Genetic regulation of fetal hemoglobin across global populations

Liam D. Cato[1,2,3,*], Rick Li[1,2,3,*], Henry Y. Lu[1,2,3,*], Fulong Yu[1,2,3], Mariel Wissman[1,2,3], Baraka S. Mkumbe[4,5,6], Supachai Ekwattanakit[7], Patrick Deelen[8,9], Liberata Mwita[10], Raphael Sangeda[4,10], Thidarat Suksangpleng[7], Suchada Riolueang[7], Paola G. Bronson[11], Dirk S. Paul[12,13], Emily Kawabata[12], William J. Astle[12,14,15,16], Francois Aguet[3], Kristin Ardlie[3], Aitzkoa Lopez de Lapuente Portilla[17,18], Guolian Kang[19], Yingze Zhang[20], Seyed Mehdi Nouraie[20], Victor R. Gordeuk[21], Mark T. Gladwin[22], Melanie E. Garrett[23], Allison Ashley-Koch[23], Marilyn J. Telen[23], Brian Custer[24,25], Shannon Kelly[24,26], Carla Luana Dinardo[27,28], Ester C. Sabino[28], Paula Loureiro[29], Anna Bárbara Carneiro-Proietti[30], Cláudia Maximo[31], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium[32], BIOS Consortium[33], Adriana Méndez[34], Angelika Hammerer-Lercher[34], Vivien A. Sheehan[35], Mitchell J. Weiss[19], Lude Franke[9,36], Björn Nilsson[3,17,18], Adam S. Butterworth[12,13,14,37,38], Vip Viprakasit[7,39], Siana Nkya[4,5,40], Vijay G. Sankaran[1,2,3,41,42, †]

* These authors contributed equally to this work

† Corresponding author, sankaran@broadinstitute.org

[1]Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, USA, [2]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA, [3]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA, [4]Sickle Cell Program, Department of Hematology and Blood Transfusion, Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania, [5]Department of Biochemistry, Muhimbili University of Health and Allied Science, Dar es Salaam, Tanzania, [6]Department of Artificial Intelligence and Innovative Medicine, Graduate School of Medicine, Tohoku University, Sendai, Japan, [7]Siriraj Thalassemia Center, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand, [8]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands, [9]Oncode Institute, Amsterdam, the Netherlands, [10]Department of Pharmaceutical Microbiology, Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania, [11]R&D Translational Biology, Biogen, Cambridge, Massachusetts, USA, [12]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK, [13]British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK, [14]National Institute for Health and Care Research Blood and Transplant Research Unit in Donor Health and Behaviour, University of Cambridge, Cambridge, UK, [15]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK, [16]NHS Blood and Transplant, Cambridge, UK, [17]Lund Stem Cell Center, Lund University, 221 84 Lund, Sweden, [18]Department of Laboratory Medicine, Lund University, 221 84 Lund, Sweden, [19]St. Jude Children's Research Hospital, Memphis, Tennessee, USA, [20]Department of Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, USA, [21]Division of Hematology and Oncology, Department of Medicine, Comprehensive Sickle Cell Center, University of Illinois at Chicago, Chicago, Illinois, USA, [22]Department of Medicine, University of Maryland School of Medicine, Baltimore, Maryland, USA, [23]Department of Medicine, Duke University Medical Center, Durham, North Carolina, USA, [24]Vitalant Research Institute, San Francisco, California, USA, [25]Department of Laboratory Medicine, UCSF, San Francisco, California, USA, [26]Division of Pediatric Hematology, UCSF Benioff Children's Hospital, Oakland, California, USA, [27]Fundacao Pro-Sangue Hemocentro de Sao Paulo, Sao Paulo, Brazil, [28]Institute of Tropical Medicine, Faculdade de Medicina da Universidade de Sao Paulo, Sao Paulo, Brazil, [29]Fundacao Hemope, Recife, Pernambuco, Brazil, [30]Fundacao Hemominas, Belo Horizonte, Brazil, [31]Fundacao Hemorio, Rio de Janeiro, Brazil, [32]The TOPMed Consortium is detailed in Supplemental Acknowledgments, [33]BIOS Consortium, [34]Institute of Laboratory Medicine, Cantonal Hospital Aarau, 5000 Aarau, Switzerland, [35]Aflac Cancer & Blood Disorders Center, Children's Healthcare of Atlanta & Department of Pediatrics, Emory University School of Medicine, Atlanta, Georgia, USA, [36]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands, [37]Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK, [38]Heart and Lung Research Institute, University of Cambridge, Cambridge, UK, [39]Department of Pediatrics, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand, [40]Tanzania Human Genetics Organisation, Tanzania, [41]Harvard Stem Cell Institute, Cambridge, Massachusetts, USA, [42]Department of Biochemistry, Muhimbili University of Health and Allied Science

1. Abstract

Human genetic variation has enabled the identification of several key regulators of fetal-to-adult hemoglobin switching, including BCL11A, resulting in therapeutic advances. However, despite the progress made, limited further insights have been obtained to provide a fuller accounting of how genetic variation contributes to the global mechanisms of fetal hemoglobin (HbF) gene regulation. Here, we have conducted a multi-ancestry genome-wide association study of 28,279 individuals from several cohorts spanning 5 continents to define the architecture of human genetic variation impacting HbF. We have identified a total of 178 conditionally independent genome-wide significant or suggestive variants across 14 genomic windows. Importantly, these new data enable us to better define the mechanisms by which HbF switching occurs *in vivo*. We conduct targeted perturbations to define BACH2 as a new genetically-nominated regulator of hemoglobin switching. We define putative causal variants and underlying mechanisms at the well-studied *BCL11A* and *HBS1L-MYB* loci, illuminating the complex variant-driven regulation present at these loci. We additionally show how rare large-effect deletions in the *HBB* locus can interact with polygenic variation to influence HbF levels. Our study paves the way for the next generation of therapies to more effectively induce HbF in sickle cell disease and β-thalassemia.

2. Introduction

During human development, there is a switch from a fetal form of hemoglobin (HbF), with the beta-subunit encoded by the *HBG1/2* genes, to the adult form of hemoglobin (HbA), with the beta-subunit encoded by the *HBB* gene, that takes place shortly after birth - a process referred to as the fetal-to-adult hemoglobin switch. Persistently increased production of HbF after infancy can ameliorate clinical symptoms in common and life-threatening disorders arising from *HBB* mutations, including sickle cell disease and β-thalassemia. While the ameliorating effect of HbF in these hemoglobin disorders have been known for several decades,[1] the underlying regulation of HbF and approaches to target this process had remained unknown. Fifteen years ago, initial genome-wide association studies (GWAS) revealed three regions of association with HbF levels - the *HBB* locus on chromosome 11, the *HBS1L-MYB* locus on chromosome 6, and the *BCL11A* locus on chromosome 2.[2–4] These studies led to functional follow up that revealed how BCL11A acted as a key and direct repressor of *HBG1/2* transcription.[5] Subsequently, additional insights have emerged from the study of rare loss-of-function mutations impacting *BCL11A*,[6–8] transcriptional regulatory elements necessary for erythroid expression of this factor,[9,10] upstream regulators of *BCL11A* expression,[11–14] and analysis of the mechanisms by which *BCL11A* alters transcription.[15–17] Suppression of *BCL11A* or its binding motif in the *HBG*1/2 promoters using genome editing or gene therapy approaches has emerged as a curative strategy for sickle cell disease and β-thalassemia.[18,19] These initial studies have also spurred further research to better define the mechanisms by which HbF is regulated in humans.[15,20,21]

Despite the substantial progress in understanding HbF regulation, much of which has largely relied upon studies in cell lines and mouse models, a number of fundamental questions about how HbF is regulated in humans *in vivo* remain unanswered. For instance, are there additional genetic variants underlying interindividual variation in HbF levels and how do these variants act? What are the identities of causal variants at known loci impacting HbF levels and how do they mechanistically function? How do common and rare variants impacting HbF levels interact to modulate HbF levels? To address these and other fundamental questions, we have performed the largest multi-ancestry genome-wide association study (GWAS) for HbF levels to date, involving 28,279 participants from a range of global populations with varied ancestries spanning five continents. Through this study, we have identified new loci impacting HbF levels and defined putative target genes/mechanisms, examined how well-studied loci can actually harbor distinct variation and mechanisms across different populations, and characterized the interface between rare large-effect mutations and polygenic variation in impacting HbF levels. These findings open the door for further insights on HbF regulation and future therapeutic advances, including improved designs for therapies inspired by insights from naturally-occurring human variation.

3. Results

3.1. HbF meta-analysis

To perform a large-scale GWAS, we included 28,279 individuals from several distinct cohorts with different ancestries (Fig. 1a, Table 1). The cohorts relied upon different selection strategies, including unselected individuals from the population (Swedish, SardiNIA,[2] INTERVAL,[22] GTEx,[23] BIOS[24]), individuals with sickle cell disease (Tanzania,[25] Walk-PHaSST,[26] OMG-SCD,[27] REDS-III Brazil[28], St. Jude Sickle Cell Clinical Research & Intervention Program (SCCRIP)/Baylor[29,30]), or individuals selected from a screened population (Thai, see Methods). Upon conducting the GWAS, there was no inflation in test statistics noted (genomic inflation factor (λ_{gc})=0.99, linkage-disequilibrium score (LDSC) intercept=0.98) (Supplementary Fig. 1). We identified 178 conditionally independent signals associated with HbF levels in 14 windows (9 windows at genome-wide significant threshold, p<5e-8, and 5 windows at the suggestive threshold, p<1e-6) (Fig. 1b, Supplementary Table 1,2). We annotated these windows with genes nominated by a combination of distance from a significant variant, long-range interaction data linking regulatory elements to genes in erythroid cells (via promoter capture Hi-C), correlations between gene expression and chromatin accessibility (RNA and ATAC-seq correlations) in hematopoietic cells, and expression quantitative trait loci (eQTLs) from whole blood (Supplementary Table 3). In addition to the previously characterized regions, we identified regions near known HbF regulators that had not previously been identified by other population-based genetic studies, including *ZBTB7A* and *KLF1*, as well as other regions that did not harbor genes previously implicated in HbF regulation.

We next conducted ancestry-specific analyses (Fig. 1c,d,e, Supplementary Table 4), including for individuals with African (AFR, n=3,963), European (EUR, n=22,882), and Thai (n=1,392) ancestry and found largely conserved association windows at the major loci identified, including at the *BCL11A*, *HBS1L-MYB*, *HBB*, and *CTC1* loci. Windows nominating *BACH2* were found to be significantly associated with HbF in the AFR and EUR populations (Supplementary Table 5,6), *PSME4* and *ABCC1* in the EUR populations alone, and *UTRN* in the Thai population alone (Supplementary Table 7). Some of the observed ancestral heterogeneity might arise from low power as a consequence of small cohort size and/or reduced genetic variation within specific populations.

3.2. SNP heritability and genetic correlations of HbF

For the whole-cohort analysis, SNP heritability estimated using the linkage disequilibrium adjusted kinships (LDAK) model was 0.164 (SD 0.015). LD score regression (LDSC) produced a similar estimate of 0.15 (SE 0.07). The EUR population summary results revealed a heritability of 0.20 (SD 0.038) (LDSC, 0.098 (SE 0.05)), AFR at 0.31 (SD 0.23) (LDSC, 0.36 (SE 0.31)), and the Thai population at 0.40 (SD 0.32) (LDSC, 0.9456 (SE 0.682)). It is important to bear in mind that while the majority of EUR populations were unselected, the AFR populations were

exclusively individuals with sickle cell disease, where a higher heritability for HbF has been inferred.[31] The Thai population was also selected with the extremes of a population distribution, which may confound these estimates. Upon analysis of heritability enrichments for different histone modifications or genomic regions, the major enrichments were seen in putative enhancer elements suggesting SNPs that reside in and potentially alter regulatory elements contribute most to the currently observed heritability in HbF (Supplementary Fig. 2a).

Given the well-powered insights from genetic analysis of blood cell phenotypes across populations,[32,33] we examined the extent to which genetic variation impacting HbF levels might also have genetic overlap with these phenotypes (Supplementary Fig. 2b, Supplementary Table 8). A number of genetic variants impacting cell phenotypes spanning the white blood cell, red cell, and platelet lineages all appeared slightly, but significantly, positively correlated with higher HbF-associated genetic variants, with the exception of mean corpuscular volume (MCV, r_g = -0.18 (SD 0.08)) and mean corpuscular hemoglobin (MCH, r_g = -0.15 (SD 0.09)) that were negatively correlated, suggesting that red blood cell size tends to be reduced (slightly) with variation that increases HbF levels, while counts of different blood cells tend to be increased. Notably, the genetic correlations for MCV and HbF match the phenotypic association seen in the Thai cohort of 1,323 individuals (rho = -0.55 ($p < 2.2e-16$)), and a more weak but consistent phenotypic correlation with mean corpuscular hemoglobin concentration (MCHC, rho = -0.098 ($p = 0.002$)).

3.3. Cellular contexts for variation associated with HbF

We next wanted to gain global insights into the cell contexts for this variation and therefore employed our recently described approach of Single Cell Analysis of Variant Enrichment through Network propagation of GEnomic data (SCAVENGE)[34] to identify relevant cell states where the fine-mapped variants showed significant co-localization with accessible chromatin across human hematopoiesis (with both bulk and single-cell assay for transpose accessible chromatin by sequencing [ATAC-seq] data). We fine-mapped each window to identify a credible set of potentially causal variants; half of the windows had 95% credible sets containing 10 or fewer variants (Supplementary Table 9), that were primarily localized to introns (Supplementary Figure 2c,d). We identified a strong enrichment in erythroid cells compared to other hematopoietic cell types (Supplementary Fig. 3a,b). At single-cell resolution, using a pseudotime projection of human erythropoiesis,[35] we found a strong enrichment at the mid-maturation of erythroid cells, peaking around the proerythroblast to basophilic erythroblast stages (Fig. 1f). Given these enrichments, we sought to define co-regulated transcription factor (TF) motifs. Spearman correlations between the SCAVENGE trait relevance score (TRS) and chromVAR TF motif enrichment scores across erythroid cells were calculated (Fig. 1g, Supplementary Table 10). Notably, KLF1 and GATA1 motifs were highlighted and both of these transcription factors are critical in HbF regulation.[36] Collectively these results highlight key differentiation stages and regulatory networks involved in HbF-associated genetic variation.

