Université de Montréal

Deciphering causal genetic determinants of red blood cell traits

par Samuel Lessard

Sciences biomédicales
Faculté de Médecine

Thèse présentée
en vue de l'obtention du grade de doctorat
en sciences biomédicales

4 avril 2017

Résumé

Les études d'association pan-génomiques ont révélé plusieurs variants génétiques associés à des traits complexes. Les mesures érythrocytaires ont souvent fait l'objet de ce genre d'études, étant mesurées de façon routinière et précise. Comprendre comment les variations génétiques influencent ces phénotypes est primordial étant donné leur importance comme marqueurs cliniques et leur influence sur la sévérité de plusieurs maladies. En particulier, des niveaux élevés d'hémoglobine fœtal chez les patients atteints d'anémie falciforme est associé à une réduction des complications et une augmentation de l'espérance de vie. Néanmoins, la majorité des variants génétiques identifiés par ces études tombent à l'intérieur de régions génétiques non-codantes, augmentant la difficulté d'identifier des gènes causaux.

L'objectif premier de ce projet est l'identification et la caractérisation de gènes influençant les traits complexes, et tout particulièrement les traits sanguins. Pour y arriver, j'ai tout d'abord développé une méthode permettant d'identifier et de tester l'effet de gènes knockouts sur les traits anthropométriques. Malgré un échantillon de grande taille, cette approche n'a révélé aucune association. Ensuite, j'ai caractérisé le méthylome et le transcriptome d'érythroblastes différentiés à partir de cellules souches hématopoïétiques et identifié plusieurs gènes potentiellement impliqués dans les programmes érythroïdes fœtaux et adultes. Par ailleurs, j'ai identifié plusieurs micro-ARNs montrant des motifs d'expression spécifiques entre les stages fœtaux et adultes et qui sont enrichis pour des cibles exprimées de façon opposée. Finalement, j'ai identifié plusieurs variants génétiques associés à l'expression de gènes dans les érythroblastes (eQTL). Cette étude a permis d'identifier des variants associés à l'expression du gène *ATP2B4*, qui encode le principal transporteur de calcium des érythrocytes. Ces variants, qui sont également associés à des traits sanguins et à la susceptibilité à la malaria, tombent dans un élément d'ADN spécifique aux cellules érythroïdes. La délétion de cet élément par le système CRISPR/Cas9 induit une forte diminution de l'expression du gène et une augmentation des niveaux de calcium intracellulaires.

En conclusion, des échantillons de génotypages exhaustifs seront nécessaires pour étudier l'effet de gènes knockouts sur les traits complexes. Les érythroblastes montrent de grandes différences au niveau de leur méthylome et transcriptome entre les différents stages développementaux. Ces différences influencent potentiellement la régulation de l'hémoglobine fœtale et impliquent de nombreux micro-ARNs et régions régulatrices non-codantes. Finalement, l'exemple d'*ATP2B4* montre qu'intégrer des études épigénomiques, transcriptomiques et des expériences d'édition de génome est une approche puissante pour caractériser des variants génétiques non-codants. Par ailleurs, ces résultats impliquent ATP2B4 dans l'hydratation des érythroblastes, qui est associé à la susceptibilité à la malaria et la sévérité de l'anémie falciforme. Cibler ATP2B4 de façon thérapeutique pourrait avoir un impact majeur sur ces maladies qui affectent des millions d'individus à travers le monde.

Mots-clés : Anémie falciforme, mesures érythrocytaires, études d'association pan-génomiques, édition de génome, malaria

Abstract

Genome-wide association studies (GWAS) have revealed several genetic variants associated with complex phenotypes. This is the case for red blood cell (RBC) traits, which are particularly amenable to GWAS as they are routinely and accurately measured. Understanding RBC trait variation is important given their significance as clinical markers and modifiers of disease severity. Notably, increased fetal hemoglobin (HbF) production in sickle cell disease (SCD) patients is associated with a higher life expectancy and decreased morbidity. Nonetheless, most variants identified through GWAS fall in non-coding regions of the human genome, increasing the difficulty of identifying causal links.

The main goal of this project was to identify and characterize genes influencing complex traits, and in particular RBC phenotypes. First, I developed an approach to identify and test potential gene knockouts affecting anthropometric traits in a large sample from the general population, which did not yield significant associations. Then, I characterized the DNA methylome and transcriptome of erythroblasts differentiated *ex vivo* from hematopoietic progenitor stem cells (HPSC), and identified several genes potentially implicated in fetal and adult-stage erythroid programs. I also identified microRNAs (miRNA) that show specific developmental expression patterns and that are enriched in inversely expressed targets. Finally, I mapped expression quantitative trait loci (eQTL) in erythroblasts, and identify erythroid-specific eQTLs for *ATP2B4*, the main calcium ATPase of RBCs. These genetic variants are associated with RBC traits and malaria susceptibly, and overlap an erythroid-specific enhancer of *ATP2B4*. Deletion of this regulatory element using CRISPR/Cas9 experiments in human erythroid cells minimized *ATP2B4* expression and increased intracellular calcium levels.

In conclusion, large and comprehensive genotyping datasets will be necessary to test the role of rare gene knockouts on complex phenotypes. The transcriptomes and DNA methylomes of erythroblasts show substantial differences correlating with their developmental stages and that may be implicated in HbF production. These results also suggest a strong implication of erythroid enhancers and miRNAs in developmental stage specificity. Finally, characterizing the erythroid-specific enhancer of *ATP2B4* suggest that integrating epigenomic, transcriptomic and

gene editing experiments can be a powerful approach to characterize non-coding genetic variants. These results implicate *ATP2B4* in erythroid cell hydration, which is associated with malaria susceptibility and SCD severity, suggesting that therapies targeting this gene could impact diseases affecting millions of individuals worldwide.

Keywords: Sickle cell disease, red blood cell traits, genome-wide association studies, genome editing, malaria susceptibility

Table of Contents

List of tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Annex 3

List of figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Annex 1

Annex 2

Annex 3

List of Abbreviations and Acronyms

3C : Chromosome Conformation Capture
5-hmC: 5-hydroxymethylcytosine
5-mC: 5-methylcytosine
AI: Allelic imbalance
ATAC-seq: Assay for Transposase-Accessible Chromatin with high throughput sequencing
BFU: Burst-forming unit-erythroid
BM: Bone marrow erythroblasts
BMI: Body mass index
bp: base pair
cDNA: complementary DNA
CFU: Colony-forming unit-erythroid
ChIA-PET: Chromatin Interaction Analysis by Paired-End Tag Sequencing
ChIP-seq: Chromatin immunoprecipitation followed by sequencing
CMP: Common myeloid progenitors
CNV: Copy number variant
CpG:Cytosine followed by a guanine
CRISPR: Clustered regularly interspaced short palindromic repeats
crRNA: CRISPR RNA
CSSCD: Cooperative study of sickle cell disease
CV: Coefficient of variation
dbGaP: database of Genotypes and Phenotypes
dCas9: dead Cas9
DE: Differential expression
DHS: DNAse hypersensitive site
DM: Differential DNA methylation
DNA: Deoxyribonucleic acid
DNAse-seq: DNase I hypersensitive sites sequencing
DSB: Double strand break
ENCODE: Encyclopedia of DNA Elements
EPO: Erythropoietin
EPOR: Erythropoietin receptor
eQTL: Expression quantitative trait loci
eSpCas9: enhanced Streptococcus pyogenes Cas9
ESP: NHLBI Exome Sequence Project
ExAC: Exome Aggregation Consortium
FAIRE-seq: Formaldehyde-Assisted Isolation of Regulatory Elements followed by sequencing
FACS: Fluorescence-activated cell sorting
FC: Fold change
FDR: False discovery rate
FL: Fetal liver erythroblasts
FPKM: Fragments per Kilobase per million mapped reads
GAGE: Generally Applicable Gene-set Enrichment
GC: Genomic control

GCTA: Genome-wide Complex Trait Analysis
GEO: Gene expression omnibus
GIANT: Genetic Investigation of Anthropometric Traits
GO: Gene ontology
GTEx: Genotype-Tissue Expression
GUIDE-Seq: genome-wide, unbiased identification of DSBs enabled by sequencing
GWAS: Genome-wide association study
h^2: heritability
HbA: Adult hemoglobin
HbC: Hemoglobin C
HbE: Hemoglobin E
HbF: Fetal hemoglobin
HBG: Hemoglobin
HbS: Sickle hemoglobin
HCT: Hematocrit
HDAC: Histone deacetylase
HDR: Homolgy directed repair
HF1 : Streptococcus pyogenes Cas9 high-fidelity variant 1
Hi-C : Chromosome Conformation Capture with high- throughput sequencing
HMM: Hidden Markov models
HPFH: Hereditary persistence of fetal hemoglobin
HPSC: Hematopoietic progenitor stem cells
HRC: Haplotype reference consortium
HSC: Hematopoietic stem cells
HUDEP-2: human erythroid precursor cell line 2
IBD: Identity by descent
ICF: Immunodeficiency, Centromeric instability and Facial anomalies
IHEC: International Human Epigenome Consortium
Indels: Insertion/deletions
IUPAC: International Union of Pure and Applied Chemistry
kb: kilobase
KEGG: Kyoto Encyclopedia of Genes and Genomes
KO: Knockout
LCR: Locus control region
LD: Linkage disequilibrium
LoF: Loss-of-function
log: logarithm
MAF: Minor allele frequency
Mb: Megabase
MCHC: Mean corpuscular hemoglobin concentration
MCH: Mean corpuscular hemoglobin
MCV: Mean corpuscular volume
MEL: Mouse erythroleukemia cells
MEP: Megakariocytes-erythroid progenitors
meQTL: methylation quantitative trait loci
MGI: Mouse Genome Informatics database

MHI: Montreal Heart Institute
miR: micro-RNA
miRNAs: micro-RNAs
MNase-seq: Micrococcal nuclease digestion followed by sequencing
mRNA: Messenger ribonucleic acid
NCBI: National Center for Biotechnology Information
NHEJ: Non-homologous end-joining
NHLBI: National Heart, Lung, and Blood Institute
NSAID: Non-steroid anti-inflammatory drugs
OMIM: Online Mendelian Inheritance in Man
Ortho EB: Orthochromatic erythroblasts
PAM: Protospacer adjacent motif
PCA: Principal component analysis
piRNA: PIWI-interacting RNA
pLI: probability of being loss-of-function intolerant
Poly EB: Polychromatic erythroblasts
P: P-value
Pro EB: Proerythroblasts
qPCR: quantitative polymerase chain reaction
QQ-plot: Quantile-quantile plot
RBC: Red blood cell
RDW: Red blood cell distribution width
REML: Restricted maximum likelihood
RETIC: Reticulocyte count
RET: Reticulocytes
RFN: RNA-guided FokI–dCas9 nuclease
RNAi: RNA interference
RNA Ribonucleic acid
RNA-seq: RNA sequencing
ROI: Region of interest
RT-qPCR: Reverse transcription quantitative polymerase chain reaction
RVIS: Residual Variation Intolerance Score
SAM: synergistic activation mediator
SCD: Sickle cell disease
SCF: Stem cell factor
SD: Standard deviation
SEM: Standard error of the mean
SE: Standard error
SFEM: serum-free expansion medium
sgRNA: single guide RNA
SKAT: Sequence kernel association test
snoRNA: small nucleolar RNAs
SNP: Single nucleotide polymorphism
SpCas9: Streptococcus pyogenes Cas9
T2D: Type 2 diabetes
TAD: Topologically associating domains

TALENS: Transcription Activator-Like Effector Nucleases
TALE: Transcription Activator-Like Effector
TFBS: Transcription factor binding site
TF: Transcription factor
tracrRNA: trans-activating crRNA
µM : microMolar
µL: mircroLiter
UTR: Untranslated region
VEP: Ensembl Variant effect predictor
WES: Whole-exome sequencing
WGS: Whole-genome sequencing
WHR: Waist-hip ratio
WT: Wild type
ZFN: Zinc-finger nuclease
ZFP: Zinc-finger protein

Acknowledgements

I would like to start by expressing my utmost gratitude to my supervisor and mentor, Dr Guillaume Lettre, for giving me the opportunity to undertake my PhD degree in his laboratory, and for giving me the opportunity to work on such a vast range of projects. I am sincerely grateful for his support and guidance, but also for all his trust in my work and ideas. Dr Lettre is an outstanding scientist and an immense inspiration to me.

I wish to thank Dr Daniel E. Bauer, Matthew C. Canver for the wonderful collaborations. I would like to offer a special thanks to Dr Bauer and members of his group for temporarily welcoming me in their lab and mentoring me on CRISPR/Cas9 experiments.

I would also like to thank all my colleagues with whom I shared the last five years: Mélissa Beaudoin, Ken Sin Lo, Geneviève Galarneau, Nathalie Chami, Cécile Low-Kam, Valérie Turcot, Yann Ilboudo, Simon Lalonde, Valérie-Anne Codina-Fauteux, and Jessica Desjardins; Thank you for all the assistance, insightful advice, and amazing lab atmosphere. I would like to offer a special thanks to Mélissa for teaching and guiding on wet lab experiments, which made these projects possible. I would also like to thank Ken for is bioinformatics advice and help with scripting languages. I extend my appreciation to Dr John Rioux and members of his group for their comments and suggestions during lab meetings, and to Sylvain Foisy for sharing the computer cluster, which I used profusely!

I extend my sincere appreciation to the members of my thesis and doctoral exam committees, Dr Gaetan Mayer, Dr Luis Barreiro, and Dr Sébastien Lemieux, for their constructive comments and suggestions. I would also like to express my gratitude to Dr Gaetan Mayer, Dr Luis Barreiro, Dr Tomi M. Pastinen and Dr Catherine Martel, for accepting to be part of my thesis jury and for agreeing to evaluate this work.

I am also thankful to Dr Nicolas Lartillot for introducing me to computational approaches and programming, and to Marie Pageau for inviting me to teach in several bioinformatics classes.

I am also very grateful for the financial support from the Canadian Institutes of Health Research (CIHR), Fonds de Recherche du Québec – Santé (FRQS), Fondation Le Grand Défi Pierre Lavoie, Faculté de Médecine de l'Université de Montréal, Fondation de l'Institut de Cardiologie de Montréal, and Fondation Desjardins.

I offer my immense gratitude to my parents, Johanne and Pierre, and to my brothers Pierre-Alexandre and David, for their support and encouragements during my studies. I also wish to extend my warm thanks to all my friends who supported me and who eagerly celebrated my accomplishments with me, including Nicolas Carignan, Karine Choquet, Alexandre Chouinard, Geneviève Côté, Guillaume Côté, Alix Salvail-Lacoste, Olivier Lamy-Canuel, Alice Lu, Corentin Monfort, Stefany Paulin, Christine Provost, Gaetan Provost, Maxime Kien Duy Quach, Matthieu Rousseau, and Olivier Roy.

Lastly, I want to express my deep appreciation to my life partner Alexandry Calisto for her amazing patience and understanding. Thank you for your immense care and encouragements, and for supporting me every moment of the way.

I want to thank everyone who showed interest in my project and who helped me throughout the course of my under and post-graduate studies.

Chapter 1: Introduction

1.1 Introduction

The first human genome sequence draft was first published in 2001 simultaneously by the International Human Genome Sequencing Consortium and Venter *et al.*[1,2] The human genome is composed of more than 3 billion deoxyribonucleic (DNA) base pairs, which encode an annotated 20,441 protein-coding genes, 20,219 non-coding genes, and 14,606 pseudogenes.[3] Accounting for the different isoforms, the human genome encompasses an annotated 198,002 different transcripts.[3] Cell type specificity is achieved through the differing activity of regulatory elements such as promoters, enhancers and insulators, which are accompanied by epigenomic and DNA tridimensional interactions.

A large portion of inter-individual variation can be explained by genetic variation. Typically, each human genome contains 4-5 million genetic variants, the vast majority of which are single nucleotide polymorphisms (SNP) – sites that differ at only one DNA base pair.[4] Other genetic variation include small insertion and deletions (indels), and structural variants such as large deletions, copy number variants, and inversions. Latest whole-genome sequencing studies estimate that each individual carries around ~38 *de novo* mutations, assuming a germline mutation rate of ~1.2×10^{-8} per bp per generation.[5-8] Most genetic variants carried by individuals are common with a minor allele frequency (MAF) > 5%, with only 1-4% of variants with a MAF < 0.05%.[4] Healthy individuals carry ~100 loss-of-function (LoF) variants predicted to disrupt the function of protein-coding genes, although most are common and found in non-essential genes.[9] Nonetheless, genetic mutations can have drastic impact on human health, as demonstrated by diseases cataloged in the Online Mendelian Inheritance in Man (OMIM) database such as sickle cell disease and cystic fibrosis.[10]

In contrast to monogenic diseases, which are caused by one specific genetic mutation, complex traits can be explained by tens or hundreds of genetic variants of generally small effects, with a sizable proportion influenced by the environment. For instance, one of the most studied trait, height, has been associated with more than 700 genetic variants.[11,12] Other generally well-studied traits include hematological traits. A recent study reported >2,500

variants associated with 36 red cells, white cells, and platelet indices.[13] Understanding the genetic architecture of complex trait can yield important insights into their biology. Blood cells are implicated in a vast set of functions such as oxygen transport, immunity and thrombosis. Because of this, dysregulation of blood cell traits can lead to many disorders such as anemia, immunodeficiency, hemophilia, and cancer. Thus, studies of complex traits can be very valuable for the identification of therapeutic targets for related human disorders.

1.2 Complex trait genetics

Gregory Mendel's famous work on laws of inheritance was rediscovered in the early 20th century. Mendelian inheritance stipulates that a given phenotype can be explained by transmission of a specific allele. This, however, could seemingly not explain quantitative traits such as height. Ronald A. Fisher 1918's paper reconciled these notions by proposing that quantitative traits can be explained by Mendelian inheritance if they were seen as the additive effect of multiple loci, each with a small effect on the phenotype.[14,15] Even when a trait is binary like in common diseases such as heart disease and type 2 diabetes, they can still be explained by the additive effect of several variants, each of which with increasing the risk of disease. Complex traits can also be influenced by non-genetic factors (environment) and non-additive effects such as gene-gene (epistasis) interactions and gene-environment interactions.

The first genetic linkage maps of *Drosophilia* phenotypes developed by Thomas Hunt Morgan and his student Alfred Sturtevant laid the foundation for genetic mapping that is the identification of genetic loci that correlates with a given phenotype.[16,17] Human linkage studies were very successful in identifying genetic markers segregating within pedigrees with monogenic diseases such as Marfan syndrome.[18] Linkage studies also were able to detect loci linked with common diseases such as inflammatory bowel disease and schizophrenia, albeit with a more limited success.[19] This is because mutations causing monogenic disorders are generally rare and have high penetrance, whereas common diseases can be explained by multiple variants with small effects and low penetrance. Thus, linkage studies are underpowered to map

variants influencing common diseases. In addition, linkage studies were often conducted with relatively low-density genetic markers (e.g. microsatellites spaced ~10 centimorgans apart). The low-resolution of linkage maps made it particularly difficult to find potential causal genes.

1.2.1 Genome-wide associations studies

An alternative to linkage studies are genetic association studies, where a genetic variant is statistically correlated with a trait or disease. When the trait is dichotomous, like in case-control studies, the frequency of the variant is compared between each group. As opposed to linkage study, association studies are usually carried in unrelated samples. Association studies can be conducted on a candidate-gene basis, but require strong hypothesis about the implication of the gene on the trait or disease of interest. Genome-wide association studies (GWAS), on the other hand, aim to survey the genome globally and thus do not make assumptions about the genes associated with the phenotype. First GWAS were made possible following the cataloging of SNPs by early sequencing studies and their decreasing genotyping cost, which made it possible to capture a large fraction of the common variation of the human genome.[19]

GWAS have been very successful in identifying genetic variants associated with traits or diseases. As of January 2017, the GWAS catalog contains 30,593 unique variant-trait associations identified from 2,701 different studies (**Figure 1**).[20] Arguably one of the first GWAS interrogated less than 100,000 SNPs for association with myocardial infarction, and successfully identified a SNP in the lymphotoxin-α (*LTA*) gene.[21] This study included 94 cases and 658 controls, which contrasts with recent GWAS which can include hundreds of thousands of samples and interrogate several million variants.[12,13]

GWAS take advantage of the local non-random correlation between genetic variants, defined as linkage disequilibrium (LD). LD is created by mutations and evolutionary forces such as genetic drift and selection. LD is broken by recombination and is inversely correlated with the distance between variants.[22] LD is what makes GWAS practical: Because variants can tag

neighboring variants, it is not necessary to genotype them all. Rather, genotyping a subset of variants is enough to tag most of the common variation.[22] Efforts such as the International HapMap project aimed at cataloging human genetic variations. Phase I of the HapMap project genotyped ~1 million SNPs in 269 samples from an African, a European, a Japanese, and a Chinese population.[23] They estimated that around 500,000 SNPs are enough to tag common variation in individuals of European descent. [22,23]

Figure 1. Genetic variants associated with traits in genome-wide association studies (GWAS).
SNPs associated with hematological measurements are highlighted in dark blue. Downloaded from www.ebi.ac.uk/gwas, accessed April 4th 2017.[20]

These reference panels can also be useful to carry genotype imputation, which aim to extend the number of variants in GWAS studies by estimating the identity of non-genotyped variants using LD. Imputation can lead to increased power to detect association, and is useful for genetic-fine mapping and meta-analyses.[24] The 1,000 genomes project phase 3 panel includes over 88 million variants from 26 populations, and includes structural variants and short indels.[4] The haplotype reference consortium (HRC) contains 64,976 human haplotypes composed of ~40 million SNPs, and can be used to accurately impute variants down to 0.1% frequency.[25] Whole-genome sequencing (WGS) studies are thus very valuable for GWAS, and

the advent of population-specific imputation panels will further increase accuracy of imputation.[26-33]

1.2.2 Heritability

Observed complex phenotypes are due to the sum of genetic and environmental factors. Thus, the phenotypic variance (σ^2_P) can be explained by the sum of genetic (σ^2_G) and environmental variance (σ^2_E):

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2$$

Heritability represents the proportion of the phenotypic variance that is explained by genetic variation. Broad-sense heritability (H^2) represent the contribution of all genetic effects on the phenotypic variance, including dominance and epistatic effects ($H^2=\sigma^2_G/\sigma^2_P$).[34] *Narrow-sense heritability* (h^2) is defined as the proportion of phenotypic variance explained by additive genetic variance only (σ^2_A; $h^2=\sigma^2_A/\sigma^2_P$). These heritability definitions do not account for gene and environment co-variation or gene-environment interactions. An example of gene-environment interaction that has been described is the interaction of diet and genetic variants on obesity measures.[35,36] Heritability is not necessarily constant and can be influence by changes in environmental factors. For instance, increased physical activity may reduce heritability of traits such as body mass index (BMI).[37] Heritability can also be modified by natural selection and inbreeding.[34]

Various approaches can be used to estimate the heritability of a trait. For example, in parent-offspring design studies, heritability is obtained by estimating the slope (β) of a linear regression between the mid-parent trait (the average of the parent's trait) and the trait of the offspring. This assumes that shared phenotype values between parent and offspring are due to genetic factors, and not to shared environment, which can inflate heritability estimates.[34] Another approach are twin studies that use monozygotic and dizygotic twins. Both are expected to share the same environment, but monozygotic twins share twice as much of their genetics.

