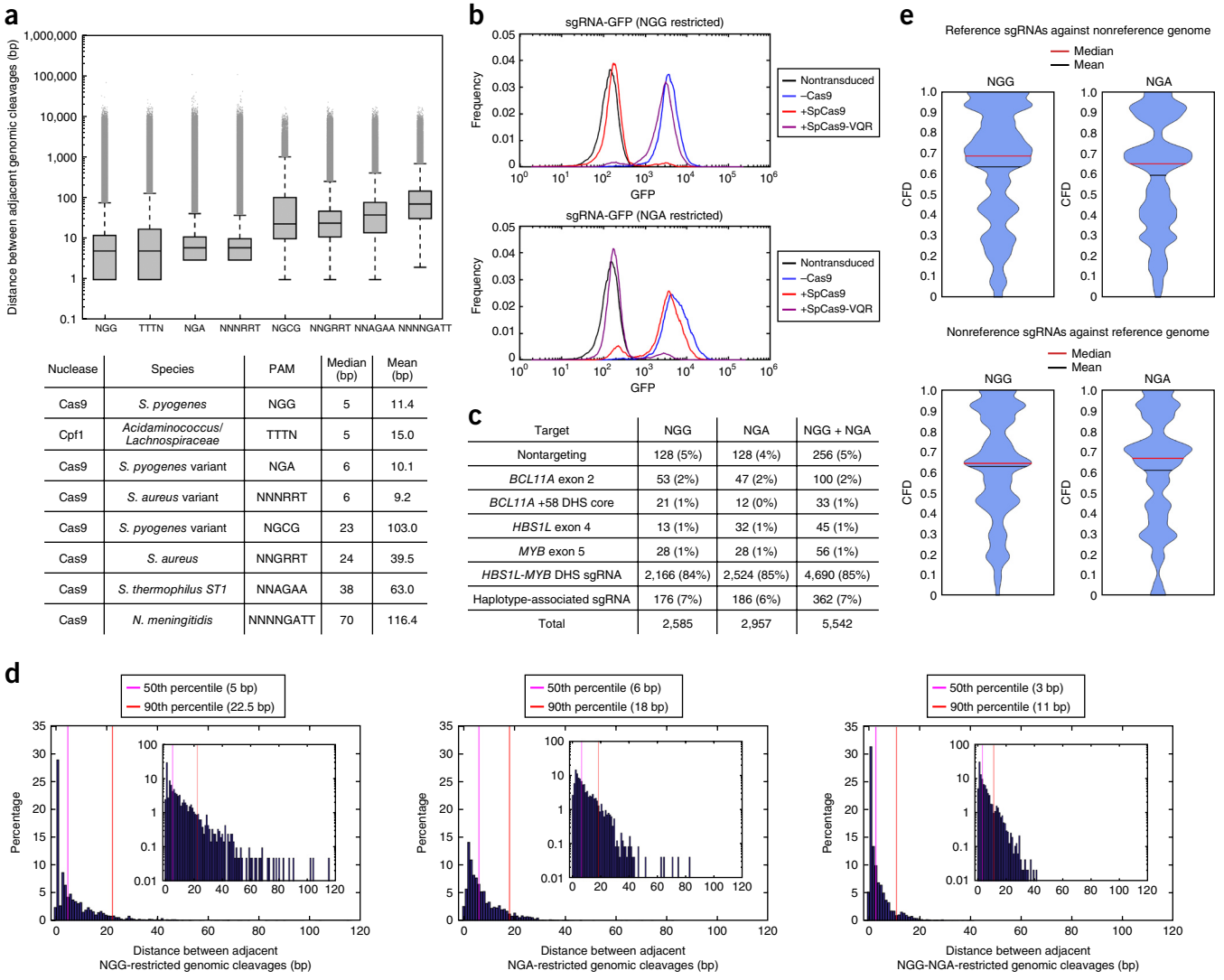
To orthogonally validate these findings, we evaluated MYB dependence in HUDEP-2 cells. Three short hairpin RNAs (shRNAs) efficiently depleted MYB and led to a cellular proliferation defect in HUDEP-2 cells, a result in agreement with the results of the CRISPR-based screen and indicative of MYB dependence (**Supplementary Fig. 11a**,**b**). We also examined the effects of MYB depletion in primary human CD34$^+$ hematopoietic stem and progenitor cells (HSPCs) from G-CSF-mobilized healthy adult donors subjected to erythroid differentiation conditions. The same shRNAs targeting *MYB* demonstrated a profound cellular proliferation defect in CD34$^+$ HSPC-derived human erythroblasts (**Supplementary Fig. 11c**). Erythroid differentiation was assessed at days 10, 14, and 18 of culture, on the basis of surface expression of the CD71 (transferrin receptor) and CD235a (glycophorin A) erythroid markers. A severe differentiation block was observed after *MYB* knockdown, in agreement with results from previous reports[33] (**Supplementary Fig. 11d**).

Introduction of an sgRNA targeting *MYB* coding sequence into HUDEP-2 cells stably expressing Cas9 resulted in an impairment of cellular proliferation, thus further indicating that HUDEP-2 cells rely on MYB for cell growth (**Supplementary Fig. 11e**). The same sgRNA targeting *MYB* also demonstrated a cell-proliferation defect in CD34$^+$ HSPC-derived human erythroblasts (**Supplementary Fig. 11f**). Notably, targeting *MYB* coding sequence in CD34$^+$ HSPC-derived human erythroblasts resulted in a significantly greater percentage of in-frame mutations than resulted from targeting of *BCL11A* and *HBS1L* coding sequences, thus suggesting strong selective pressure against loss-of-function *MYB* alleles (**Supplementary Fig. 11g**). Furthermore, targeting *MYB* led to a decrease in MYB expression (**Supplementary Fig. 11h**). Together our findings suggested that shRNA-mediated knockdown and CRISPR-mediated knockout of *MYB* resulted in proliferation defects in both HUDEP-2 cells and CD34$^+$ HSPC-derived erythroblasts, thus indicating the MYB dependence of these cells. These data additionally suggested that the *HBS1L-MYB* DHS CRISPR-based screen data could be analyzed on the basis of cellular dropout as opposed to HbF enrichment as the phenotype. Analysis of the library for dropout demonstrated that the majority of sgRNAs in both the NGG- and NGA-restricted libraries did not drop out, thereby suggesting a neutral effect on cell growth (**Fig. 4c**,**d**). Notably, specific sgRNAs with significant dropout were identified in both libraries (**Fig. 4c**,**d**).

Variant-aware high-resolution saturating mutagenesis of the *HBS1L-MYB* interval

The presence of multiple colocalizing top-scoring sgRNAs within *in situ* saturating-mutagenesis screens suggests the position of minimal functional sequences[3]. After mapping of the library sgRNAs to their associated genomic loci, the most potent dropout sgRNAs colocalized to discrete loci for both the NGG- and NGA-restricted libraries. A hidden Markov model (HMM) segmentation with three states (neutral, repressive, and active) was applied to the merged NGG and NGA dropout scores to identify functional sequence (**Fig. 4e**). The HMM analysis identified multiple regions of regulatory potential. These DHSs were termed −126, −83, −71, −36 (composed of two adjacent DHSs), and −7, on the basis of their distance from the *MYB* TSS (**Fig. 4e** and **Supplementary Figs. 12–15**).