3.4. Identifying BACH2 as a genetically-nominated regulator of HbF

Having shown at a global level that much of the genetic variation impacting HbF levels mapped to the intermediate stages of human erythropoiesis and to transcriptional regulatory elements, we wondered whether new mechanistic insights could emerge from these findings. While a number of previously undescribed regions were identified through our GWAS (Fig. 1b), a notable region contained a lead variant within the gene *BACH2* (rs2325259). This was compelling, as *BACH2* encodes a transcriptional factor that can compete with NFE2 and other related proteins for binding to small Maf proteins and can thereby alter gene expression at a number of loci.[37] The complexes of NFE2 and NRF2 play a critical role in the transcriptional regulation at the β-globin genes and in HbF expression, suggesting potentially relevant mechanisms for the observed association.[38–45] By fine-mapping, we identified two putative causal variants (rs1010473 and rs1010474) in tight linkage disequilibrium with the lead variant (D`>0.98, R^2>0.97 in both EUR and AFR populations) that overlapped a region of accessible chromatin in human hematopoietic stem and progenitor cells (HSPCs), whose accessibility was rapidly lost with erythroid differentiation (Fig. 2a). We targeted this region using CRISPR/Cas9 genome editing to excise the full 0.6 kb element in primary adult human $CD34^+$ HSPCs (Fig. 2b, Supplementary Table 11). Three days following editing that excised the enhancer in ~40% of alleles (Supplementary Fig. 4), we found that the expression of *BACH2* was selectively reduced (Fig. 2c), but importantly, several other genes in the topologically-associated domain containing this regulatory element were not impacted (Supplementary Fig. 5). These findings suggested that the removal of a variant-harboring regulatory element appeared to selectively impact *BACH2*, which thereby might regulate HbF levels. To directly test this and given challenges in effectively perturbing *BACH2* by genome editing of HSPCs, we increased expression of BACH2 in HSPCs through lentiviral expression (Fig. 2d) and fluorescence activated cell sorted (FACS) the top (BACH2-GFP^{hi}) and bottom (BACH2-GFP^{lo}) 30% of GFP^+ transduced cells (Fig. 2e-f) (Supplementary Fig. 6a). By segregating cells that either had a low or high levels of GFP expression, which is linked on the same transcript to the human *BACH2* cDNA through an internal ribosomal entry site, we found a dosage-dependent repression of HbF levels as assessed by both measurement of *HBG1/2* mRNA levels (Fig. 2g) and flow cytometric assessment of cells with HbF present (Fig. 2h) (Supplementary Fig. 6b) with a concurrent increase in *HBB* mRNA levels (Supplementary Fig. 6c). These observations held true across erythroid differentiation (Supplementary Fig. 6d-e). These changes in HbF levels were accompanied by only a slight delay in differentiation that was most notable in the cells with higher BACH2 expression, as assessed by analysis of the cell surface markers CD235a and CD71, as well as by morphological assessment (Fig. 2i) (Supplementary Fig. 7-8). While further studies are needed to define underlying mechanisms for HbF regulation, these initial findings demonstrate how through our GWAS, we have defined a previously undescribed genetically-nominated factor, BACH2, that regulates HbF and which might prove to be an important therapeutic target.

3.5. Previously described associations at BCL11A and HBS1L-MYB result from multiple variants that vary across ancestries

Significant advances in our understanding of how HbF is regulated have arisen from prior genetic studies that have identified the *BCL11A* and *HBS1L-MYB* loci. However, despite progress made in understanding the function of the genes within these loci, the precise causal variants and the underlying mechanisms by which these variants act have remained unknown. Indeed, while early studies had suggested that HbF-associated variation within the *BCL11A* locus might impact an erythroid regulatory element,[9] mapping of this enhancer has suggested that the most potent elements that are necessary for gene expression within this enhancer occur within regulatory motifs that are invariant in humans.[10] Therefore, even in this well understood case, the precise variants underlying this association signal that has motivated therapeutic efforts have remained undefined. We reasoned that the increased power through our large GWAS and the availability of data across multiple ancestry groups would provide an opportunity to define causal variants in these previously identified regions.

Using conditional analyses at these loci, we identified 46 independent signals within 1 Mb of *BCL11A* in a mixed ancestry analysis (chr2:59,450,520-61,554,467) (31 significant before adjustment; then, 18 in AFR only analysis, 21 EUR only, 11 Thai only) (Fig. 3a) and 31 independent signals within 1 Mb of *HBS1L-MYB* (chr6:133,960,378-136,540,310) (21 significant before adjustment; then 14 AFR only analysis, 24 EUR only, 23 Thai only) (Fig. 3b). Remarkably, at both loci, there were few independent signals that overlapped, suggesting distinct mechanisms of variation at these loci across different ancestries (Fig. 3c, d, Supplementary Table 12, Supplementary Fig. 9). We then examined how many of these variants overlapped regions of accessible chromatin in HSPCs undergoing erythroid differentiation.[46] While a number of overlaps were noted suggesting potential alteration of transcriptional regulation at the *BCL11A* and *HBS1L-MYB* loci, there was an even further restriction of overlap across ancestries (Fig. 3c, d). Interestingly, while the one fine-mapped variant that did demonstrate overlap with accessible chromatin and across ancestries at the *BCL11A* locus was the previously reported rs1427407 polymorphism,[9] each ancestry group had a distinct set of conditionally independent variants that would collectively impact *BCL11A*, suggesting significant and previously unappreciated complexity in the genetic variation at this locus. Similar observations were also present at the *HBS1L-MYB* locus, where functional fine-mapping of causal variants has also been attempted with earlier and more limited genetic data.[47] These findings emphasize two critical concepts: (1) the signals at these loci are likely attributable to multiple independent variants that collectively contribute to the robust variation in HbF levels and (2) these loci were fortuitously identified in early studies that were conducted in different ancestry groups, but these signals likely arose from distinct variants across ancestries. An important implication of these findings is that multiplexed targeting of these loci

may be an ideal approach, which would mimic nature, for more effective HbF induction than current therapeutic approaches.[19]

3.6. Variable HbF phenotype in known rare deletions and common variant influence

While new insights have emerged from the study of population-based variation, a chasm in human genetic studies of HbF has emerged between the findings from rare variant studies that have highlighted large-effect structural and single nucleotide variants that are rarely found in individuals, and more common polygenic variation, as we identify through our GWAS. The design of the Thai cohort enabled us to assess both of these types of variation simultaneously, as this cohort was selected from extremes of a screened population of ~86,000 individuals. We found that a number of individuals with higher HbF levels harbored substantial increases that were likely due to large effect variants. Using a variety of mapping approaches (see Methods), we identified deletions in the cohort known to cause hereditary persistence of fetal hemoglobin or variant forms of thalassemia (associated with high HbF) in individuals with elevated HbF (Fig. 4a, Supplementary Table 13). While a wide distribution was seen in HbF levels across any specific deletion (Fig. 4b), we found that incorporation of polygenic variation (Supplementary Fig. 10a, b) using a phenotype score stratified the impact on HbF levels for most of the deletions we identified (Fig. 4b). Interestingly, we observed genetic interactions between the rare deletions and the common variant polygenic scores for the Thai $(\delta\beta)^0$-thal, 3.48 kb Thai $(\beta)^0$, and the negative 10 deletion set, with borderline significance for Filipino-type β^0-thal (Supplementary Fig. 11). Remarkably, these deletions that demonstrated interactions with the polygenic scores were those that maintained the region upstream of the δ-globin gene, which we have suggested might be critical for HbF silencing by BCL11A through long-range interactions (Fig. 4a).[15,48] This nexus of common and rare variation we identify for HbF illuminates a key opportunity as population studies expand in size, which is to decipher the interactions between the full allelic spectrum impacting disease-relevant phenotypes. These findings also suggest that current efforts to mimic such variation for therapeutic purposes or target key regulators of HbF without accounting for polygenic background might result in more limited or variable HbF induction than desired.

4. Discussion

Tremendous progress in our understanding of HbF regulation and switching has emerged from examining human genetic variation. However, the studies of genetic variation to date have suffered from limitations. Population-based studies of common genetic variation have been restricted in scale to several thousands of individuals at most and have typically been focused on specific ancestry groups.[25,49,50] Concomitantly, studies of rare individuals with substantially elevated HbF levels have revealed rare genetic variation at the β-globin gene locus and in other genes, including *BCL11A*, *KLF1*, and *ZBTB7A*, which leads to more considerable increases in HbF levels.[6,15,51,52] Here, by conducting the largest multi-ancestry GWAS of HbF levels to date we have uncovered new loci underlying variation in HbF levels, including the identification of BACH2 as a new genetically-nominated regulator of HbF. We have also defined how polygenic variation can interact with rare large-effect alterations to modify HbF levels in a population stratified across extremes of the HbF distribution. Importantly, the finding of interactions in some cases suggests distinct biological and mechanistic overlap between pathways involved in HbF induction, including the role of BCL11A in silencing HbF through long-range interactions,[15] which will be an important avenue for future mechanistic studies. These observations might provide guidance for combination approaches to achieve optimal therapeutic induction of HbF.

Even at extensively studied loci that are already the targets of therapeutic approaches, including the *BCL11A* locus, there is substantial complexity, with many more conditionally-independent causal variants than has been appreciated from prior smaller genetic studies. Additionally, these causal variants appear to vary by ancestry, suggesting a fortuitous mechanistic overlap resulting from distinct signals at each of the previously described loci at *BCL11A* and *HBS1L-MYB*. These findings warrant further functional dissection, particularly using systematic mapping and mutagenesis approaches that enable comprehensive interrogation of regulatory elements at high-resolution[53–55]. Our findings not only motivate further genetic and functional mapping at these well-studied loci, but also suggest that existing therapeutic approaches could be substantially improved by mimicking the multiplexed approach that nature has employed to alter HbF levels.

In summary, we have demonstrated how by studying the genetic basis of variation in HbF levels across diverse populations, we could uncover unappreciated genetic variation and new biological insights, including a role for BACH2 in regulating HbF. This highlights the importance of conducting increasingly larger genetic studies involving diverse populations. This will enable further insights into the genetic complexity of even seemingly well-understood phenotypes like HbF and the additional mechanistic insights that will emerge are likely to be even more notable in complex human diseases.

5. Figures

Fig. 1 GWAS of HbF across global populations with enrichment of fine-mapped variants in erythroid cells. | a, Population geography of included studies and associated sample numbers. † , the Thai population had a proportion of individuals selected for elevated HbF and thus is not a general population. *, cohorts that employed gene expression measurements. **b,** Combined meta-analysis of fetal hemoglobin details several unexplored loci. Gene symbols shown are the most likely impacted gene nominated using several approaches (Supplementary Table 3). Window boxes are drawn over significant and suggestive signals identified via conditional analysis (*Methods*). Colored shading represents significant windows (p<5e-8), while gray represents a suggestive signal (p<1e-6). **c, d, e,** show ancestry specific analyses conducted using MAMA (*Methods*) for African (AFR), European (EUR), and Thai ancestry backgrounds, respectively. Areas of differential signal indicate potential ancestry-specific effects on HbF, y-axis was limited to p>1e-100. **f,** SCAVENGE analysis using scATAC-seq data, within the enriched erythroid population there is particular enhancement of the trait relevance score (TRS) in the mid-late erythroid population. **g,** Spearman correlations between SCAVENGE TRS and chromVAR TF motif enrichment scores across erythroid cells to identify co-regulated transcription factor motifs.

Fig. 2 Defining BACH2 as a genetically-nominated regulator of HbF. | a, The BACH2 locus, with sentinel variant rs* shown as a purple diamond, and LD R^2 colored from red (high) to yellow (low). Two variants rs1010473 and rs1010474 display high LD (both in EUR and AFR populations) with the sentinel and are positioned in a peak of accessible chromatin in HSC cells, tracks of bulk ATAC-seq for erythroid relevant trajectories are shown below. **b,** HSC chromatin accessibility around rs1010473 and rs1010474 and the two CRISPR– Cas9 guide RNA pairs (ENH1 and ENH2) used to delete this region. sgRNA, single-guide RNA. **c,** Expression of *BACH2* transcript in bulk human primary $CD34^+$ hematopoietic stem and progenitor cells (HSPCs) three days after deletion of the *BACH2* enhancer (n=6) compared to *AAVS1* editing (n=3). ENH1 and ENH2 gRNA results were combined due to similar editing efficiencies. **d,** Schematic representation of lentivirus-mediated increased expression of *BACH2* in HSPCs. Transduced HSPCs were sorted on the top and bottom 30% of GFP^+ cells (GFP^{hi} and GFP^{lo}, respectively) and subjected to erythroid differentiation and functional evaluation. **e,** BACH2-GFP expression in FACS sorted populations across erythroid differentiation. Mean fluorescence intensities are indicated. **f,** Relative *BACH2* transcript abundance on days 7 and 13 of erythroid differentiation in transduced HSPCs. **g,** Proportion of *HBG1/2* expression relative to overall *HBG1/2*+*HBB* expression on days 7 and 13 of erythroid differentiation in transduced HSPCs. **h,** Frequency of F-cells across erythroid differentiation in transduced HSPCs quantified by intracellular staining of fetal hemoglobin (HbF). $\%HbF^+$ and HbF^- are indicated. **i,** Erythroid differentiation status of transduced HSPCs on days 6 and 10 of erythroid

differentiation culture as assessed by surface expression of CD71 and CD235a. Note in (d-h) Each data point is representative of individual transductions.

Fig. 3 Overlap of potentially causal variants at known HbF loci. | Comprehensive study of the **a,** *BCL11A* and **b,** *HBS1L-MYB* loci shows many potential ancestry-specific causal effects (diamond shaped points) overlapping with accessible chromatin at various cell stages of human erythropoiesis. These regions are linked to other regions via erythroid promoter-capture Hi-C interactions. Finemapped signals in the fixed-effects analysis are highlighted by a triangle. Significant conditionally independent SNPs found in significant ATAC peaks are highlighted by a colored circle. Colored ranges above ATAC tracks correspond to statistically significant peaks. **c, d,** Overlap between ancestry groups of independent sentinel variants, those in accessible chromatin and 95% credible set fine mapped variants in accessible chromatin are shown at the **c,** *BCL11A* and **d,** *HBS1L-MYB* loci. Specific variants are described in Supplementary Table 9.

Fig. 4 Stratification of impact on HbF levels by large-effect structural variants by polygenic variation. | a, Rare deletions identified previously in individuals of Thai ancestry from case studies are shown in relation to affected genes, on hg38 coordinates. These deletions were identified in the included Thai population using a combination of mapping approaches. **b,** Within each known deletion category, individuals carrying these deletions show variable effects on HbF (%) levels and polygenic trait scores (PRS) derived from common single nucleotide variation, low and high were determined by less than or greater than median global PRS value, respectively.

6. Supplementary Figures

Supplementary Figure 1. QQ-plots for meta-analysis and ancestry-specific analyses from MAMA.

Supplementary Figure 2. a, Heritability enrichments from LDAK for 64 functional categories in the BLD model. **b,** Genetic correlations between HbF with various red and white blood cell parameters. **c,** Conditionally independent analysis revealed a number of potential lead variants per locus, and after fine-mapping, 95% credible sets are shown. **d,** Functional consequences of the fine-mapped variants are shown.

Supplementary Figure 3. a, UMAP projection SCAVENGE cell-stage enrichment results from single cell ATAC-seq data. **b,** Data in bulk.

Supplementary Figure 4. Enhancer perturbations were quantified by qPCR and shown as a percentage of total alleles in the bulk population. Plotted are wild-type alleles, inversions, and deletions for the *AAVS1* negative control and both enhancer deletion pairs.

Supplementary Figure 5. *MDN1, CASP8AP2, and MAP3K7* transcript abundance normalized to *ACTB* in the *AAVS1* negative control and enhancer deletion samples.

Supplementary Figure 6. a-e, Lentivirus increased expression of *BACH2* in primary human $CD34^+$ hematopoietic stem and progenitor cells sorted on the top (hi) and bottom (lo) 30% of GFP^+ cells subjected to erythroid differentiation. Each point represents an independent transduction. **a,** *BACH2* **b,** *HBG1/2* and **c,** *HBB* transcript abundance across erythroid differentiation in each sorted population. **d,** % *HBB* and *HBG1/2* transcripts across erythroid differentiation. **e,** Frequency of F-cells detected across erythroid differentiation.

Supplementary Figure 7. a-b, Lentivirus increased expression of *BACH2* in primary human $CD34^+$ hematopoietic stem and progenitor cells sorted on the top (hi) and bottom (lo) 30% of GFP^+ cells subjected to erythroid differentiation. Each point represents an independent transduction. **a-b,** Frequency of $CD71^-CD235a^-$, $CD71^+CD235a^-$, $CD71^+CD235a^+$, $CD71^-CD235a^+$ cells on **a,** day 6 and **b,** day 10 of erythroid differentiation. **a-b,** Each point represents an independent transduction.