Hence, heritability can be estimated by taking twice the difference of the correlation between monozygotic (r_{MZ}) and dizygotic (r_{DZ}) twins[38]:

$$heritability = 2 \times (r_{MZ} - r_{DZ})$$

For example, twin studies estimate the heritability of height at around 80%.[39] A meta-analysis of 2,748 twin studies including over 14 million subjects reported that the average heritability of 17,804 human traits was around 49%, the most heritable being ophthalmological, dermatological, and skeletal traits (mean h^2>0.59).[40] For most traits (69%), the correlation between twins were consistent with most of the heritability being explained by additive factors.[40]

Heritability can also be estimated using identity by descent (IBD). IBD corresponds to part of the genome that are shared between individuals and that have been inherited by a common ancestor. The idea behind this method is that siblings will share around 50% of their genome IBD with some variation (~4%) due to random segregation and recombination. Siblings with higher IBD are expected to be more similar to siblings with lower shared IBD. Heritability can thus be estimated by using the correlation between phenotypic similarity and the proportion of the genome shared IBD.[34,41] Tools like Genome-wide Complex Trait Analysis (GCTA) use the correlation between unrelated samples to estimate heritability.[42,43] This method uses a linear model to fit the GWAS SNP data to the phenotype and uses restricted maximum likelihood (REML) to estimate the variance explained by these variants. These methods require large datasets to obtain estimates with reasonable precision, and are limited to the variants tagged by the GWAS dataset. The LD score regression method estimates heritability from summary statistics and can be used to partition heritability on function annotations.[44]

1.2.3 Missing heritability

GWAS of height in 253,288 individuals identified ~700 variants associated with this trait.[11] Even though a large number of variants were identified, they explained ~16% of the phenotypic variance, leaving a large gap with the estimated heritability of height of 80%.[11,42]

This “missing” heritability can be in part explained by variants that do not reach statistical significance. Including the ~9,500 most strongly associated variants increased the variance explained to 29%, and all common variants capture ~60% of the heritability, suggesting that even larger sample size will be required to identify common variants with very small effects.

Rare and low-frequency variants may also explain a part of the missing heritability. Although these variants are expected to have greater effect sizes, very large study are needed to capture enough samples with the low-frequency or rare variants to have power to detect the association. A recent study using the ExomeChip genotyping array in >700,000 samples identified 606 independent variants associated with height, including 83 with low frequency (minor allele frequency (MAF) <5%).[12] New loci identified by this study explained an additional 4.1% of the heritability.[12]

Imprecise phenotyping can also influence heritability measures, and better phenotyping can lead to increased power to detect associations. Power can also be increased by using quantitative phenotypes instead of categorical phenotypes, for example by replacing type 2 diabetes status by blood glucose levels.[45] Joint testing of multiple correlated phenotypes can also increase power.[46] Missing heritability can also reside in variants not captured by GWAS studies such as structural variants and mitochondrial DNA variants.[45] Genetic fine-mapping can reveal new local associations and increase the explained heritability of GWAS loci.[47]

1.3 Red blood cell traits

The adult bone marrow produces 2.4 million red blood cells (RBC) per seconds.[48] RBCs, also called erythrocytes, are the most common cell type in the human body. They are responsible for the transport of oxygen (O_2) from the lungs to the tissues, and for the transport of carbon dioxide (CO_2) back from the tissues to the lungs. Mature RBCs are rich in hemoglobin, a biomolecule responsible for carrying O_2 and CO_2. Erythropoiesis is the process of proliferation and differentiation of hematopoietic progenitor stem cells (HPSC) into mature RBCs. It is a

highly regulated process, as it has to accommodate the limited life span of erythrocytes (100-120 days), as well as physiological and pathological stresses such as changes in altitude or bleeding. Aside from their vital role in oxygen delivery to the organs, RBCs also serve roles in vascular tonus regulation, pathogen resistance, inflammatory response, and thrombosis.[49-54]

Although erythropoiesis is a well-regulated process, deviation from normal RBC indices can reflect conditions such as anemia, which may in turn reflect disorders such as cardiovascular or liver diseases. Thus, it is not surprising that blood indices are routinely monitored to assess the health of patients. Several indices are routinely measured in a standard complete blood count (**Table 1**).[55] The number of RBCs is measured with RBC counts ($\times 10^6$ cells/μL). The hematocrit (HCT) represents the fraction of blood occupied by RBCs (% volume) whereas the relative amount of hemoglobin is measured with hemoglobin concentration (HBG, g/dL). At the cellular level, mean corpuscular hemoglobin (MCH, pg), mean corpuscular volume (MCV, fL), and MCH concentration (MCHC, g/dL) represent the amount of hemoglobin, the average volume, and hemoglobin concentration of RBCs respectively. Finally, the RBC distribution width (RDW), presented as the coefficient of variation (RDW-CV, %) or the standard deviation (RDW-SD, fL), represents the range of the RBC volume distribution.

RBC indices are highly heritable (40-90%). Because RBC traits are routinely and accurately measured, they are amenable to genetic association studies and can be straightforwardly applied to large cohorts.[55] Moreover, functional characterization is possible in model organisms and human erythroid cell culture systems as these assays are well developed and phenotypes are usually cell-autonomous. [55-61] Large-scale genetic association studies have been performed.[13,61,62] A GWAS in 135,367 individuals of European and South Asian ancestry identified 75 loci associated with RBC indices, explaining ~4-9% of the phenotypic variation.[61] Individuals with a high genetic risk score derived from these SNP were more likely to have high hemoglobin levels associated with adverse outcomes in individuals with various conditions such as neurological and cardiovascular diseases.[61,63,64] Chami *et al* reported 16 additional variants associated with RBC traits in 130,273 samples, including 23,896 individuals of non-European

descent.[54] This study focused on rare coding variants, and identified genes implicated in monogenic disorders such as *ALAS2*, where mutations cause X-linked sideroblastic anemia and erythropoietic protoporphyria.[10,54] More recently, Astle *et al* reported 2,706 loci, defined as high LD groups, associated with hematological traits in 173,480 individuals of European ancestry. This GWAS used comprehensive imputation panels to test ~29.5 million markers down to a minor allele frequency of 0.01%.[13] This study estimated that common variants captured 10-28% of the variance in RBC indices, and identified variants associated with pleiotropic effects on multiple hematological traits. For example, *SH2B3* was associated with platelet, red blood cell and white blood cell phenotypes.[13] Although comprehensive, these studies focused mainly of individuals of European descent. The largest study of RBC traits in individuals of African descent included 16,500 samples.[62] Because of positive selection, individuals in malaria-stricken regions have high frequency of variants that modify the susceptibility to this disease, such as mutations causing glucose-6-phosphate dehydrogenase (*G6PD*) deficiency, which can strongly influence RBC traits.[62,65] This highlights the importance of conducting association studies in different populations, especially since GWAS discoveries are generally transferable across ethnicities.

1.3.1 Erythropoiesis and hemoglobin synthesis

The first organ responsible for erythropoiesis is the yolk sac, which releases primitive, macrocytic, RBCs in the developing blood stream.[48] The source of RBC eventually switches to the fetal liver (FL), which produces definitive RBCs.[66] In contrast to primitive RBCs, definitive RBC progenitors proliferate before entering the blood stream enucleated.[67] The months following birth, the site of erythropoiesis switches again to the bone marrow (BM), and to a lesser extent, the spleen. The latter being mainly active as an additional erythropoietic resource under stress conditions like anemia.[68]

Table 1. Red blood cell traits and examples of related conditions[69,70]

Trait	Description	Unit	Normal Range	Conditions
Red blood cell (RBC) count	Count of red blood cell per unit	$\times 10^6$ cells/μL	Female: 4.2-5.0 Male:4.6-6.0	
Hemoglobin (HBG)	Hemoglobin concentration	g/dL	Female: 12.0-15.8 Male: 13.3-16.2	Anemia, erythrocytosis
Hematocrit (HCT)	Fraction of blood volume occupied by RBCs	%	Female: 35.4-44.4 Male: 38.8-46.4	
Mean corpuscular hemoglobin (MCH)	Amount of HBG per RBC	pg	26.7-31.9	Micro-, normo-, or macrocytic anemia
Mean corpuscular volume (MCV)	Average RBC volume	fL	79-98	
MCH concentration (MCHC)	HBG divided by HCT	g/dL	32-36	Hypo- or normochromic anemia; spherocytosis.
RBC distribution width (RDW)	Variation of RBC volume in standard deviation (SD) or coefficient of variation(CV)	fL or %	SD: 35-45 CV: 11.5-14.5	Anisocytosis; nutritional deficiencies
Reticulocyte counts (RETIC)	Fraction of reticulocytes	%	0.8-2.3	Hemolytic anemia, aplastic anemia
Hemoglobin fractions	Fraction of HbA, HbA_2, HbF, and hemoglobin variants).	%	HbA: 96.5-98.5 HbA_2: 1.5-3.5 HbF: 0-1	Hereditary persistence of fetal hemoglobin (HPFH); Sickle cell disease; β-thalassemia.

The switch of organ carrying erythropoiesis is correlated with a switch of the main type of hemoglobin produced. Hemoglobin is a tetramer composed of two beta and two alpha chains. Each of these globin chains contain a heme prosthetic group covalently bound by a histidine.[48] Heme is composed of protoporphyrin IX with an iron atom at its center. The iron can bind O_2

and CO_2 when in its ferrous state (Fe^{2+}).[48] Embryonic hemoglobin is mainly composed of the α-chain ζ, and the β-chain ε ($\zeta_2\varepsilon_2$).[71] Hemoglobin produced by fetal and adult erythrocytes are fetal hemoglobin (HbF; $\alpha_2\gamma_2$) and adult hemoglobin (HbA, $\alpha_2\beta_2$), respectively. The production of γ-globin is characteristics of anthropoid primates, and may confer a competitive advantage to the fetus as HbF has a higher affinity to oxygen than HbA, produced by the mother. Other types of hemoglobin are produced at lower levels in each developmental stages. For instance, HbA_2 is composed δ-globin instead of β-globin, and comprises around 2% of adult hemoglobin. HbF is also found in healthy adults at usually less than 1% (**Figure 2**).[72]

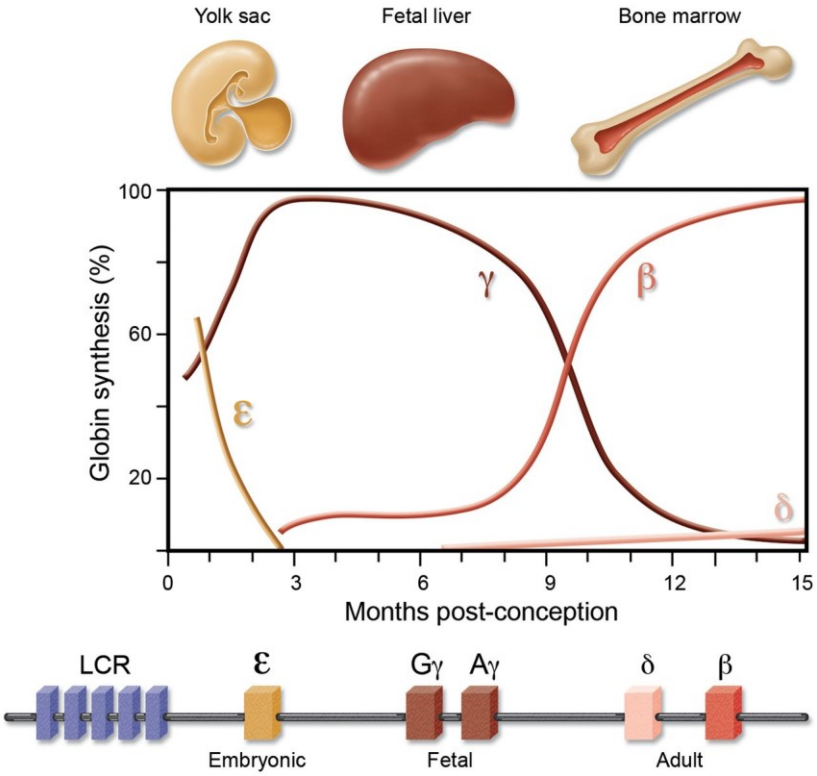


Figure 2. Switch of hemoglobin β-chains
Embryonic ε-globin, encoded by the *HBE* gene, is replaced by the fetal γ-globin at the first semester post-conception. Two types of γ-globin exists (Gγ and Aγ), which are encoded by *HBG2* and *HBG1*. Around birth, β-globin, encoded by *HBB*, becomes the major β-chain. Adults also produce low levels of δ-globin, encoded by the *HBD* gene. Expression of the globin genes is regulated by of the locus control region (LCR). α-chains (not shown) are produced by a different locus on chromosome 16. The α-chain ζ-globin is expressed during the early embryonic stage, then permanently replaced by α-globin. Adapted from Bauer *et al.*[73]

The α- and β-*globin* genes are found in two separate gene clusters (**Figure 2**). The β-*globin* cluster contains the ε, γ, δ, and β globins. Note that two different γ-*globins* exist, denoted

as Gγ and Aγ, which are encoded by two different genes (*HBG2* and *HBG1* respectively). Gγ-globin predominates on Aγ-globin. The β-*globin* cluster is under the control of the locus control region (LCR), an element that confers erythroid-specific and high-level expression of the β-*globin* genes. It is stipulated that the transcription of β-*globin* genes is activated by looping of the LCR to the promoter of the gene to be expressed – each gene competing with each other.[71,74]

HSC progenitors are directed to the myeloid lineage by giving rise to common myeloid progenitor (CMP) cells.[75-77] CMP are then transformed to megakaryocytes-erythroid progenitors (MEP), which give rise to committed erythroid progenitor cells (**Figure 3**).[75-77] Erythroid progenitors can be subdivided as burst-forming unit-erythroid (BFU-E), which differentiate to colony-forming unit-erythroid (CFU-E).[71,78] The latter finally differentiate into erythroblasts, which enucleate and give rise to erythrocytes.[71,78] The growth and differentiation of BFU-E to CFU-E and latter stages is dependent on multiple growth factors, including stem cell factor (SCF), thrombopoietin (TPO), interleukins 3, 6 and 11 (IL-3, IL-6, IL-11), insulin-like growth factor 1 (IGF1), glucocorticoids, and erythropoietin (EPO). [71,78]

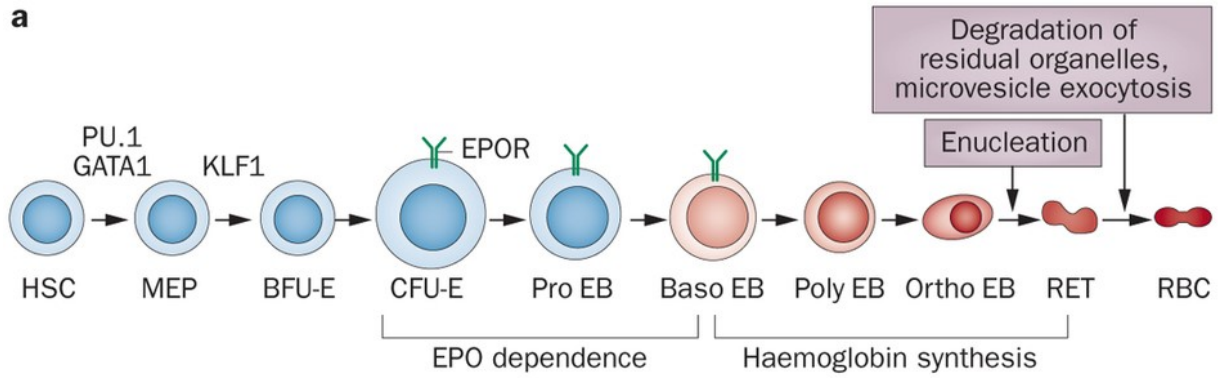


Figure 3. Stages of erythroid differentiation.
PU.1, GATA1 and KLF1 are involved in erythroid cell fate determination from hematopoietic stem cells (HSC). Erythropoietin (EPO) dependence ends at the early basophilic erythroblast (Baso EB) phase, and is followed by the initiation of hemoglobin synthesis. MEP: Myeloid early progenitors; BFU-E: Burst-forming unit-erythroid; CFU-E: Colony-forming unit-erythroid; Pro EB: Proerythroblasts; Baso-EB: Basophilic erythroblasts; Poly EB: Polychromatic erythroblasts; Ortho EB: Orthochromatic erythroblasts; RET: Reticulocytes; RBC: Red blood cells. Adapted from Koury & Haase.[79]

RBC production is tightly regulated. Under hypoxia, levels of EPO increases. Binding of EPO to its receptor (EPO receptor, EPOR) to erythroid progenitor cells leads to increased expression of erythroid program genes, stimulating differentiation. Production and differentiation of RBCs are also tightly linked with heme production and thus to iron availability. For example, transcripts implicated in heme biosynthesis, such as the mRNA of δ-aminolevulinate synthase 2 (*ALAS2*), contain iron response elements bound by iron response protein that block translation until sufficient intracellular iron concentration.[80] Because of its importance, iron is highly conserved and recycled, leading to daily losses of iron of only about 1-2mg representing 0.1% of total iron in the adult human body.[81] Production of hepcidin by hepatocytes limits availability of transferrin-bound plasma iron, which is used for erythropoiesis. Hepcidin acts by lowering absorption of dietary iron, recycling of iron by macrophages, and increasing storage of iron in hepatocytes via the degradation of ferroportin, responsible for iron efflux. Conversely, increased erythropoietic demands leads to secretion of erythroferrone into the plasma, repressing liver hepcidin production, and increased plasma iron availability.[82]

1.3.2 Diseases related to red blood cell traits

Given the necessarily stringent regulation and coordination of erythropoiesis, it is not surprising that several disorders are associated with changes in RBC traits (**Table 1**). Erythrocytosis is a condition characterized by a higher than normal hematocrit, and can be caused by inherited mutations causing increased response of erythroid progenitor cells to circulating cytokines such as erythropoietin (EPO) (https://omim.org/entry/133100). Increased RBC can also indicate myeloproliferative disorders due to somatic mutations in erythroid progenitor cells such as polycythemia vera, which is caused by mutations in *JAK2* that promote clonal proliferation (https://omim.org/entry/133100). Increased RBC counts can also indicate chronic hypoxemia due to lung disease or congenital heart defects.

Conversely, low RBC, HCT, or HBG indicates anemia, which can be due to multiple factors including inherited genetic defects, leukemia, kidney diseases, and nutritional

deficiencies limiting hemoglobin production such as vitamin B12 or iron deficiency.[55,83,84] Anemia can either result or be a consequence of cardiovascular diseases.[83] Congestive heart failure can be caused by severe anemia. Conversely, anemia in patients with chronic heart failure is associated with increased comorbidities and can be caused by renal dysfunction, increased proinflammatory cytokines, and impaired iron absorption.[83] Gene mutations leading to anemia can hinder the production of RBCs such as in Diamond-Blackfand anemia which is due to mutations in genes encoding ribosomal proteins.[10] Other mutations lead to decrease hemoglobin production. This is the case of genes implicated in heme biosynthesis such as *ALAS2*. Mutation in this gene leads to X-linked sideroblastic anemia, characterized by decreased heme production and iron overload.[85-87] Other examples are α- and β-thalassemia, caused by mutation in *HBA1*/*HBA2* and *HBB* respectively. Moreover, MCV, MCH, MCHC, and RDW can distinguish different types of anemia (**Table 1**). For example, iron deficiency is characterized by microcytic (low MCV) and hypochromic (low MCHC) RBCs.[84] Increased RDW is also an independent predictor of mortality due to cardiovascular or kidney diseases, obesity, and cancer.[83,88-91] Even in healthy individuals, high RDW is associated with higher mortality.[92] Finally, anemia can results from misshapen erythrocytes. This is the case of hereditary spherocytosis, which can be caused by mutations in *ANK1,* or sickle cell disease (SCD) cause by a mutation in *HBB.*[10,93]

1.3.3 Malaria

In 2015, the World Malaria Report estimated 212 million cases of malaria infection with 438,000 related deaths. Although the number of cases is decreasing, it is still heavy in developing regions, and especially in Africa where ~90% of all deaths occurred in 2015. The disease is most fatal in children, as their immunity is not fully developed: 70% of deaths were from children under 5. The disease is spread by common females *Anopheles* mosquitoes, making around 3.2 billion individuals at risk. It is caused by protozoans of the *Plasmodium* genius, in particular *P. falciparum* and *P. vivax*.[94]

The life cycle of these parasites is complex (**Figure 4**). Briefly, they first enter the bloodstream in the form of sporozoites following an *Anopheles* mosquito bite. The sporozoites

are transported to the liver where they infect hepatocytes and undergo schizogony – asexual reproduction of the protozoan. This results in the production of tens of thousands of merozoites, which are released back to the bloodstream in packets called merosomes. There, they start a cycle where they infect RBC, continue schizogony, and are released through RBC lysis. Inside RBCs, some merozoites break this cycle and undergo gametocytogenesis. For *P. falciparum*, gametocytes sequester to the bone marrow where they mature into gametes, which are released to the peripheral circulation where they can be ingested by feeding mosquitoes. Finally, gametes mate and produce new sporozoites in the infected mosquitoe.[94]

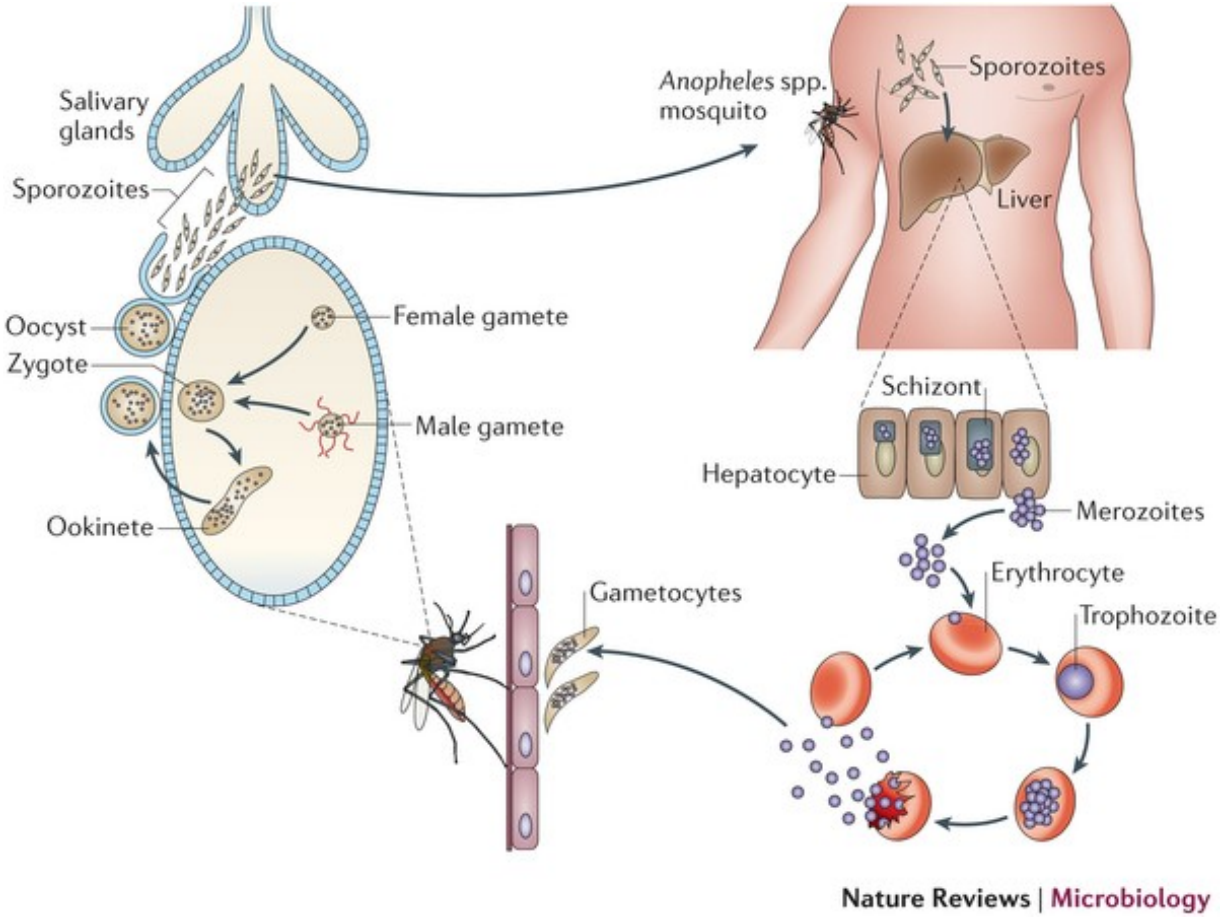


Figure 4. Life cycle of malaria parasites.
Malaria parasites infect the human body through bites of *Anopheles* mosquitoes. Sporozoites are transported to the liver to undergo schizogony. Schizonts release tens of thousands of merozoites into the bloodstream where they infect erythrocytes and undergo gametocytogenesis. Gametes are eventually ingested by feeding mosquitoes, where they are fertilized and become zygotes, which matures into oocyst. Occyst eventually burst and release sporozoites into the salivary glands of the mosquito. Adapted from de Koning-Ward *et al*.[95]