Notably, the utilization of SpCas9 and SpCas9-VQR species together enhanced resolution at these DHSs by decreasing the gaps between adjacent genomic cleavages (**Supplementary Fig. 16**). In addition, a higher sgRNA density enhanced the reliability of functional sequence detection by HMM analysis. Notably, the −83 and −71 DHSs fell



Nuclease	Species	PAM	Median (bp)	Mean (bp)
Cas9	*S. pyogenes*	NGG	5	11.4
Cpf1	*Acidaminococcus/ Lachnospiraceae*	TTTN	5	15.0
Cas9	*S. pyogenes* variant	NGA	6	10.1
Cas9	*S. aureus* variant	NNNRRT	6	9.2
Cas9	*S. pyogenes* variant	NGCG	23	103.0
Cas9	*S. aureus*	NNGRRT	24	39.5
Cas9	*S. thermophilus ST1*	NNAGAA	38	63.0
Cas9	*N. meningitidis*	NNNNGATT	70	116.4

c

Target	NGG	NGA	NGG + NGA
Nontargeting	128 (5%)	128 (4%)	256 (5%)
BCL11A exon 2	53 (2%)	47 (2%)	100 (2%)
BCL11A +58 DHS core	21 (1%)	12 (0%)	33 (1%)
HBS1L exon 4	13 (1%)	32 (1%)	45 (1%)
MYB exon 5	28 (1%)	28 (1%)	56 (1%)
HBS1L-MYB DHS sgRNA	2,166 (84%)	2,524 (85%)	4,690 (85%)
Haplotype-associated sgRNA	176 (7%)	186 (6%)	362 (7%)
Total	2,585	2,957	5,542

Figure 3 Pooled saturating-mutagenesis screening of the *HBS1L-MYB* region by using NGG- and NGA Cas9s and variants from 1000 Genomes haplotypes. (**a**) Distances between adjacent genomic cleavages to assess genome-wide PAM availability and distribution. For each box plot, the three lines of the box represent the 25th, 50th and 75th percentiles. The upper and lower whiskers represent the 99th and first percentiles, respectively. Outliers, defined as above the 99th percentile or below the first percentile, are plotted as individual points. Lower whiskers are omitted if the first percentile is 0. *S*, *Streptococcus*; *N*, *Neisseria*. (**b**) Cells stably expressing SpCas9 (red) or SpCas9-VQR (purple), or lacking Cas9 (blue) were transduced with a Cas9-activity reporter containing *GFP* and either an NGG- (top) or NGA-restricted (bottom) *GFP*-targeting sgRNA. A nontransduced sample (black) was included as a negative control. (**c**) Library composition for NGG-restricted sgRNA library only, NGA-restricted sgRNA library only, and NGG- and NGA-restricted sgRNA libraries together. (**d**) For the *HBS1L-MYB* intergenic region DHSs, the genomic cleavage density when NGG-only (left), NGA-only (middle), and NGG and NGA combined (right) libraries were used. (**e**) Violin plots of CFD analysis for haplotype-associated sgRNAs with reference genomic sequence and for nonvariant sgRNAs with haplotype variants present.

within an annotated super-enhancer region, and each of these five DHSs colocalized with GATA1 and/or GATA1–TAL1 binding (**Fig. 4e** and **Supplementary Figs. 12** and **13**). These identified DHSs suggest regulatory potential for *MYB* expression.

Previous reports have nominated possible causal variants within the −84 and −71 DHSs that influence *MYB* expression[15]. Although saturating mutagenesis identified the −71 DHS as containing functional sequence, it suggested functional sequence localized to the −83 DHS as opposed to the −84 DHS (**Fig. 4e** and **Supplementary Figs. 13** and **14**). rs9389268, which is highly associated with erythroid traits, is located within the −83 DHS (**Fig. 5** and **Supplementary Fig. 13**). Interestingly, the 545-bp interval between −83 and −84 (chromosome (chr) 6, 135418850–135419395, hg19) has several HbF- and erythroid-associated SNPs (**Fig. 5a**,**c** and **Supplementary Fig. 17**). This region is DNase I insensitive in erythroid cells, so it was not included in the library design, although functional elements that lack epigenetic or chromatin characteristics typical of regulatory regions have recently been identified by CRISPR-based mutagenesis[34]. The top-scoring sgRNAs at the −71 element specified a cleavage ~200 bp from the peak of DNase I sensitivity and GATA1–TAL1 binding[34] (**Fig. 4e** and **Supplementary Fig. 12**). The highly trait-associated SNP within the −71 DHS that disrupts a GATA1 motif, rs9494142, also known as rs11154792 (ref. 15; denoted rs9494142/rs11154792 herein), localizes approximately 100 bp closer to the peak of DNase I sensitivity, as compared with the putative functional sequence (**Fig. 5b,d** and **Supplementary Fig. 12**). rs66650371 is a 3-bp indel that disrupts a

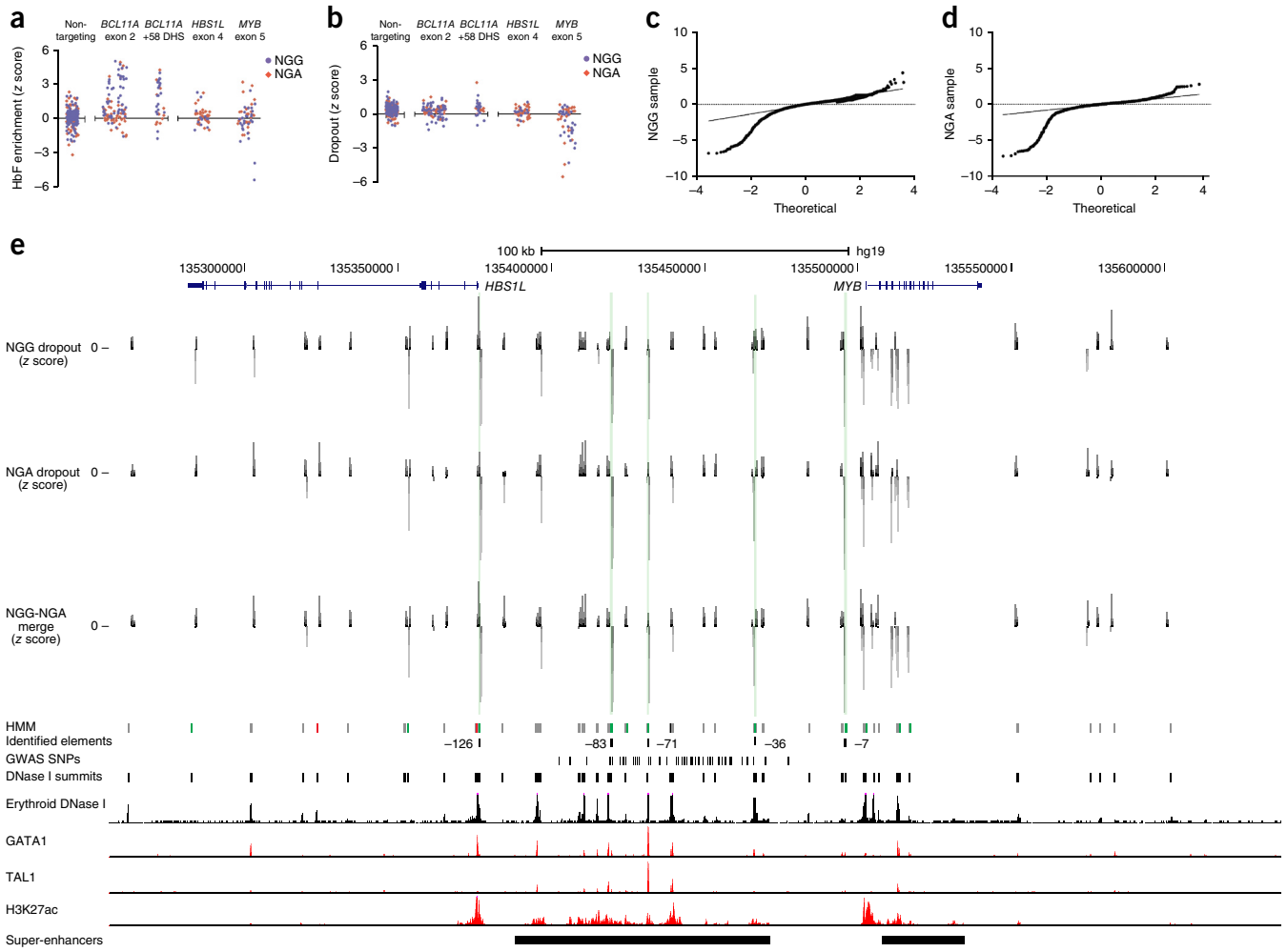


Figure 4 Mapping NGG- and NGA-restricted sgRNA dropout scores to genomic cleavage position identifies putative functional elements. (**a**) Mapping HbF enrichment scores to associated genomic loci. Nontargeting sgRNAs are pseudomapped with 5-bp spacing. (**b**) Mapping dropout scores to associated genomic loci. Nontargeting sgRNAs are pseudomapped with 5-bp spacing. (**c**,**d**) Quantile–quantile plots of NGG (**c**) and NGA (**d**) sgRNA-library dropout scores. (**e**) Mapping NGG-restricted and NGA-restricted dropout scores to associated genomic loci identifies functional elements. The elements with the highest dropout scores, –126, –83, –71, –36, and –7, are indicated by green highlighting. Erythroid DNase I hypersensitivity, H3K27ac, GATA1 binding, and TAL1 binding are shown. HMM designations as active (green), repressive (red), and neutral (gray) are shown for each DHS. The hg19 coordinates for each DHS in the *HBS1L-MYB* interval on chr 6 are: –126 DHS, 135376369–135376770; –83 DHS, 135419396–135419797; –71 DHS, 135431355–135431756; –7 DHS, 135495667–135496068; –84 DHS, 135418448–135418849; –36 DHS, 135466090–135466491 and 135466671–135467072 (comprising two DHSs).