Supplementary Figure 8. Lentivirus increased expression of *BACH2* in primary human $CD34^+$ hematopoietic stem and progenitor cells sorted on the top (hi) and bottom (lo) 30% of GFP^+ cells subjected to erythroid differentiation. Cells were harvested on day 11 of erythroid differentiation culture and prepared by cytospin. Cell morphology was assessed by May

Grunwald-Giemsa staining and imaging on a Nikon II Eclipse E800 microscope at 60x magnification. Shown is representative of three independent transductions across 30 fields.

Supplementary Figure 9. A zoomed locus plot of part of BCL11A intron 2 region from Figure 3a. Showing the well known BCL11A variant rs1427407, and previously described DHS sites at +55, +58 and +62 kBp from the TSS[9] in the context of ancestry and other potential causative loci. Diamond shaped points show ancestry based conditionally independent signals. Erythroid promoter-capture Hi-C interactions are shown, followed by ATAC-seq data for erythroid lineages. Finemapped signals in the fixed-effects analysis are highlighted by a triangle. Significant conditionally independent SNPs found in significant ATAC peaks are highlighted by a colored circle. Colored ranges above ATAC tracks correspond to statistically significant peaks.

Supplementary Figure 10. **a,** Calculated polygenic risk scores (PRS) perform well in HbF discrimination in a test set **b,** PRS shows spread of distribution by deletion.

Supplementary Figure 11. Genetic interactions between common variant polygenic risk score (PRS) and deletions on normalized HbF in Thai population in a Generalized Additive Model, corrected for age, sex and principal components of ancestry. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

7. Tables

Table 1. Details of the included populations in HbF meta-analysis.

Supplementary Table 1. Windows created based on whole-cohort meta-analysis conditionally independent signals.

Supplementary Table 2. Conditionally-independent identified signals from whole-cohort meta-analysis.

Supplementary Table 3. Basis for gene nominations for windows.

Supplementary Table 4. Signal windows across all populations.

Supplementary Table 5. Conditionally-independent identified signals from AFR population MAMA meta-analysis.

Supplementary Table 6. Conditionally-independent identified signals from EUR population MAMA meta-analysis.

Supplementary Table 7. Conditionally-independent identified signals from THAI population MAMA meta-analysis.

Supplementary Table 8. Results of genetic correlation analysis.

Supplementary Table 9. Fine-mapped posterior probabilities from FINEMAP (see Methods) for fixed-effects meta-analysis results.

Supplementary Table 10. Trait Relevance Score (TRS) and Transcription Factor Motif correlation on the Erythroid trajectory of SCAVENGE.

Supplementary Table 11. Deletion sgRNA sequences.

Supplementary Table 12. Variant overlap in erythroid lineage accessible chromatin at known loci.

Supplementary Table 13. Definitions of rare deletions in thai ancestry populations, coordinates are in hg38.

Supplementary Table 14. Enhancer deletion qPCR primers.

Supplementary Table 15. Gene expression qPCR primers.

8. Methods

8.1. Individual GWAS study methods and quality control

An overview of the included studies is in Table 1. Most included samples had HbF measured in the traditional way using high performance liquid chromatography, however two of the included cohorts (BIOS and GTEx) derived the HbF phenotype from expression data (in TPM units) as a ratio of gene expression (HBG1 + HBG2) / HBB. We found this to faithfully replicate expected results from the traditionally measured HbF cohorts. In addition, we found the ratio approach to traditionally measured approach in a EUR subset to be genetically correlated r_g = 0.59 (SE 0.317). Selected studies were included from the BIOS; LifeLines DEEP (LL), The Leiden Longevity Study (LLS_660Q), Netherlands Twin Register (NTR), PAN, The Rotterdam Study (RS). The included Swedish and Thai populations are previously undescribed cohorts, and were analyzed specifically for this study.

All GWAS summary statistics were lifted-over from their respective genome builds to reference genome hg38. Alleles were flipped according to the hg38 build reference allele, and if neither allele was present the variant was removed. Strand ambiguous and non-biallelic SNPs were removed. Minor allele frequency was filtered to >=0.1%. RSIDs were assigned using dbSNP version 144. All models included adjustment for at least, age, sex and top 10 principal components of ancestry. In addition, for the SCD cohort analyzes there was appropriate adjustment for SCD genotype and hydroxyurea use. In all cohorts, the same transformation (inverse-normalization) was performed on HbF to produce a normalized response variable.

The Thai cohort was a unique study design, because from a large general population we intentionally sampled individuals with HbF>2% for array genotyping in order to gain maximal power and to identify individuals suitable for WGS with the intent of elucidation of structural variation. Low HbF control samples were also included and the resultant HbF was transformed using an inverse normalization transformation and checked for normality to represent a normal distribution before GWAS analysis.

8.2. Meta-analysis

Fixed effects meta-analysis (FEMA) was performed using METAL r2020-05-05 (github.com/statgen/METAL). Multi-ancestry Meta analysis (MAMA) provides improved power in meta-analysis of different populations with low type 1 error rates.[56] We used MAMA per population using LD reference panels derived from a combination of 1000 genomes data, and Thai population whole genome sequenced (WGS) samples.

8.3. Identifying conditionally independent loci and fine-mapping

External LD reference panels were created from AllOfUs (v5) data for European (EUR, n=51125), African (AFR, n=22837) and a combination of East-Asian, South Asian and Thai

population-specific WGS data (EAS-SAS-THAI, n=3788). Some analyses used panels limited to 10,000 individuals for computational efficiency. GCTA-COJO v1.94.0 [57] was used to identify conditionally independent loci that became our LD-sentinel markers. Each LD-sentinel marker was treated as a 1MBp window which was labeled with the nearest or, if known, biologically relevant gene. Once each region was determined, fine-mapping was performed using FINEMAP[58].

8.4. Heritability and genetic correlation analyses

LD scores were established from external LD panels as described above, using maximum 1cM window positions. LDSC was performed on summary statistics restricted to high quality, HapMap 3 variants. LDAK was also used to estimate heritability using the thin and BLD models appropriate for ancestry. Genetic correlations with blood cell traits were estimated using LDAK, using the thin model. Summary statistics for blood cell traits were obtained from the published BCX2 consortium summary statistics available at http://www.mhi-humangenetics.org/en/resources/.

8.5. Enrichment and in-silico functional study

SCAVENGE[34] was performed using fine-mapped statistics derived as described above. scATAC data from 10 individuals representing 33,819 cells from 23 cell populations were used.[59] Derivation and preparation of bulk ATACseq data is described elsewhere.[46] Peak calling was performed using MACS2. Hi-C data was acquired from a previously described dataset.[60]

8.6. Primary cell culture

CD34+ HSPCs were thawed into a maintenance medium consisting of a StemSpan II base (StemCell Technologies), CC100 (StemCell Technologies), 50 ng/mL human TPO (Pepro Tech), and 1% penicillin/streptomycin (Life Technologies).[61,62] Cells treated with RNP complexes for enhancer deletions were electroporated 48 hours after thawing and collected 72 hours post-nucleofection. Cells treated with lentivirus were transduced 24 hours after thawing, sorted 72 hours after thawing, and moved to erythroid media 96 hours after thawing.

After the maintenance phase, $CD34^+$ HSPCs were differentiated using the three-phase culture system previously described.[63,64] First, a base erythroid medium was created by supplementing IMDM with 2% human AB plasma, 3% human AB serum, 3 U/mL heparin, 10 µg/mL insulin, 200 µg/mL holo-transferrin, and 1% penicillin/streptomycin. From days 1-7 in erythroid media, this base medium was further supplemented with 3 U/mL EPO, 10 ng/mL human SCF, and 1 ng/mL IL-3. From days 7-12, this base medium was further supplemented with 3 U/mL EPO and 10 ng/mL human SCF. After day 12, the base medium was supplemented with 1 mg/mL total of holo-transferrin and 3 U/mL of EPO.

8.7. Electroporation of primary cells

Two days after thawing, RNP complexes were electroporated into $CD34^+$ HSPCs using a P3 Primary Cell 4D-Nucleofector X Kit S on the Lonza 4D Nucleofector system. Complexes were formed by combining 50 pmol of Cas9 nuclease (IDT) and 100 pmol total of sgRNAs (Synthego). Cells were treated either with two pairs of guides targeting the putative *BACH2* enhancer or a negative control targeting *AAVS1* (Supplementary Table 11). Electroporated cells were harvested three days post-nucleofection for genomic DNA and RNA extraction using the AllPrep DNA/RNA Micro Kit (QIAGEN) according to kit instructions. Deletions, inversions, and wild-type alleles were quantified in the bulk population using qPCR (Supplementary Table 14).

8.8. Lentiviral increased expression

The human *BACH2* coding sequence was synthesized by Azenta and inserted into the HMD lentiviral vector[65]. Lentiviral particles were produced as previously described.[8] Briefly, 293T cells cultured in DMEM supplemented with 10% FBS were co-transfected with packaging vectors pVSVG and pΔ8.9, and the expression vectors HMD-empty vector or HMD-BACH2. DMEM was replaced with erythroid differentiation base media 24 h later and supernatant containing lentivirus were collected, filtered with a 0.45 µm filter, and concentrated by ultracentrifugation (24,000 rpm, 2 h, 4^0C). Concentrated virus was used to transduce HSPCs in the presence of 8 µg/mL polybrene (Millipore) by spinfection (2,000 rpm, 1.5 h, RT). Transduced cells were sorted based on the top and bottom 30% of GFP^+ cells by fluorescence activated cell sorting (FACS) before erythroid differentiation and subsequent functional analyses.

8.9. RT-qPCR

RNA was collected from cultured cells using the RNAqueous Total RNA Isolation Kit (Invitrogen) or the AllPrep DNA/RNA Micro Kit (QIAGEN) according to kit instructions. Isolated RNA was inputted into the iScript cDNA synthesis kit (BioRad) following kit instructions in order to create cDNA. RT-qPCR was run on the CFX96 Real Time System (BioRad) using iQ SYBR Green Supermix (BioRad) following kit instructions. Primer pairs for RT-qPCR are listed by gene in (Supplementary Table 15). Transcript levels are expressed as fold change using the delta-delta-Cq quantification strategy and normalized to the expression of housekeeping gene *ACTB*.

8.10. Flow cytometry

The frequency of F-cells in transduced HSPCs undergoing erythroid differentiation was quantified as previously described.[8] Briefly, cells were fixed in 0.05% glutaraldehyde for 10 min, permeabilized with 0.1% Triton X-100 (Life Technologies) for 5 min, and stained with an anti-HbF APC antibody (Invitrogen) for 30 min. Cells were subsequently washed, acquired on an Accuri C6 flow cytometer (BD Biosciences), and analyzed using FlowJo software (v.10.8.1, BD Biosciences).

To assess the impact of increased BACH2 expression on erythroid differentiation, transduced HSPCs undergoing erythroid differentiation were stained with a combination of anti-CD71 APC and CD235a PE (all from BD Biosciences) for 10 min. Stained cells were acquired on an Accuri C6 flow cytometer and analyzed using FlowJo software.

8.12. Erythroid morphology assessment

The morphology of erythroid cells was assessed as previously described.[66] Briefly, cells were harvested from culture, washed, and cytospinned using a Shandon Cytospin 4 (Thermo Fisher) onto polysine slides (Epredia). Slides were stained with May Grunwald and Giemsa stains (both from Sigma-Aldrich) according to manufacturer's recommendations. Stained slides were dried, mounted with coverslips using permount (Fisher Scientific), and imaged on a Nikon II Eclipse E800 microscope at 60x magnification.

8.11. CNV and structural variant calling

CNVs were called using an ensemble approach utilizing PennCNV, QuantiSNP, and iPattern from array data. Subsequent results were limited to *HBB* and *HBD* genes and with additional manual confirmation, individuals were identified as carrying one or more previously described HPFH deletions (Supplementary Table 13). Structural variants were also called using manta in 197 WGS individuals configured for germline analysis.

8.12. Polygenic risk score calculation

We used LDpred2,[67,68] to calculate a polygenic risk score based on summary statistics excluding the target thai population data that we wished to predict results for. Analysis was restricted to high quality HapMap3 SNPs. Infinitesimal modeling was performed, and predictions were made upon the array calls from Thai population individuals. Linear regression was then performed with predictions, age, sex, and ten principal components upon the response variable of measured HbF with the predictions performing significantly ($p<2e-16$).

9. Acknowledgements

We are grateful to members of the Sankaran laboratory for valuable comments and discussion, as well as S. Orkin and D. Nathan for their inspiration and guidance. We acknowledge early work on these cohorts by J. Verboon, A. Cheng, and C. Fiorini. We thank V. Kuchroo, A. Schnell, and Y. Hou for generously sharing mouse Bach2 constructs. This work was supported by the New York Stem Cell Foundation (to V.G.S), NIH grants R01 DK103794 and R01 HL146500 (to V.G.S). V.G.S is a New York Stem Cell–Robertson Investigator.

Molecular data for the TOPMed program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: walk_PHaSST" (phs001514) was performed at the Baylor Genomics Center (HHSN268201500015C). WGS for "NHLBI TOPMed: Whole Genome Sequencing and Related Phenotypes in the Outcome Modifying Genes in Sickle Cell Disease Study" (OMG-SCD) (phs001608) was performed at the Baylor College of Medicine Human Genome Sequencing Center (HHSN268201600033I). WGS for "NHLBI TOPMed: REDS-III_Brazil" (phs001468) was performed at the Baylor Genomics Center (HHSN268201600033I & HHSN268201500015C). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The OMG-SCD study was administrated by Marilyn J. Telen, M.D. and Allison E. Ashley-Koch, Ph.D. from Duke University Medical Center and collection of the data set was supported by grants HL068959 and HL079915 from the National Heart, Lung, and Blood Institute (NHLBI) of the National Institute of Health (NIH). We thank Dr. Mark Gladwin and the investigators of the Walk-PHasst study and the patients who participated in the study. We also thanks the walk-PHaSST clinical site team: Albert Einstein College of Medicine: Jane Little and Verlene Davis; Columbia University: Robyn Barst, Erika Rosenzweig, Margaret Lee and Daniela Brady; UCSF Benioff Children's Hospital Oakland: Claudia Morris, Ward Hagar, Lisa Lavrisha, Howard Rosenfeld, and Elliott Vichinsky; Children's Hospital of Pittsburgh of UPMC: Regina McCollum; Hammersmith Hospital, London: Sally Davies, Gaia Mahalingam, Sharon Meehan, Ofelia Lebanto, and Ines Cabrita; Howard University: Victor Gordeuk, Oswaldo Castro, Onyinye Onyekwere, Vandana Sachdev, Alvin Thomas, Gladys Onojobi, Sharmin Diaz, Margaret Fadojutimi-Akinsiku, and Randa Aladdin; Johns Hopkins University: Reda Girgis, Sophie Lanzkron and Durrant Barasa; NHLBI: Mark Gladwin, Greg Kato, James Taylor, Wynona Coles, Catherine Seamon, Mary Hall, Amy Chi, Cynthia Brenneman, Wen Li, and Erin Smith; University of Colorado: Kathryn Hassell, David Badesch, Deb McCollister and Julie McAfee; University of Illinois at Chicago: Dean Schraufnagel, Robert Molokie, George Kondos, Patricia Cole-Saffold, and Lani Krauz; National Heart & Lung Institute, Imperial College London: Simon Gibbs. Thanks

also to the data coordination center team from Rho, Inc.: Nancy Yovetich, Rob Woolson, Jamie Spencer, Christopher Woods, Karen Kesler, Vickie Coble, and Ronald W. Helms. We also thank Dr. Yingze Zhang for directing the Walk-PHasst repository and Dr. Mehdi Nouraie for maintaining the Walk-PHasst database and Dr. Jonathan Goldsmith as a NIH program director for this study. Special thanks to the volunteers who participated in the Walk-PHaSST study. This project was funded with federal funds from the NHLBI, NIH, Department of Health and Human Services, under contract HHSN268200617182C. This study is registered at www.clinicaltrials.gov as NCT00492531. Detail description of the study was published in Blood, 2011 118:855-864, Machado et al "Hospitalization for pain in patients with sickle cell disease treated with sildenafil for elevated TRV and low exercise capacity".