Malaria infection causes a large range of symptoms including fevers, headaches, abdominal pain, diaphoresis, vomiting, tachycardia, jaundice, and hepatomegaly. Symptoms usually appear after 6-14 days after infection by the parasite.[96] It can develop into severe malaria, which is associated with complication such as cerebral malaria, pulmonary edema, severe anemia, bleeding, and renal failure.[96] If untreated, the complications can develop rapidly, and translate to death within hours or days. Since 2000, it is estimated that there was a reduction of 60% of malaria-related deaths, thanks to prevention measures, mosquito control, and the development and use of anti-malarial drugs, such as quinine and artemisinin-based therapies. However, development of new therapeutic strategies are needed because of the emergence of treatment-resistant parasites.[94]

Malaria is notable because of its heavy evolutionary pressure on the human genome. Indeed, several mutations have been identified that confer some forms of resistance to infection. For instance, mutations in *CD36* or *ICAM1* limit adherence of infected RBC to the endothelium or to other cells.[97-99] Mutations that affect the immune system like *HLA* variants are also common. Other mutations affect different proprieties of RBCs. A notable example is a mutation in the GATA1-binding site of promoter of the Duffy antigen/receptor for chemokines (*DARC*) gene, which encodes the Duffy antigen. Duffy-negative individuals are resistant to infection by *P. vivax,* as binding to this antigen is essential for erythrocyte invasion of the parasite.[100] Mutations in *SLC4A1, FECH* and *GYPA* are also associated with reduced parasite invasion or growth.[101-103] Deficiency of the glucose-6-phosphate dehydrogenase enzyme due to mutations in the *G6PD* gene causes increased hemolysis due to increased oxidative stress, which protects against severe malaria by limiting parasite replication.[98,99,104] Finally, regulatory and structural mutations in the globin genes can lead to important changes in the RBC environment where the parasite lives. α^{+}-thalassemia, caused by the disruption of either *HBA1* or *HBA2*, protects against severe malaria.[105] More remarkably, mutations in the *HBB* gene, can lead to different version of globins produced, which are termed HbC, HbE, and HbS. These mutations have distinct regional localization, and their consequences have different severities. HbC heterozygotes and homozygotes are protected from severe malaria, which may be through increased splenic clearance of RBCs. HbE carriers are resistant to *P. falciparium* infection. Finally, heterozygotes

for the HbS allele are ~10 times more protected from severe malaria.[98,106] The mechanism of action might be through increased clearance of infected RBCs or suppression of parasite growth in cells. GWAS have identified variants in the *G6PD*, *ABO*, *HBB*, *ATP2B4,* and *CD40LG* loci associated with malaria susceptibility.[107]

1.3.4 Sickle cell disease

SCD is a recessive genetic disease caused by a mutation in the β-globin gene *HBB* (**Figure 5**). Specifically, SCD is due to an A→T transversion, which changes the glutamic acid at position 6 of the β-globin protein by a valine resulting in sickle hemoglobin (HbS). SCD can also be caused by compound heterozygous mutations that includes β^S, such as a combination of β^S with β-thalassemia (β^S/β^0 or β^S/β^+) or with other mutations such as hemoglobin C (β^S/β^C).[93] A damaging consequence of this substitution is polymerization of HbS when deoxygenated, disrupting the shape of erythrocytes which adopt a characteristic "sickled" shape. Polymerization of HbS is also linked to cation homeostasis dysregulation and RBC dehydration, which in turn increases hemoglobin content and exacerbates polymerization.[108] Sickled RBC are more fragile: Normal RBC have lifespan of ~3-4 months, whereas sickled RBCs can last <20 days due to increased hemolysis, resulting in anemia.[109] The increased stress related to higher RBC production to account for the rapid loss of RBCs leads to an increased number of premature RBCs in the blood stream. These "stress reticulocytes" and sickle RBCs both tend to adhere to blood vessels, which leads to occlusion of capillaries and local hypoxia.[110]

Because of the selective advantage of heterozygote carriers of the SCD mutation in malaria stricken regions, this disease has a very high prevalence, affecting 20-25 million individuals worldwide, particularly in sub-Saharan Africa. An estimated 300,000 individuals are born with SCD each year.[111] In Africa, the mortality rates of children under 5 affected by the disease can reach 50-90%, most of which succumb to infections.[112] In high-income countries, SCD patients have higher life expectancy due to earlier diagnosis, better disease management, education, and greater access to care. Unfortunately, the burden of the disease in sub-Saharan Africa is expected to rise in the coming decades, although it could be reduced by

better health care management, such as the implementation of screening programs and better education.[111]

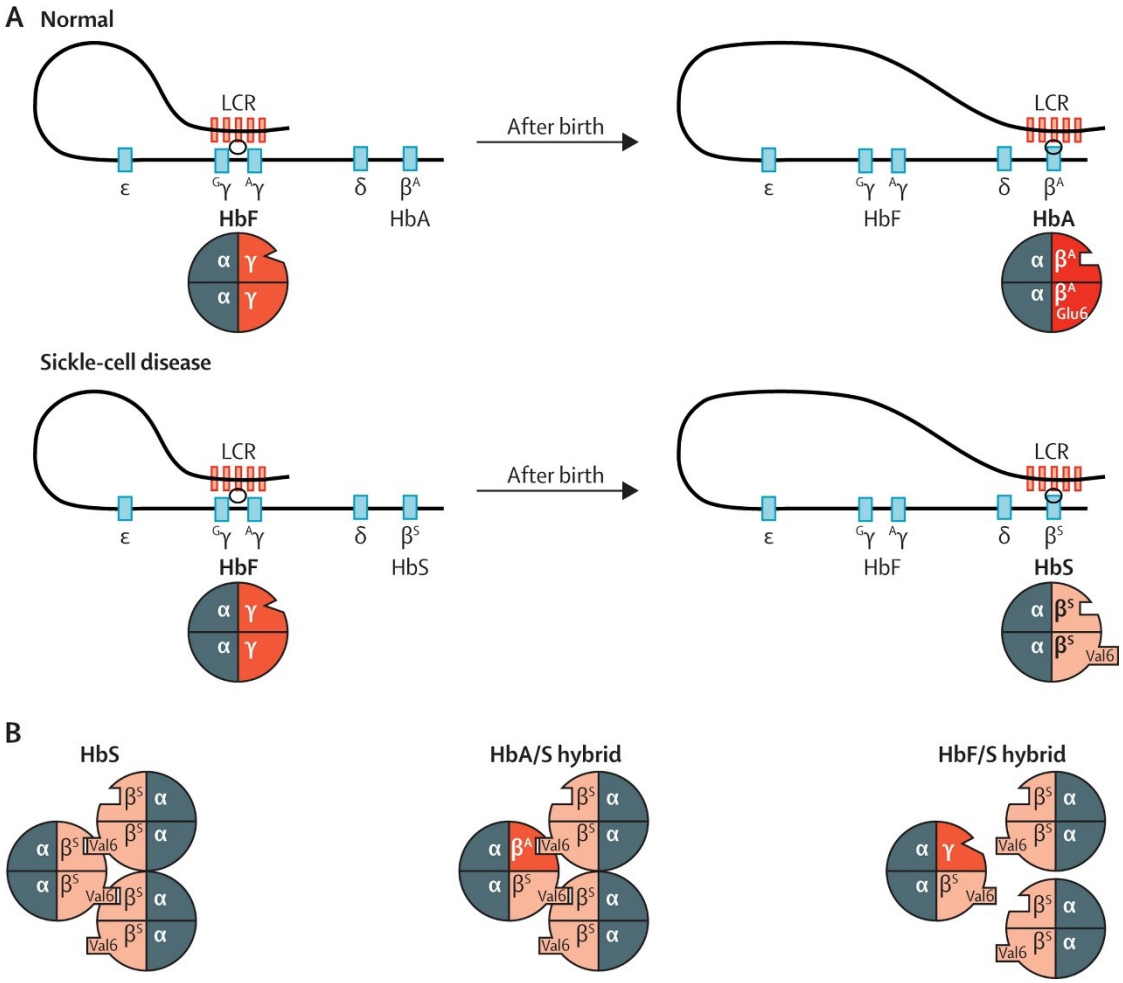


Figure 5. Hemoglobin production and sickle cell disease.
(**A**) Fetal hemoglobin (HbF, $\alpha_2\gamma_2$) is the main type of hemoglobin before birth γ-globin constitutes the β-chain of HbF, and is encoded by the $^{G}\gamma$- $^{A}\gamma$-globin genes. After birth, healthy individuals produce adult hemoglobin (HbA, $\alpha_2\beta_2$), which contains β-globins. In sickle cell disease patients, the glutamic acid at position 6 (Glu6) is mutated into a valine (Val6) to form sickle hemoglobin (HbS, $\alpha_2\beta^S{}_2$). (**B**) HbS tends to polymerise under deoxygenated conditions. γ-globin inhibits polymerisation. Its anti-sickling effect is greatly higher than that of β-globin. Adapted from Lettre & Bauer.[93]

The morbidities associated with SCD are broad and heterogeneous. Pain crises, characterized by excruciating musculoskeletal pain, are the most common complication of SCD and are due to vaso-obstruction of capillaries.[113] 25% of SCD will suffer from strokes.[109] Acute chest syndrome is a pulmonary complication defined by new pulmonary infiltrates due to vaso-

occlusion of blood vessels. It is responsible for 25% of SCD deaths.[114,115] Infection, such as by *Streptococcus pneumoniae*, are very common particularly because of the loss of function of the spleen, and are a major cause of mortality, especially in children.[109] Other complications include osteonecrosis, aplastic crises, priapism, leg ulcers, and kidney failure.[109]

Most treatments of SCD aim at reducing its complications. Hydroxyurea is the only drug approved by the US Food and Drug Administration to treat SCD. This drug increases HbF levels in SCD patients, reducing the number of complications and increasing life expectancy. The mechanism of action of hydroxyurea on HbF is incompletely understood. Its main HbF-inducing effect is probably through increased stress erythropoiesis, which results from the cytotoxic suppression of erythroid progenitors in the bone marrow.[116] Indeed, hydroxyurea inhibits the ribonucleotide reductase, which catalyzes the formation deoxyribonucleotides from ribonucleotides, resulting in decreased DNA synthesis and cytotoxicity. Although this drug is usually well tolerate and have generally mild side effects, it usually doesn't increase HbF levels enough to completely abrogate SCD complications.[93] SCD patients can routinely receive blood transfusion, which can correct anemia and reduce vaso-occlusion-related complications such as stroke and acute chest syndrome.[117] However, transfusions can lead to complications such as iron overload. SCD patients can be treated for pain crises with opioids or non-steroid anti-inflammatory drugs (NSAID). Anti-malarial drugs, vaccines and antibiotics can reduce the risk of infections.[109,118] The only known cure for SCD is bone marrow transplant, which is associated with 6-year event-free survival rate of >90% for HLA-matched donors.[119] However, it is not a feasible health care strategy in low-income countries where the burden of SCD is severe.

1.3.5 Modifiers of β-hemoglobinopathy severity

Individuals with β-hemoglobinopathies present highly heterogeneous clinical manifestations. For example, a study reported that ~40% of SCD patients followed during 9 years never suffered from painful vaso-occlusive crises, whereas ~5% had 3-10 episodes per year.[113] In β-thalassemia, excess α-globin chains form insoluble aggregates, damaging cell membranes. This leads to inefficient erythropoiesis and hemolysis, and ultimately to anemia and

to complications such as splenomegaly, skeletal deformities, and iron overload.[120] Consistently, α-thalassemia reduces the severity of β-thalassemia by restoring α- and β-globin chains balance. α-thalassemia also reduces SCD complications by decreasing the concentration of HbS, which is then less likely polymerize.[121] Conversely, duplication of α-globin chain in heterozygote carriers of a β-thalassemia mutation can cause a β-thalassemia-like syndrome.[122] Sickle red cells become dehydrated and have high MCHC, which increases the rate of HbS polymerization. The ratio of dense RBCs, which have high hemoglobin concentration, is associated with SCD complications such as leg ulcers and renal dysfunction.[123] Variants in the promoter of *UGT1A1* have been associated with the risk of formation of gallstones, and variants in the *MYH9-APOL1* are with proteinuria and decreased glomerular filtration rate in SCD patients, which reflects risk of kidney disease.[124]

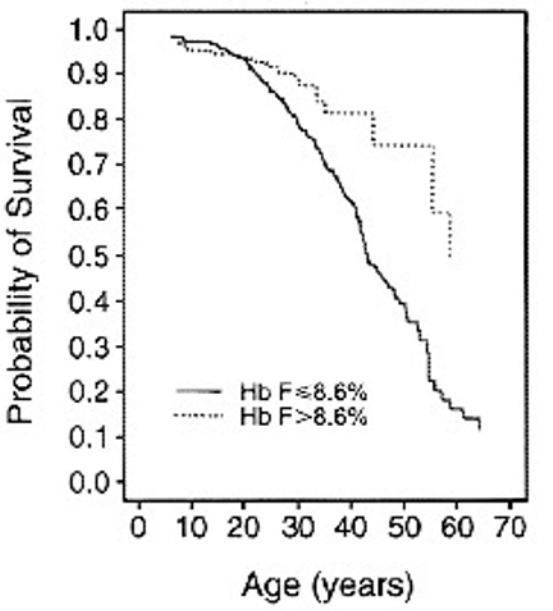


Figure 6. Survival of sickle cell disease patients in the cooperative study of sickle cell disease (CSSCD)
Patients (N=2542) who have HbF levels higher than 8.6% (75th percentile) have a higher survival probability. Adapted from Platt *et al.*[115]

Perhaps the most important modifier of severity of SCD is HbF (**Figures 5-6**).[93] In 1948, Janet Watson observed that infants did not display SCD complications before reaching 6 months of age, and speculated that this was due to a fetal-specific form of hemoglobin, now known as

HbF.[125] Furthermore, it was observed that individuals with SCD of Arab-Indian descent had increased HbF levels and a milder form of the disease.[93,126,127]. Since, HbF has been recognized as a major modifier of SCD severity, increasing survival and reducing the risk of complications such as pain crises, acute chest syndrome, leg ulcers, osteonecrosis, priapism, dactylitis, acute splenic sequestration and priapism (**Figure 6**).[113,115,128-133] Mechanistically, HbF have a low probably of polymerizing with HbS, and thus reduces the speed at which HbS polymers are created in RBCs.[93]

Although HbF levels in healthy individuals are low (<1%), this trait is highly heritable, and individuals with SCD have HbF levels that can reach >5%.[93,134,135]. Mutations at the β-globin locus can cause an asymptomatic Mendelian condition called hereditary persistence of fetal hemoglobin (HPFH), where individuals display high levels of HbF (10-40%).[121,136] Notably, individuals who are homozygotes for the 7.2 kb Corfu deletion, which encompasses a putative regulatory element necessary for HbF silencing, have HbF levels close to 100%.[137] Candidate-gene DNA sequencing, linkage scans, and genome-wide association studies (GWAS) have associated HbF with genetic variants at the *HBB*, *BCL11A*, and the *HBSIL-MYB* loci, which have been well replicated.[47,138-141] Together, these loci explain 50% of the heritability of HbF, and correlate with decreased rate of pain crises.[47,139] *BCL11A* encodes a zinc-finger transcription factor that repressed γ-globin, and that is under the control of an erythroid stage-specific enhancer.[56,58,142,143] Deletions of *BCL11A* in a SCD murine model rescues the disease phenotype.[143] *MYB* is a key regulator of erythropoiesis. Its overexpression in human erythroid cells is associated with decreased HbF, and *MYB* knockdown leads to proliferation and differentiation defects.[73,144,145] Recently, WGS in the founder Sardinian population has revealed an association with a rare intronic genetic variant in the *NFIX* gene, but this association remains to be replicated.[146] *NFIX* is a transcription factor part of the Nuclear Factor 1 family. In mice, it is required for the survival of hematopoietic stem and progenitor cells during stress hematopoiesis.[147] Rare mutations in *BCL11A, MYB* and *KLF1* can also lead to increased HbF levels in adults.[47,148-150] *KLF1* is an erythroid transcription factor that regulates gene implicated in terminal differentiation of erythroid progenitors.[93] Homozygous mutations in this gene can lead to serious disorders such as transfusion-dependent hemolytic anemia.[151] Finally, increased

expression of microRNA (miR)-15a and miR-16-1 in individuals with trisomy 13 is associated with increased HbF, and may act by targeting *MYB*.[152]

1.4 Identifying causal genes influencing human complex phenotypes by omics approaches

The major challenge of GWAS is the difficulty of establishing a causal link between a variant and its associated phenotype. Finding the causal gene is challenging, especially for non-coding variants. Pinpointing potential causal variants is also not straightforward because of high linkage disequilibrium between variants that can span large regions of the genome. Nonetheless, it has been shown that causal variants are more likely to fall in regulatory regions such as enhancers. Several resources have been created to catalog the genomic and epigenomic status of different cell types and tissues.

1.4.1 Coding variants

Coding variants can dramatically alter the encoded protein. For example, nonsense mutations, which create a premature stop codon, can lead to truncated proteins and loss-of-function resulting from nonsense mediated decay.[9] Missense mutations lead to change in amino acid and can alter the function of proteins. Mutation at the exon-intro boundaries can affect splice-site and modify transcript usage or lead to loss-of-function (LoF). Even synonymous mutations can have deleterious effects by modulating RNA processing and translation.[153]

Coding mutations can lead to dramatic phenotypes and pathologies, as exemplified by 4,966 genes in the Online Mendelian Inheritance in Man (OMIM) database with phenotype-causing mutations.[10] Identifying the gene affected by these mutations is important to increase our comprehension of the pathophysiology, and can lead to the discovery of novel pathways implicated in related phenotypes. However, this can be a challenging task because these are usually very rare events and relatively large families may be needed to identify the causal

mutation with confidence. Moreover, patients can have similar or identical clinical manifestation, but be affected by different mutations in the same or in a different gene. Rare coding mutation may help identify the causal genes affecting complex traits. For example, if a gene in a GWAS locus harbors mutations causing related phenotypes, than this gene represent a good candidate for causality. Whole-exome sequencing (WES), WGS, and exome genotyping studies have implicated rare coding variants associated with complex traits.[12,54,154-158]

A limitation of rare-variants association studies is the need of large sample sizes because of their low frequency. Gene-based approaches are an alternative approach to single-variant association analyses, and can increase power to find associations. These tests aggregate multiple variants in a defined genomic unit such as a gene and test these variants as a group. This effectively increases power by reducing the multiple testing burden to ~20,000 tests in gene-based studies. Several methods have been proposed to carry gene-based analyses. Perhaps the simplest is the burden test, which tests if having more rare variants (higher burden) is associated with a trait. The sequence kernel association test (SKAT) is another widely used gene-based approach.[159,160] SKAT is a generalization of the C-alpha test that can account for covariates and epistatic effects. SKAT first regresses to phenotype on the genotypes using a linear mixed model and aggregates weighted variance component scores through a kernel matrix. The main advantages of SKAT over burden tests is that it can account for variants with no or opposite effects inside the same genomic unit, and can incorporate a weight for each variants, for example so that variants of higher frequency are expected to have lower effects.

Another way to investigate the role of rare variants on complex traits is through the study of human gene knockouts. Rare LoF variants have been implicated in complex traits such as height, hematological traits, inflammatory diseases, and cardiovascular diseases.[9,12,13,54,161,162] Gene knockouts arises when both copies of the genes are inactivated due to LoF variants. They can have major impact on human health, as demonstrated by Mendelian disorders such as sickle cell disease and cystic fibrosis.[10] The study of these conditions can yield significant insight into the function of the mutated genes. However, these event are generally rare and private to

families, making it challenging to study their polygenic effects.[163,164] Variants that induce a premature stop codon, that disrupt a splice-site, or indels that create a frameshift are often considered LoF. Several studies have started to catalog human LoF variants and gene knockouts.[9,154,158,161,163-173] In a study of 60,706 whole-exomes, the number of LoF variants was around 85 per individual, with 35 at the homozygote state.[163] However, when considering only rare variants (frequency <1%), the average number of homozygous LoF variants per individual was around 0.19. Most LoF alleles per individual reflect common variants in the population, consistent with most LoF disrupting non-essential genes such as genes encoding taste and olfactory receptors.[164,168] Around 72% of genes predicted to be intolerant to LoF and carrying such alleles had no known human disease phenotype.[163] Sequencing of founder or consanguineous populations can be useful to identify gene knockouts.[27,158,164-167,170,174] For example, 7.7% of 2,636 Icelanders where predicted to be complete knockouts for one of 1,171 different genes (LoF variants frequency <2%).[165] This population shows increased homozygosity for rare protein-coding variants.[30] Overall, founder or consanguineous populations have decreased diversity of LoF variants, although those that are present reach higher frequency.[158,170] Targeted or random mutagenesis in animal and cell models have had (and continue to have) success in elucidating the function of genes. This is reflected by the rising interest toward CRISPR-Cas9 knockout screens.[175-181] Nonetheless, these assays are generally aimed towards a phenotype of interest, and may not capture the pleiotropic effect of these knockouts, especially considering the complex human physiology and interactions with the environment. Thus, studying the effect of human gene knockouts on complex traits in the population may yield significant insight into their biology. Studies have started to identify the phenotypic impact of rare gene knockouts from large sequencing studies. Examples include the switch of PRDM9-dependant recombination hotspots in *PRDM9* knockouts, low lipoprotein(a) levels in *LPA* knockouts offering potential cardiovascular protective effects, and *C1QTNF8* knockouts with elevated serum magnesium levels.[158,161,170] Individuals with autism spectrum disorders have twice as many knockout genes than healthy controls.[154] Recently, a study of 10,503 participants from South Asia reported gene knockouts of 7 different genes associated with different phenotypes.[169] These studies are however often limited by the low number of individuals with complete knockouts and by the difficulty to assess whether variants are true LoF.[9,169] Indeed, LoF variants can be rescued by nearby variants or exon skipping, and nonsense

and frameshift indels at the 3' end of a gene may be tolerated.[9] Nonetheless, the study of rare gene knockouts may be particularly useful to assist drug design, as demonstrated by therapies aiming to block PCSK9 or ANGPTL3 in order to lower low-density lipoprotein cholesterol levels.[182,183]

1.4.2 Epigenomics

In order to give sense to non-coding variation, it is essential to understand how these variants might act. These variants may disrupt regulatory elements, which will be reflected by the underlying epigenetic structure. However, epigenetic marks vary not only between individuals, but also between each cell type and tissues of an individual. It is also influenced by exterior stimuli and physiological changes. To address this, resources such as ENCODE and the Roadmap Epigenomics Project have assembled large collections of epigenomic data of different cell types and tissues, such as chromatin accessibility, transcription factor binding sites, and chromatin interactions.[184,185]