TAL1-binding motif within the –84 DHS and localizes to the peak of DNase I sensitivity. However, application of the HMM designated the entire DHS as neutral (**Fig. 4e** and **Supplementary Fig. 14**).

Stratification by off-target scores alters identification of functional sequences and implicates –36 and –84 DHSs

Recent studies have suggested a correlation between genomic copy number and dropout after Cas9 targeting of protein-coding sequences[35,36]. Genomic copy number was evaluated for all sgRNAs in the SpCas9 and SpCas9-VQR associated libraries. This analysis identified highly repetitive sequence within the *HBS1L-MYB* interval DHS that produced a wide distribution of the number of genomic matches for each sgRNA (**Fig. 6a**,**b** and **Supplementary Fig. 18**). shRNA-mediated knockdown of *MYB* expression demonstrated that loss of MYB decreases cellular fitness in HUDEP-2 and CD34$^+$ HSPC-derived erythroblasts (**Supplementary Fig. 11b**,**c**). This finding was further supported by sgRNAs targeting *MYB* exon 5, all of which had a single genomic match and induced dropout and a decrease in *MYB* expression (**Fig. 4b**, **Supplementary Fig. 11e–h**, **Fig. 6a**,**b**, and **Supplementary Fig. 18**). However, increased genomic matches for a given sgRNA have also been predicted to decrease cellular fitness[35,36]. Our data suggest a correlation between the number of genomic matches and dropout. However, this trend was incompletely predictive, because numerous sgRNAs with ten or more genomic matches did not result in dropout (sgRNAs with ten or more genomic matches and dropout score; R^2 = 0.076) (**Fig. 6b**). This result might reflect sgRNA-specific variation in editing or cellular responses.

Off-target scores were calculated, as previously described, except all possible 20-mers upstream of an NG motif were used, thus leading to a decrease in the overall scores as compared with published values[9,37] (**Supplementary Fig. 19**; additional details in Online Methods). Off-target scores determined through this methodology ranged from 0 to 100, with a higher score signifying fewer predicted off targets. Stratification of the library sgRNAs on the basis of

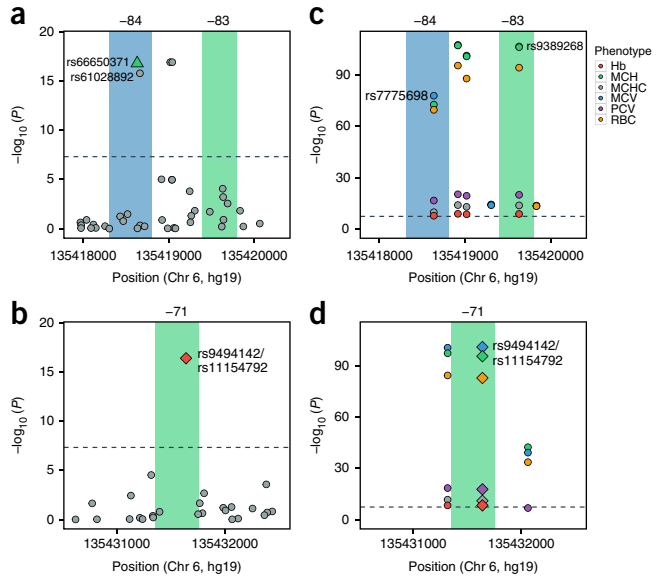


Figure 5 Trait-associated SNPs mark essential enhancer elements. (**a**) Genome-wide HbF-associated SNPs localize to the –83 and –84 DHSs. Genome-wide significant SNPs ($P < 5 \times 10^{-8}$) are indicated: rs66650371 (–84 DHS, green triangle) and rs61028892 (–84 DHS, gray circle). rs66650371 has previously been associated with altering *MYB* expression[15]. –84 DHS (chr 6,135418307–135418807, hg19) is highlighted in blue. –83 DHS (chr 6,135419396–135419797, hg19) is highlighted in green. (**b**) Genome-wide HbF-associated SNPs localize to the –71 DHS. Genome-wide significant SNPs ($P < 5 \times 10^{-8}$) are indicated: rs9494142/rs11154792 (–71 DHS, red diamond). rs9494142/rs11154792 has previously been associated with altering *MYB* expression[15]. –71 DHS (chr 6,135431355–135431756, hg19) is highlighted in green. (**c**) Genome-wide RBC-associated SNPs localize to the –83 and –84 DHSs. Genome-wide significant SNPs ($P < 5 \times 10^{-8}$) are indicated: rs7775698 (–84 DHS) and rs9389268 (–83 DHS). rs7775698 and rs9389268 are associated with all six RBC traits at genome-wide significance ($P < 5 \times 10^{-8}$). –84 DHS (chr 6,135418307–135418807, hg19) is highlighted in blue. –83 DHS (chr 6,135419396–135419797, hg19) is highlighted in green. (**d**) Genome-wide RBC-associated SNPs localize to the –71 DHS. Genome-wide significant SNPs ($P < 5 \times 10^{-8}$) are indicated: rs9494142/rs11154792 (–71 DHS, red diamond). rs9494142/rs11154792 is associated with all six RBC traits at genome-wide significance ($P < 5 \times 10^{-8}$). rs9494142/rs11154792 has previously been associated with altering *MYB* expression[15]. –71 DHS (chr 6,135431355–135431756, hg19) is highlighted in green. Hb, hemoglobin; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume; PCV, packed-cell volume; RBC, red-blood-cell count.

off-target scores >10 abolished the dropout signal from the –71 and –7 DHSs; however, the signal was retained at the –126, –83, and –36 DHSs (**Supplementary Fig. 20**). To validate the filtered screen data, we focused on the –36 DHS site, because it had lower off-target potential than did the –83 and –126 sites (**Supplementary Figs. 12–15**). We used sgRNA 1910, which had the maximal off-target score (indicative of lower off-target potential) in the –36 region (**Supplementary Fig. 15a** and **Supplementary Table 6**). Editing with sgRNA 1910 resulted in lower *MYB* expression and decreased proliferation in HUDEP-2 cells (**Fig. 6c**,**d**), in agreement with *MYB* regulatory potential within the –36 DHS. sgRNA 1910 did not overlie a predicted GATA1-binding motif (**Supplementary Fig. 15**), and its target sequence lacked GATA1 binding, as determined by chromatin immunoprecipitation (ChIP)–qPCR (data not shown).

In addition, we sought to evaluate the sequences flanking the previously implicated SNPs in the –71 and –84 DHSs[15,38]. We used an NGG-restricted guide targeting the –71 DHS (sgRNA 1582) that produced a DSB directly adjacent to the rs9494142/rs11154792/GATA1 motif (**Supplementary Fig. 21**). Targeting this motif in CD34$^+$ HSPC-derived erythroblasts resulted in successful mutagenesis (**Supplementary Fig. 21a–d**) but did not alter cellular proliferation or *MYB* expression (**Supplementary Fig. 21e**,**f**).

An NGA-restricted guide targeting the –84 DHS (sgRNA 1500) was used with a DSB position 1 bp from the implicated TAL1-binding motif[15]. Targeting this motif in CD34$^+$ HSPC-derived erythroblasts resulted in moderate levels of editing (**Supplementary Fig. 22a–d**) but did not alter cellular proliferation (**Supplementary Fig. 22e**). *MYB* expression trended toward a decrease; however, this effect did not reach statistical significance (**Supplementary Fig. 22f**). Notably, sgRNA 1500 resulted in a predominance of indels sparing the adjacent TAL1- and GATA1-binding motifs (**Supplementary Fig. 22c**,**d**). It is possible that selection against alleles disrupting key binding sites may have limited overall functional effects.