The fetal hemoglobin measurements in INTERVAL were funded by Biogen. Participants in the INTERVAL randomised controlled trial were recruited with the active collaboration of NHS Blood and Transplant England (www.nhsbt.nhs.uk), which has supported field work and other elements of the trial. DNA extraction and genotyping were co-funded by the National Institute for Health and Care Research (NIHR), the NIHR BioResource (bioresource.nihr.ac.uk) and the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014) [*]. The academic coordinating centre for INTERVAL was supported by core funding from the: NIHR Blood and Transplant Research Unit in Donor Health and Genomics (NIHR BTRU-2014-10024), NIHR BTRU in Donor Health and Behaviour (NIHR203337), UK Medical Research Council (MR/L003120/1), British Heart Foundation (SP/09/002; RG/13/13/30194; RG/18/13/33946) and NIHR Cambridge BRC (BRC-1215-20014) [*]. A complete list of the investigators and contributors to the INTERVAL trial is provided in reference [22]. The academic coordinating centre would like to thank blood donor centre staff and blood donors for participating in the INTERVAL trial. This work was supported by Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome. The authors would like to thank colleagues at the University of Cambridge and the National Haemoglobinopathy Reference Laboratory for their help with preparing and shipping INTERVAL samples and conducting HbF assays. *The views expressed are those of the author(s) and not necessarily those of the NIHR, NHSBT or the Department of Health and Social Care.

The All of Us Research Program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2 OD026551; 1 OT2 OD026555; IAA #: AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research Center: 5 U2C OD023196; Biobank: 1 U24

OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; 1 OT2 OD025276. In addition, the All of Us Research Program would not be possible without the partnership of its participants.

10. Author Contributions

V.G.S., L.D.C., R.L., H.Y.L. conceptualized and designed the study. L.D.C., R.L., H.Y.L., F.Y., M.W., B.M., S.E., P.D., L.M., R.S., T.S., S.R., P.G.B., D.S.P., E.K., W.J.A., F.A., K.A., A.L.L.P, G.K., Y.Z., S.N., V.R.G., M.T.G, M.E.G, A.A., M.J.T., B.C., S.K., C.D., E.C.S., P.L., A.C., C.M., T.M., A.M., A.H., G.T., K.S., U.T., V.A.S., M.J.W., L.F., B.N., A.S.B., V.V., S.N., V.G.S. obtained and provided cohort data. L.D.C., R.L., H.Y.L., F.Y., M.W., B.M., P.D., S.E., F.A. performed functional studies and computational analyses. V.G.S., L.D.C., R.L., H.Y.L. wrote the original draft with input from all authors. V.G.S. provided overall study oversight. All authors were involved in reviewing and editing the manuscript. All authors read and approved the final version of the manuscript.

11. Competing Interests

During the drafting of the manuscript, D.S.P. became a full-time employee of AstraZeneca. F.A. is an employee and shareholder of Illumina, Inc. M.T.G. serves as a consultant for Actelion, Bayer Healthcare, Pfizer, Forma, and Fulcrum Therapeutics. A.H.L. reports speakers honoraria from Siemens Healthineers and Beckman Diagnostics, as well as participation on an advisory board of Roche Diagnostics, all unrelated to the present work. A.S.B. reports institutional grants from AstraZeneca, Bayer, Biogen, BioMarin, Bioverativ, Novartis, Regeneron and Sanofi. V.G.S. serves as an advisor to and/or has equity in Branch Biosciences, Ensoma, Novartis, Forma, and Cellarity, all unrelated to the present work.

12. Additional Information and Correspondence

Correspondence and requests for materials should be addressed to Dr. Vijay G. Sankaran sankaran@broadinstitute.org.

13. Reference List

1. Sankaran VG, Orkin SH. The switch from fetal to adult hemoglobin. *Cold Spring Harb Perspect Med*. 2013;3(1):a011643.

2. Uda M, Galanello R, Sanna S, 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 U S A*. 2008;105(5):1620-1625.

3. Lettre G, Sankaran VG, Bezerra MAC, 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 U S A*. 2008;105(33):11869-11874.

4. Menzel S, Garner C, Gut I, et al. A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat Genet*. 2007;39(10):1197-1199.

5. Sankaran VG, Menne TF, Xu J, et al. Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. *Science*. 2008;322(5909):1839-1842.

6. Basak A, Hancarova M, Ulirsch JC, et al. BCL11A deletions result in fetal hemoglobin persistence and neurodevelopmental alterations. *J Clin Invest*. 2015;125(6):2363-2368.

7. Dias C, Estruch SB, Graham SA, et al. BCL11A Haploinsufficiency Causes an Intellectual Disability Syndrome and Dysregulates Transcription. *Am J Hum Genet*. 2016;99(2):253-274.

8. Shen Y, Li R, Teichert K, et al. Pathogenic BCL11A variants provide insights into the mechanisms of human fetal hemoglobin silencing. *PLoS Genet*. 2021;17(10):e1009835.

9. Bauer DE, Kamran SC, Lessard S, et al. An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science*. 2013;342(6155):253-257.

10. Canver MC, Smith EC, Sher F, et al. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature*. 2015;527(7577):192-197.

11. Basak A, Munschauer M, Lareau CA, et al. Control of human hemoglobin switching by LIN28B-mediated regulation of BCL11A translation. *Nat Genet*. 2020;52(2):138-145.

12. Huang P, Peslak SA, Ren R, et al. HIC2 controls developmental hemoglobin switching by repressing BCL11A transcription. *Nat Genet*. 2022;54(9):1417-1426.

13. Lee YT, Terry Lee Y, de Vasconcellos JF, et al. LIN28B-mediated expression of fetal hemoglobin and production of fetal-like erythrocytes from adult human erythroblasts ex vivo. *Blood*. 2013;122(6):1034-1041. doi:10.1182/blood-2012-12-472308

14. de Vasconcellos JF, Tumburu L, Byrnes C, et al. IGF2BP1 overexpression causes fetal-like hemoglobin expression patterns in cultured human adult erythroblasts. *Proc Natl Acad Sci U S A*. 2017;114(28):E5664-E5672.

15. Shen Y, Verboon JM, Zhang Y, et al. A unified model of human hemoglobin switching through single-cell genome editing. *Nat Commun*. 2021;12(1):4991.

16. Liu N, Hargreaves VV, Zhu Q, et al. Direct Promoter Repression by BCL11A Controls the Fetal to Adult Hemoglobin Switch. *Cell*. 2018;173(2):430-442.e17.

17. Martyn GE, Wienert B, Yang L, et al. Natural regulatory mutations elevate the fetal globin gene via disruption of BCL11A or ZBTB7A binding. *Nat Genet*. 2018;50(4):498-503.

18. Esrick EB, Lehmann LE, Biffi A, et al. Post-Transcriptional Genetic Silencing of BCL11A to Treat Sickle Cell Disease. *N Engl J Med*. 2021;384(3):205-215.

19. Frangoul H, Altshuler D, Cappellini MD, et al. CRISPR-Cas9 Gene Editing for Sickle Cell Disease and β-Thalassemia. *N Engl J Med*. 2021;384(3):252-260.

20. Lu HY, Orkin SH, Sankaran VG. Fetal Hemoglobin Regulation in Beta-Thalassemia. *Hematol Oncol Clin North Am*. 2023;37(2):301-312.

21. Crossley M, Christakopoulos GE, Weiss MJ. Effective therapies for sickle cell disease: are we there yet? *Trends Genet*. 2022;38(12):1284-1298.

22. Di Angelantonio E, Thompson SG, Kaptoge S, et al. Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors. *Lancet*. 2017;390(10110):2360-2371.

23. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020;369(6509):1318-1330.

24. Võsa U, Claringbould A, Westra HJ, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet*. 2021;53(9):1300-1310.

25. Mtatiro SN, Singh T, Rooks H, et al. Genome wide association study of fetal hemoglobin in sickle cell anemia in Tanzania. *PLoS One*. 2014;9(11):e111464.

26. Gladwin MT, Barst RJ, Gibbs JSR, et al. Risk factors for death in 632 patients with sickle cell disease in the United States and United Kingdom. *PLoS One*. 2014;9(7):e99489.

27. Ashley-Koch AE, Elliott L, Kail ME, et al. Identification of genetic polymorphisms associated with risk for pulmonary hypertension in sickle cell disease. *Blood*. 2008;111(12):5721-5726.

28. Kleinman S, Busch MP, Murphy EL, et al. The National Heart, Lung, and Blood Institute Recipient Epidemiology and Donor Evaluation Study (REDS-III): a research program striving to improve blood donor and transfusion recipient outcomes. *Transfusion* . 2014;54(3 Pt 2):942-955.

29. Hankins JS, Estepp JH, Hodges JR, et al. Sickle Cell Clinical Research and Intervention Program (SCCRIP): A lifespan cohort study for sickle cell disease progression from the

pediatric stage into adulthood. *Pediatr Blood Cancer*. 2018;65(9):e27228.

30. Rampersaud E, Kang G, Palmer LE, et al. A polygenic score for acute vaso-occlusive pain in pediatric sickle cell disease. *Blood Adv*. 2021;5(14):2839-2851.

31. Bao EL, Lareau CA, Brugnara C, et al. Heritability of fetal hemoglobin, white cell count, and other clinical traits from a sickle cell disease family cohort. *Am J Hematol*. 2019;94(5):522-527.

32. Vuckovic D, Bao EL, Akbari P, et al. The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell*. 2020;182(5):1214-1231.e11.

33. Chen MH, Raffield LM, Mousas A, et al. Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell*. 2020;182(5):1198-1213.e14.

34. Yu F, Cato LD, Weng C, et al. Variant to function mapping at single-cell resolution through network propagation. *Nat Biotechnol*. 2022;40(11):1644-1653.

35. Caulier AL, Sankaran VG. Molecular and cellular mechanisms that regulate human erythropoiesis. *Blood*. 2022;139(16):2450-2459.

36. Barbarani G, Fugazza C, Strouboulis J, Ronchi AE. The Pleiotropic Effects of GATA1 and KLF1 in Physiological Erythropoiesis and in Dyserythropoietic Disorders. *Front Physiol*. 2019;10:91.

37. Oyake T, Itoh K, Motohashi H, et al. Bach proteins belong to a novel family of BTB-basic leucine zipper transcription factors that interact with MafK and regulate transcription through the NF-E2 site. *Mol Cell Biol*. 1996;16(11):6083-6095.

38. Brand M, Ranish JA, Kummer NT, et al. Dynamic changes in transcription factor complexes during erythroid differentiation revealed by quantitative proteomics. *Nat Struct Mol Biol*. 2004;11(1):73-80.

39. Sun J, Brand M, Zenke Y, Tashiro S, Groudine M, Igarashi K. Heme regulates the dynamic exchange of Bach1 and NF-E2-related factors in the Maf transcription factor network. *Proc Natl Acad Sci U S A*. 2004;101(6):1461-1466.

40. Andrews NC, Kotkow KJ, Ney PA, Erdjument-Bromage H, Tempst P, Orkin SH. The ubiquitous subunit of erythroid transcription factor NF-E2 is a small basic-leucine zipper protein related to the v-maf oncogene. *Proc Natl Acad Sci U S A*. 1993;90(24):11488-11492.

41. Andrews NC, Erdjument-Bromage H, Davidson MB, Tempst P, Orkin SH. Erythroid transcription factor NF-E2 is a haematopoietic-specific basic-leucine zipper protein. *Nature*. 1993;362(6422):722-728.

42. Sawado T, Igarashi K, Groudine M. Activation of beta-major globin gene transcription is associated with recruitment of NF-E2 to the beta-globin LCR and gene promoter. *Proc Natl*

Acad Sci U S A. 2001;98(18):10226-10231.

43. Zhu X, Li B, Pace BS. NRF2 mediates γ-globin gene regulation and fetal hemoglobin induction in human erythroid progenitors. *Haematologica*. 2017;102(8):e285-e288.

44. Zhu X, Oseghale AR, Nicole LH, Li B, Pace BS. Mechanisms of NRF2 activation to mediate fetal hemoglobin induction and protection against oxidative stress in sickle cell disease. *Exp Biol Med* . 2019;244(2):171-182.

45. Krishnamoorthy S, Pace B, Gupta D, et al. Dimethyl fumarate increases fetal hemoglobin, provides heme detoxification, and corrects anemia in sickle cell disease. *JCI Insight*. 2017;2(20). doi:10.1172/jci.insight.96409

46. Ludwig LS, Lareau CA, Bao EL, et al. Transcriptional States and Chromatin Accessibility Underlying Human Erythropoiesis. *Cell Rep*. 2019;27(11):3228-3240.e7.

47. Stadhouders R, Aktuna S, Thongjuea S, et al. HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers. *J Clin Invest*. 2014;124(4):1699-1710.

48. Sankaran VG, Xu J, Byron R, et al. A functional element necessary for fetal hemoglobin silencing. *N Engl J Med*. 2011;365(9):807-814.

49. Danjou F, Zoledziewska M, Sidore C, et al. Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels. *Nat Genet*. 2015;47(11):1264-1271.

50. Liu L, Pertsemlidis A, Ding LH, et al. Original Research: A case-control genome-wide association study identifies genetic modifiers of fetal hemoglobin in sickle cell disease. *Exp Biol Med* . 2016;241(7):706-718.

51. Perkins A, Xu X, Higgs DR, et al. Krüppeling erythropoiesis: an unexpected broad spectrum of human red blood cell disorders due to KLF1 variants. *Blood*. 2016;127(15):1856-1862.

52. von der Lippe C, Tveten K, Prescott TE, et al. Heterozygous variants in ZBTB7A cause a neurodevelopmental disorder associated with symptomatic overgrowth of pharyngeal lymphoid tissue, macrocephaly, and elevated fetal hemoglobin. *Am J Med Genet A*. 2022;188(1):272-282.

53. Hanna RE, Hegde M, Fagre CR, et al. Massively parallel assessment of human variants with base editor screens. *Cell*. 2021;184(4):1064-1080.e20.

54. Goel VY, Huseyin MK, Hansen AS. Region Capture Micro-C reveals coalescence of enhancers and promoters into nested microcompartments. *bioRxiv*. Published online July 14, 2022:2022.07.12.499637. doi:10.1101/2022.07.12.499637

55. Hua P, Badat M, Hanssen LLP, et al. Defining genome architecture at base-pair resolution. *Nature*. 2021;595(7865):125-129.

56. Turley P, Martin AR, Goldman G, et al. Multi-Ancestry Meta-Analysis yields novel genetic

discoveries and ancestry-specific associations. *bioRxiv*. Published online April 24, 2021:2021.04.23.441003. doi:10.1101/2021.04.23.441003

57. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. 2011;88(1):76-82.

58. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*. 2016;32(10):1493-1501.

59. Granja JM, Klemm S, McGinnis LM, et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat Biotechnol*. 2019;37(12):1458-1465.

60. Javierre BM, Burren OS, Wilder SP, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*. 2016;167(5):1369-1384.e19.

61. Bao EL, Nandakumar SK, Liao X, et al. Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells. *Nature*. 2020;586(7831):769-775.