DNA methylation

At the base pair level, DNA can be modified via methylation of carbon 5 (C5) of cytosines at CpG sites (**Figure 7**). CpG sites are generally depleted in the human genome as they promote mutations as ^{5m}C can spontaneously deaminate into thymines. Nonetheless, CpG-rich regions termed CpG island occur in the genome, the majority of which are demethylated and are associated with gene promoters (>50%).[186,187] DNA methylation has been associated with transcriptional repression. For example, the promoters of the γ-globin genes are demethylated in fetal RBC progenitors and hyper-methylated in adults. [188,189] Gene bodies of highly transcribed genes show inverse trend, being highly methylated. Here, ^{5m}C could be implicated in splicing or repression of alternative promoters.[190,191] At a larger scale, this epigenetic mark is implicated in processes such as genomic imprinting and chromosome X inactivation. The importance of DNA methylation is reflected by the range of human disease

where it is involved. Loss of imprinting can lead to cancers or pathologies such as Prader–Willi and Angelman syndromes.[192] Mutation of the DNA methyltransferase gene *DNMT3B* can lead to the Immunodeficiency, Centromeric instability and Facial anomalies (ICF) syndrome. 5-methyl-cytosines can be oxidized by the TET family of enzymes by a stepwise process into 5-hydroxymethylcytosine, 5-formylcytosine, and 5-carboxylcytosine. The latter two are recognized by the Thymine DNA Glycosylase (TDG) to restore cytosines into its un-methylated form. Evidence shows that these intermediates may have regulatory roles.[193] DNA methylation is usually assessed using bisulfite DNA sequencing or DNA methylation arrays.[194]

Histone modifications

At a higher level, DNA is packaged into more or less compact regions defined by nucleosomes, which are composed of histones (**Figure 7**). Tight assembly of nucleosomes effectively obstructs transcription by blocking RNA polymerase II's access to the promoters. Around 147bp of DNA are bound around a single nucleosome.[195] The core human histones H2A, H2B, H3, and H4 are arranged into octamers to form a single nucleosome unit.[196] These histone can be replaced by a collection of histone variants such as histone H3.3, which is associated with a higher histone turnover at promoters, enhancers, and gene bodies.[197] Moreover, histones can be modified at more than 130 different sites by several enzymatically catalyzed reactions such as acetylation, methylation, and phosphorylation. The diversity of nucleosome dictates the chromatin state which reflects the activity of underlying elements. The chromatin state is highly dynamic, being influenced by transcription factors, nucleosome modifiers and chromatin remodelers. Since chromatin state is reflected by histone modifications and variants, determining these can be key to identify active elements of a specific cell type in defined conditions. Histone states are usually determined by chromatin immunoprecipitation followed by sequencing (ChIP-seq). Hidden Markov Models (HMM) can be used to class the different combinations into interpretable regulatory units.[198] At a general level, histone 3 lysine 4 trimethylation (H3K4me3) is linked to higher transcription levels and active promoters.[199] On the other hand, repressed promoters are marked by H3K27me3 and H3K9me3. Bivalent promoters are marked by both H3K4me3 and repressive marks, which could reflect promoters that need to be turn on or off rapidly, for example during cell differentiation. Gene bodies that

are transcribed are generally marked with H3K79me2 and H3K36me3, with exons showing a higher enrichment for the latter mark, which might play a role in splicing. Finally, active enhancers are enriched for H3K4me1, H3K4me2, H3K27ac, but depleted for H3K4me3.[199]

Chromatin accessibility and binding

Active regions of the genome such as enhancer and promoters are bound by transcription factors and the transcription machinery. Thus, these region must be "unpacked", that is free from nucleosomes, in order to allow access to the underlying DNA. Nucleosome repositioning or eviction correlates with transcriptional activity. Hence, methods to measure accessible chromatin can be used to identify regulatory regions. Several techniques exist to identify accessible regions. For example, MNase-seq is used to measure regions of the DNA that are bound by nucleosomes, and thus gives an indirect estimate of accessible regions.[200] DNAse-seq uses digestion of accessible DNA by DNAse I to measure DNAse hypersensitive sites (DHS). Formaldehyde-Assisted Isolation of Regulatory Elements followed by sequencing (FAIRE-seq) is another approach which uses phenol-chloroform to separate nucleosome-bound regions (crosslinked with DNA) from non-bound regions. The main advantages of this method is that it enriches for active regions of the DNA and is generally a less laborious method than DNAse-seq. However, it is limited by high background signal. [200] A more recent approach is the Assay for Transposase-Accessible Chromatin followed by sequencing (ATAC-seq). This approach uses a hyperactive Tn5 transposase, which preferentially integrates into accessible DNA. The transposase is loaded with sequence adapters so that the integrated regions can be amplified and sequenced. This method is gaining popularity due to its simple protocol, high sensitivity, and requires a low number of cells (50 to 50,000).[200] These approach are often used in combination with sequence motif analysis or ChIP-seq to identify regions bound by transcription factors.

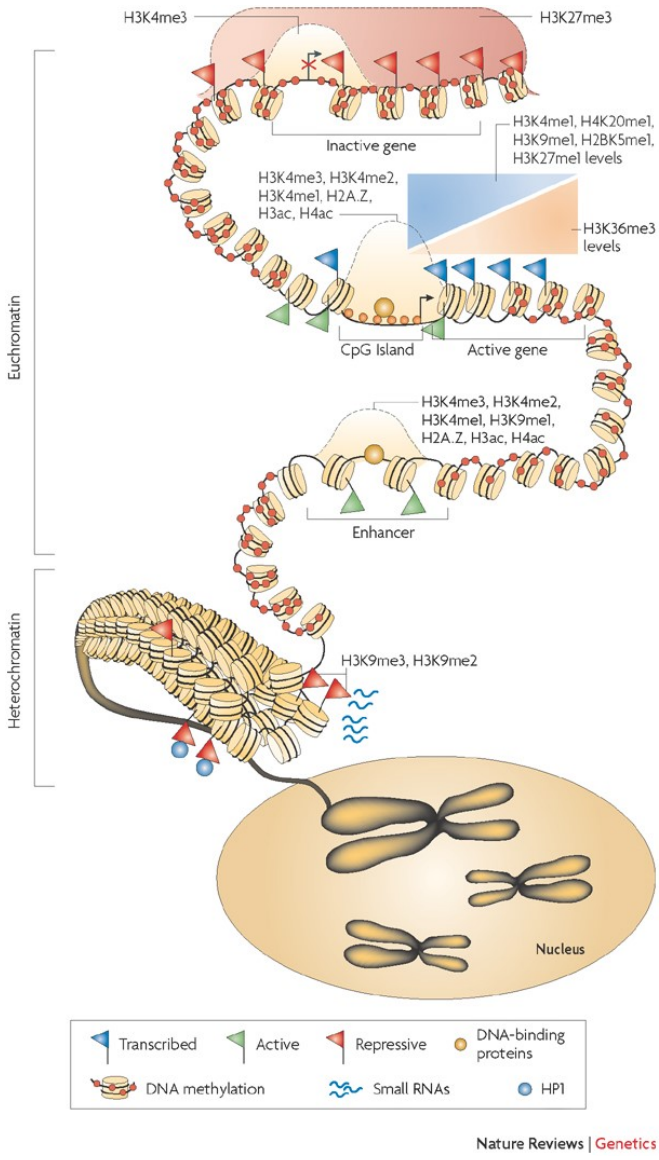


Figure 7. Epigenetic modifications

Histone modifications and DNA methylation mark the activity of genes and regulatory elements such as promoters and enhancers. At the euchromatin level, DNA is decorated by DNA methylation, which is only absent in open regions like enhancers, promoters and CpG islands, which are accessible to DNA-binding proteins. Inactive genes are marked by the histone modifications H3K4me3 and H3K27me3. Several modifications mark active genes such as H3K4me3, H3K4me2, H3K4me1, H3ac, and the histone variant H2A.Z, which cluster at the transcription start site of genes. Transcribed regions are marked by H3K4me1, H4K20me1, H3K9me1, H2BK5me1, and H3K27me1. Transcribed genes are also marked by H3K36me3 at their 3' end. Enhancers are marked by H3K4me3, H3K4me2, H3K4me1, H3K9me1, H2A.Z, H3ac, and H4ac. Small miRNAs are implicated in the regulation of heterochromatin, which is bound by heterochromatic protein 1 (HP1) and is marked by H3K9me3 and H3K9me2. Figure adapted from Schones & Zhao[201].

Chromatin interactions and domains

It can be difficult to link active regulatory elements to specific genes. Enhancer don't necessarily interact with the closest gene, and interactions can span large distances.[202-204] Moreover, genes tend to cluster into groups of active or inactive transcriptional activity, and thus studying chromosome architecture can reveal groups of co-regulated genes within a specific cell context.[205] Methods have been developed to measure these interactions, in particular chromosome conformation capture (3C)-based methods. These methods use crosslinking with formaldehyde to covalently link interacting DNA regions. Different 3C-based methods give different level of information: 3C is classically used to detect interaction between two specific products by PCR. Hi-C, on the other hand, give a complete "all-by-all" interaction map, but resolution is limited by sequencing depth. Promoter Capture Hi-C involves using probes to enrich for promoter interactions, and increase their resolution.[203,206] Other variants of these methods include chromatin interaction analysis by paired-end tag sequencing (ChIA-PET), which uses chromatin immunoprecipitation to select fragments bound by a protein of interest.[207]

Studies of chromatin architecture have reveal that chromosome looping, which is classically exemplified by the looping of the locus control region of the β-globin genes, is a ubiquitous event in the genome.[202,207] Promoters can interact with several distal elements, and active enhancers tend to have an additive effect on gene expression.[206] Chromosomes are discretely separated by topologically associating domains (TAD), which span around ~185 kb.[208-211] Genes within a TAD tend to have coordinated gene expression, and promoters tend to interact with enhancer inside the same TAD.[209,212] Promoter-interacting regions are enriched for eQTLs for the promoter's associated gene, and thus may help prioritize genes affected by non-coding trait-associated variants, and identify regulatory regions that are not defined as classical enhancers.[206]

Resources

As mentioned above, epigenetic modifications are highly cell-specific. Hence, for a given trait, it is optimal to assay a related cell-type. Studies such as the Encyclopedia of DNA Elements (ENCODE) are gathering large collections of cell and tissue-specific epigenetic information. ENCODE was the first large-scale international consortium of epigenetic data and pioneered several epigenetic profiling techniques.[185] ENCODE is now part of the International Human Epigenome Consortium (IHEC) which is an umbrella organization containing projects such as the Roadmap Epigenomics Project and BLUEPRINT.[213-215] The latter is mainly focused on healthy and malignant hematopoietic primary cells, including erythroblasts and erythroid progenitor cells. As of April 2017, IHEC contains 8,753 datasets from 1,380 epigenomes. In addition to contributing datasets, these studies have developed novel tools to study epigenomes, and provided insight on the role of epigenomics on cell fate, diseases and variant-associated traits. Importantly, these studies reinforce the idea that genetic variation act thought cell-specific events. Trait-associated SNPs are generally enriched in enhancers.[216-219] Moreover, SNPs associated with platelet or RBC traits are predominantly enriched in promoter-associated regions of myeloid lineages.[206] Several bioinformatics tool have been developed to integrate the different levels of epigenetic marks[198,220-222], link GWAS variants to active regions[223], and finally link these regions to genes.[219,224-229]

1.4.3 Transcriptomics

Non-coding variant can impact complex traits through modulation of gene expression. Gene expression programs can be highly cell-specific and diverse. Over 95% of human genes are predicted to encode more than 1 isoform.[230] RNA-sequencing (RNA-seq) has emerged as a powerful tool to assess transcriptomes, as it can be used to discover novel transcripts and isoforms, can detect large differences in gene expression, and RNA-seq reads can be used to genotype variants which can be useful for allelic imbalance studies.

RNA-seq consists of fragmenting the purified RNA sample into small sequences (30-400bp), convert it to cDNA, add sequencing adaptors, and sequence the fragments using a next-generation sequencing platform.[231,232] Reads are then aligned to a reference genome or transcriptome using aligners such as Tophat2.[233] Reads counts need to be normalized to account for library size and reduce the impact of highly expressed genes which can reduce the sampling of other transcripts.[234] Measures such as fragments (or reads) per kilobase per millions (FPKM) also account for gene length allowing different genes can be compared. Transcriptomic studies are often designed to test for differential expression between two sets of samples. Statistical tools such as Cuffdiff2, EdgeR, and DESeq2 are used to test for differential expression.[235-238] These tools use extension of the Poisson distribution to model RNA-seq reads such as the negative binomial distribution. Moreover, these tools borrow information from all transcripts to test differential expression of a gene, which serves at increases the power to detect differences when the sample size is low. Differential expression experiments are expected to identify a large number of differentially expressed genes. Thus, multiple testing correction that are more liberal such as the false discovery rate (FDR) are generally used.

Small RNAs can also be detected by RNA-sequencing, and can be used to measure expression of microRNAs (miRNAs), small nucleolar RNAs (snoRNAs), and PIWI-interacting RNAs (piRNAs). miRNAs are small RNAs of ~22 nucleotides of length. Precursor miRNAs are processed by the double stranded RNA-specific ribonuclease DROSHA, which is part of the Microprocessor complex, and exported to the cytoplasm where they are processed into mature miRNAs by the endoribonuclease DICER. Mature miRNAs associate with the Argonaute (AGO) proteins to form miRNA-induced silencing complexes (RISC). miRNAs usually bind 3′ untranslated region (UTR) of mRNAs, where they induce translational repression, mRNA deadenylation and mRNA decay.[239,240] Software such as miRDeep2 can be used to detect novel miRNAs from small RNA-seq data.[241]

eQTL mapping

Genetic variants can be linked to gene expression with expression quantitative trait loci (eQTL) mapping. These studies are important given that eQTLs are enriched in GWAS variants.[13,242-246] Together with the statistically significant co-localization of trait-associated variants in regulatory regions[247], this suggests that non-coding variants influence phenotypes through changes of gene expression regulation. Nonetheless, the co-localization of GWAS and eQTL signals is often not perfect. The top signal may not be the causal variant, and different signals may exist in the same locus, which may be challenging to disentangle. Moreover, with the increased number of eQTL datasets available, the chance of random overlap increases. Bayesian models have been developed to predict likely localization events between eQTLs and GWAS variants.[248-250] The tissue or cell model used to discover eQTL is crucial, as the causal signal may only exist in a specific cell type.[251-255] For instance, a study comparing eQTLs in B-lymphocytes and peripheral blood monocytes showed that 80% of eQTLs signals were specific to either cell type, with 31 genes with opposite eQTL effects.[251] Consistently, eQTLs that were GWAS variants were associated with different trait depending on the cell type. Moreover, ~5% of these GWAS variants were associated with different genes depending on the cell type. Thus, large collections of eQTL in different cells or tissues such as the Genotype-Tissue Expression (GTEx) or BLUEPRINT databases can provide important data from GWAS loci characterization.[246,250] The current release of GTEx contains eQTL analyses of 44 tissues from 449 donors (V6p). Other types of QTL analyses such as RNA splicing QTLs, DNA methylation QTLs, transcription factor binding QTLs, and histone modification QTLs, can also be useful to characterize GWAS variants.[223,250] Of note, Chen *et al* found that GWAS variants often co-localized with histone QTLs without associated change in gene expression, which could reflect a poised or primed state of underlying regulatory elements.[250]

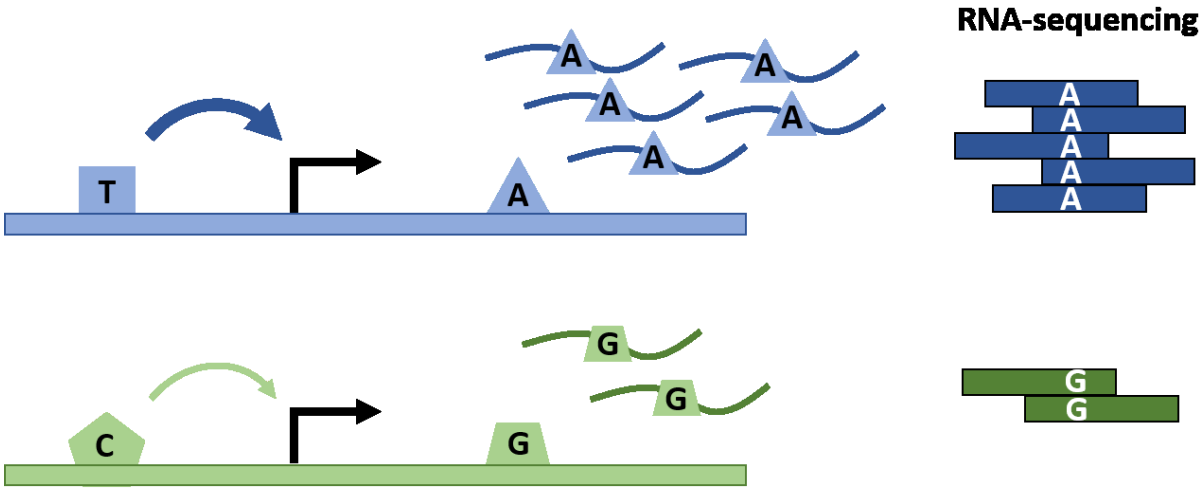


Figure 8. Principle of allelic imbalance

In this example, the T→C single nucleotide polymorphism (SNP) reduces binding of transcription factors in a regulatory element, resulting in lower expression of a proximal gene. This gene harbors an exonic silent A→G SNP, which will mark expressed transcripts. RNA-sequencing of an heterozygote for both SNP will result in allelic imbalance due to the modulating effects of the T→C SNP, and reflected in an higher proportion of reads containing the exonic A allele than reads harboring the G allele. In a population, the T→C SNP can be observed as an expression quantitative trait loci (eQTL) of the gene.

Allelic imbalance

eQTL studies are limited in power by the number of sample assayed. However, variant effects on gene expression can still be observed in a single heterozygote sample through allelic imbalance (AI, **Figure 8**). Indeed, at the heterozygote state for the causal variant, the expression of the affected gene will differ between the paternal and maternal chromosomes. The differential contribution of both chromosome can be measured by RNA-sequencing if at least one heterozygote exonic variant is present in the gene: the number of reads containing each allele will differ from the expected 50/50 ratio. AI studies can be particularly useful to detect the effect of nonsense variants.[9,250,256,257] Moreover, AI can detect effect of variants on exon usage that may not impact overall gene expression.[257] It is also possible to detect AI of other regulatory marks such as transcription factor binding, histone marks, or DNAse hypersensitivity.[250,258-260] Finally, it is possible to combine eQTL analysis with AI analyses to improve the detection of potential causal regulatory variants affecting gene expression.[246,250,257,261,262]

1.5 Validating causal genes using genome editing

The large genetic, epigenomic, and transcriptomic data collections can help make strong hypotheses about the effect of a variant on a phenotype of interest, which must be validated. In recent years, genome editing technologies – in particular the clustered regularly interspaced short palindromic repeats (CRISPR) - *Cas9* system – are emerging has powerful and compelling approaches to validate these causal genetic relationships (**Figure 9**).

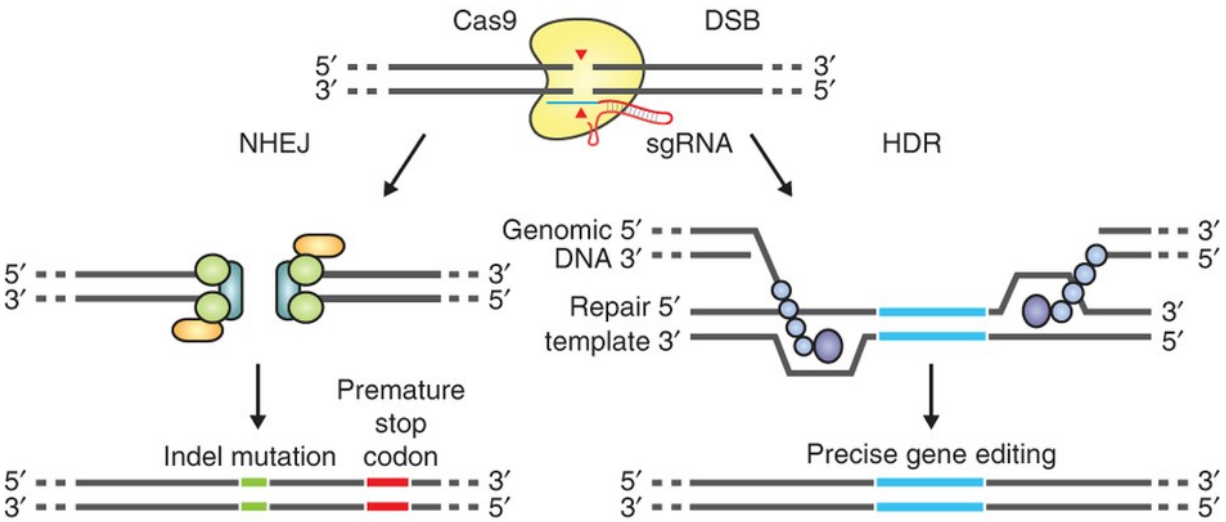


Figure 9. CRISPR/Cas9 genome editing
Cas9 targets a genomic region complementary to a single guide RNA (sgRNA), and induces double strand breaks (DSB). This break is repaired by either non-homologous end-joining (NHEJ) or homology directed repair (HDR). Repair by NHEJ often creates small indels, which can create premature stop codon in genes. Repair by HDR requires a template, which can be used to insert DNA elements, genes or specific mutations at the targeted genomic loci. Adapted from Ran *et al.* [263].

1.5.1 Genome editing

The concept of genome editing is not new. In 1979, Scherer and Davis showed that deletions or insertion of foreign DNA sequences could be engineered in yeast by taking advantage of the homologous recombination pathway, a form of homology directed repair (HDR).[264] HDR is used in cells to repair double strand breaks (DSB), which would be lethal otherwise (**Figure 9**). HDR copies information from a homologous DNA template – such as a homologous chromosome – to repair the break. DNA differences contained in the template will be copied on the target DNA, inducing mutations.[265] DSBs can also be repaired by another

pathway: non-homologous end joining mechanism (NHEJ). In the NHEJ pathway, the DSB ends are ligated together without the help of a homologous template, but relies on microhomologies between broken ends (**Figure 9**).[266] DSBs often leads to small single-strand DNA overhangs at both DSB ends, which will be resected or filled in by polymerases to create micro-homologies that facilitate ligation. Thus, this repair mechanism can lead to small insertion and deletions. In humans, NHEJ is preferred to HDR, as NHEJ is carried more rapidly and efficiently.[267] Moreover, HDR is mostly restricted to when DNA is duplicated during the S and G2 phases of the cell cycle.[267]

Genome editing relies on inducing double-strand breaks (DSB) at a specific genomic locus, then taking advantage of the cell repair pathways to either induce small deletions, insert novel DNA elements, or engineer specific mutations (**Figure 9**). Earlier methods relying on DNA base pairing to induce DSBs at specific loci involved small oligonucleotides or molecules coupled to chemical cleavage or cross-linking reagents.[268] Other methods involved self-splicing introns or homing nucleases.[269] However, these early methods lacked robustness or could only target a limited number of loci.