Finally, we used an NGG-restricted sgRNA (sgRNA 1321) with a DSB position within the –84 DHS directly adjacent to a GATA1-binding motif. In addition, the DSB position was 3 bp upstream of rs61028892 (seventh-highest association with HbF levels from the HbF meta-analysis) (**Supplementary Fig. 9**); this sgRNA demonstrated significant dropout in the saturating-mutagenesis screen (**Supplementary Figs. 14** and **22**, and **Supplementary Table 6**). Notably, this GATA1-binding motif corresponds to the peak of GATA1 binding at this DHS and is 14 bp downstream from the previously implicated TAL1- and GATA1-binding motifs (**Supplementary Figs. 9**, **14** and **22**). Targeting of this motif resulted in downregulation of *MYB* expression and decreased proliferation in HUDEP-2 cells (**Fig. 6c**,**d**). Furthermore, mutagenesis resulted in a decrease in GATA1 binding in HUDEP-2 cells, as determined by ChIP–qPCR (**Fig. 6e**). Together, these data suggest MYB regulatory potential in the –84 DHS mediated by GATA1 and also demonstrate the utility of multiple species of Cas9, thus allowing for more precise mutagenesis of motifs and putative causal variants. The lack of identification of –84 DHS in the screen may suggest that this element has a modest effect on MYB expression or a narrow region of regulatory DNA, which would require an even higher density of colocalizing dropout sgRNAs for detection by HMM analysis.

Putative *MYB* enhancer activity of –126, –83, –71, and –7 DHSs confounded by off-target effects

The saturating-mutagenesis screen suggested that –126, –83, –71, and –7 may potentially contain functional sequence. HMM segmentation further identified subregions within these four DHSs with dropout scores significantly diverging from the baseline, thus suggesting potential discrete minimal active sequences (**Fig. 4e** and **Supplementary Figs. 12** and **13**). All four of these DHSs contain repetitive sequence (**Supplementary Figs. 12**, **13** and **20**). We chose individual sgRNAs targeting –126, –83, –71, and –7, which exhibited the most significant dropout but also had poor off-target scores (sgRNA 0841 in –126, sgRNA 1449 in –83, sgRNA 5093 in –71, and sgRNA 2281 in –7). A set of negative-control sgRNAs (sgRNA 5430 at DHS –49, and *HBS1L*-targeting and *BCL11A*-targeting sgRNAs) were also included.

HUDEP-2 and CD34$^+$ HSPCs were transduced with CRISPR–Cas9 components and subjected to erythroid differentiation conditions. Targeting the –126, –83, –71, and –7 DHSs led to a severe proliferation defect in HUDEP-2 cells (**Supplementary Fig. 23a**). Similarly, a cellular proliferation defect was observed in the CD34$^+$ HSPC-derived erythroblasts (**Supplementary Fig. 23b**). Targeting *MYB* coding sequence had an intermediate phenotype. Targeting *HBS1L* and *BCL11A* coding sequence, –84 DHS (1329), -49 DHS (5430), and –71

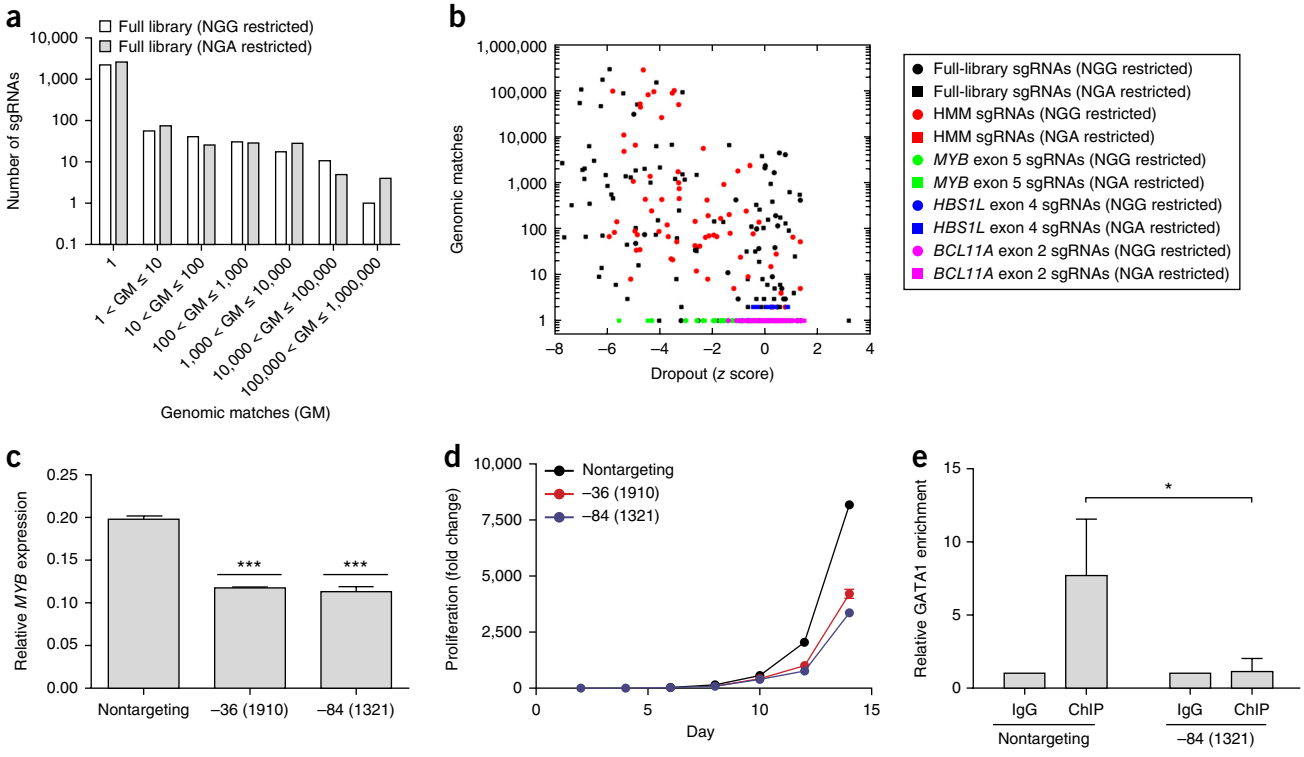


Figure 6 The *HBS1L-MYB* intergenic region contains highly repetitive genomic sequences. (**a**) Histogram of the number of genomic matches for each sgRNA in the full library. (**b**) Correlation between the number of genomic matches and dropout score. HMM sgRNA (red) indicates sgRNAs located in regions designated as active by HMM analysis. (**c**) *MYB* expression in HUDEP-2 cells after 14 d of culture (normalized to *GAPDH* expression). (**d**) Proliferation rates of HUDEP-2 cells with sgRNAs targeting *MYB* enhancer elements. (**e**) GATA1 binding in HUDEP-2 cells, determined by ChIP–qPCR after 6 d of culture. Error bars, s.d. ($n = 3$ independent experiments). Samples were compared with unpaired two-sided *t*-tests. $^{*}P < 0.01$; $^{**}P < 0.001$; $^{***}P < 0.0001$.

DHS (1582) had no effect on cellular proliferation (**Supplementary Fig. 23b**). After 18 d of erythroid differentiation, *MYB* levels were significantly decreased after targeting of the four enhancer elements and *MYB* coding sequence, in agreement with the observed cellular proliferation defects (**Supplementary Fig. 23c**). *HBS1L* expression levels were unchanged (**Supplementary Fig. 23d**). A moderate differentiation block was also observed after targeting of the −126, −83, −71, and −7 DHSs (**Supplementary Fig. 23e**).

Decreased *MYB* expression after targeting the sequences within the −126, −83, −71, and −7 DHSs, as implicated by the saturating-mutagenesis screen, suggested that these regions may contain *MYB* enhancer activity; however, these results were confounded by the increased off-target cleavage potential caused by the repetitive sequences. Therefore, the importance of these regions remains unclear. Current genome-editing technology has limited ability to unambiguously target a single site when an sgRNA has multiple genomic matches.

DISCUSSION

The functional sequences responsible for most GWAS-identified trait associations have remained unclear, owing to the paucity of methods to interrogate the function of noncoding sequences in a high-throughput manner. Comprehensive mutagenesis by HDR, introducing every possible base within a segment, may be the most stringent test of the functional effects of individual variants[39]; however, this approach is limited by throughput and efficiency. We propose that high-resolution, variant-informed, CRISPR-based saturating mutagenesis is a powerful tool with which to investigate variant-decorated regulatory DNA. Notably, previous studies of the *HBS1L-MYB* intergenic region associated with the HbF level and other erythroid traits have focused on two functional regions, −71 and −84 (ref. 15). Our approach allowed for high-resolution functional mapping of all DHSs in an ~300-kb locus, which identified multiple putative functional regions. This analysis suggested *MYB* enhancer function in the previously known −84 DHS and identified a novel *MYB* enhancer at −36. Furthermore, we identified potential function for the −7, −71, −83, and −126 elements. Our data confirmed the genetic association of the −84 DHS region with *MYB* expression levels and suggested rs61028892 as a potential causal variant.