62. Voit RA, Tao L, Yu F, et al. A genetic disorder reveals a hematopoietic stem cell regulatory network co-opted in leukemia. *Nat Immunol*. 2023;24(1):69-83.

63. Giani FC, Fiorini C, Wakabayashi A, et al. Targeted Application of Human Genetic Variation Can Improve Red Blood Cell Production from Stem Cells. *Cell Stem Cell*. 2016;18(1):73-78.

64. Nandakumar SK, McFarland SK, Mateyka LM, et al. Gene-centric functional dissection of human genetic variation uncovers regulators of hematopoiesis. *Elife*. 2019;8. doi:10.7554/eLife.44080

65. Abdulhay NJ, Fiorini C, Verboon JM, et al. Impaired human hematopoiesis due to a cryptic intronic GATA1 splicing mutation. *J Exp Med*. 2019;216(5):1050-1060.

66. Ludwig LS, Lareau CA, Bao EL, et al. Congenital anemia reveals distinct targeting mechanisms for master transcription factor GATA1. *Blood*. 2022;139(16):2534-2546.

67. Ni G, Zeng J, Revez JA, et al. A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts. *Biol Psychiatry*. 2021;90(9):611-620.

68. Privé F, Arbel J, Vilhjálmsson BJ. LDpred2: better, faster, stronger. *Bioinformatics*. Published online December 16, 2020. doi:10.1093/bioinformatics/btaa1029

Table 1. Details of the included populations in HbF meta-analysis.

Study	Inferred ancestry background	Participants	Percent total	Additional cohort notes:
St Jude	AFR	526	1.86%	SCD cohort
GTEx	EUR	670	2.37%	Expression ratio phenotype
walk_PHaSST	AFR	408	1.44%	SCD cohort
OMG_SCD	AFR	253	0.90%	SCD cohort
REDS-III_Brazil	AFR	1589	5.63%	SCD cohort
Tanzania	AFR	1187	4.20%	SCD cohort
Thai	THAI	1392	4.93%	Selected from extremes of distribution from larger population
Sardinia	EUR	6305	22.33%	
BIOS	EUR	1872	6.63%	Expression ratio phenotype
Sweden	EUR	3031	10.73%	
Interval	EUR	11004	38.97%	
Total	ALL	28237		

Figure 1

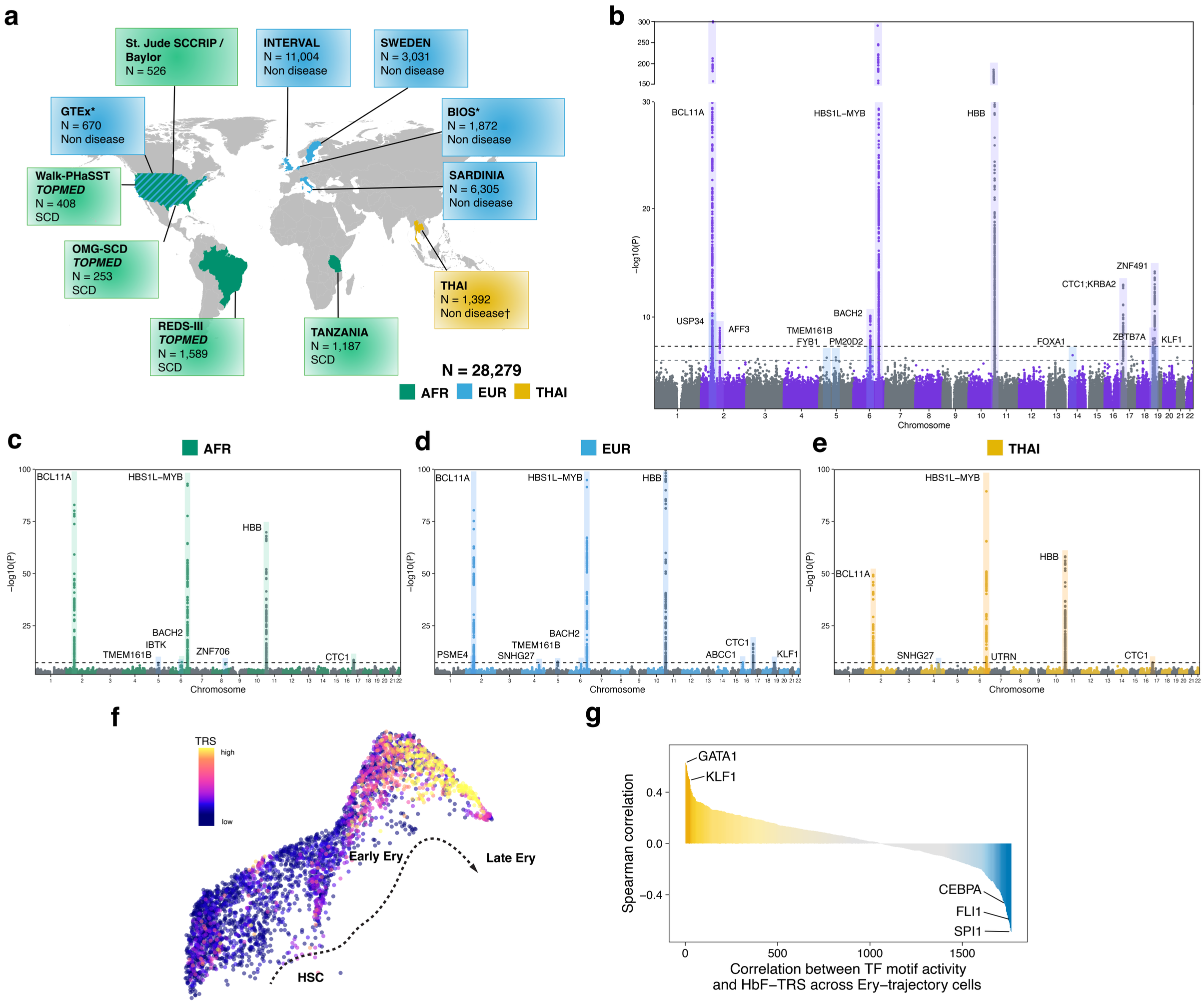

a
St. Jude SCCRIP / Baylor
N = 526
INTERVAL
N = 11,004
Non disease
SWEDEN
N = 3,031
Non disease
GTEx*
N = 670
Non disease
BIOS*
N = 1,872
Non disease
Walk-PHaSST
TOPMED
N = 408
SCD
SARDINIA
N = 6,305
Non disease
OMG-SCD
TOPMED
N = 253
SCD
THAI
N = 1,392
Non disease†
REDS-III
TOPMED
N = 1,589
SCD
TANZANIA
N = 1,187
SCD
N = 28,279
AFR
EUR
THAI
b
−log10(P)
Chromosome
BCL11A
HBS1L−MYB
HBB
USP34
AFF3
TMEM161B
FYB1
PM20D2
BACH2
FOXA1
CTC1;KRBA2
ZNF491
ZBTB7A
KLF1
c
AFR
BCL11A
HBS1L−MYB
HBB
BACH2
IBTK
TMEM161B
ZNF706
CTC1
d
EUR
BCL11A
HBS1L−MYB
HBB
PSME4
SNHG27
TMEM161B
BACH2
ABCC1
CTC1
KLF1
e
THAI
HBS1L−MYB
HBB
BCL11A
SNHG27
UTRN
CTC1
f
TRS
high
low
Early Ery
Late Ery
HSC
g
GATA1
KLF1
CEBPA
FLI1
SPI1
Spearman correlation
Correlation between TF motif activity
and HbF−TRS across Ery−trajectory cells

Figure 2

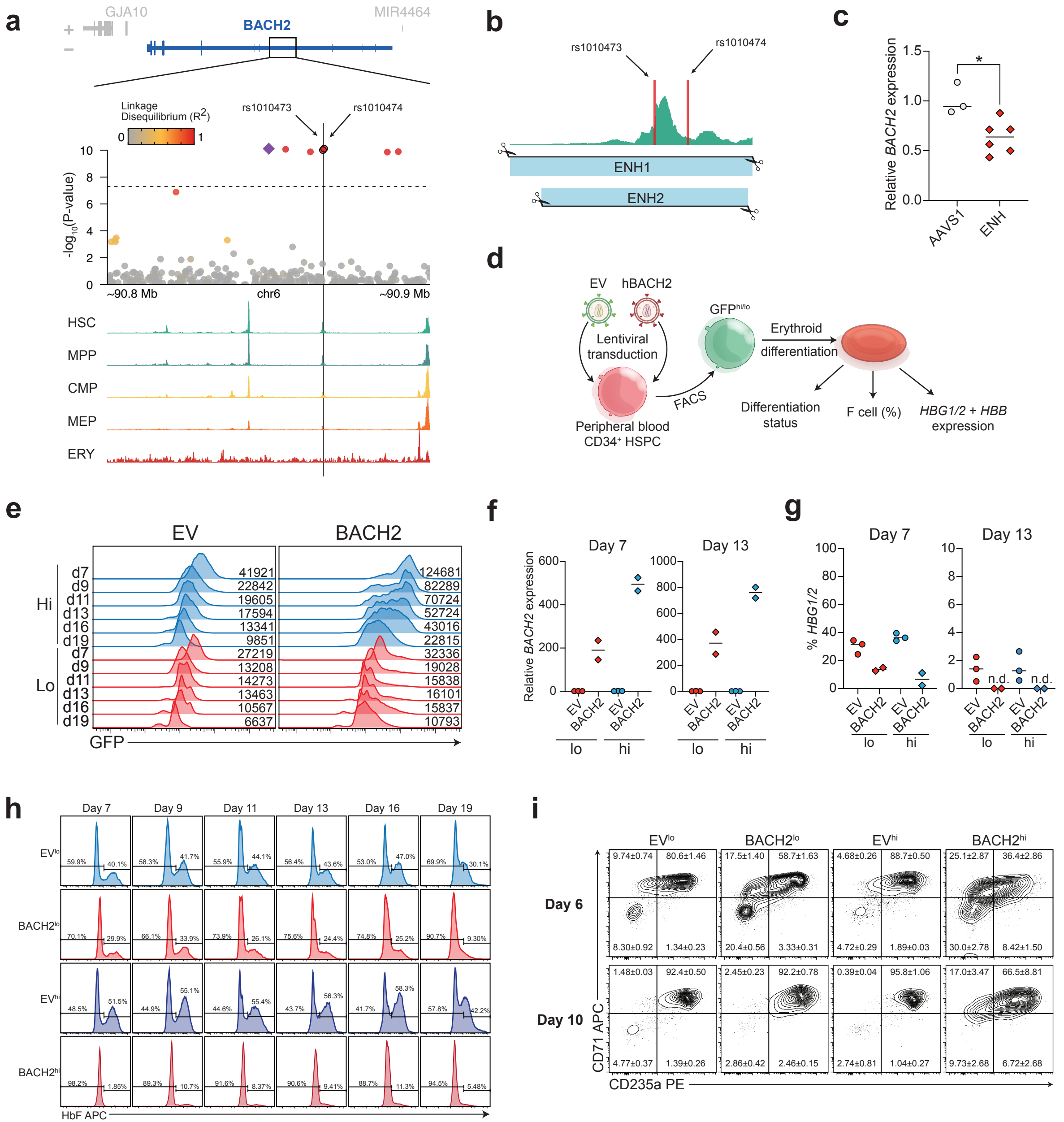

a
GJA10
MIR4464
BACH2
Linkage Disequilibrium (R^2)
rs1010473
rs1010474
-log10(P-value)
~90.8 Mb
chr6
~90.9 Mb
HSC
MPP
CMP
MEP
ERY
b
rs1010473
rs1010474
ENH1
ENH2
c
Relative BACH2 expression
AAVS1
ENH
d
EV
hBACH2
Lentiviral transduction
GFPhi/lo
Erythroid differentiation
FACS
Peripheral blood CD34+ HSPC
Differentiation status
F cell (%)
HBG1/2 + HBB expression
e
EV
BACH2
Hi
Lo
GFP
f
Day 7
Day 13
Relative BACH2 expression
lo
hi
g
Day 7
Day 13
% HBG1/2
n.d.
h
Day 7
Day 9
Day 11
Day 13
Day 16
Day 19
EVlo
BACH2lo
EVhi
BACH2hi
HbF APC
i
EVlo
BACH2lo
EVhi
BACH2hi
Day 6
Day 10
CD71 APC
CD235a PE