Zinc-finger nucleases (ZFNs) offered a more efficient and versatile alternative. The nuclease domain of the restriction enzyme *FokI* from *Flavobacterium okeanokoites* is independent from the recognition domain. [265] Hence, by fusing the nuclease domain to a tandem of different DNA-binding zinc finger proteins (ZFPs), which recognizes 3-4 nucleotides each, it is possible to induce DSB at different DNA sequences. [270] However, creating efficient ZFPs has proven challenging, as most combination of ZFPs will yield inefficient or non-specific proteins.[265] Transcription Activator-Like Effector Nucleases (TALEN) represent a more flexible and efficient method. TALENs use the *FokI* nuclease fused to Transcription Activator-Like Effector (TALEs) proteins.[271] Each TALE recognizes only one nucleotide, facilitating the design of TALENs. Nonetheless, TALENs also can be difficult to synthesize, as the binding domains are highly homologous and can recombine together in cells.[272]

1.5.2 The CRISPR/Cas9 system

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) were first described in 1987 by a Japanese group investigating the genomic locus encoding the alkaline phosphatase isozyme in *Escherichia coli*.[273] The group described a region of unknown significance containing five homologous 29bp DNA sequence repeats separated by 32bp spaces at the 3'-end of the gene. CRISPR genes were since found in several prokaryotes.[274] Evidence that the genes are transcribed and that the protospacers contain sequences seemingly derived from viral and plasmid origin hinted that CRISPR were involved in immunity.[275-279] This was further evidenced by the characterization of a helicase and nuclease domain in CRISPR-associated (*Cas*) proteins, and that bacteria were more resistant to infection when their protospacer sequences contained DNA from the infecting pathogen.[280,281]

Adaptive immunity from the CRISPR/Cas9 system is conferred when short fragments from the pathogen's genome are inserted as a spacer sequence into the CRISPR array.[282] These sequences are then transcribed and matured into CRISPR RNA (crRNA), which associated with *Cas* proteins and bind to the homologous region of the invading pathogen, inducing cleavage of the region. Three different types of CRISPR/Cas systems have been described. Type I and III system are phylogenetically more related and use large complex of *Cas* proteins.[268,282] In type II system, which has been prominently described in *Streptococcus pyogenes, Cas9* is the only protein needed for DNA sequence recognition and cleavage.[283] In this system a third component, a trans-activating crRNA (tracrRNA) forms a complex with the crRNA and mediate its maturation.[284] The crRNA-tracrRNA complex associates with Cas9 to induce cleavage.[285] A requirement of this system is the ~20bp complementary crRNA sequence followed by the protospacer adjacent motif (PAM), which is 5'-NGG in the case of *S. pyogenes*.[286,287] The RuvC-like and HNH domains are responsible for the nuclease activity of *Cas9* by inducing single strand breaks at opposite sides of the target DNA.[285,286]

The target specificity of the Cas9 nuclease is conferred by simple DNA pairing complementary to the dual tracrRNA:crRNA. The latter can be simplified into a single guide

RNA (sgRNA) (**Figure 9**).[285] Unlike ZFNs and TALENs, the nuclease domain of the system is independent of component that confers sequence-specificity. Thus, the CRISPR/Cas9 system offers a flexible solution to genome editing. Because the Cas9 protein does not change, only the design and production of sgRNA is needed, which can be prepared by cloning or *in vitro* transcription.[271,288] The flexibility and relatively easy design of the system is reflected by its numerous applications. Apart from inducing small indels or inserting specific elements, the system can be used to create large biallelic deletions when combining two different sgRNA targeting the flanking regions of the desired deletion.[289] This system can be used in knockout screens, which represent an attractive alternative to RNA interference screens.[176,178,180,181,290,291] The Cas9 protein can also be modified by fusing different domains to an inactivated "dead" Cas9. For example, VP16 transactivation or KRAB domains can be fused to Cas9 for knockdown or gain-of-function screens.[181,292-294] Cas9 can be used to induce chromosomic rearrangements.[295] Finally, the CRISPR/Cas9 system also promises several clinical application. Its potential use as gene therapy to correct genetic defects is heavily considered.[296] Encouragingly, genome therapy aiming at correcting the Duchenne muscular dystrophy (*DMD*) gene in mice partially restored muscle function.[297-299]

1.5.3 Using genome editing to identify causal loci

The previous sections have highlighted several bioinformatic dataset and prediction tools to help and identify potential causal regulatory variants associated with complex traits. Genome editing technologies such as CRISPR/Cas9 makes it increasingly easier to confirm to causal relationships of genetic variants on these phenotypes. In this section, I will highlight two studies that successfully identify casual regulatory elements using genome editing (**Figures 10-11**).[56,300]

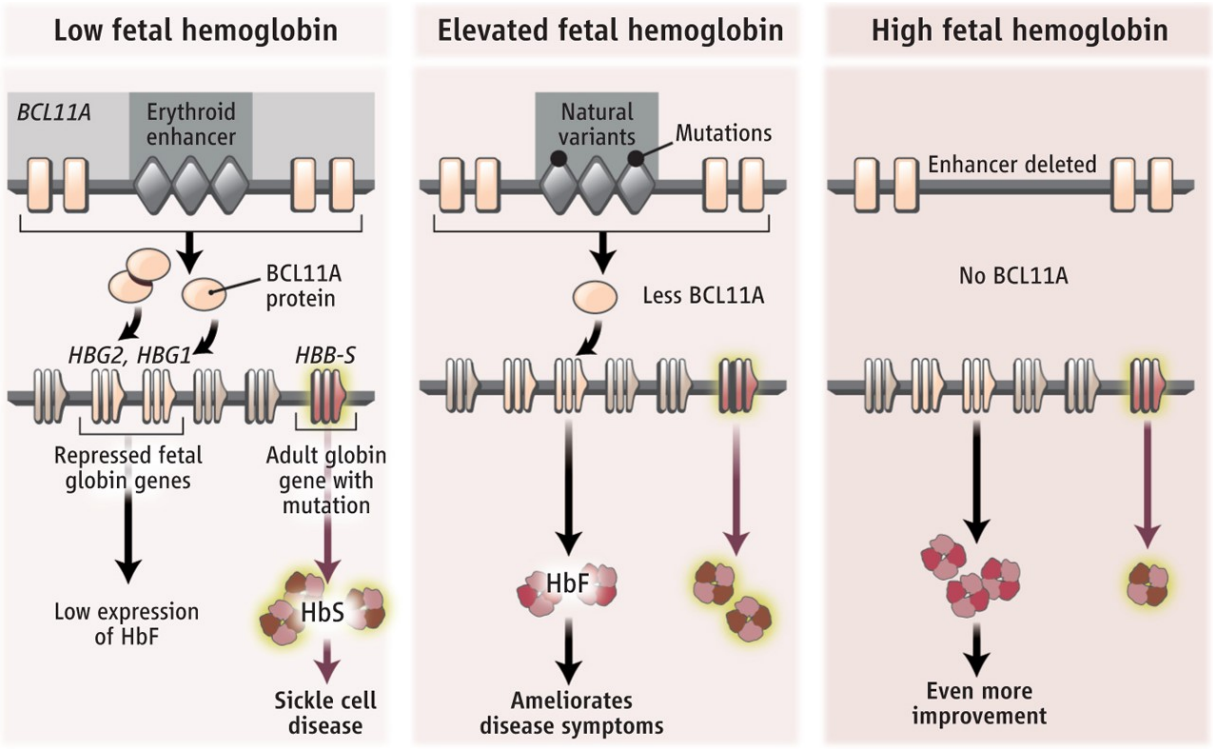


Figure 10. Modulation of *BCL11A* by genetic variants in an erythroid enhancer
Under high BCL11A production, fetal γ-globin genes (*HBG1* and *HBG2)* are repressed and the adult β-globin gene (*HBB*) is highly expressed (left panel). When *HBB* harbors the sickle cell mutation (annotated as *HBB-S*), this produces sickle hemoglobin (HbS), and cause sickle cell disease (SCD) in homozygotes. Variants in the stage-specific erythroid enhancer of *BCL11A* reduces its production, which is associated with more fetal hemoglobin (HbF; middle panel), and less SCD complications. Deletion of the enhancer increases HbF production even more, and could lead to therapeutic benefits in SCD patients. Adapted from Hardison & Blobel.[301]

The first example is the identification of non-coding variants affecting HbF production through modulation of an erythroid-specific expression in *BCL11A,* a known repressor of HbF (**Figure 10**).[56] Genetic fine-mapping of the region identified at least 2 independent genetic variants - rs1427407 and rs7606173 - in the intron 2 of *BCL11A* associated with HbF levels in African-Americans. Three putative regulatory regions were first identified in erythroblasts using ChIP-seq, chromosome conformation capture, and DNAse I hypersensitivity assays. Erythroblasts heterozygotes for rs1427407 demonstrated allelic imbalance of GATA1 and TAL1 binding at this site, as well as imbalance of mRNA expression. The exon-2 *BCL11A* enhancer induced expression of a *lacZ* reporter specifically in erythroid cells. Using TALENs, biallelic deletion of the region was engineered is mouse erythroleukemia (MEL) cells and pre-B lymphocytes. *BCL11A* expression was completely disrupted in MEL cells, but not in pre-B

lymphocytes, demonstrating the specificity and necessity of this enhancer for erythroid expression of *BCL11A*.

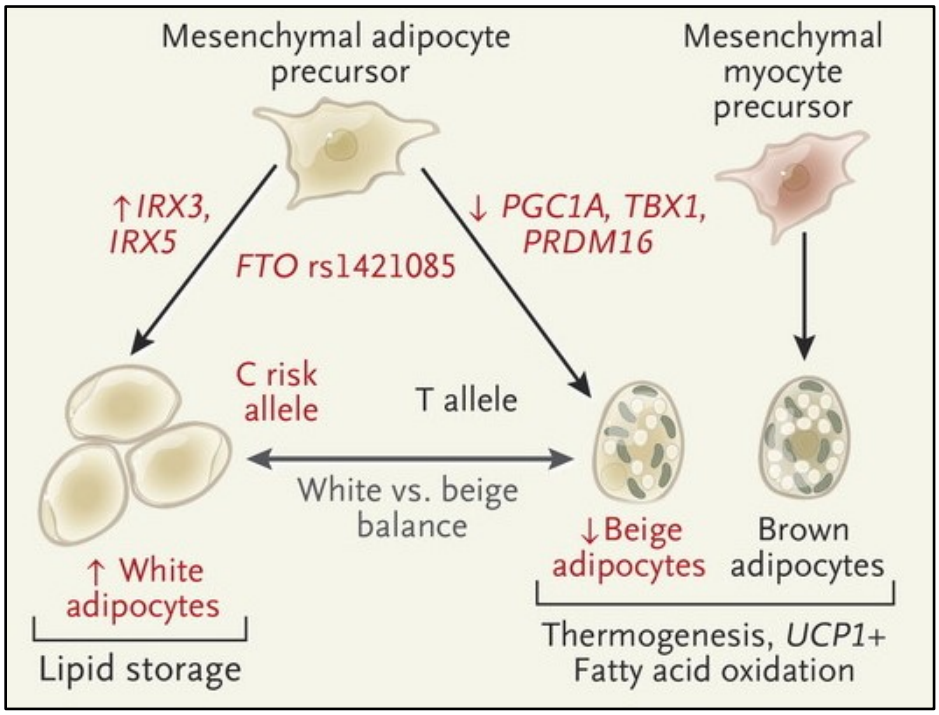


Figure 11. Role of the *FTO* rs1421085 variants on obesity.
The C-allele of the variant increases the number of white adipocytes and decreases the number of beige adipocytes derived from mesenchymal adipocyte precursors via increased expression of the transcription factors *IRX3* and *IRX5*. This is associated with a decreased expression of genes encoding regulators of mitochondrial function and thermogenesis: *UCP1*, *PGC1A, PRDM16,* and *TBX1.* The altered white and beige adipocyte balance increases lipid storage, and decreases mitochondrial thermogenesis and oxygen consumption. In contrast to white and beige adipocytes, brown adipocytes derive from mesenchymal myocyte precursors. Adapted from Claussnitzer *et al.* [300]

Another example is the dissection of the obesity-associated *FTO* locus (**Figure 11**). This particular study addresses a limitation of the previous study, namely that the specific potential causal genetic variants were not directly assayed. The obesity-associated variants are clustered inside the first and second introns of the *FTO* gene in a region of high LD. The activity of a super-enhancer in intron-1 of *FTO* is modulated by the obesity-associated haplotype in adipocytes. Authors showed that two genes, *IRX3* and *IRX5,* interacted with the enhancer and were eQTL targets of the *FTO* variants in pre-adipocytes. Pre-adipocytes with the risk allele had increase *IRX3* and *IRX5* expression, which was accompanied with increased adipocyte size, reduced mitochondrial thermogenesis, and decreased oxygen consumption. Knockdown of the

candidate genes in these adipocytes restored oxygen consumption and thermogenesis response, whereas overexpression of the gene in non-risk cells had the opposite effect. Knockdown of these genes in mice resulted in animals with anti-obesity characteristics, such as decreased fat mass, white and brown fat depots, and adipocyte size, as well as increased energy expenditure and oxygen consumption. Authors identified rs1421085 as the potential causal variant, which disrupted an *ARIDB5* binding site, and found that the risk allele induced increased luciferase activity when implemented into the non-risk haplotype background. Using CRISPR/Cas9, the risk allele was introduced into pre-adipocytes with the non-risk haplotype using HDR, which resulted in higher expression of *IRX3* and *IRX5* and reduced metabolic rate. Importantly, re-editing the non-risk allele rs1421085 rescued the non-risk phenotype, thus providing strong evidence that rs1421085 is the causal variant.

In conclusion, these studies demonstrate that integrating genetic and function genomic data with genome editing experiments can successfully identify causal genetic variants. These studies can be crucial to identify the gene influenced by non-coding variants – especially in region overlapping multiple credible candidates. As in the *BCL11A* example, this studies can lead to the identification of regulatory regions specific to particular cell type, paving the way to very specific genome editing therapies.

1.6 Research Questions and Thesis Outline

Monitoring RBC traits such as RBC count and mean corpuscular volume can reveal underlying pathologies, which can range from nutritional deficiencies to inherited disorders. Genome-wide association studies (GWAS) have identified several loci associated with these traits. These loci may present attractive therapeutic targets for blood-related diseases, and can be predictors of severity of pathologies such as sickle cell disease and malaria infection. However, like most GWAS, the majority of these variants are non-coding, making it challenging to attribute causal genetic mechanisms.

Hypothesis: We can identify genes affecting RBC traits using genomic, transcriptomic, and epigenomic approaches.

Main goal: Identify and characterize genes affecting red blood cell traits.

Specific objectives: First, I developed a method to test the association between predicted human gene knockouts and complex traits (chapter 2). Then, I compared DNA methylation (chapter 3) and transcriptomic (chapter 4) patterns of fetal and adult RBC progenitors. Finally, I used expression quantitative trait loci (eQTL) mapping followed by gene editing experiments in human erythroblasts to identify and characterize causal genetic variants affecting RBC traits (chapter 5).

Expected impact:

- Provide transcriptomic and epigenomic resources for RBC studies and GWAS loci characterization
- Provide a better understanding of the underlying genetic mechanisms regulating RBC traits
- Identify novel therapeutic targets to treat RBC-related diseases such as SCD or malaria infection

Chapter 2: Testing the Role of Predicted Gene Knockouts in Human Anthropometric Trait Variation

Authors

Samuel Lessard, Alisa K. Manning, Cécile Low-Kam, Paul L. Auer, Ayush Giri, Mariaelisa Graff, Claudia Schurmann, Hanieh Yaghootkar, Jian'an Luan, Tonu Esko, Tugce Karaderi, NHLBI GO Exome Sequence Project, GOT2D, T2D-GENES, GIANT Consortium, Erwin P. Bottinger, Yingchang Lu, Chris Carlson, Mark Caulfield, Marie-Pierre Dubé, Rebecca D. Jackson, Charles Kooperberg, Barbara McKnight, Ian Mongrain, Ulrike Peters, Alex P. Reiner, David Rhainds, Nona Sotoodehnia, Joel N. Hirschhorn, Robert Scott, Patricia B. Munroe, Timothy M. Frayling, Ruth J.F. Loos, Kari E. North, Todd L. Edwards, Jean-Claude Tardif, Cecilia M. Lindgren, Guillaume Lettre

Reference

Lessard S, Manning AK, Low-Kam C, Auer PL, Giri A, Graff M, et al. *Testing the role of predicted gene knockouts in human anthropometric trait variation*. Human molecular genetics. 2016;25(10):2082-92.

2.1 Context

A potential approach to characterize genome-wide association study (GWAS) loci is through rare coding variant studies. If in a specific loci a gene contains rare variant associated with a complex trait of interest, then one can hypothesize that (independent) neighboring non-coding variants associated with the same trait act through modulation of that gene's regulation. A related approach consist of gene knockout studies, where loss-of-function mutation are created in genes in order to characterize their function. Although humans are not amenable to such approaches, the study of naturally occurring gene knockout is still a powerful approach to study gene functions, as evidenced studies of monogenic diseases such as cystic fibrosis. Although these genetic mutations can have large phenotypic effects, they are usually very rare and are unlikely to contribute to trait variation in the general population. The goal of this study was to identify gene knockout events with modest phenotypic effect and determine whether they contribute to human trait variation. Although we tested the method on human anthropometric traits, the goal was to determine if current sequencing and genotyping methodologies could capture such events using large samples. Moreover, the method proposed here can be extended to red blood cell traits (or any complex trait).

2.2 Abstract

Although the role of complete gene inactivation by two loss-of-function mutations inherited in *trans* is well-established in recessive Mendelian diseases, we have not yet explored how such gene knockouts (KOs) could influence complex human phenotypes. Here, we developed a statistical framework to test the association between gene KOs and quantitative human traits. Our method is flexible, publicly available, and compatible with common genotype format files (*e.g.* PLINK and vcf). We characterized gene KOs in 4,498 participants from the NHLBI Exome Sequence Project (ESP) sequenced at high coverage (>100X), 1,976 French Canadians from the Montreal Heart Institute Biobank sequenced at low coverage (5.7X), and >100,000 participants from the GIANT Consortium genotyped on an exome array. We tested associations between gene KOs and three anthropometric traits: body mass index (BMI), height

and BMI-adjusted waist-to-hip ratio (WHR). Despite our large sample size and multiple datasets available, we could not detect robust associations between specific gene KOs and quantitative anthropometric traits. Our results highlight several limitations and challenges for future gene KO studies in humans, in particular when there is no prior knowledge on the phenotypes that might be affected by the tested gene KOs. They also suggest that gene KOs identified with current DNA sequencing methodologies probably do not strongly influence normal variation in BMI, height, and WHR in the general human population.

2.3 Introduction

The identification of complete loss-of-function (LoF) alleles (*i.e.* genetic null or amorphic alleles) is a powerful strategy to characterize gene functions through random (*e.g.* chemical mutagenesis) or targeted (*e.g.* knockout (KO) methodology in the mouse, RNAi) genetic experiments. In contrast to model organisms, humans are not amenable to such genetic manipulations. Yet, there is tremendous biomedical interest in understanding how the complete disruption of both copies of a gene may impact human biology [302]. Our complex physiology, interactions with our environment, and gene redundancy within our genome are only few of the reasons highlighting the importance of describing the phenotypic consequences of gene inactivation in humans. From a drug development perspective, the identification of humans with gene KOs also offers naturally occurring genetic experiments to assess the potential pleiotropic effects of candidate target genes [182].

Mendelian diseases, such as sickle cell anemia [MIM 603903] and cystic fibrosis [MIM 219700], offer an entry point into the study of gene functions in humans. Indeed, the study of these conditions continues to yield important insights into human biology in health and disease [303]. But only a limited number of genes have been implicated in Mendelian diseases: as of October 13 2015, there were 4,651 genes in the Online Mendelian Inheritance in Man (OMIM) database with phenotype-causing mutations. Furthermore, these mutations are often rare such that it is difficult to assemble sufficiently large cohorts of patients to study their pleiotropic effects. Gene KOs can have strong phenotypic effects on anthropometric traits in the context of

Mendelian disorders or syndromes, as evidenced by mutations causing early-onset morbid obesity (*PCSK1*, *LEPR*) or dwarfism (*GH1, GHR, ATR*) [304-306]. These mutations are rare (often private) and unlikely to be involved in anthropometric trait variation in the general population. However, the possibility that gene KOs of more subtle effect might influence normal variation in anthropometric traits remains to be investigated.

Large-scale whole-exome and -genome sequencing projects are beginning to systematically catalogue coding genetic variation in the human genome, including predicted LoF variants [164,171,173,307,308]. On average, there are ~100-200 LoF variants per individual, resulting in ~20 genes that are inactivated through homozygosity or compound heterozygosity [9]. These numbers include mostly common variants, which are more likely to be phenotypically neutral given the effect of purifying selection [309]. Limiting to variants with a minor allele frequency (MAF) <0.5%, the 1000 Genomes Project estimated that there are 10-20 LoF variants per individual [307]. LoF variants are usually defined as variants that truncate protein sequences (nonsense and frameshift insertion-deletion (indel)) or that abrogate splice sites or stop codons (stop-loss) [9]. Using this definition of LoF variant, and limiting their analyses to variants with a MAF <2%, Sulem et *al.* found that ~8% of 104,220 Icelanders carry at least one complete gene KO, and that most gene KOs are seen in <5 individuals [165].

Recently, several studies have explored the link between gene KOs and human complex phenotypes, such as chronic diseases [9,158,161,166] and autism [154]. As mentioned above, it is well-established that rare gene inactivation can cause extreme anthropometric phenotypes in several human recessive disorders. The goal of our study is to extend this observation and determine whether gene KOs of modest phenotypic effect also contribute to anthropometric trait variation in the general human population. We developed a statistical method to test for association between predicted gene KOs and quantitative human phenotypes and characterized the distribution of predicted gene KOs in 2,772 European Americans and 1,726 African Americans from the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequence Project (ESP). We then applied our method to detect associations between gene KOs and three quantitative anthropometric traits (body mass index (BMI), adult height, and BMI-adjusted waist-to-hip ratio (WHR)) using high coverage whole-exome sequence (WES) data from 4,498 ESP participants,

low coverage whole-genome sequence (WGS) data from 1,969 French Canadians, and >100,000 participants from the GIANT Consortium genotyped on an exome array.

2.4 Results

2.4.1 Number and distribution of predicted gene KOs in ESP

We identified 18,137 and 21,935 LoF variants in 1,726 African Americans and 2,772 European Americans from ESP, respectively (**Table 1**, **Table S1**). These LoF variants included protein truncating (nonsense, frameshift indel), stop-loss and splice site variants. On average, we found 65 and 39 rare or low-frequency LoF variants (MAF <5%) per African-American and European-American ESP participant, respectively (**Table 1**). These numbers are higher than some of the previous estimates [9,154,158], mostly because we included frameshift indels in our analyses. When excluding frameshift indels, we found on average 26 and 16 LoF variants with MAF <5% per ESP African American and European American, respectively. Descriptive statistics on the number of LoF variants in ESP after excluding frameshift indels are available in **Table S2**. We screened the ESP dataset for individuals who are homozygous or compound heterozygous for LoF variants, and are therefore predicted KOs for a given gene. To detect compound heterozygosity, we used phased genotype information generated with the software Beagle to distinguish between LoF variants inherited in *cis* or *trans* (**Table 1**) [310]. The identification of LoF variants depends on the gene annotation used. To address this concern, we re-analyzed the ESP WES data using the GENCODE basic transcripts annotation instead of RefSeq, and only considered variants that fell within all transcripts for a given gene. We obtained very similar association results between the two annotation software (**Figure S1**). We present below results generated with the RefSeq annotation.

Common LoF variants are responsible for most predicted gene KOs (**Figure 1,** and **Figure S2** for distributions without frameshift indels). For instance, in ESP African Americans, we found on average 25.9 and 2.5 predicted gene KOs per individual when analyzing all or rare/low-frequency LoF variants, respectively (**Table 1**). The corresponding numbers in European Americans are 23.2 and 1.1 for all and rare/low-frequency LoF variants (**Table 1**).