Intriguingly, the screen identified the −71 DHS as a site for potential *MYB* enhancer activity. Notably, mutagenesis of the GATA1-binding motif modified by the genetically implicated rs9494142/rs11154792 did not alter *MYB* expression. However, although its importance remains unclear, the identified repetitive region in proximity to rs9494142/rs11154792 may be essential for *MYB* regulation in this region. Our data identifying repetitive elements in proximity to genetically implicated variants suggest that the unique context of a repetitive sequence may influence its function.

This work highlights the challenge posed by repetitive sequences present in noncoding regions. Experimental methods to circumvent the issue of targeting a repetitive sequence are limited. One possibility is to engender deletion of an entire repetitive region; however, this approach has the drawbacks of low throughput and low resolution. Our results suggest that genomic match and off-target analysis should be considered in execution of noncoding dropout screens, to rule out off-target cleavages as a source of cellular toxicity. In addition, it may be important to consider that SNPs present in cell lines used for study may create novel

off-target genomic matches[40]. Our data suggest that thorough off-target analysis can decrease ambiguity and allow for reliable assignment of regulatory potential, even in the setting of repetitive regions.

We created *DNA Striker* to streamline the design of variant-aware saturating-mutagenesis libraries by using single or multiple nucleases and present a computational algorithm to calculate off-target scores for these sgRNA libraries. Together, our data establish a methodology for high-resolution, variant-informed, off-target-aware, saturating mutagenesis as a powerful and high-throughput approach for identification of functional sequences at disease- and trait-associated regulatory DNA.

URLs. *DNA Striker*, https://github.com/mcanver/DNA-Striker/; CRISPR Off-Target Tool, http://www.mhi-humangenetics.org/en/resources; 1000 Genomes Project, http://www.internationalgenome.org/; R Statistical Computing and Graphics, https://cran.r-project.org/; CRISPResso, http://crispresso.rocks/; MATLAB, https://www.mathworks.com/; PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/; Minimac3, http://genome.sph.umich.edu/wiki/Minimac3; Raremetals, http://genome.sph.umich.edu/wiki/RareMETALS; RVtests, http://genome.sph.umich.edu/wiki/RvTests/; Off-target formula, http://crispr.mit.edu/about.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank Z. Herbert, M. Berkeley, and M. Vangala (Dana-Farber Cancer Institute Molecular Biology Core Facility) for sequencing, F. Lu at the HHMI Sequencing facility, and members at the Hematologic Neoplasia Flow Cytometry and the Flow Cytometry Core facilities at the Dana-Farber Cancer Institute for cell-sorting. We also thank J. Doench, M. Haeussler, J.-P. Concordet, R. Barretto, V. Sankaran, and J. Xu for helpful discussions. M.C.C. is supported by a National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) award (F30DK103359-01A1). L.P. is supported by a National Human Genome Research Institute (NHGRI) Career Development Award (K99HG008399). S.L. is funded by a Canadian Institutes of Health Research Banting Doctoral Scholarship. E.N.S. is supported by a Hematology Opportunities for the Next Generation of Research Scientists (HONORS) Award from the American Society of Hematology. G.C.Y. is supported by awards from the National Heart, Lung, and Blood Institute (NHLBI) (R01HL119099). G.L. is funded by the Canada Research Program, the Montreal Heart Institute Foundation, and the Canadian Institute of Health Research (MOP123382). A portion of the DNA genotyping was performed as part of the Biogen Sickle Cell Disease Consortium. D.E.B. is supported by NIDDK (K08DK093705, R03DK109232), NHLBI (DP2OD022716), the Burroughs Wellcome Fund, a Doris Duke Charitable Foundation Innovations in Clinical Research Award, an ASH Scholar Award, a Charles H. Hood Foundation Child Health Research Award, and a Cooley's Anemia Foundation Fellowship. S.H.O. is supported by an award from the NHLBI (P01HL032262) and an award from the NIDDK (P30DK049216, Center of Excellence in Molecular Hematology).

AUTHOR CONTRIBUTIONS

M.C.C., D.E.B., and S.H.O. conceived this study. M.C.C. developed the *DNA Striker* computational tool and performed computational analysis of degrees of PAM saturation. M.C.C., Y.W., E.N.S., A.J.N., D.D.C., P.P.D., M.A.C., and J.Z. performed the experiments. S.L., Y.I., F.G., C.B., A.K., C.M., M.R., and G.L. performed the genotyping and genetic analysis. R.K. and Y.N. provided the HUDEP-2 cell line. M.C.C., S.L., Y.I., L.P., G.-C.Y., and G.L. performed computational and statistical analysis. D.E.B. and S.H.O. supervised this work. M.C.C., D.E.B., and S.H.O. wrote the manuscript with input from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
2. Bauer, D.E. *et al.* An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science* **342**, 253–257 (2013).
3. Canver, M.C. *et al. BCL11A* enhancer dissection by Cas9-mediated *in situ* saturating mutagenesis. *Nature* **527**, 192–197 (2015).
4. Canver, M.C. *et al.* Characterization of genomic deletion efficiency mediated by clustered regularly interspaced palindromic repeats (CRISPR)/Cas9 nuclease system in mammalian cells. *J. Biol. Chem.* **289**, 21312–21324 (2014).
5. Corradin, O. *et al.* Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.* **24**, 1–13 (2014).
6. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
7. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
8. Ran, F.A. *et al.* Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380–1389 (2013).
9. Hsu, P.D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).
10. Uda, M. *et al.* Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. *Proc. Natl. Acad. Sci. USA* **105**, 1620–1625 (2008).
11. Lettre, G. *et al.* DNA polymorphisms at the BCL11A, HBS1L-MYB, and beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. *Proc. Natl. Acad. Sci. USA* **105**, 11869–11874 (2008).
12. Thein, S.L. *et al.* Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults. *Proc. Natl. Acad. Sci. USA* **104**, 11346–11351 (2007).
13. Galarneau, G. *et al.* Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat. Genet.* **42**, 1049–1051 (2010).
14. Farrell, J.J. *et al.* A 3-bp deletion in the HBS1L-MYB intergenic region on chromosome 6q23 is associated with HbF expression. *Blood* **117**, 4935–4945 (2011).
15. Stadhouders, R. *et al.* HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers. *J. Clin. Invest.* **124**, 1699–1710 (2014).
16. Mtatiro, S.N. *et al.* Genome wide association study of fetal hemoglobin in sickle cell anemia in Tanzania. *PLoS One* **9**, e111464 (2014).
17. Bae, H.T. *et al.* Meta-analysis of 2040 sickle cell anemia patients: BCL11A and HBS1L-MYB are the major modifiers of HbF in African Americans. *Blood* **120**, 1961–1962 (2012).
18. Ganesh, S.K. *et al.* Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nat. Genet.* **41**, 1191–1198 (2009).
19. Soranzo, N. *et al.* A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat. Genet.* **41**, 1182–1190 (2009).
20. Kamatani, Y. *et al.* Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat. Genet.* **42**, 210–215 (2010).
21. Menzel, S., Garner, C., Rooks, H., Spector, T.D. & Thein, S.L. HbA2 levels in normal adults are influenced by two distinct genetic mechanisms. *Br. J. Haematol.* **160**, 101–105 (2013).
22. van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* **492**, 369–375 (2012).
23. Chen, Z. *et al.* Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network. *Hum. Mol. Genet.* **22**, 2529–2538 (2013).
24. Esvelt, K.M. *et al.* Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat. Methods* **10**, 1116–1121 (2013).
25. Mali, P., Esvelt, K.M. & Church, G.M. Cas9 as a versatile tool for engineering biology. *Nat. Methods* **10**, 957–963 (2013).
26. Ran, F.A. *et al. In vivo* genome editing using *Staphylococcus aureus* Cas9. *Nature* **520**, 186–191 (2015).
27. Kleinstiver, B.P. *et al.* Broadening the targeting range of *Staphylococcus aureus* CRISPR-Cas9 by modifying PAM recognition. *Nat. Biotechnol.* **33**, 1293–1298 (2015).
28. Kleinstiver, B.P. *et al.* Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481–485 (2015).
29. Zetsche, B. *et al.* Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**, 759–771 (2015).
30. Doench, J.G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
31. Kurita, R. *et al.* Establishment of immortalized human erythroid progenitor cell lines able to produce enucleated red blood cells. *PLoS One* **8**, e59890 (2013).
32. Canver, M.C. & Orkin, S.H. Customizing the genome as therapy for the β-hemoglobinopathies. *Blood* **127**, 2536–2545 (2016).
33. Sankaran, V.G. *et al.* MicroRNA-15a and -16-1 act via MYB to elevate fetal hemoglobin expression in human trisomy 13. *Proc. Natl. Acad. Sci. USA* **108**, 1519–1524 (2011).
34. Rajagopal, N. *et al.* High-throughput mapping of regulatory DNA. *Nat. Biotechnol.* **34**, 167–174 (2016).