Figure 3

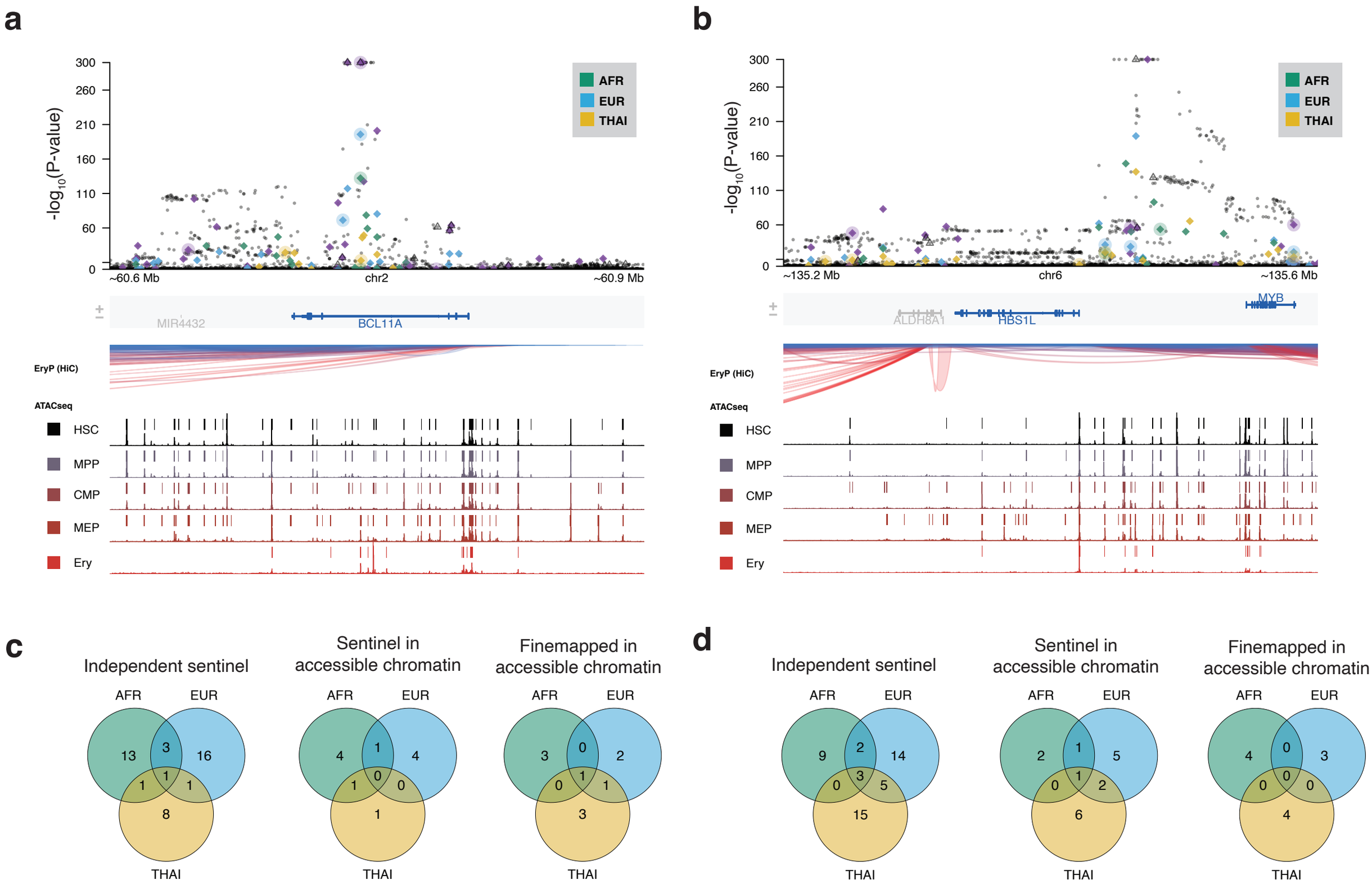
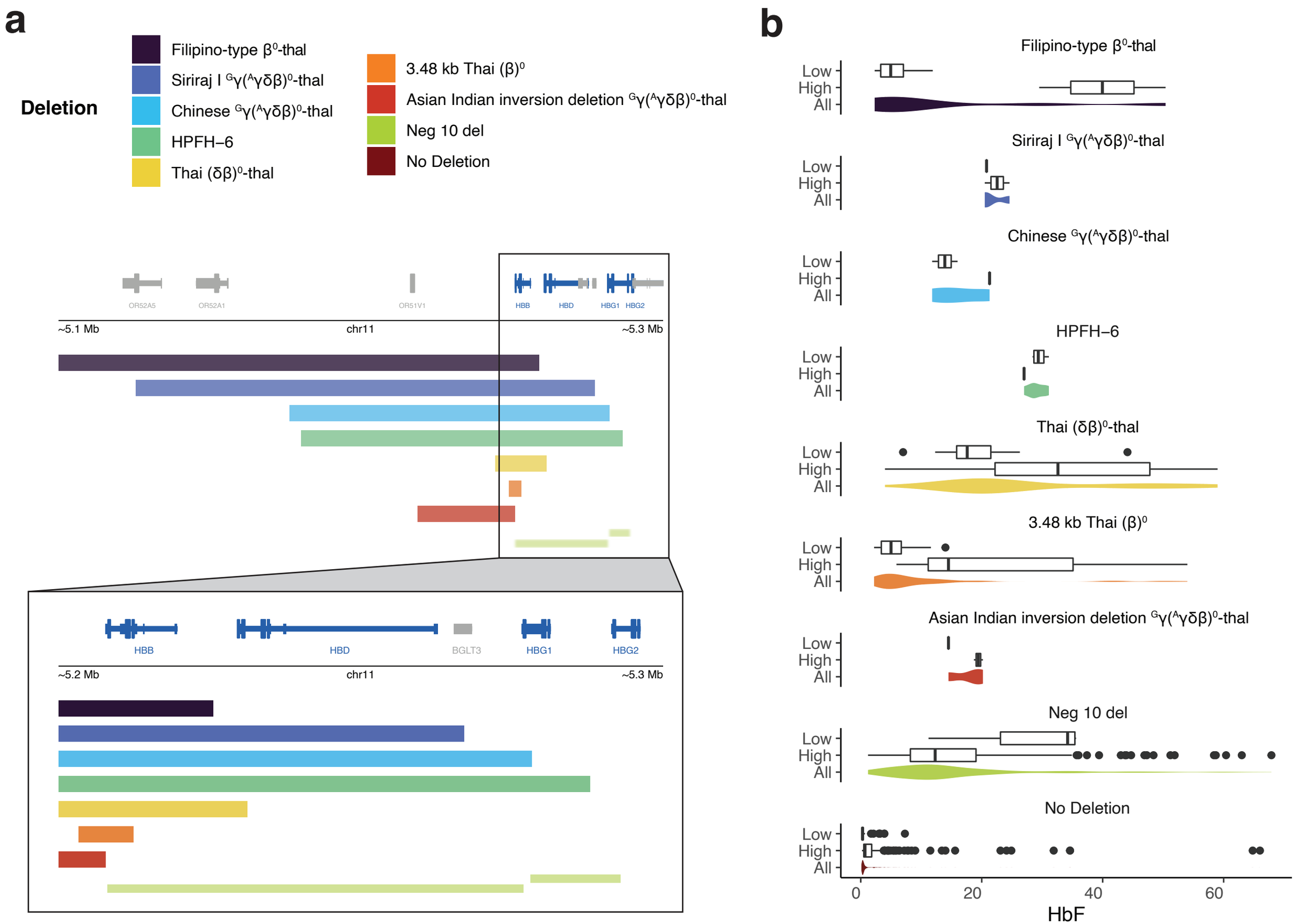
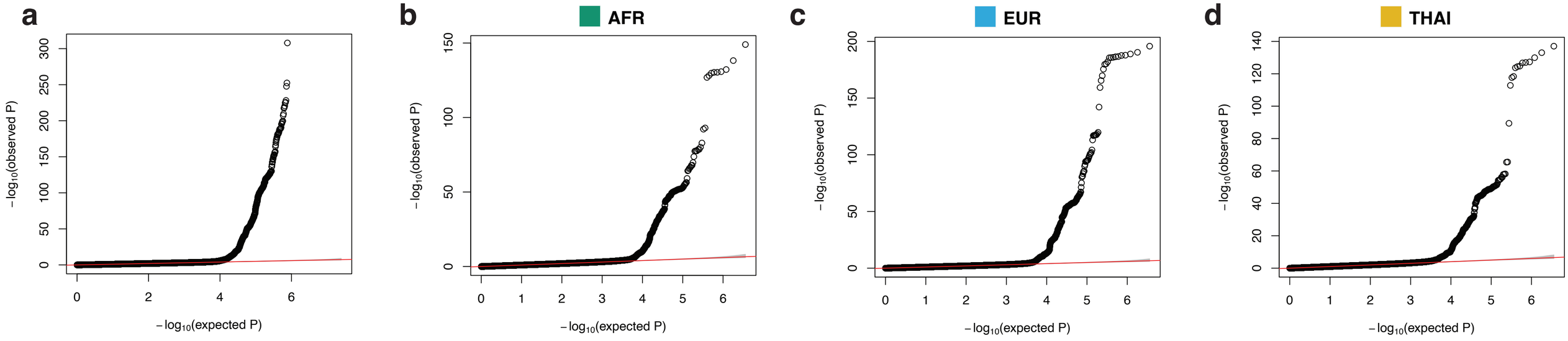

a
b
c
d
-log10(P-value)
AFR
EUR
THAI
~60.6 Mb
chr2
~60.9 Mb
MIR4432
BCL11A
~135.2 Mb
chr6
~135.6 Mb
ALDH8A1
HBS1L
MYB
EryP (HiC)
ATACseq
HSC
MPP
CMP
MEP
Ery
Independent sentinel
Sentinel in accessible chromatin
Finemapped in accessible chromatin

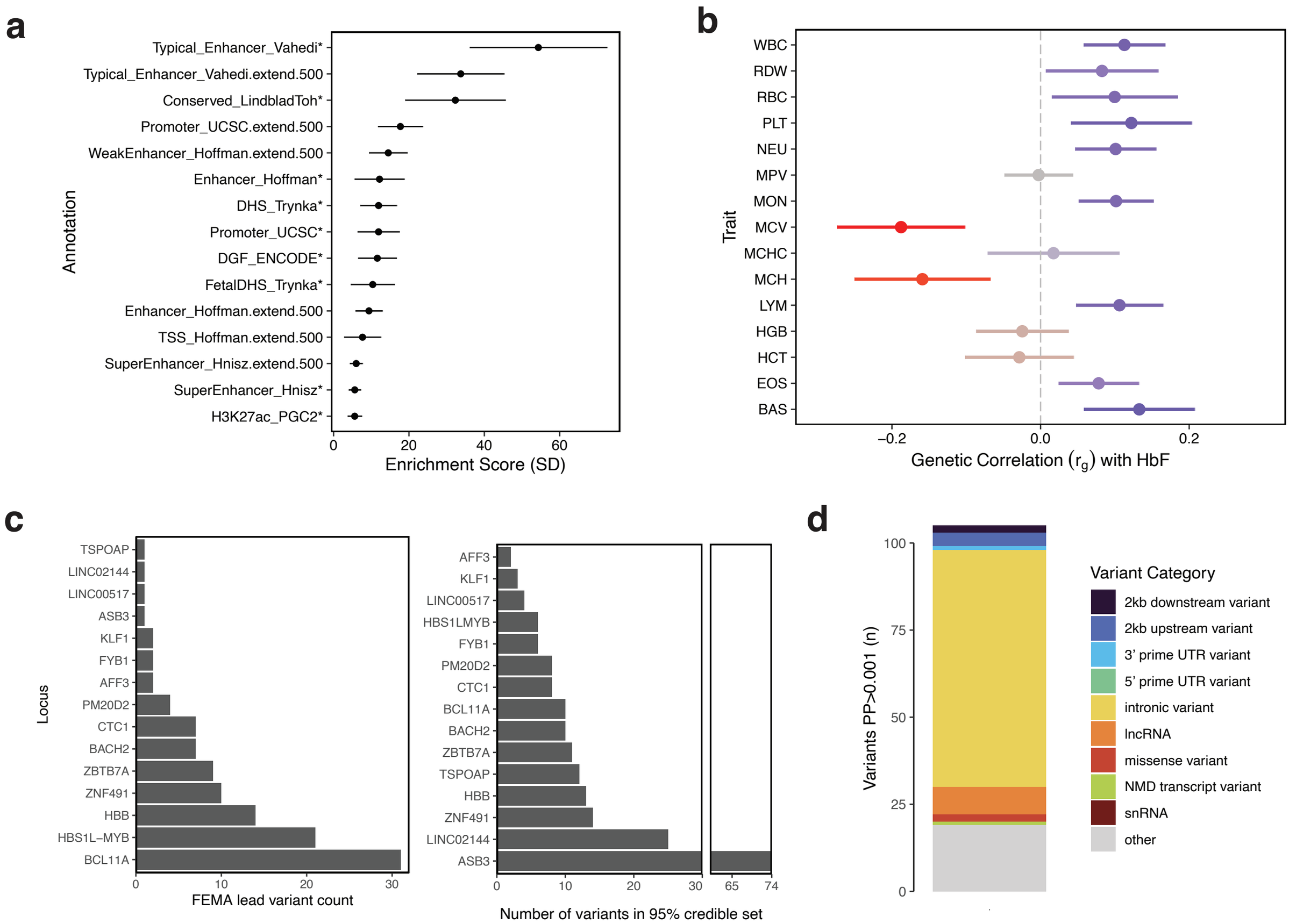
Figure 4


a
Deletion
Filipino-type β⁰-thal
Siriraj I Gγ(Aγδβ)⁰-thal
Chinese Gγ(Aγδβ)⁰-thal
HPFH–6
Thai (δβ)⁰-thal
3.48 kb Thai (β)⁰
Asian Indian inversion deletion Gγ(Aγδβ)⁰-thal
Neg 10 del
No Deletion
OR52A5
OR52A1
OR51V1
HBB
HBD
HBG1
HBG2
~5.1 Mb
chr11
~5.3 Mb
BGLT3
~5.2 Mb
b
Low
High
All
0
20
40
60
HbF