While this article was under review, the Exome Aggregation Consortium (ExAC) reported an average of 35 homozygous protein-truncating variants per individual. This number is higher than the average number of homozygous LoF variants that we found in ESP (~21-23/participant, **Table 1**) [163]. This difference might simply reflect increased power in ExAC to discover rare mutations owing to its larger sample size (N=60,706 *vs* N=4,498 for ESP). Because common LoF are more likely to be phenotypically neutral [309], we focused all subsequent analyses on LoF with MAF <5% within ethnic group or sub-study. In the ESP dataset, we found 2,071 and 1,433 genes with both alleles inactivated by such LoF variants in at least one African American or one European American, respectively (**Table 1**). The higher number of predicted gene KOs in African Americans has been previously observed and is consistent with increased genetic diversity in African-ancestry populations [9]. Overall, very few individuals shared the same gene KOs, most of them being found in only one individual (**Figure 1**). Homozygosity of LoF variants is responsible for the majority of these KO events as we only found (after taking phase information into account) compound heterozygous individuals for ~8% of the genes with at least one gene KO (**Table 1**). Stop-loss variants might not be as detrimental as other categories of LoF variants, but they are implicated in less than 0.9% of all gene KOs identified in ESP.

2.4.2 Predicted gene KO associated with anthropometric traits in ESP

We tested our newly developed method (**Figure 2**) on three anthropometric traits (BMI, height, and WHR) that are available in a large number of ESP participants. We stratified our analyses by ethnicity and meta-analyzed association results (**Figure 3**). Assuming that most genes are independent and given the number of genes for which we could find at least one predicted knocked out individual, we used the following Bonferroni-corrected significance threshold to declare significance: $\alpha=2\text{x}10^{-5}$. No single genes reached this significance threshold for any of the three tested anthropometric traits after meta-analysis.

To increase statistical power, we attempted to replicate genes with a nominal $P<0.05$ in the ESP dataset using the WGS data from the MHI Biobank (N=1,976). We limited our analysis to genes with at least two KO individuals. Although the MHI Biobank dataset results from low-pass WGS, the number of identified LoF variants and gene KOs was similar to the number

observed in ESP (**Table S1**), suggesting that the data is sufficiently comprehensive to support these analyses. We found that 30-40% of gene KOs in ESP were also knocked out in the MHI Biobank, highlighting the challenge to replicate such studies in humans. This might particularly be true for gene KOs observed only in ESP African Americans given that the MHI Biobank includes individuals of European ancestry. We combined the ESP and MHI Biobank results but we did not observe any significant associations with quantitative anthropometric traits (**Table S3**). We report results with a meta-analysis $P<0.005$ in **Table 2**. The most promising gene KO association that we found is between *PKHD1L1* and lower BMI: we found 20 KO individuals for this gene who have on average a BMI that is 0.8 standard deviation (SD) below the population mean (corresponding to ~-3.6 kg/m^2). *PKHD1L1* may play a role in immunity [311].

While examining the top candidate genes, we noticed that *PKHD1L1* is a large gene (78 exons, coding sequence is ~14 kilobases), raising the possibility that our method could favor longer genes. In the ESP dataset, we found, as expected, that the number of LoF variants in a given gene is strongly correlated with the length of the coding sequence or the number of exons (all $P<1x10^{-67}$). However, the number of individuals who carry a rare gene KO is not correlated with the length of the coding sequence or the number of exons of the gene (all $P>0.2$), except for a weak correlation observed in ESP African Americans with the length of the coding sequence (Pearson's $r=0.066$, $P=0.003$). To exclude the possibility that gene length may influence our results, we tested correlations with association results from the ESP and MHI Biobank combined analyses. With one exception (among 12 correlation tests performed), we found no significant correlations between the length of the coding sequence or the number of exons and association *P*-values for BMI, height, and WHR (all $P>0.25$). In ESP African Americans, there was a weak correlation between the length of the coding sequence and the BMI *P*-values (Pearson's $r=0.069$, $P=0.002$), but it was in the opposite direction from our expectations (shorter genes have slightly more significant *P*-values). Together, these analyses suggest that our method to test association between gene KOs and human quantitative traits is largely insensitive to gene length.

2.4.3 Gene KO identification and association testing using exome array data

Recognizing that the main limitation of our analysis is sample size, we contacted studies that are involved in the GIANT Consortium. Although WES or WGS data is not readily available for most of these studies, they all have genotyped their participants using an exome array that targets 250,000 – mostly coding – variants. We reasoned that the large sample size offered by the GIANT Consortium could compensate for the limited number of variants present on the exome array. We recruited 22 studies, totaling >102,000 individuals (BMI and height available for all, WHR available for >62,000 individuals). Each study ran the method locally, stratifying all analyses by ethnicity, and we then combined results using meta-analysis methodology [312]. The frequency of KO events was similar in ESP and the GIANT studies. However, there were more singletons (genes with a single KO individual) observed in European-ancestry individuals from the GIANT studies because of the very large sample size (**Figure S3**).

We present the BMI, height, and WHR meta-analysis results for the GIANT studies in **Figure 3**. As reported above for the WES sequence datasets, and despite a sample size that is >10-times larger, we could not detect significant associations between gene KOs and quantitative anthropometric traits after accounting for the number of tests performed (**Table 3**). The most interesting finding pertains to the association between height and inactivation of *GRHPH*: autosomal recessive Mendelian mutations in this gene cause primary hyperoxaluria type 2 [MIM 260000][313]. Primary hyperoxaluria type 1 [MIM 259900], a more severe form of the disease caused by mutations in *AGXT*, is characterized by very severe growth failure [314]. However, the connection between primary hyperoxaluria type 2 caused by recessive mutations in *GRHPH* and growth in humans has not been as clearly documented, although there is one case report of a child with this disease and short stature [315].

2.4.4 Prioritizing gene KOs using a candidate-gene approach

We next asked whether we would increase power to detect associations between gene KO and anthropometric traits by restricting our analyses to strong candidate genes. We focused on

subsets of genes that are associated with any phenotypes in OMIM, or genes that are intolerant to LoF mutations based on the Residual Variation Intolerance Score (RVIS) or the probability of being LoF Intolerant (pLI) score [163,316]. We observed several genes that deviate from the null when restricting our analyses to these candidate genes, especially for the OMIM genes in the larger GIANT datasets for BMI and WHR (**Figure 4**). We also reasoned that the Mouse Genome Informatics (MGI) database might be a good source of candidate genes for our human KO experiment. We retrieved the human homologues of genes from 30 MGI phenotype categories, and tested them against anthropometric traits (**Figure S5**). Again, we observed inflation of the KO association results when compared to the null distribution, suggesting that some of these genes might influence anthropometric traits when completely inactivated. The most noticeable result was the distribution of test statistics in the GIANT BMI analysis for genes related to taste and olfaction (**Figure S5**). Genes related to this category were significantly enriched for genes with a BMI *P*-value<0.05 in GIANT (Fisher's exact test *P*=0.008).

Table 1. Number and frequency of predicted gene knockouts (KO) in 1,727 African Americans and 2,772 European Americans from the NHLBI Exome Sequence Project (ESP).

					NOT PHASED		PHASED	
			Variants /individuals	Variants /gene	Gene KOs /individuals	Number of KO genes	Gene KOs /individuals	Number of KO genes
African Americans	All LoF (N=18,137)		237	0.92	33.7	2,530	25.9	2,429
		Homozygotes			23.2	2,384	23.2	2,384
		Compound heterozygotes			10.4	601	2.6	334
	Rare LoF (N=17,446)		65	0.89	4.2	2,174	2.5	2,071
		Homozygotes			2.3	2,028	2.3	2,028
		Compound heterozygotes			1.9	381	0.2	155
European Americans	All LoF (N=21,935)		197	1.12	28.8	1,844	23.2	1,741
		Homozygotes			21.3	1,694	21.3	1,694
		Compound heterozygotes			7.6	487	1.9	247
	Rare LoF (N=21,351)		39	1.09	1.8	1,538	1.1	1,433
		Homozygotes			1	1,390	1.01	1,390
		Compound heterozygotes			0.8	318	0.09	124

For this loss-of-function (LoF) variant analysis, we consider autosomal nonsense, stop-loss and splice site variants, as well as frameshift insertion-deletions (indels). Rare LoF variants have a minor allele frequency <5%. In the absence of phasing information, we assume that rare LoF are inherited in *trans*. As expected, considering phased genotype information significantly impacts the number of gene KOs that we can detect due to compound heterozygosity.

Table 2. Association of gene knockouts (KOs) with anthropometric traits in the Exome Sequence Project (ESP) and Montreal Heart Institute (MHI) Biobank DNA sequencing datasets.

		ESP					MHI			Combined	
Trait	**Gene**	**Mean EA (real units)**	**N_{KO} EA**	**Mean AA (real units)**	**N_{KO} AA**	**P**	**Mean (real units)**	**N_{KO}**	**P**	**Weighted average (real units)**	**P**
BMI	*PKHD1L1*	0.7 (+3.2 kg/m^2)	11	0.5 (+2.3 kg/m^2)	6	0.009	1.6 (+7.2 kg/m^2)	3	0.009	0.8 (+3.6 kg/m^2)	0.0002
	PLIN4	2.7 (+12.2 kg/m^2)	1	3.1 (+14.0 kg/m^2)	1	5x10^{-5}	-0.2 (-0.9 kg/m^2)	2	0.67	1.4 (+6.3 kg/m^2)	0.002
Height	*RMDN2*	NA	0	-1.1 (-7.0 cm)	4	0.03	-1.6 (-10.2 cm)	2	0.02	-1.3 (-8.3 cm)	0.002
	ASIC4	3.6 (23.0 cm)	1	1.5 (9.6 cm)	2	5x10^{-5}	-0.4 (-2.6 cm)	2	0.56	1.2 (+7.7 cm)	0.002
	SH2B2	-1.6 (-10.2 cm)	2	NA	0	0.02	-1.9 (-12.2 cm)	1	0.06	-1.7 (-10.9 cm)	0.003
WHR	*C1QTNF5*	0.6 (+0.04)	1	1.8 (+0.13)	2	0.04	1.5 (+0.11)	2	0.03	1.4 (0.10)	0.003

We attempted to replicate gene KO associations from the ESP whole-exome DNA sequencing dataset in the MHI Biobank whole-genome DNA sequencing dataset. We tested for replication genes with P<0.05 in the ESP dataset and at least two KO individuals in the MHI Biobank. We report genes with a combined *P*<0.005. We provide the mean gene KO effect size in standard deviation (SD) and metric units, assuming that 1 SD corresponds to 4.5 kg/m^2, 6.4 cm, and 0.07 for BMI, height, and WHR respectively. N_{KO}: number of individuals that are KO for the given gene. EA: European-ancestry; AA: African-ancestry.

Table 3. Top association results between anthropometric traits and predicted gene knockouts (KOs) identified using ExomeChip data from 22 studies participating in the GIANT Consortium.

Trait	Gene	N_{KO}	N_{study}	Weighted mean (SD)	P
BMI	*CYP20A1*	100	15	-0.35	0.001
	ME2	2	2	-1.90	0.002
	KIAA1024	7	5	-0.75	0.002
	TBC1D5	4	3	1.05	0.003
	LRRC39	147	16	0.23	0.003
	TAS1R1	191	6	0.15	0.004
	LAMA3	9	2	-1.02	0.004
	KIAA0391	3	2	-1.64	0.004
	TAS2R60	2	2	-2.04	0.005
Height	*GRHPR*	2	2	-2.28	0.0001
	ABCB7	365	10	-0.12	0.0003
	ZDHHC14	3	2	-2.01	0.0003
	ZFPM1	21	3	-0.60	0.0008
	DHX57	2	2	-2.04	0.0009
	CD8A	2	2	2.35	0.001
	CDC42BPA	4	2	1.70	0.001
	NSUN4	13	3	-0.78	0.002
	ARPC5L	6	2	-1.25	0.002
	CCDC125	45	9	0.35	0.002
	BOK	27	4	0.61	0.003
	NSRP1	9	1	-1.00	0.003
	TEX13A	2	1	2.00	0.004
	RPGRIP1	10	4	-0.72	0.004
	SCGN	6	5	-0.96	0.005
WHR	*C18orf56*	7	1	1.39	0.0002
	AARS2	3	2	-1.78	0.001
	C18orf34	6	3	1.27	0.002
	CCDC68	13	1	0.83	0.002
	HRG	3	2	-1.52	0.004
	SPTA1	2	2	1.86	0.004
	SPTBN5	191	11	0.15	0.005

We only report genes with $P<0.005$ and at least two KO individuals. The weighted mean corresponds to the average phenotype (in standard deviation units) of individuals that are KO for this gene. N_{KO}: number of individuals with a KO gene; N_{study}: number of studies with at least one KO individual for a given gene.

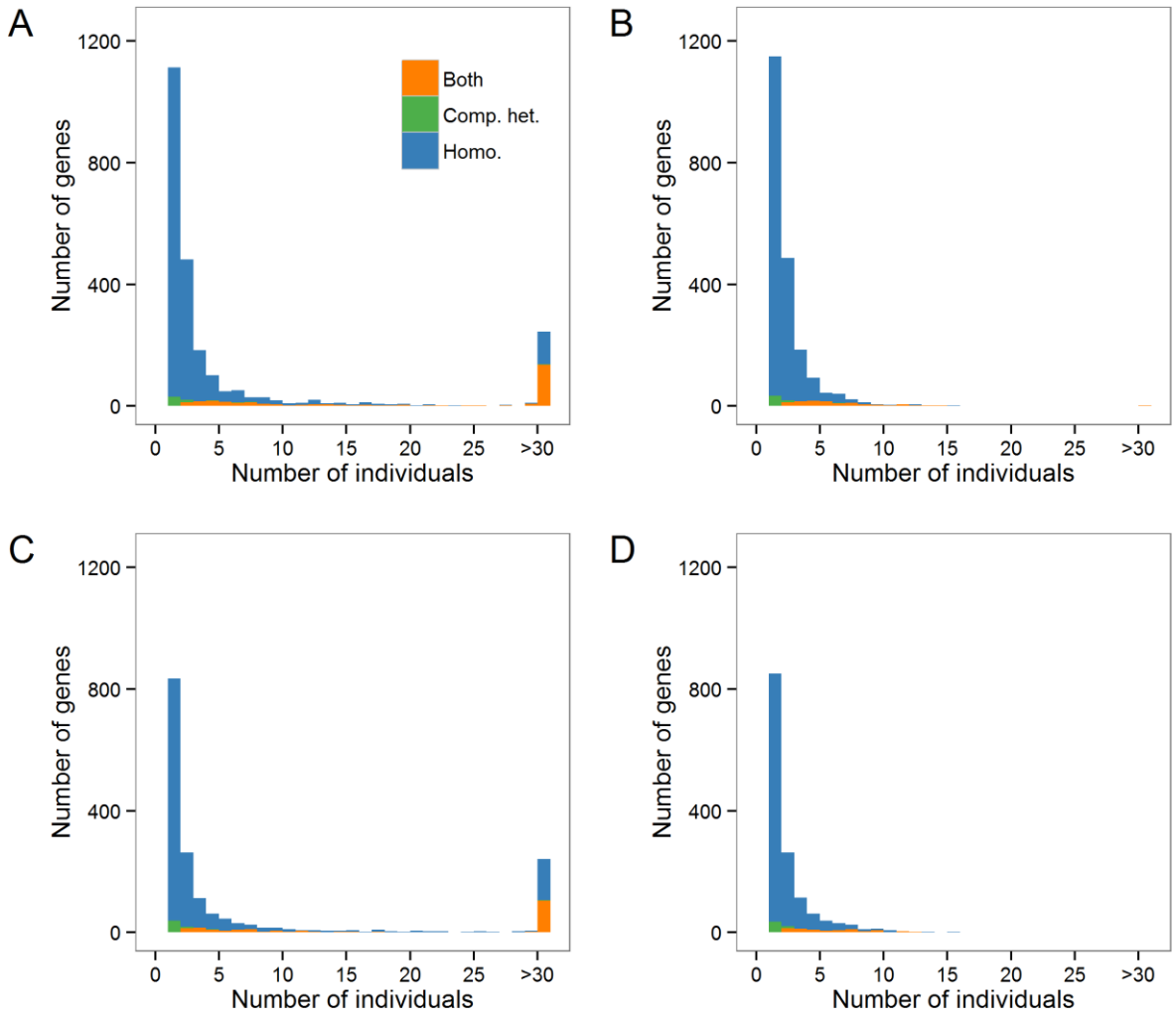


Figure 1. Distributions of the number of NHLBI Exome Sequence Project (ESP) participants with predicted gene knockouts (KOs)

We present distributions in African Americans (**A** and **B**) and European Americans (**C** and **D**). We include all loss-of-function (LoF: nonsense, stop-loss, splice site, frameshift indel) variants in (A) and (C), whereas only rare/low-frequency LoF variants (minor allele frequency <5%) are included in (B) and (D). Homo., gene KO due to homozygosity; Comp. het., gene KO due to compound heterozygosity; Both, genes with homozygous and compound heterozygous LoF variants.

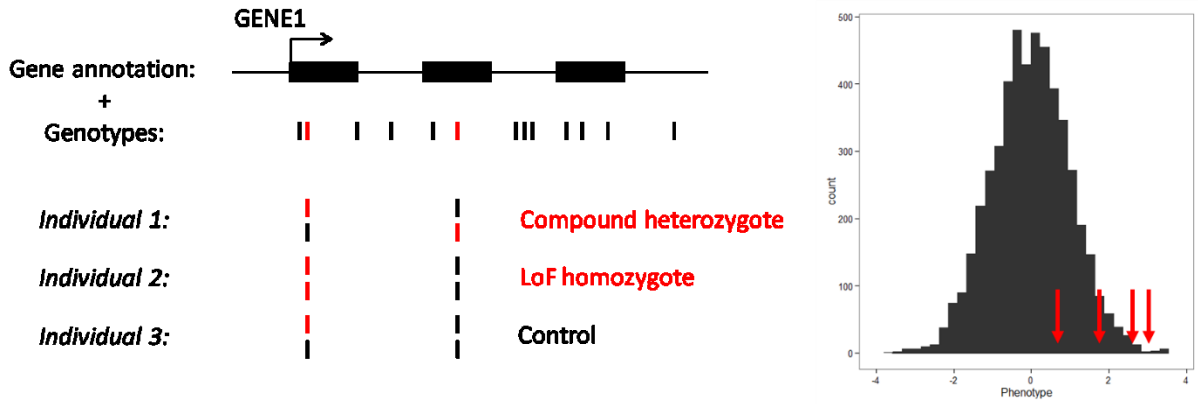


Figure 2. Schematic representation of the method to detect association between gene knockouts (KOs) and human quantitative variation.

This example depicts a fictive gene with three exons (*GENE1*) that contains several SNPs. Our analytical framework only considers loss-of-function (LoF) variants (shown in red). *GENE1* KOs are individuals who are either compound heterozygous or homozygous for LoF variants (individual 1 and 2). The histogram shows the distribution of a normalized human quantitative trait. Our method tests whether individuals that are KOs for a given gene (red arrows) have on average more extreme phenotypes than the rest of the individuals.

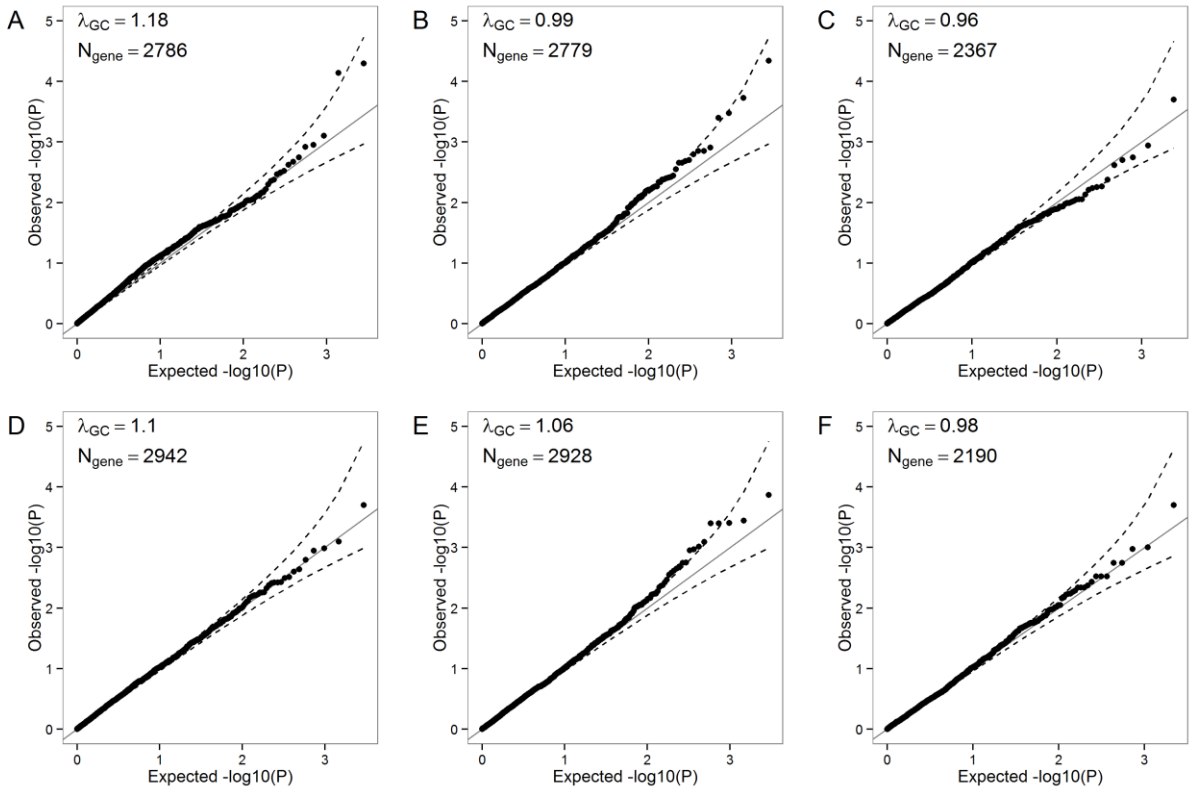


Figure 3. Quantile-quantile (QQ) plots of association results between predicted gene knockouts (KOs) and anthropometric traits in the (A–C) NHLBI Exome Sequence Project (ESP) and (D–F) GIANT ExomeChip datasets

In these datasets, we only considered loss of function (LoF) variants (nonsense, stop-loss, splice site, frameshift indels (ESP only)) with a minor allele frequency (MAF) <5%. We analyzed three anthropometric traits: (A) body mass index (BMI) ($N_{participants}$ = 4475), (B) height ($N_{participants}$ = 4423) and (C) waist-to-hip ratio (WHR) ($N_{participants}$ = 2973). We performed these analyses stratified by ethnicity, and then combined the European American and African American results using meta-analysis methodology. We analyzed the same traits in the GIANT dataset: (D) BMI ($N_{participants}$ = 103 838), (E) height ($N_{participants}$ = 102 775) and (F) WHR ($N_{participants}$ = 62 355). Results are not corrected for the genomic inflation factor. The dash lines correspond to the 95% confidence interval. λ_{GC}, genomic inflation factor; N_{gene}, number of genes with at least one participant that carries two LoF alleles.