35. Munoz, D.M. *et al.* CRISPR screens provide a comprehensive assessment of cancer vulnerabilities but generate false-positive hits for highly amplified genomic regions. *Cancer Discov.* **6**, 900–913 (2016).
36. Aguirre, A.J. *et al.* Genomic copy number dictates a gene-independent cell response to CRISPR-Cas9 targeting. *Cancer Discov.* **6**, 914–929 (2016).
37. Sanjana, N.E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).
38. Pinello, L. *et al.* Analyzing CRISPR genome-editing experiments with CRISPResso. *Nat. Biotechnol.* **34**, 695–697 (2016).
39. Findlay, G.M., Boyle, E.A., Hause, R.J., Klein, J.C. & Shendure, J. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* **513**, 120–123 (2014).
40. Yang, L. *et al.* Targeted and genome-wide sequencing reveal single nucleotide variations impacting specificity of Cas9 in human stem cells. *Nat. Commun.* **5**, 5507 (2014).

ONLINE METHODS

No statistical methods were used to predetermine sample size.

HUDEP-2 cell culture. HUDEP-2 cells were used as previously described[3,31] and tested negative for *Mycoplasma* contamination. HUDEP-2 cells were expanded in SFEM (Stem Cell Technologies) supplemented with 100 ng/mL stem-cell factor (R&D), 3 IU/mL erythropoietin (Amgen), 10^{-6} M dexamethasone (Sigma), 1 μg/mL doxycycline (Sigma), and 2% penicillin/streptomycin (Thermo Fisher). HUDEP-2 cells were differentiated in Iscove's modified Dulbecco's medium (IMDM) supplemented with 330 μg/mL holo-human transferrin (Sigma), 10 μg/mL recombinant human insulin (Sigma), 2 IU/mL heparin (Sigma), 5% solvent/detergent-treated pooled human AB plasma (Rhode Island Blood Center), 3 IU/mL erythropoietin (Amgen), 100 ng/mL human stem-cell factor (SCF) (R&D), 1 μg/mL doxycycline (Sigma), 1% L-glutamine (Life Technologies), and 2% penicillin/streptomycin (Life Technologies).

HUDEP-2 SpCas9 and HUDEP-2 SpCas9-VQR cells. NGG Cas9 lentivirus was prepared as described below, with LentiCas9-Blasticidin plasmid (Addgene plasmid no. 52962). Cells were transduced with LentiCas9-Blasticidin lentivirus and maintained with 10 μg/mL blasticidin (Sigma). The LentiCas9-Blasticidin plasmid was modified to include the VQR mutations, as described in Kleinstiver *et al.*[28] (Addgene plasmid no. 87155). SpCas9-VQR lentivirus was prepared as described below by using VQR-modified LentiCas9-Blasticidin plasmid. Cells were transduced with VQR-modified LentiCas9-Blasticidin and maintained with 10 μg/mL blasticidin (Sigma).

SpCas9 and SpCas9-VQR Cas9-activity reporters. To assess Cas9 activity, lentiviral reporters were used that included a green fluorescent protein sequence (*GFP*) and either an NGG-restricted or NGA-restricted sgRNA targeting the *GFP* sequence. The NGG Cas9-activity reporter has previously been described[41] (**Supplementary Table 8**). To construct an NGA Cas9-activity reporter, pLentiGuide-Puromycin (Addgene plasmid no. 52963) was modified to express *GFP* and an NGA-restricted sgRNA targeting the *GFP* sequence (Addgene plasmid no. 87156; **Supplementary Table 8**).

Lentivirus production. HEK293T cells were cultured with Dulbecco's modified Eagle's medium (DMEM) (Life Technologies) supplemented with 10% FBS (Omega Scientific) and 2% penicillin/streptomycin (Life Technologies). HEK29T were transfected at 80% confluence in 15-cm tissue-culture Petri dishes with 16.25 μg psPAX2, 8.75 μg VSV-G, and 25 μg of the lentiviral construct plasmid of interest, with 150 μg of branched polyethylenimine (Sigma). Medium was refreshed 16–24 h after transfection. Lentiviral supernatant was collected at 48 and 72 h after transfection. Viral supernatants were concentrated by ultracentrifugation (24,000 r.p.m. for 2 h at 4 °C; Beckman Coulter SW 32 Ti rotor).

Design of nontargeting sgRNAs and calculation of off-target scores. To design sgRNAs that do not target the human (hg19) and mouse (mm9) genomes, we first extracted all possible 20-bp sequences immediately preceding the NG PAM motifs in both genomes. We created 5,000 random 20-base sgRNA sequences, which we compared with all 20-bp reference sequences. We calculated a targeting score dependent on the number and position of mismatches between both sequences, by using the methodology of Sanjana *et al.*[37]. The score ranged from 0 (nontargeting) to 1 (perfect match). We assigned a score of 0 to sequences with more than four mismatches. Reference sequences with scores >0 were considered to be potential off targets. For each random guide, we derived an aggregated score from all possible off targets, per Sanjana *et al.*[37]:

$$S_{\text{guide}} = \frac{100}{100 + \sum_{i=0}^{n} S_{\text{hit}}(h_i)}$$

where n is the number of potential off-target 'hits', and $S_{\text{hit}}(h_i)$ is the targeting score of the possible off-target sequence h_i. In this situation, an aggregated score of 100 corresponds to no possible targets in the genome. Multiple off targets or the presence of h_i-scoring off targets lowers the score toward 0. We defined guides with an aggregated score >90 as being nontargeting (n = 128). This formula was also applied to all sgRNAs in both NGG- and NGA-restricted libraries to calculate a predicted off-target score. This procedure produced scores between 0 and 100, and a higher score indicated a decreased probability of off-target effects. This tool is publically available for download.

Design of a pooled CRISPR–Cas9 library for high-resolution variant-informed functional mapping of the *HBS1L-MYB* intergenic region. The summit of every DNase I–hypersensitive site (DHS) within the *HBS1L-MYB* region (n = 98) was identified from fetal- and adult-derived $CD34^+$ HSPCs subjected to erythroid differentiation[2]. The regions of DHS summits ±200 bp were chosen for saturating mutagenesis, on the basis of previous work suggesting that functional sequence tends to be located within 200 bp of the peak of DNase I hypersensitivity[3]. Using the *DNA Striker* tool, we identified every 20-mer sequence upstream of an NGG or NGA PAM sequence on the sense or antisense strand for each *HBS1L-MYB*-region DHS as well as *BCL11A* exon 2, the core of the +58 DHS within the *BCL11A* enhancer[3], *HBS1L* exon 4, and *MYB* exon 5 (**Fig. 3c**; **Supplementary Tables 6** and **7**). Phased variants within these regions were taken from the 1000 Genomes Project database in VCF file format, including all individuals available in August 2015 (2,504 individuals; 5,008 alleles)[42]. For the 1000 Genomes variants, the variants feature within *DNA Striker* was used to identify sgRNAs altered by variants or new sgRNAs resulting from PAM sequences created by variants. Variant-associated sgRNAs were included in the library if they were present at a frequency ('guide frequency') ≥1% (**Supplementary Fig. 10a,b**). Guide frequency was used as a surrogate for variant frequency. After nonunique sgRNAs were filtered out, the NGG library comprised 2,166 sgRNAs targeting *HBS1L-MYB* DHS, 176 variant-associated sgRNAs, 13 sgRNAs targeting *HBS1L* exon 4, 28 sgRNAs targeting *MYB* exon 5, 21 sgRNAs targeting the *BCL11A* enhancer +58 DHS core, 53 sgRNAs targeting *BCL11A* exon 2, and 128 nontargeting sgRNAs, for a total of 2,585 sgRNAs. After filtering of nonunique sgRNAs, the NGA library contained 2,524 sgRNAs targeting *HBS1L-MYB* DHS, 186 variant-associated sgRNAs, 32 sgRNAs targeting *HBS1L* exon 4, 28 sgRNAs targeting *MYB* exon 5, 12 sgRNAs targeting the *BCL11A* enhancer +58 DHS core, 47 sgRNAs targeting *BCL11A* exon 2, and 128 nontargeting sgRNAs, for a total of 2,957 sgRNAs. Each of these 20-mer oligonucleotides was synthesized as previously described[3,37,43,44] and cloned with Gibson Assembly master mix (New England BioLabs) into pLentiGuide-Puromycin (Addgene plasmid no. 52963). Plasmid libraries were deep-sequenced to confirm representation.