Extended Data Figure 1

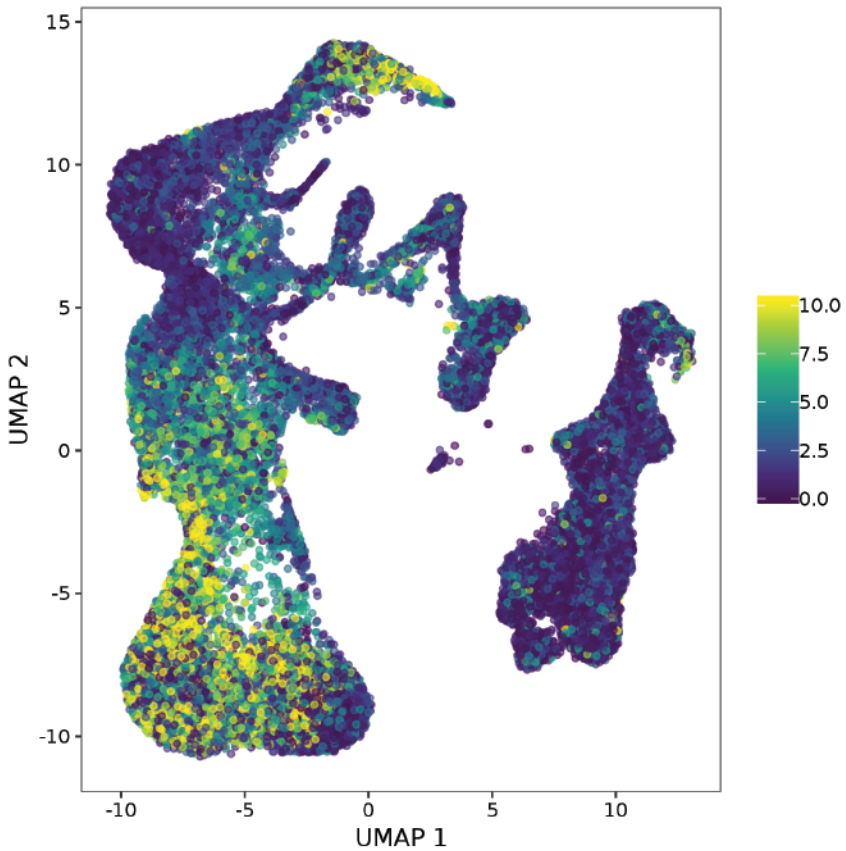


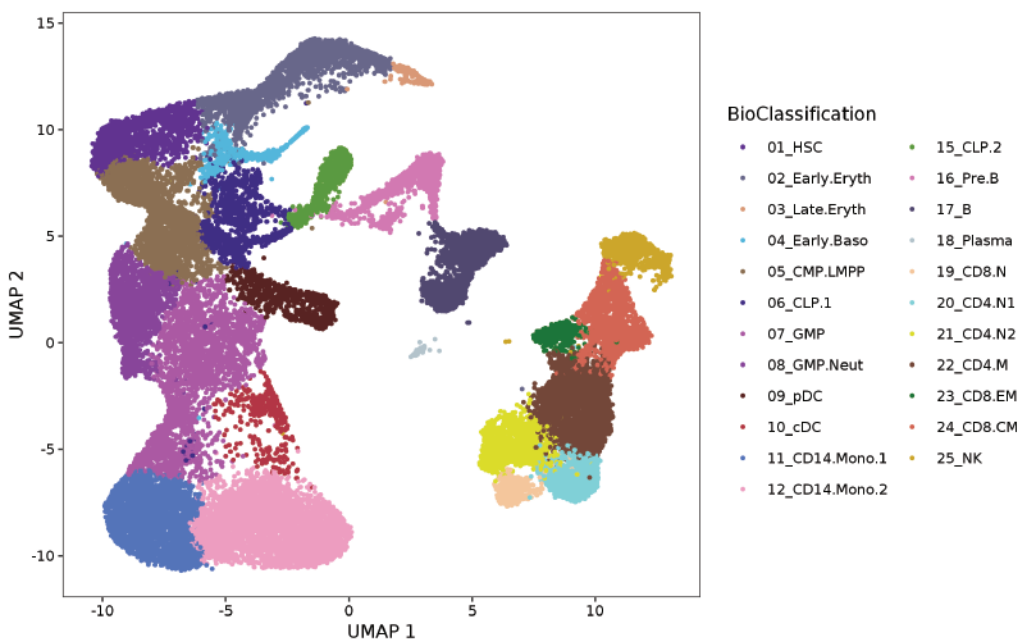
Extended Data Figure 2



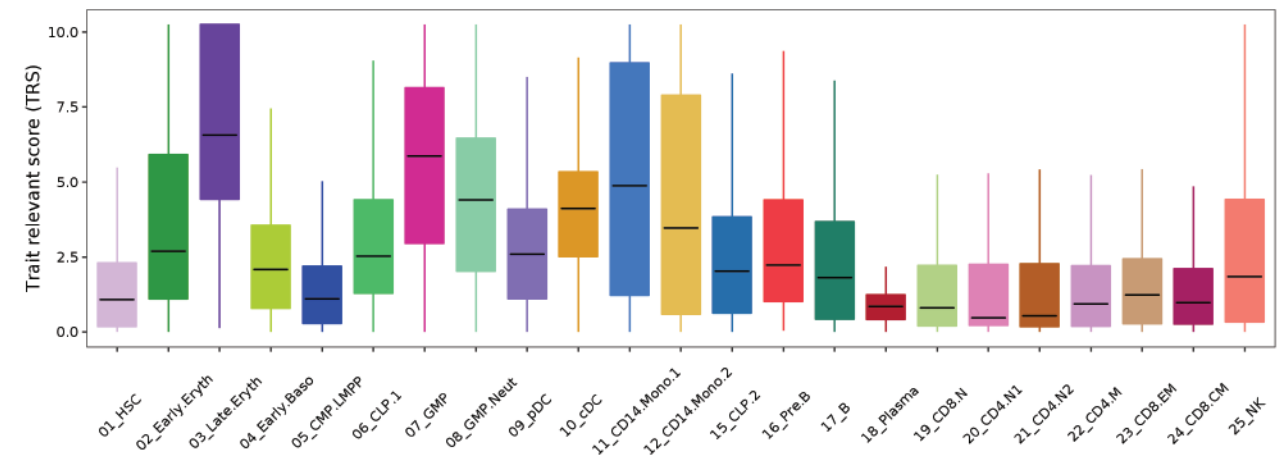
Extended Data Figure 3

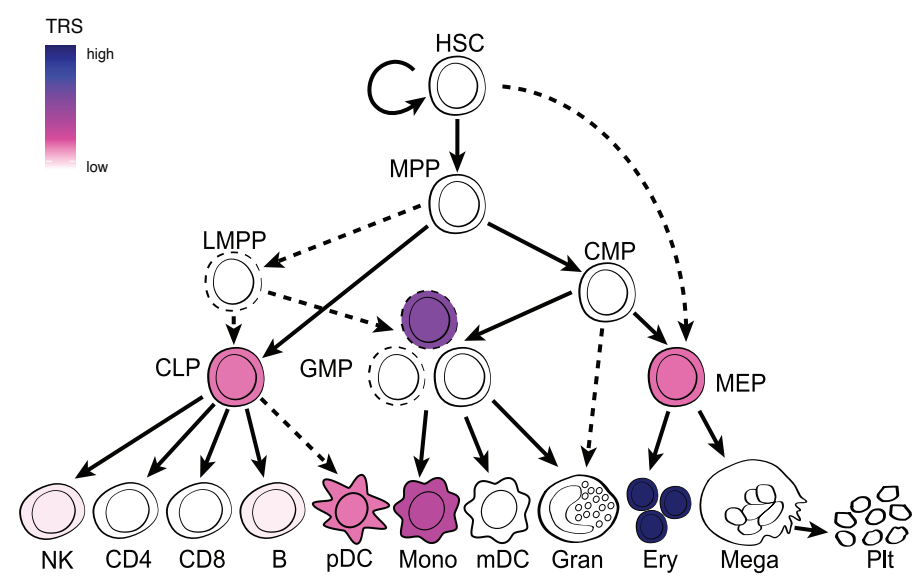
a



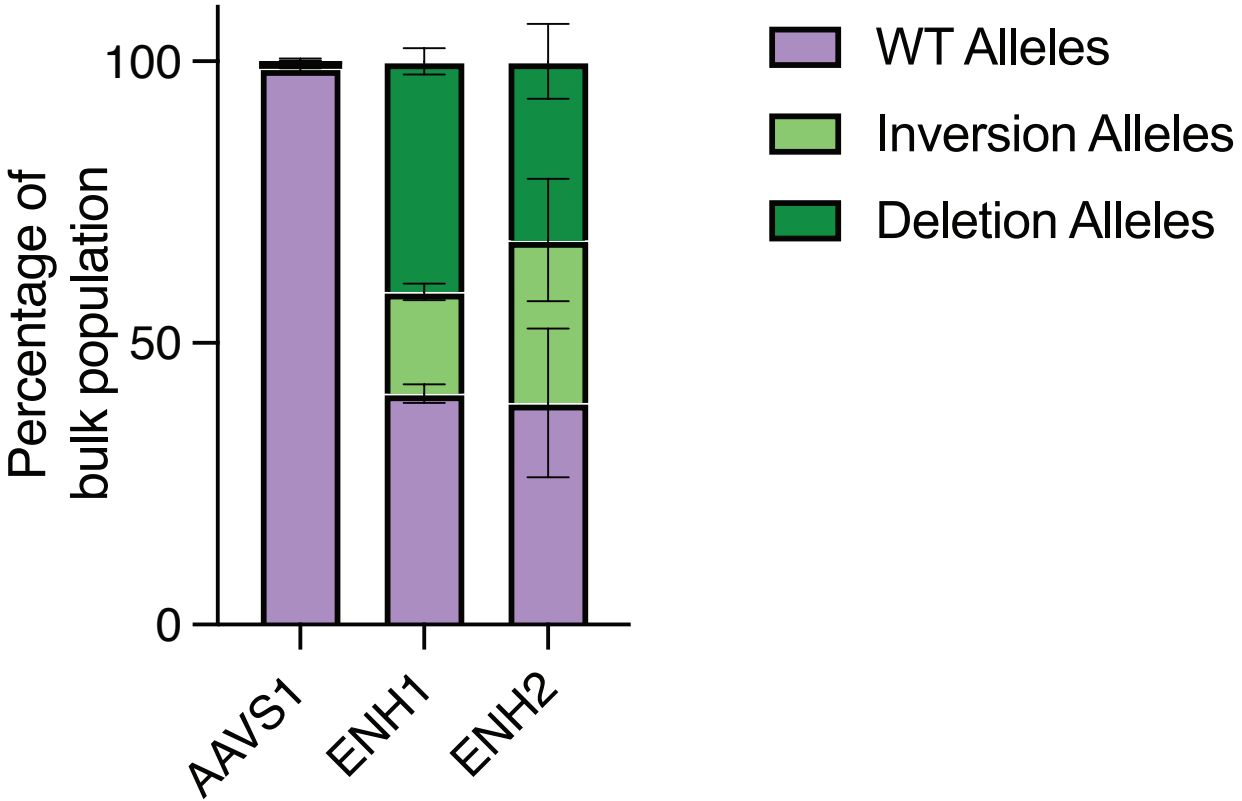


b

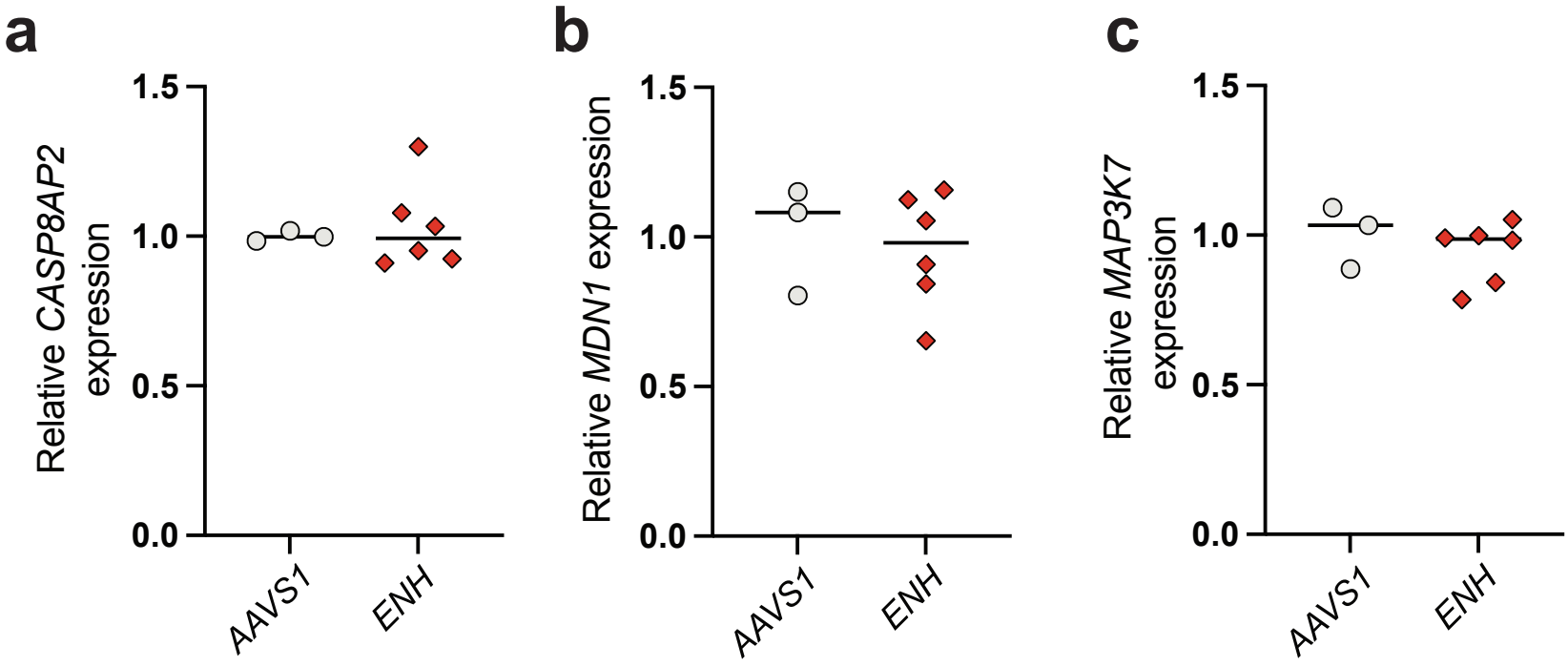




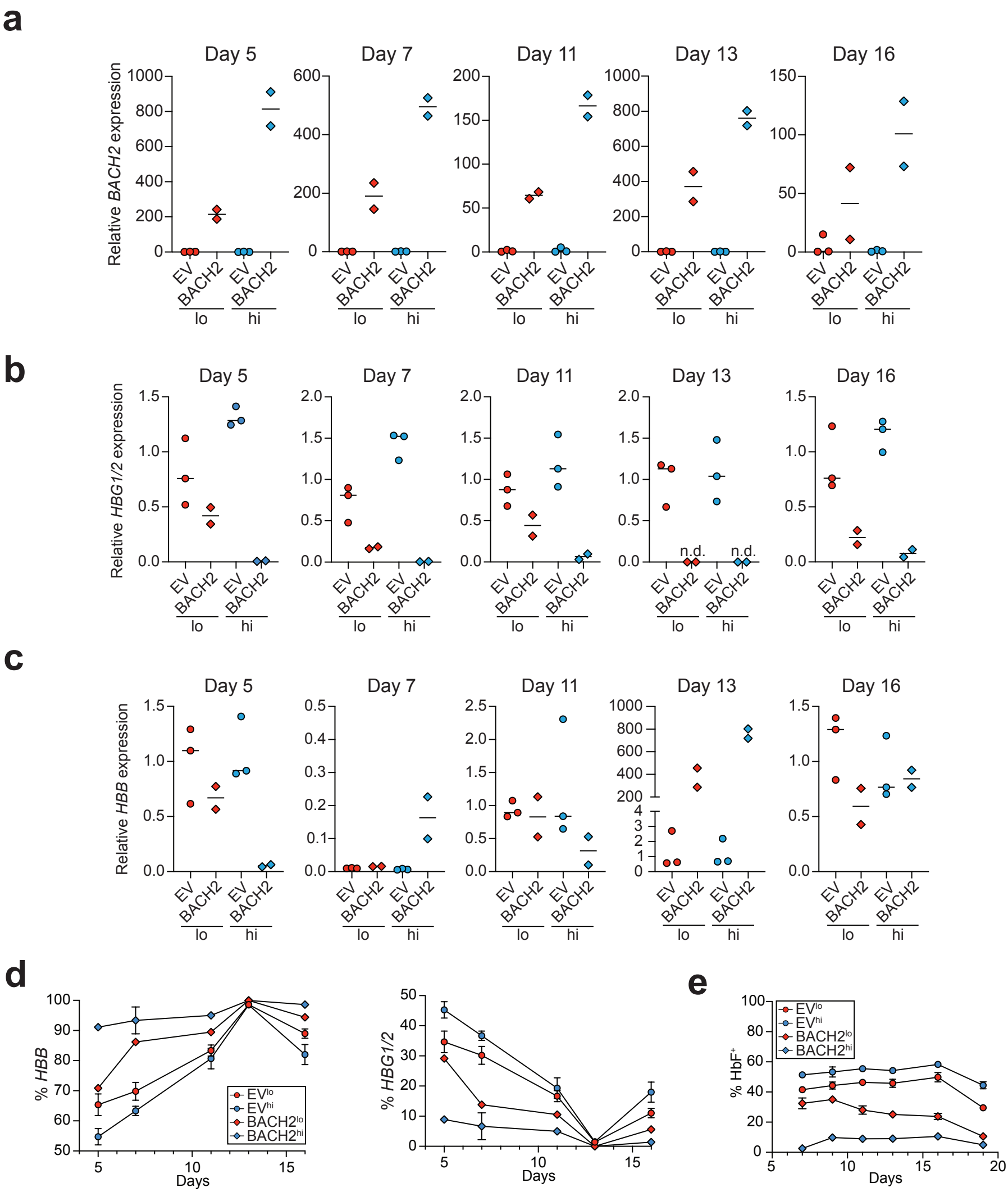
Extended Data Figure 4



Extended Data Figure 5