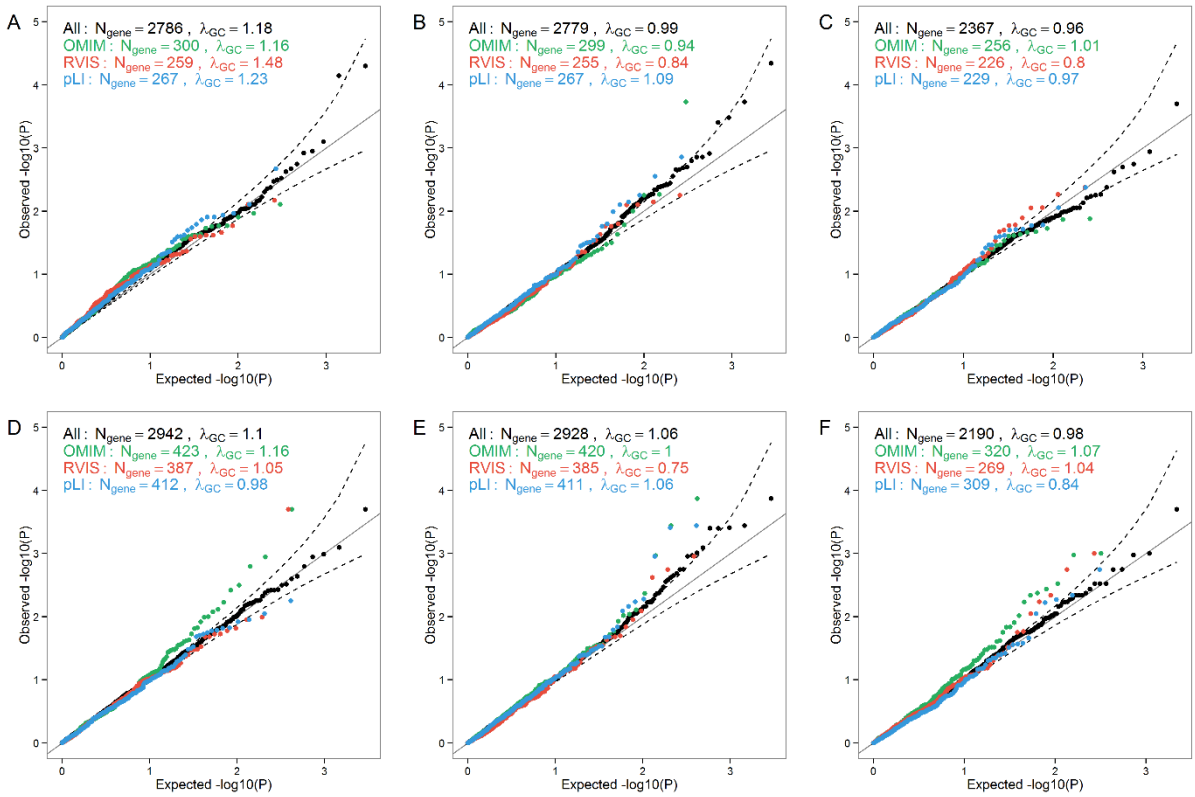


Figure 4. Quantile-quantile (QQ) plots of association results between predicted gene knockouts (KOs) in candidate-genes and anthropometric traits

We restricted these analyses to OMIM disease-causing genes (green), genes with Residual Variation Intolerance Score (RVIS) score <15% of RVIS scores for all genes in the human genome (red), or genes with a probability of being loss-of-function intolerant (pLI) score >0.9 (blue). We report results for three anthropometric traits in the NHLBI Exome Sequence Project (ESP): (**A**) body mass index (BMI) ($N_{participants}$ = 4475), (**B**) height ($N_{participants}$ = 4423) and (**C**) waist-to-hip ratio (WHR) ($N_{participants}$ = 2973). We also report results for the same traits in the GIANT ExomeChip datasets: (**D**) BMI ($N_{participants}$ = 103 838), (**E**) height ($N_{participants}$ = 102 775), and (**F**) WHR ($N_{participants}$ = 62 355). Results are not corrected for the genomic inflation factor. The dash lines correspond to the 95% confidence interval. λ_{GC}, genomic inflation factor; N_{gene}, number of genes with at least one participant that carries two LoF alleles.

2.5 Discussion

We developed a simple statistical method to test the association between predicted gene KOs and human quantitative traits. We tested our method on three quantitative anthropometric traits (BMI, height, and WHR) in large DNA sequencing (ESP and MHI Biobank, >6,400 individuals) and genotyping (22 participating GIANT studies, >102,000 individuals) datasets. Despite this large sample size, we did not identify significant genetic associations with predicted gene KOs, although the association between *PKHD1L1* and BMI or *GRHPH* and height are interesting and should be tested for replication. Within the limitations of our study design (sample size, incomplete catalogue of LoF variants), our results suggest that there are no predicted gene KOs with modest-to-large effect size on anthropometric trait variation in the general population. This conclusion is largely consistent with results from a recent study of homozygous LoF variants in 1,432 individuals [166].

Importantly, our approach and results can guide future gene KO studies in humans. First, our method assumes that all LoF alleles for a given gene will shift the phenotypic mean in the same direction. Although it is a sensitive approach in this first large-scale gene KO experiment for quantitative traits, alternative methods could explore effect on phenotypic variance. Second, in order to maximize our sample size, we combined datasets from different technologies (WES, WGS, exome array). Although we accounted for this technical heterogeneity – gene KO statistics were similar across datasets – this approach could have introduced unanticipated biases. Ideally, high coverage WGS data would be available for gene KO studies. Third, haplotype phasing of DNA sequence data from unrelated individuals (ESP and MHI Biobank), and the lack of phase information for the GIANT ExomeChip studies, has limited our ability to identify compound heterozygous individuals. This could impact our results as nearly 20% of all gene KOs identified in this study were due to compound heterozygosity. We note, however, that excluding compound heterozygotes from the ESP analyses had very limited impact on our results (**Figure S6**). Fourth, we only considered nonsense, splice site, stop-loss and frameshift indels as LoF variants to identify gene KOs. Some of these variants are likely neutral: for instance, genes are more tolerant to non-synonymous variants at the 3' end of a gene, and nearby

variants can rescue the effect of LoF alleles [9]. Furthermore, we excluded missense variants from our analyses, although functional characterization can lead to the identification of missense alleles with strong phenotypic effect on human complex phenotypes [317,318].

The main limiting factors of gene KO studies in humans are the sample size and the depth of genetic information available. We have shown that even when the sample size is very large, most gene KOs are identified in single individuals (**Figure S3**). To be successful, we will need to develop tools to prioritize genes or increase the number of gene KOs. One possibility may be to consider only genes expressed in a tissue that is relevant for the phenotype of interest (*e.g.* genes expressed in growth plates for height). Another promising solution may be to consider KOs in biological pathways instead of single genes as the testing unit. For instance, a researcher interested in blood lipid genetics could pool together all individuals that carry a gene KO in any of the enzymes or transporters implicated in lipid metabolism. We illustrated this candidate-gene approach by prioritizing OMIM disease-causing genes, genes intolerant to LoF mutations, and genes with relevant mouse KO phenotypes. In particular for the BMI analysis, the enrichment of genes with mouse homologues that disrupt taste or olfaction when inactivated is of interest (**Figure S5**). Reverse genetic strategies – finding a function to a gene by first disrupting it – have been very successful in model organisms. Despite early challenges, the large-scale identification of LoF variants and characterization of gene KOs promise to also yield interesting insights into human biology.

2.6 Materials and methods

2.6.1 Ethics statement

This project was approved by the Ethics Committee of the Montreal Heart Institute (#11-1333, #2013-297, #2013-1438).

2.6.2 NHLBI Exome Sequence Project (ESP)

We conducted our initial analysis on the final whole-exome ESP dataset, which is described elsewhere [173]. This dataset was generated from high coverage WES (median depth >100X)[173]. All study participants in each of the component studies provided written informed consent for the use of their DNA in studies aimed at identifying genetic risk variants for disease and for broad data sharing. Institutional certification was obtained for each sample to allow deposition of phenotype and genotype data in dbGaP and BAM files in the short-read archive. We excluded individuals based on sex mismatch between clinical database and genotype-inferred sex (N=13), high homozygosity (N=1), high genotyping missing rate (>10%)(N=1), if they appear as population outliers in principal component analyses (N=30), low concordance to genome-wide association study data (N=4), or unresolved participant identifiers (N=4). Moreover, we randomly excluded one member of each pair of duplicates (N=16), and of first- and second-degree relatives (N=108). We also removed individuals with chronic obstructive pulmonary disease or asthma, as these conditions could influence anthropometric traits (N=688). Finally, we removed participants from the CARDIA (N=201) and MESA (N=406) studies, as requested by investigators from these studies. We kept individuals aged between 21 and 80 years old, height between 125 and 240 cm, BMI <75 kg/m^2, and WHR <1.5. In total, we analyzed anthropometric traits in 1,726 African Americans and 2,772 European Americans (**Table S1**).

2.6.3 Variant quality-control and annotation

We phased variants in the ESP dataset using Beagle 4.0 and the default parameters [310]. We define LoF variants as variants that create or remove stop codons (nonsense and stop-loss) that disrupt essential splice sites (two intronic bases at exon-intron boundaries), or that change the reading frame (frameshift indel). We annotated single-base pair variants using in-house custom scripts and build 37.1 of the human genome reference sequence. We annotated frameshift indels using SeattleSeq (hg19, v.9.03, http://snp.gs.washington.edu/SeattleSeqAnnotation138/). We included in our analyses only frameshift indel variants that fall within validated RefSeq genes (release 69). After filtering out variants with a call rate <95% or a Hardy-Weinberg P<1x10^{-6}, we retained in our analyses

18,137 and 21,935 LoF variants in African- and European-ancestry individuals, respectively (**Table S1**). For comparison, we also annotated ESP variants using Ensembl's Variant Effect Predictor (VEP) module and basic transcripts from GENCODE. We obtained very similar results (**Figure S1**).

2.6.4 Replication cohorts with WGS or WES data available

We used low-pass WGS data (mean coverage 5.7X) from 2,002 French-Canadian participants recruited by the Montreal Heart Institute (MHI) Biobank. Genotypes were imputed and phased using Beagle 4.0 using the default parameters [310]. Individuals were removed due to low or high inbreeding coefficient (N=4). Variants with Hardy-Weinberg $P<1x10^{-8}$ were excluded. In total, 1,976 MHI Biobank participants with anthropometric traits available were included in the replication analyses (**Table S1**).

2.6.5 GIANT Consortium ExomeChip datasets

We analyzed Illumina ExomeChip genotype data from 22 studies that are members of the Genetic Investigation of ANthropometric Traits (GIANT) Consortium (**Table S1**). In total, 103,838, 102,775, and 62,355 individuals were included in the BMI, height and BMI-adjusted WHR analyses, respectively. Individuals were from European- (N=90,927; 19 studies), African- (N=7,576; 2 studies), and Hispanic-ancestry (N=5,335; 1 study). To increase the number of LoF variants available on the ExomeChip, we broaden our definition of splice-site variants to include variants located two base pairs on either side of exon-intron boundaries. This is the splice-site definition implemented by dbNSFP [319] and used by GIANT across the Consortium's ExomeChip effort. Using the most severe annotation from ENSEMBL's VEP tool, we found that 17.8% (797/4483) of these splice site-variants disrupt a canonical splice-site, 46.7% (2094/4483) are missense variants, and 31.6% (1419/4483) affect a nucleotide around the splice-site (1-3 bases within exon or 3-8 bases within intron). Phasing information was not available for the GIANT exome array data. Because we focused on rare variants, we assumed that when two rare LoF variants were observed in the same gene in the same individual, they were inherited in *trans* to create a compound heterozygous gene KO.

2.6.6 Statistical analyses

We developed a flexible method to determine if the complete inactivation of genes by LoF variants is associated with human quantitative traits (**Figure 2**). For each gene, our method searches for individuals that are either homozygotes or compound heterozygotes for LoF variants; we refer to these individuals as predicted KOs. For X-linked markers that fall outside of the pseudoautosomal regions, we consider predicted gene KOs in men if they carry a single LoF variant. For compound heterozygosity, we use phase information to distinguish LoF variants that segregate on the same haplotype (in *cis*) or on different haplotypes (in *trans*). When phasing information is not available (*e.g.* GIANT ExomeChip data), we assume that rare LoF variants segregate on different haplotypes. The method then calculates for each gene the phenotypic mean in predicted KO individuals. Finally, it computes statistical significance using phenotype permutations, as follows:

$$P_{left} = \frac{\sum_{i=1}^{n} \mathbb{I}_{m_i \leq m}}{n} \; ; \; P_{right} = \frac{\sum_{i=1}^{n} \mathbb{I}_{m_i \geq m}}{n}$$

$$P_{final} = 2 \times \text{minimum}(P_{left}, P_{right}),$$

where $\mathbb{I}$ is the indicator function, m is the observed mean phenotype in predicted KO individuals, m_i is the i^{th} permuted mean, n is the number of permutations, P_{left} and P_{right} are the left- and right-tail P-values, and P_{final} is the reported two-tailed P-value. Using simulated null phenotypes and the ESP dataset, we showed that the test is well-calibrated (**Figure S4**). This method assumes that gene inactivation results in the same phenotypic effect (increase or decrease trait value) in all predicted KO individuals for a given gene. The current implementation of our method also currently assumes that tested individuals are unrelated and that the phenotypic distributions are symmetrical. It is compatible with standard genotype file formats (*e.g.* PLINK, vcf). The scripts to run our method are publicly available at: http://www.mhi-humangenetics.org/en/resources.

2.6.7 Association of rare predicted gene KOs with anthropometric traits

We analyzed BMI, adult height and BMI-adjusted WHR. We stratified all our analyses by ethnic group, and we only considered rare or low-frequency LoF variants with MAF <5%.

We used 10,000 permutations to assess statistical significance. For genes with an empirical $P<2x10^{-4}$ (*i.e.* permuted means were never higher (or lower) than the observed mean among 10,000 permutations), we re-ran the analysis using 100,000 permutations: only two genes fell in that category (*BRPF1* $P_{height}=1.8x10^{-4}$; *SPZ1* $P_{WHR}=2.2x10^{-4}$). For ESP samples, we corrected anthropometric traits for sex, age, ESP phenotype groups, exon capture reagents and the first three principal components, as recommended by the ESP investigators. We then applied inverse normal transformation on the residuals from the previous correction. For the MHI Biobank, and the GIANT studies, each anthropometric trait was corrected for sex, age, age-squared and the first ten principal components, and we normalized the resulting residuals using inverse normal transformation. Taking into account the direction of the effect, we combined results across studies using a weighted *Z*-score meta-analysis method implemented in the software METAL, where the weight is the sample size of the corresponding study [312]. To estimate statistical power of our approach, we modeled the effect of a recessive LoF variant on a normally distributed quantitative trait, as previously described [320]. This is a simplistic model as we ignore the presence of additional LoF variants in the same gene, which are considered in our method because they can lead to additional individuals that have a predicted gene KO. We assume that the variant has a MAF=5%, explains 1% of the genetic variance, and used a sample size of N=4,500 (corresponding to ESP), $\alpha=2x10^{-5}$ (Bonferroni correction for the number of genes with KOs), and 5,000 simulations to perform power calculations. Under this scenario, our gene KO approach would have 95% power to detect the association. Alternatively, testing the association while assuming that the variant has an additive effect would result in only 3% power. Using the same assumptions, we estimated 64% and 1% power for a variant that explains 0.5% of the variance when tested using our gene KO methodology or a simple additive model, respectively.

2.6.8 Candidate-gene enrichment analyses

We explored whether prioritizing gene KOs into different categories could increase the chance to reveal an association. First, we investigated whether the gene was an OMIM disease-causing gene, as defined elsewhere [316]. Next, we considered whether the genes were LoF intolerant by either having a Residual Variation Intolerance Score (RVIS) < 15% of the RVIS scores for all genes in the human genome (release 0.3) or a probability of being LoF intolerant

(pLI) score > 0.9 [163,316]. We looked for enrichment by overlapping the QQ-plots of genes belonging to these different categories separately on the QQ-plot containing all genes. We also created subsets of genes based on 30 phenotype categories from the Mouse Genome Informatics (MGI) Database [321]. We tested the enrichment using Fisher's exact test.

2.7 Acknowledgements

We thank all participants involved in this project, and Ekat Kritikou for comments on the manuscript. PBM and MC acknowledge this work forms part of the research program of the NIHR Barts Cardiovascular Biomedical Research Unit. MC is a senior National Institute for Health Research Investigator. Sequencing of the MHI Biobank samples was performed at the McGill University and Génome Québec Innovation Centre. The authors wish to acknowledge the support of the National Heart, Lung, and Blood Institute (NHLBI) and the contributions of the research institutions, study investigators, field staff and study participants in creating this resource for biomedical research. SL is funded by a Canadian Institutes of Health research Banting doctoral scholarship. GL is funded by Genome Canada and Génome Québec; the Canada Research Chair program; and the Montreal Heart Institute Foundation. CML is supported by Wellcome Trust [grant numbers 086596/Z/08/Z, 086596/Z/08/A]; and the Li Ka Shing Foundation. NS is funded by National Institutes of Health [grant numbers HL088456, HL111089, HL116747]. The Mount Sinai BioMe Biobank Program is supported by the Andrea and Charles Bronfman Philanthropies. GO ESP is supported by NHLBI [RC2 HL-103010 to HeartGO, RC2 HL-102923 to LungGO, RC2 HL-102924 to WHISP]. The ESP exome sequencing was performed through NHLBI [RC2 HL-102925 to BroadGO, RC2 HL-102926 to SeattleGO]. EGCUT work was supported through the Estonian Genome Center of University of Tartu by the Targeted Financing from the Estonian Ministry of Science and Education [grant number SF0180142s08]; the Development Fund of the University of Tartu [grant number SP1GVARENG]; the European Regional Development Fund to the Centre of Excellence in Genomics (EXCEGEN) [grant number 3.2.0304.11-0312]; and through FP7 [grant number 313010]. EGCUT were further supported by the US National Institute of Health [grant number R01DK075787].

2.8 Supplementary information

Table S1. Descriptive statistics for the different studies analyzed in this project.

Study (ethnicity)	Sample size (men / women)	Age (men / women)	Height (cm) (men / women)	BMI (kg/m^2) (men / women)	WHR (men / women)	Number of loss-of-function (LoF) variants (all / rare)			
						Nonsense	Stop-loss	Splice site	Frameshift indel
				NHLBI Exome Sequence Project (ESP)					
ESP (AA)	372 / 1354	54.6 / 58.2	176.3 / 162.7	28,9 / 35.2	0.95 / 0.84	5787 / 5611	288 / 249	2705 / 2584	9357 / 9002
ESP (EA)	994 / 1778	50.5 / 58.9	176.6 / 161.7	28.4 / 28.2	0.97 / 0.84	7773 / 7623	274 / 242	3500 / 3398	10388 / 10088
				MHI Biobank whole-genome sequencing					
MHI (EA)	1436 / 531	65.1 / 69.7	172.7 / 157.4	29.4 / 28.9	0.99 / 0.88	3399 / 3361	211 / 187	1761 / 1710	3397 / 3174
				GIANT ExomeChip studies					
ARIC (AA)	1271 / 2086	53.9 / 53.5	176.1 / 163.1	27.7 / 31.0	0.94 / 0.91				NA
ARIC (EA)	5102 / 5771	54.7 / 54.0	176.2 / 162.0	27.4 / 26.6	0.97 / 0.89				NA
ASCOT-SC (EA)	1833 / 629	62.5 / 64.4	176 / 160	28.6 / 28.9	NA				NA
ASCOT-UK (EA)	2656 / 587	65/63	174 / 161	29 / 28	NA				NA
BioME (AA)	1530 / 2682	52.8 / 54.1	177.9 / 163.9	28.5 / 31.6	NA				NA
BioME (EA)	1475 / 1135	66.3 / 66.0	177.2 / 162.5	27.7 / 26.2	NA				NA
BioME (HA)	1980 / 3355	54.3 / 56.2	172.1 / 159.4	28.9 / 29.8	NA				NA
BIOVU (EA)	9535 / 11391	58.1 / 55.4	178.5 / 163.3	28.3 / 27.9	NA				NA
BRIGHT (EA)	308 / 480	57.4 / 58.6	173.4 / 160.8	27.7 / 27.4	0.94 / 0.83				NA
EFSOCH (EA)	726 / 794	32.9 / 30.4	177.7 / 165.0	26.6 / 25.0	0.89 / 0.81				NA
EGCUT (EA)	2041 / 2384	48.6 / 49.2	178.2 / 164	28.5 / 28.7	0.95 / 0.82				NA
Endo (EA)	0 / 533	NA / 37.7	NA / NA	0 / 24.4	NA / NA				NA
EXTEND (EA)	610 / 971	58.4 / 55.33	175.9 / 163.2	28.1 / 26.8	0.94 / 0.81	5658 / 2136	244 / 93	10667 / 4483	NA
InterAct – Cases (EA)	1289 / 1352	55.0 / 55.8	173.5 / 160.3	29.3 / 30.3	0.98 / 0.85				NA
InterAct – Subcohort (EA)	1644 / 3085	51.5 / 51.7	174.0 / 161.4	26.4 / 25.5	0.93 / 0.80				NA
MHI (EA)	5629 / 3956	64.4 / 62.8	172.4 / 159.2	28.9 / 28.0	0.98 / 0.87				NA
PICOS (EA)	0 / 582	NA / 32.1	NA / 165.0	NA / 28.0	NA / 0.80				NA
PIVUS – men (EA)	487 / 0	70.1 / NA	175.9 / NA	27.0 / NA	0.94 / NA				NA
PIVUS – women (EA)	0 / 474	NA / 70.3	NA / 162.3	NA / 27.1	NA / 0.86				NA
ULSAM	1102 / 0	71.0 / NA	174.9 / NA	26.2 / NA	0.94 / NA				NA
RISC (EA)	156 / 157	44.7 / 45.8	178.5 / 164.9	26.0 / 25.2	0.93 / 0.82				NA
WHI (EA)	0 / 22125	NA / 66.2	NA / 161.5	NA / 28.3	NA / 0.82				NA

For age, height, body mass index (BMI) and waist-hip ratio adjusted for BMI (WHR), we provide the mean trait value by sex. For the GIANT ExomeChip studies, we provide the maximum number of loss-of-function (LoF) variants available based on the array content. Rare LoF variants have a minor allele frequency <5% (for the ExomeChip datasets, we used allele frequencies from the MHI Biobank to fill in this table). AA, African Americans; EA, Europeans or individuals of European-ancestry. NA, not availabl

Table S2. Number and frequency of predicted gene knockouts (KOs) in 1,785 African Americans and 2,896 European Americans from the NHLBI Exome Sequence Project (ESP)

				NOT PHASED		PHASED	
		Variants /individuals	***Variants /gene***	***KO Events /Individuals***	***Number of KO Genes***	***KO Events /Individuals***	***Number of KO Genes***
	All LoF (N=8,802)	112.4	0.45	12.9	497	12.3	474
	Homozygotes			11.6	447	11.6	447
African-	Compound heterozygotes			1.32	156	0.71	112
Americans	Rare LoF (N=8,505)	25.7	0.43	0.41	275	0.29	255
	Homozygotes			0.24	230	0.24	230
	Compound heterozygotes			0.17	70	0.05	47
	All LoF (N=11,687)	92.1	0.60	12.0	342	11.7	328
	Homozygotes			11.1	317	11.1	317
European-	Compound heterozygotes			0.89	103	0.59	74
Americans	Rare LoF (N=11,437)	15.5	0.58	0.15	171	0.14	158
	Homozygotes			0.13	151	0.13	151
	Compound heterozygotes			0.03	37	0.01	20

In comparison with **Table 1**, we consider in this loss-of-function (LoF) variant analysis autosomal nonsense, stop-loss and splice site variants, but excluded frameshift insertion-deletions (indels). Rare LoF variants have a minor allele frequency <5%. In the absence of phasing information, we assume that rare LoF are inherited in *trans*. As we can see, considering phased genotype information significantly impacts the number of gene KOs that we can detect due to compound heterozygosity.