***DNA Striker* computational tool.** *DNA Striker* allows users to create high-resolution variant-aware saturating-mutagenesis libraries and allows for quantification of the degree of saturation and visualization of the distribution of sgRNAs across the region(s) of interest. *DNA Striker* includes support for any combination of 3′-PAM sequences, such as those used for Cas9 from various species (such as SpCas9, SaCas9, and NmCas9), or 5′-PAM sequences, such as those used for the Cpf1 nuclease[6,7,29]. Briefly, uploaded DNA sequence(s) are analyzed for all selected PAM sequences through a sliding-window approach. The sgRNA length can be customized for each PAM sequence in the library, given that the optimal sgRNA length varies for different CRISPR-associated nucleases[6,7,24,26,29]. Variant-aware sgRNA library design involves identifying sgRNAs altered by variants and novel sgRNAs resulting from PAM sequences created by the presence of variants (**Fig. 2**).

Variant analysis for WGS or a custom list of variants occurs by creating multiple versions of the sliding window: the nonvariant version, versions with each variant in the window inserted in isolation (and all combinations of up to three variants in each window for custom variant lists). Variant analysis for haplotype data occurs by creating each individual allele present in the haplotype data provided. The output includes a list of oligonucleotides for full library design and two figures demonstrating the distribution of cleavages within the uploaded sequence(s) (**Supplementary Fig. 2**).

Cutting-frequency determination (CFD). CFD scores were calculated to evaluate the effects of mismatches on sgRNA activity. Published CFD scores were obtained from Doench *et al.*[30], which provides tables of CFD for all possible combinations of sgRNA and DNA single mismatches. For the calculation of >1 mismatches, single-mismatch CFD scores were multiplied together.

Pooled CRISPR–Cas9 screen for high-resolution variant-informed functional mapping of the *HBS1L-MYB* intergenic region. HUDEP-2 cells with stable SpCas9 or SpCas9-VQR Cas9 expression were transduced at low multiplicity with the corresponding NGG or NGA sgRNA-library lentivirus pool in expansion medium (NGG and NGA screens were performed independently). 10 μg/mL blasticidin (Sigma) and 1 μg/mL puromycin (Sigma) were added 24 h after transduction to select for lentiviral library integrants in cells with Cas9. The screens for fetal hemoglobin expression in HUDEP-2 cells were performed as previously described[3]. Briefly, HUDEP-2 cells were differentiated and intracellularly stained for HbF (anti-HbF-1, clone HbF-1 with APC conjugation; Life Technologies; validation available on manufacturer's website). 0.2 μg anti-HbF was used per 500,000–5 million cells. An HbF-stained non-targeting sgRNA sample was used as a negative control to set a sorting gate for the HbF-high population (approximately the top 5% of HbF-expressing cells). A corresponding percentage of cells from the HbF-low population were also sorted. After sorting into HbF-high and HbF-low pools, library preparation and deep sequencing were performed as previously described[3,45]. 6.6 μg of DNA per sample was subjected to Illumina MiSeq paired-end sequencing with Nextera sequencing primers. Guide sequences present in the HbF-high and HbF-low pools were enumerated. HbF enrichment was determined as the $\log_2$ transformation of the median number of occurrences of a particular sgRNA in the HbF-high pool divided by the median number of occurrences of the same sgRNA in the HbF-low pool across the three independent experiments for each PAM-restricted library. Dropout scores were calculated as the ratio of normalized reads in the cells at the end of the experiment (average of reads in the HbF-high and HbF-low pools) to reads in the plasmid pool for the median of the three independent experiments for each PAM-restricted library, and the data were then $\log_2$ transformed. Enrichment and dropout scores were converted to z scores by using the z-score function in MATLAB software. sgRNA sequences were mapped to the human genome (hg19). The plasmid library was deep-sequenced to confirm representation through the same methodology. A quantile–quantile (Q–Q) plot was made in MATLAB software by using the dropout scores before z-score normalization with a line fitted through the first and third quantiles.

Determination of PAM distributions. Repeat-masked regions of the human genome (hg19) were removed. Non-repeat-masked repeats were parsed out separately to avoid creating false genomic junctions. PAMs were identified, and the associated DSB site for each potential sgRNA was determined. sgRNAs with DSB positions outside of these regions were excluded from analysis. DSB positions were compiled from sgRNAs on both the plus and minus strands. The differences between adjacent genomic DSB sites were calculated. Promoters (transcriptional start site ±2 kb), exons, and introns were determined from RefSeq annotations. Enhancer and DHS sequences for GM12878, H1 hESC, HepG2, HMEC, HSMM, HUVEC, K562, NHEK, and NHLF cell lines were taken from publically available databases[46]. Repressed regions were used from previously published data[47].

Super-enhancer analysis. The ROSE algorithm was used to perform super-enhancer analysis[48].

GATA1–TAL1 chromatin immunoprecipitation sequencing (ChIP–seq) and chromatin immunoprecipitation quantitative PCR (ChIP–qPCR). ChIP–seq data were obtained from primary human erythroblasts from $CD34^+$ HSPCs subjected to erythroid differentiation conditions with anti-GATA1 (ab11852; Abcam), anti-TAL1 (clone C-21; Santa Cruz), and anti-H3K27ac (ab4729; Abcam). Antibody validation is available on the manufacturers' websites. ChIP–qPCR data were obtained from HUDEP-2 cells 6 d after lentiviral transduction with CRISPR–Cas9 reagents.

Erythroid DNase I hypersensitivity. Erythroid DNase I–hypersensitivity data were obtained from a previously published data set[2].

Analysis of transcription-factor-binding motifs. Motif analysis was performed with FIMO software to scan for putative transcription-factor-binding sites within the identified elements within the *HBS1L-MYB* intergenic region (*P* value cutoff of 10^{-4})[49]. The most recent version of the JASPAR database with hg19 sequences was used for the analysis[50].

Hidden Markov model (HMM) analysis. HMM analysis to identify repressive, active, and neutral sequences was performed as previously described[3].

Red-blood-cell trait meta-analysis. Red-blood-cell-associated SNPs were taken from a previously published meta-analysis[22]. Only SNPs with $P < 10^{-6}$ are publically available.

Genotyping of individuals with SCD. Briefly, genotyping of 1,139 African Americans from the Cooperative Study of Sickle Cell Disease (CSSCD) was performed on Illumina Human610-Quad arrays, as previously described[51]. We further genotyped 353 independent samples from the CSSCD, 57 samples from the Multicenter Study of Hydroxyurea in Sickle Cell Anemia (MSH) study, 398 samples from GENMOD, 186 from the Sickle Cell Center at Georgia Health Sciences University, and 89 from the Jamaica Sickle Cell Cohort Study (JSCCS), by using Illumina Infinium HumanOmni2.5Exome-8v1.1 arrays. We performed quality control with PLINK, removing SNPs with Hardy–Weinberg $P < 1 \times 10^{-7}$ and genotyping rate <90%. After quality control, a total of 1,083 samples with available HbF measures and genotyping success rate >99.8% remained. We conducted genotype imputation on 1000 Genomes Project (phase 3) haplotypes (version 5, hg19) with Minimac3 (v1.0.11). After imputation, both data sets contained ~47 million markers. We restricted the analysis to markers with an imputation r^2 >0.3 and falling inside the *HBS1L-MYB* intergenic region (chr 6, 135281517–135540311, hg19). In total, 2,763 markers were included in the analysis. We transformed HbF measures to z scores corrected for age and sex. We derived HbF-association *P* values independently for both data sets with RVtests (v.20140416), further correcting for the top ten principal components. We performed meta-analysis of *P* values with Raremetals (v.6.0).