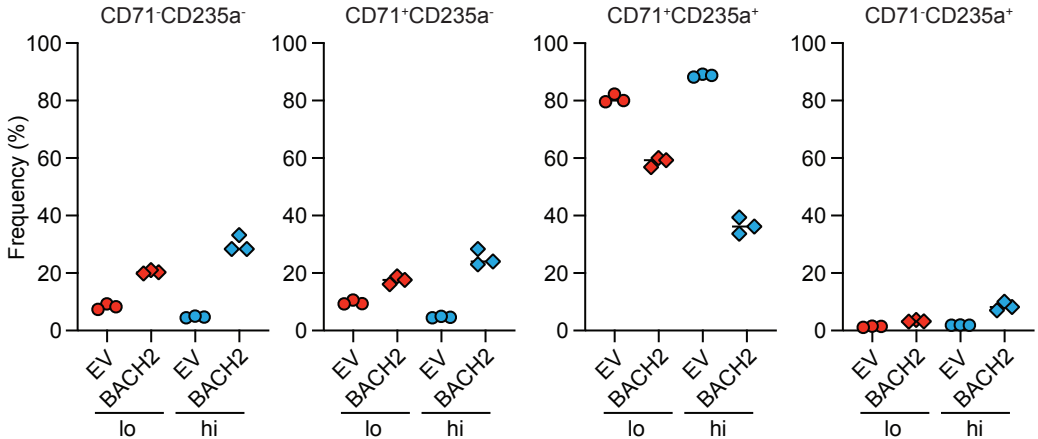


Extended Data Figure 6

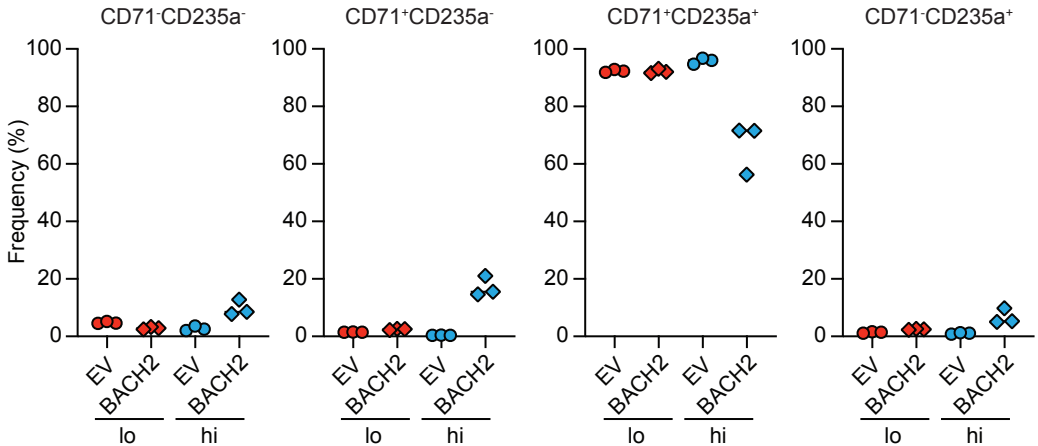


Extended Data Figure 7

Day 6

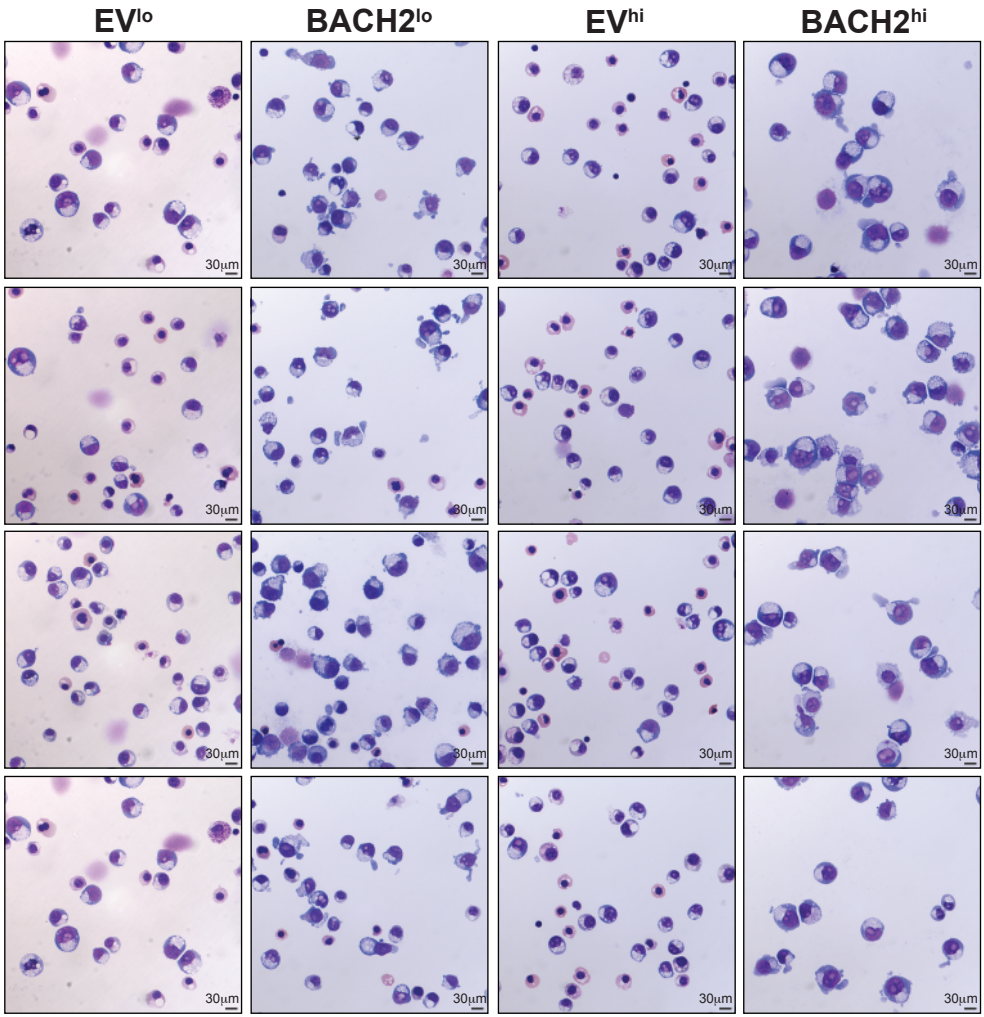


Day 10

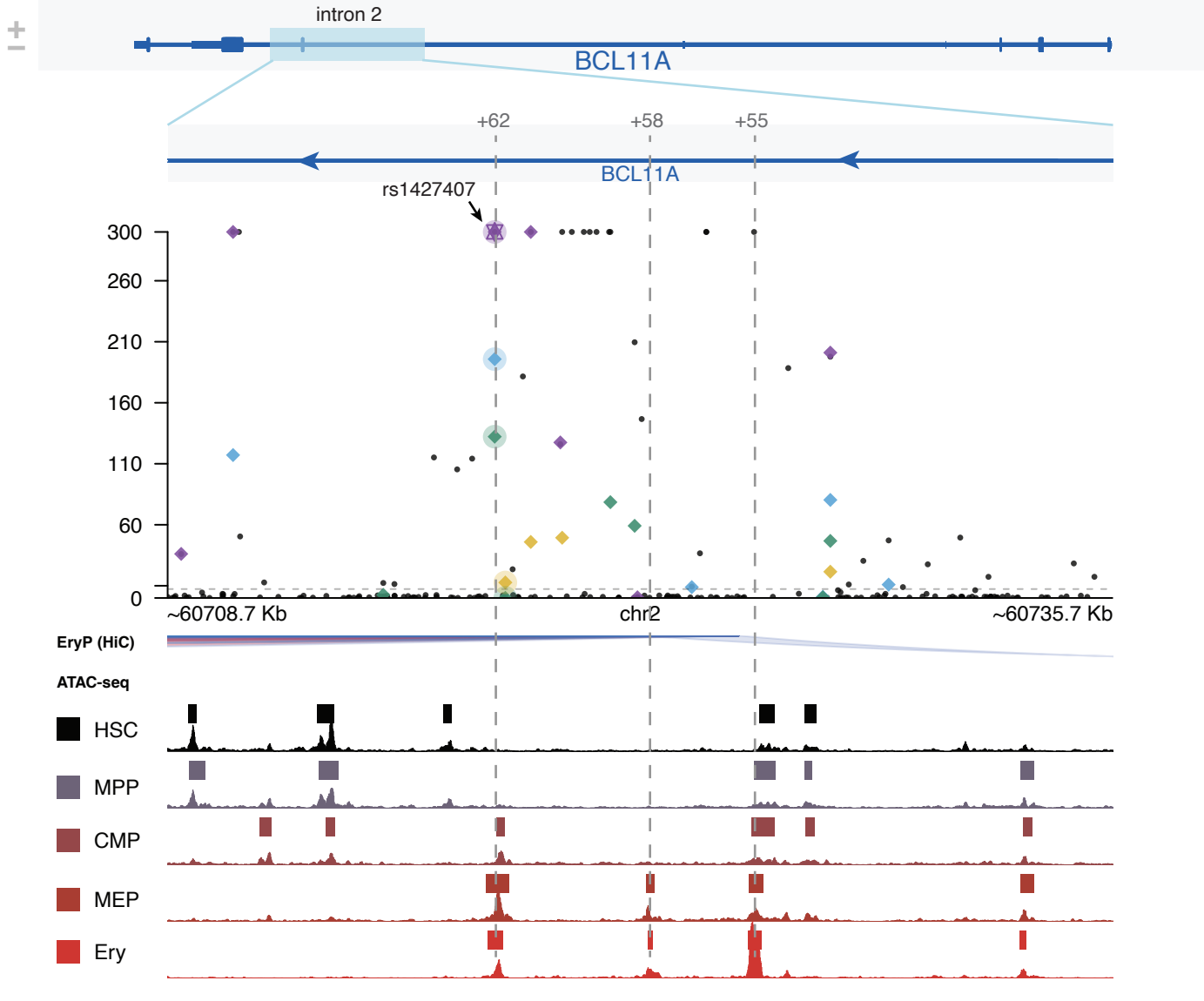


Extended Data Figure 8

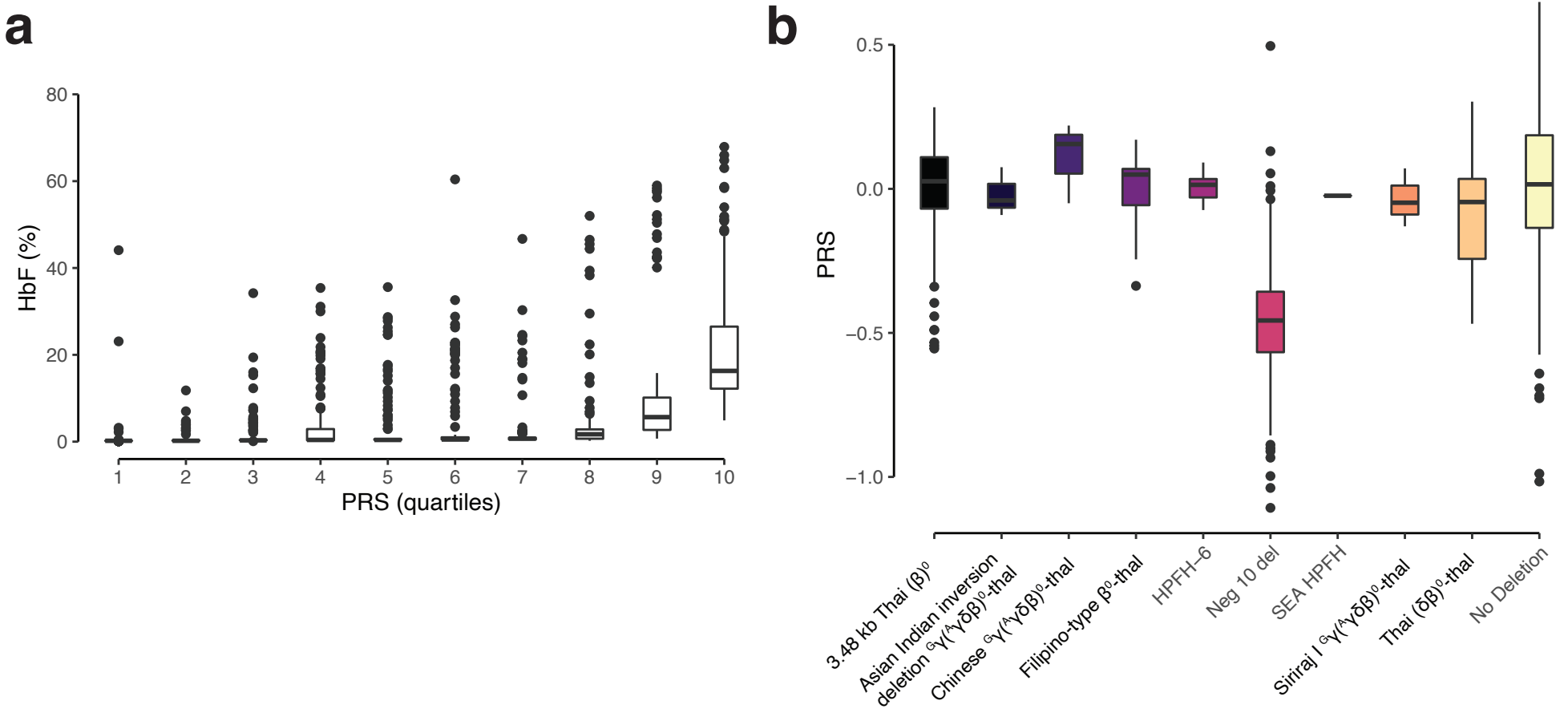
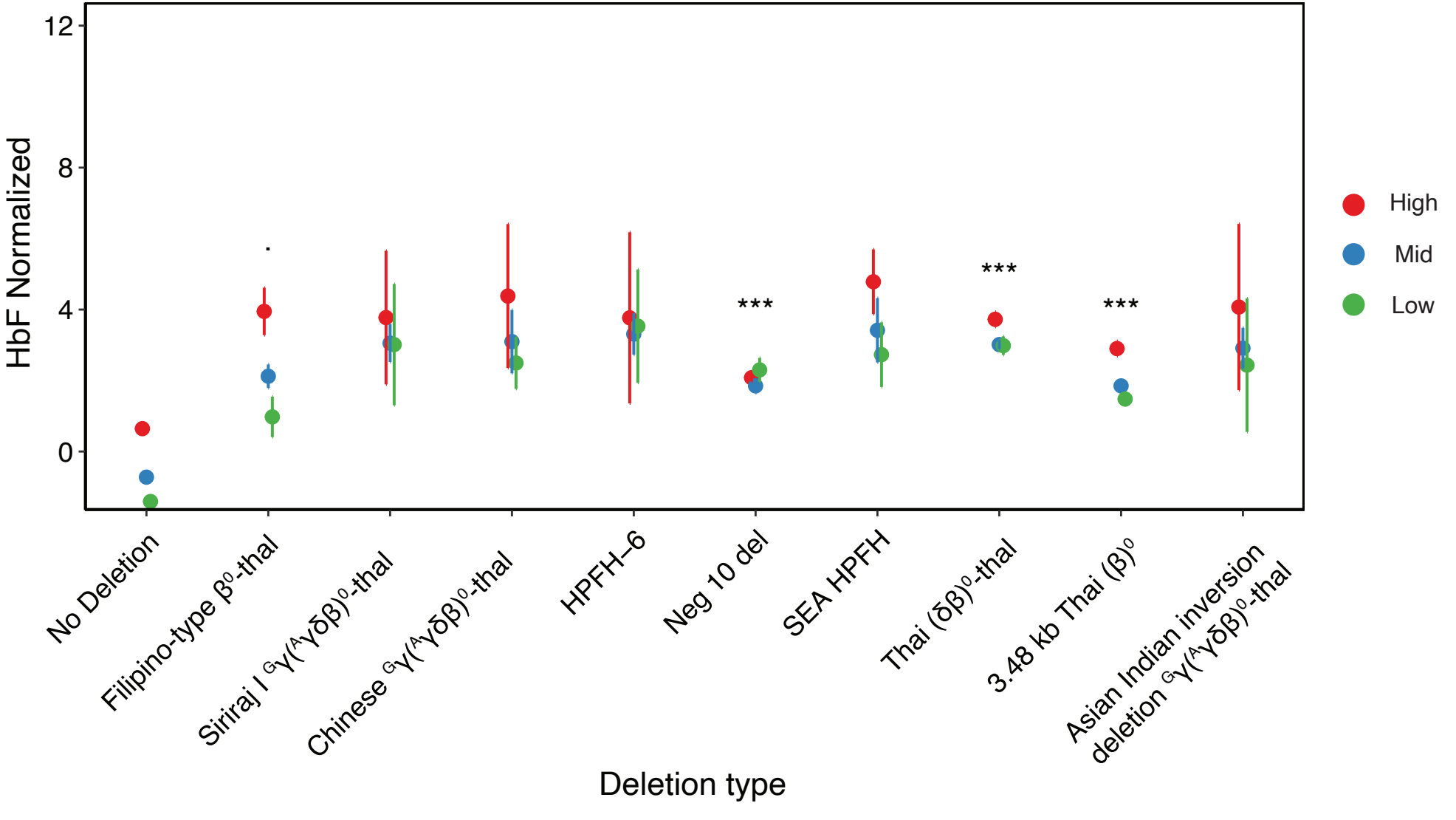
Day 11



Extended Data Figure 9



Extended Data Figure 10



Extended Data Figure 11