Conditional analysis. Stepwise conditional analysis was performed until the top SNP had a $P < 3.15 \times 10^{-5}$. This *P* value represents the Bonferroni-corrected *P* value for the number of independent SNPs in the *MYB* region. The number of independent SNPs in the African 1000 Genomes Project data was calculated with the PLINK option --indep 200 5 2, which identified 1,587 independent SNPs from a total of 2,743 SNPs.

Deep-sequencing indel quantification and frameshift analysis. Locus-specific deep sequencing was performed through a two-PCR strategy, as previously described[3,45]. Briefly, genomic DNA was extracted with a Qiagen Blood and Tissue kit. For PCR 1, Herculase PCR reactions (Agilent) were performed with locus-specific primers that included Illumina Nextera handle sequences. The PCR reactions contained Herculase II reaction buffer (1×), forward and reverse primers (0.5 μM each), DMSO (8%), dNTPs (0.25 mM each), and Herculase II Fusion DNA polymerase (0.5 reactions), and the following PCR cycling parameters were used: 95 °C for 2 min; 20 cycles of 95 °C for 15 s, 60 °C for 20 s, 72 °C for 30 s; 72 °C for 5 min. For PCR 2, the PCR 1 reaction product was diluted (1:10) and subjected to PCR with handle-specific primers to add adaptors and indexes to each sample[3,45]. The reactions contained Herculase II reaction buffer (1×), forward and reverse primers (0.5 μM each), dNTPs (0.25 mM each), and Herculase II Fusion DNA polymerase (0.5 reactions), and the following cycling parameters were used: 95 °C for 2 min; 25 cycles of 95 °C for 15 s, 60 °C for 20 s, 72 °C for 30 s; 72 °C for 5 min. Products of the expected size from PCR 2 were gel-purified and subjected to Illumina MiSeq 150-bp paired-end sequencing. Quantification of indels and analysis of frameshift and in-frame mutations from the deep-sequencing data were performed with CRISPResso[38].

Sequencing. Sanger sequencing of the −126, −84, −83, −71, −36, and −7 DHSs identified a single variant in HUDEP-2 cells, which exhibited heterozygosity for rs144062313 in −126 DHS. rs144062313 has a minor allele frequency <1% in the 1000 Genomes Project database and hence was not included in library design.

shRNA-mediated knockdown of *MYB*. shRNA constructs cloned into the pLKO.1-puromycin lentiviral vector were acquired from the Sigma Mission shRNA library. Three shRNAs targeted against *MYB* were obtained (**Supplementary Table 9**): *MYB* shRNA 1 (TRCN0000295917), *MYB* shRNA

doi:10.1038/ng.3793

2 (TRCN0000040058), and *MYB* shRNA 3 (TRCN0000040060). A scrambled-sequence shRNA was used as a nontargeting control. Lentiviruses for each shRNA were produced as described above. MYB knockdown was confirmed by western blotting for three shRNA constructs in HEK293T cells, because MYB is not required for cellular fitness. HEK293T cells were transduced with lentiviruses for shRNA expression. Successful transductants were selected with 1 μg/mL puromycin for 24 h after lentivirus administration. Western blots were performed with anti-MYB antibody (1:1000; EP769Y; Abcam) and anti-GAPDH (1:2,000 dilution; FL-335; Santa Cruz). Validation is available on the manufacturers' websites.

Erythroid differentiation of primary human CD34⁺ hematopoietic stem and progenitor cells (HSPCs). Primary human CD34⁺ HSPCs from deidentified, healthy adult donors after G-CSF mobilization were acquired from the Center for Excellence in Molecular Hematology at the Fred Hutchinson Cancer Research Center (Seattle, Washington). These studies with anonymous, deidentified samples were conducted with IRB exemption by the Boston Children's Hospital IRB. CD34⁺ HSPCs were subjected to erythroid differentiation conditions in a three-phase culture system, as previously described[3,52]. The erythroid differentiation medium (EDM) was IMDM (CellGro) supplemented with 330 μg/mL holo-human transferrin (Sigma), 10 μg/mL recombinant human insulin (Sigma), 2 IU/mL heparin (Sigma), 5% human solvent/detergent-treated pooled human AB plasma (Rhode Island Blood Center), 3 IU/mL erythropoietin (Amgen), 1% L-glutamine (Life Technologies), and 2% penicillin/streptomycin (Life Technologies). The phase I medium consisted of EDM supplemented with 10^{-6} M hydrocortisone (Sigma), 100 ng/mL human SCF (R&D), and human IL-3 (R&D). The phase II medium consisted of EDM supplemented with 100 ng/mL SCF. The phase III medium consisted of EDM without additional supplementation. CD34⁺ HSPCs were thawed into phase I medium and were maintained in that medium for the first 7 d of culture. Cells were switched to phase II medium for days 7–11 of culture. Cells were switched to phase III medium for days 11–18 of culture.

Transduction of CD34⁺ HSPCs with CRISPR–Cas9. CD34⁺ HSPCs were thawed into phase I medium on day 0. On day 1, 10 μM prostaglandin E2 (PGE2) (Cayman Chemical) was added to culture medium in conjunction with Cas9 lentivirus (LentiCas9-Blasticidin; Addgene plasmid no. 52962). On day 2, the medium was refreshed, and 10 μM prostaglandin E2 (PGE2) (Cayman Chemical) was added to the fresh phase I culture medium in conjunction with sgRNA lentivirus (LentiGuide-Puromycin; Addgene plasmid no. 52963). On day 3, medium was refreshed, and fresh phase I medium was supplemented with 10 μg/mL blasticidin (Invivogen) and 1 μg/mL puromycin (Sigma) to select for successful transductants. Blasticidin selection persisted for 5 d, and puromycin selection persisted for 14 d.

Transduction of CD34⁺ HSPCs with shRNA. CD34⁺ HSPCs were thawed into phase I medium on day 0. On day 2, 10 μM prostaglandin E2 (PGE2) (Cayman Chemical) was added to culture medium in conjunction with shRNA lentivirus. On day 3, medium was refreshed, and fresh phase I medium was supplemented with 1 μg/mL puromycin (Sigma) to select for successful transductants. Puromycin selection continued for 14 d.

Assessment of erythroid differentiation. Success of erythroid differentiation of CD34⁺ HSPCs was assessed at three time points during the 18-d three-phase culture (days 10, 14, and 18) through staining for the transferrin receptor (anti-CD71; clone OKT9 with FITC conjugation; eBioscience) and glycophorin A (anti-CD235; clone HIR2 with PE conjugation; eBioscience). Antibody validation is available on the manufacturers' websites.

Assessment of cellular proliferation. Cell proliferation was assessed with a Countess automated cell counter (Invitrogen) with trypan-blue exclusion.

Statistical tests. Unpaired two-sided Mann–Whitney testing was used to compare dropout between SpCas9 and SpCas9-VQR species ($\alpha = 0.05$). All other statistical testing was performed with unpaired two-sided *t*-tests ($\alpha = 0.05$).

Code availability. *DNA Striker* was developed in MATLAB software. The MATLAB .m file and a stand-alone version (.exe) for *DNA Striker* are available for download along with user instructions and example input/output data sets.

Data availability. GATA1, TAL1, and H3K27ac ChIP–seq experiments are publicly available from the Gene Expression Omnibus database under accession code GSE93372.

41. Doench, J.G. *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).
42. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
43. Shalem, O., Sanjana, N.E. & Zhang, F. High-throughput functional genomics using CRISPR-Cas9. *Nat. Rev. Genet.* **16**, 299–311 (2015).
44. Chen, S. *et al.* Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *Cell* **160**, 1246–1260 (2015).
45. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
46. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
47. Pinello, L., Xu, J., Orkin, S.H. & Yuan, G.-C.Analysisof chromatin-state plasticity identifies cell-type-specific regulators of H3K27me3 patterns. *Proc. Natl. Acad. Sci. USA* **111**, E344–E353 (2014).
48. Whyte, W.A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
49. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
50. Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42**, D142–D147 (2014).
51. Solovieff, N. *et al.* Fetal hemoglobin in sickle cell anemia: genome-wide association studies suggest a regulatory region in the 5′ olfactory receptor gene cluster. *Blood* **115**, 1815–1822 (2010).
52. Giarratana, M.C. *et al.* Proof of principle for transfusion of in vitro-generated red blood cells. *Blood* **118**, 5071–5079 (2011).