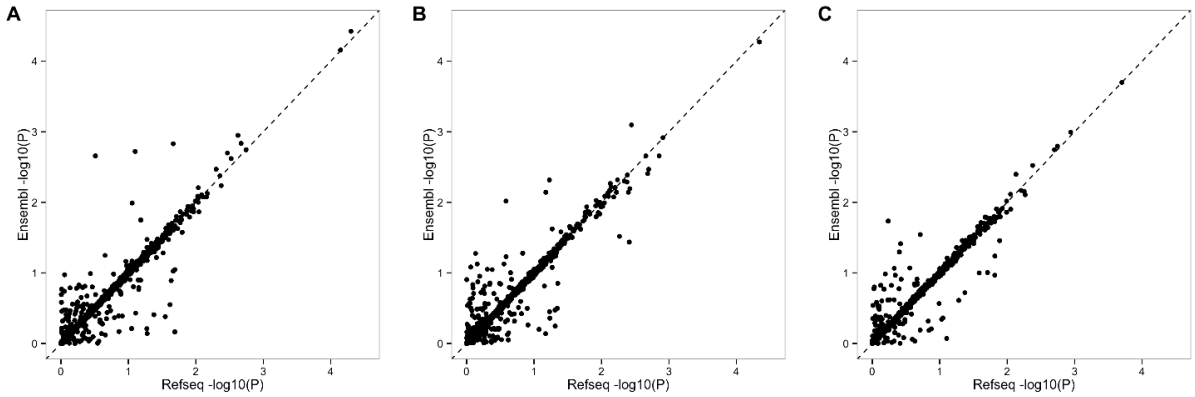


Figure S1. Comparison of association *P*-values using different variant annotations in the ESP dataset for (A) BMI, (B) height, and (C) WHR.

On the *x*-axis, the analysis was performed on variants that fell within validated Refseq genes (release 69) annotated as a frameshift indel, as creating or removing a stop codon (nonsense or stop-loss), or disturbing an essential splice-site. These results are equivalent to the analyses reported in **Figure 3** (**A-C**) of the main manuscript. On the y-axis, variants were chosen to fall within GENCODE basic transcripts (GENCODE version 19), and annotated as having a high impact using the ENSEMBL variant effect predictor tool. Only those predicted as high confidence by the Loss-Of-Function Transcript Effect Estimator (LOFTEE) plugin were kept. In all cases, we observed high correlation between *P*-values generated using both annotations (BMI r^2=0.929, $P<2x10^{-16}$; Height r^2=0.930, $P<2x10^{-16}$; WHR r^2=0.931, $P<2x10^{-16}$). No genes reached statistical significance using the GENCODE annotation.

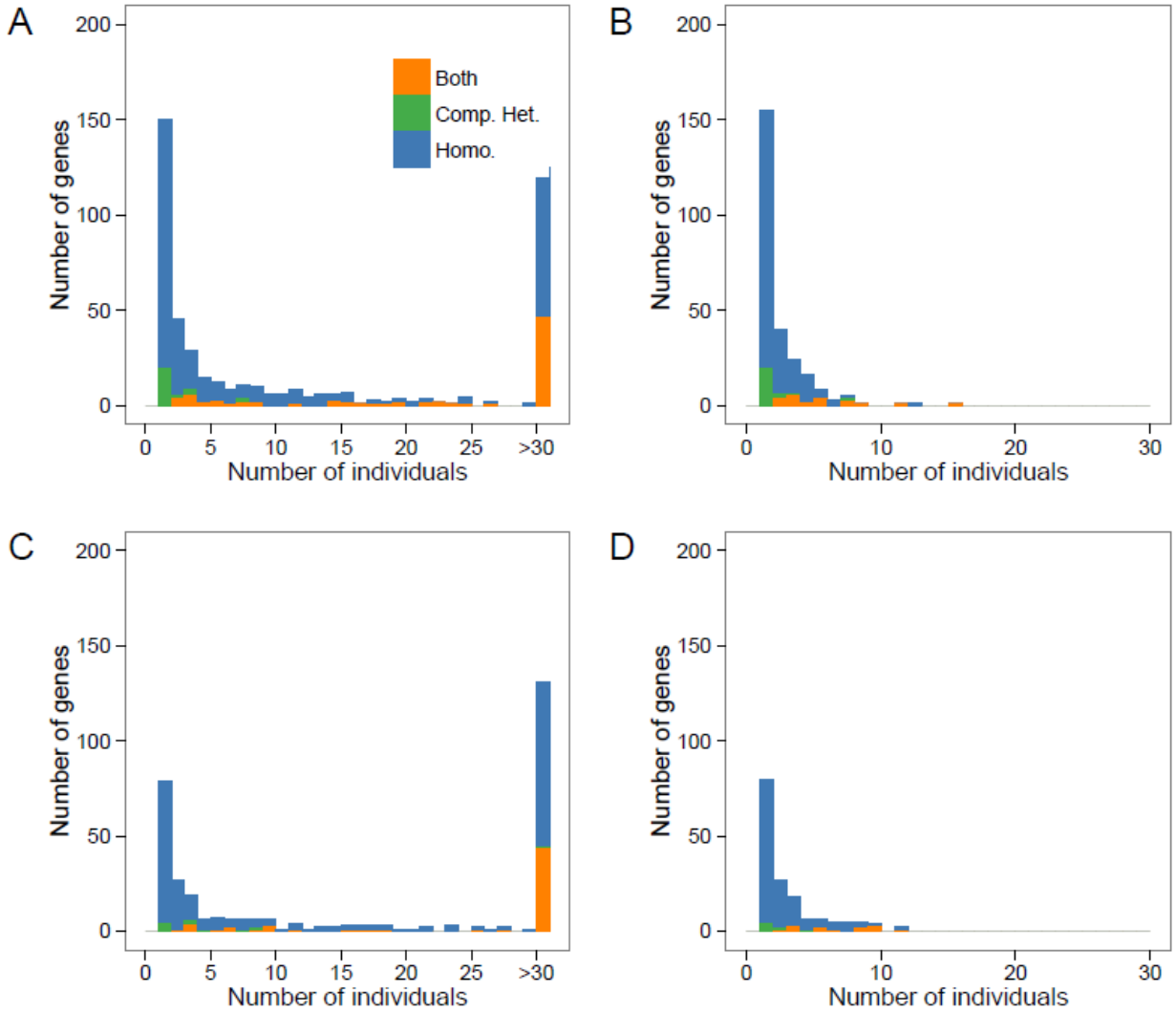


Figure S2. Number of predicted knockout (KO) genes per NHLBI Exome Sequence Project participant.

We present distributions in African Americans (**A** and **B**) and European Americans (**C** and **D**) separately. For these analyses, we did not include frameshift indels. We include common and rare loss-of-function (LoF: nonsense, stop-loss, splice site) variants in **A** and **C**, whereas only rare LoF variants (minor allele frequency <5%) are included in **B** and **D**. Homo.: gene KO due to homozygosity; Comp. het.: gene KO due to compound heterozygosity; Both: genes with homozygous and compound heterozygous LoF variants.

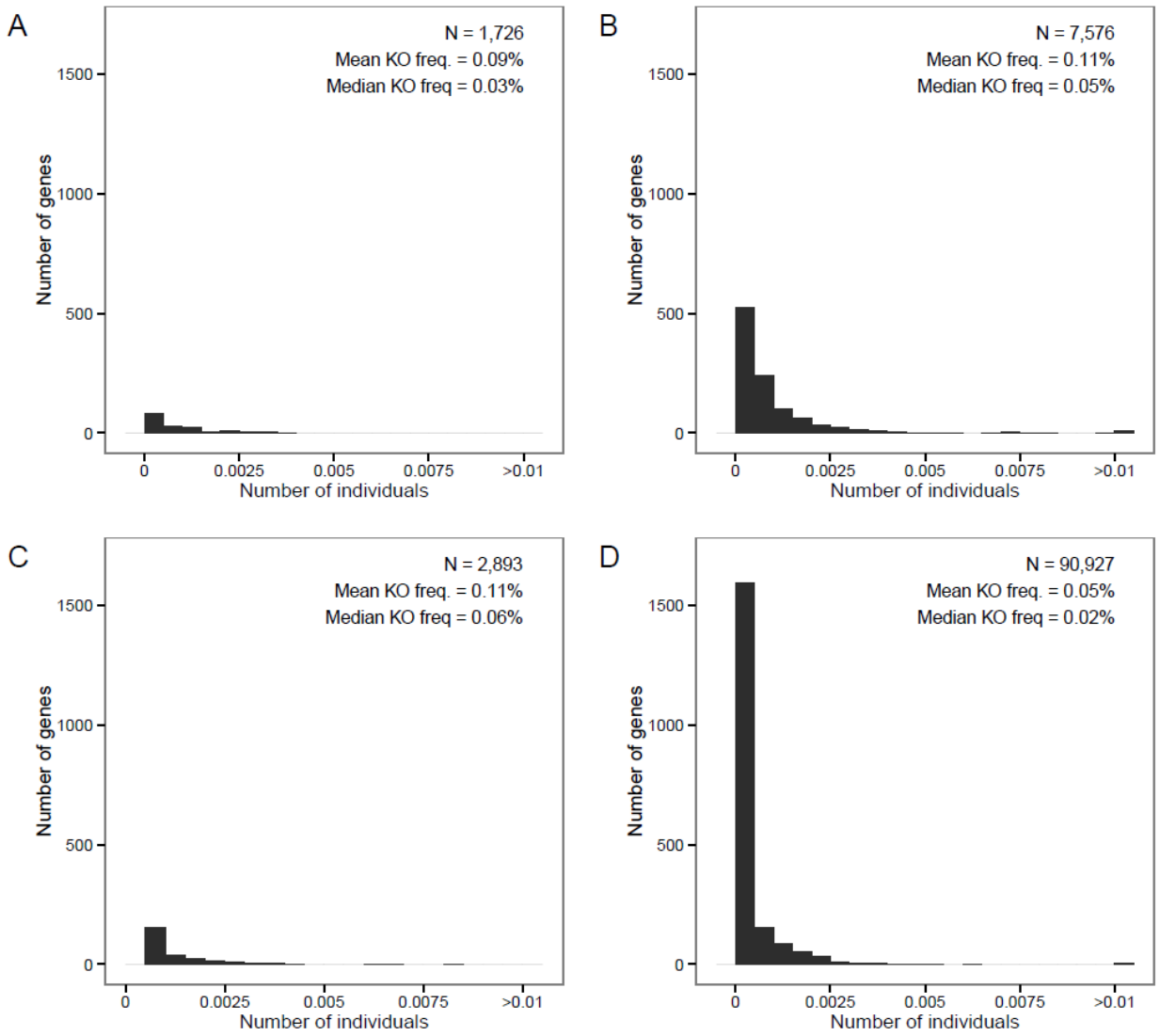


Figure S3. Frequency of KO events in the ESP and GIANT datasets.
Frequency of KOs in individuals of African ancestry are shown for the ESP (A) and GIANT (B) datasets. Frequency of KOs in individuals of European ancestry are shown for the ESP **(C)** and GIANT **(D)** datasets. As we can see in Europe-ancestry individuals from the GIANT studies, increasing the sample size mostly increases singletons that is genes that are KO in a single individual. N: number of individuals.

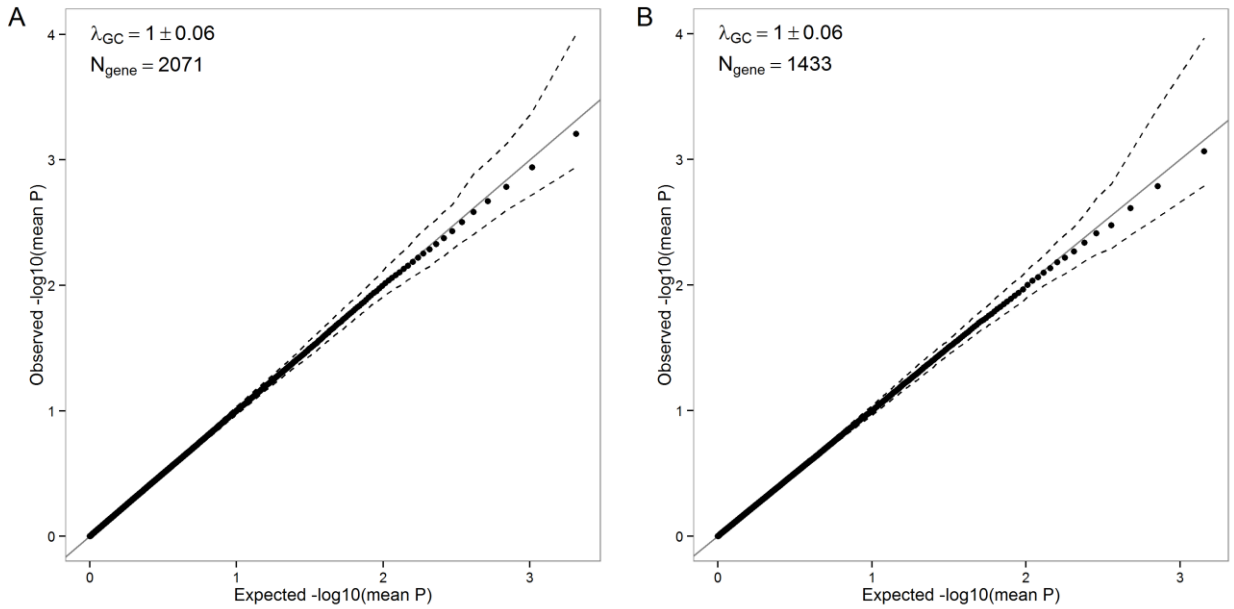


Figure S4. Calibration of our statistical method using the NHLBI Exome Sequence Project (ESP) dataset.

We present quantile-quantile (QQ) plots of associations between predicted KO genes in African Americans (A) and European Americans (B), and 100 randomly generated quantitative phenotypes (mean=0, standard deviation (SD)=1). For each simulation, we ranked P-values and calculated the mean of each rank across simulations. The dots represent this mean, whereas the dotted lines are mean ± 1 SD. We also calculated the mean GC ± SD. As expected, the genomic inflation factor (λ_{GC}) is ~1.

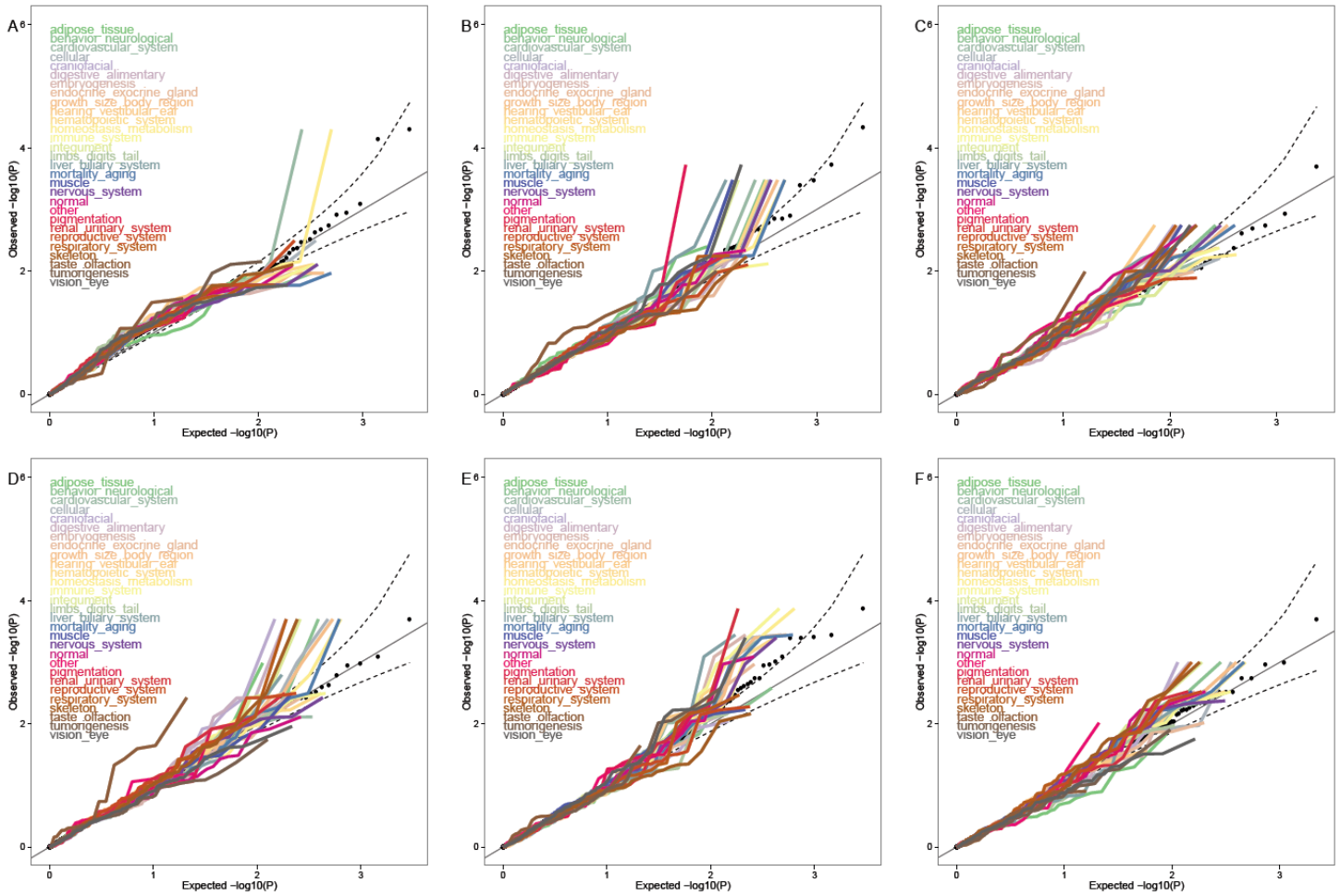


Figure S5. Quantile-quantile (QQ) plots of association results between predicted gene knockouts (KOs) and anthropometric traits restricted to 30 phenotype categories from the Mouse Genome Informatics (MGI) database in the (A-C) NHLBI Exome Sequence Project (ESP) and (D-F) GIANT ExomeChip datasets.

We analyzed three anthropometric traits: (**A**) BMI ($N_{participants}$=4,475), (**B**) height ($N_{participants}$=4,423), and (**C**) WHR ($N_{participants}$=2,973). We analyzed the same traits in the GIANT dataset: (**D**) BMI ($N_{participants}$=103,838), (**E**) height ($N_{participants}$=102,775), and (**F**) WHR ($N_{participants}$=62,355). Results are not corrected for the genomic inflation factor. The dash lines correspond to the 95% confidence interval. λ_{GC}, genomic inflation factor; N_{gene}, number of genes with at least one participant that carries two LoF alleles.

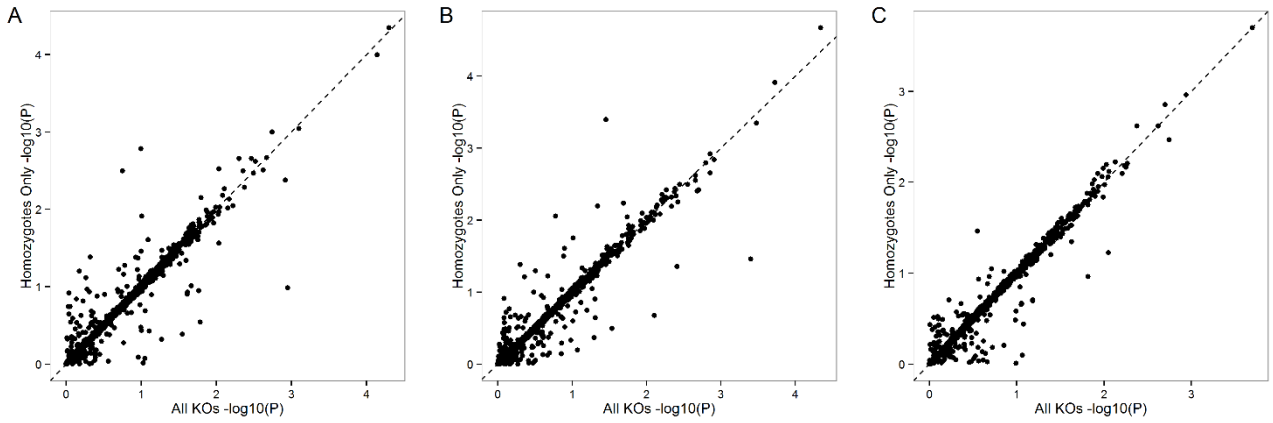


Figure S6. Comparison of association *P*-values using different variant annotations in the ESP dataset for (A) BMI, (B) height, and (C) WHR.

On the *x*-axis, the analysis includes all KO events (compound heterozygotes and homozygotes). On the *y*-axis, compound heterozygotes were excluded. These results are equivalent to the analyses reported in **Figure 3** (**A-C**) of the main manuscript. In all cases, we observed high correlation between *P*-values generated using both annotations (BMI r^2= 0.944, $P<2x10^{-16}$; Height r^2=0.943, $P<2x10^{-16}$; WHR r^2=0.942, $P<2x10^{-16}$).

Chapter 3: Comparison of DNA methylation profiles in human fetal and adult red blood cell progenitors

Authors

Samuel Lessard, Mélissa Beaudoin, Karim Benkirane, Guillaume Lettre

Reference

Lessard S, Beaudoin M, Benkirane K, Lettre G. *Comparison of DNA methylation profiles in human fetal and adult red blood cell progenitors*. Genome medicine. 2015;7(1):1.

Author's contribution

SL performed cell culture and molecular biology experiments, carried out bioinformatic and statistical analyses, and wrote the manuscript. MB performed cell culture experiments. KB quantified hemoglobin in primary erythroblasts using capillary electrophoresis. GL conceived the study, participated in its design and coordination and wrote the manuscript. All authors read and approved the final manuscript.

3.1 Context

Around birth, fetal hemoglobin (HbF) is gradually replaced by adult hemoglobin (HbA). This transcriptionally regulated event can have major health impact, especially in individuals carrying homozygote mutations of the adult β-globin gene leading to disorders like sickle cell disease (SCD). The fetal-to-adult hemoglobin switch is accompanied by DNA hypermethylation of the promoter of the fetal γ-globin gene, but less is known about the global DNA methylation patterns of erythroid cells marking this switch. Understanding the epigenomic mechanisms regulating this developmental event is important since reactivating HbF production is still the most promising therapy for SCD. Here, we characterize global DNA methylation differences between fetal and adult erythroblasts. These DNA methylomes may assist future genome-wide association studies of red blood cell traits – in particular of HbF levels – by pinpointing candidate genes or regions of interest.

3.2 Abstract

Background

DNA methylation is an epigenetic modification that plays an important role during mammalian development. Around birth in humans, the main site of red blood cell production moves from the fetal liver to the bone marrow. DNA methylation changes at the β-globin locus and a switch from fetal to adult hemoglobin production characterize this transition. Understanding this globin switch may improve the treatment of patients with sickle cell disease and β-thalassemia, two of the most common Mendelian diseases in the world. The goal of our study was to describe and compare the genome-wide patterns of DNA methylation in fetal and adult human erythroblasts.

Methods

We used the Illumina HumanMethylation 450k BeadChip to measure DNA methylation at 402,819 CpGs in ex vivo-differentiated erythroblasts from 12 fetal liver and 12 bone marrow CD34+ donors.

Results

We identified 5,937 differentially methylated CpGs that overlap with erythroid enhancers and binding sites for erythropoiesis-related transcription factors. Combining this information with genome-wide association study results, we show that erythroid enhancers define particularly promising genomic regions to identify new genetic variants associated with fetal hemoglobin (HbF) levels in humans. Many differentially methylated CpGs are located near genes with unanticipated roles in red blood cell differentiation and proliferation. For some of these new candidate genes, we confirm the correlation between DNA methylation and gene expression levels in red blood cell progenitors. We also provide evidence that DNA methylation and genetic variation at the β-globin locus independently control globin gene expression in adult erythroblasts.

Conclusions

Our DNA methylome maps confirm the widespread dynamic changes in DNA methylation that occur during human erythropoiesis. These changes tend to happen near erythroid enhancers, further highlighting their importance in erythroid regulation and HbF production. Finally, DNA methylation may act independently of the transcription factor BCL11A to repress fetal hemoglobin production. This provides cues on strategies to more efficiently re-activate HbF production in sickle cell disease and β-thalassemia patients.

3.3 Background

DNA methylation is a dynamic epigenetic mark mostly found on cytosine residues of certain CpG dinucleotides in mammalian genomes. In humans, it is involved in gene imprinting, X-chromosome inactivation and transposable element suppression [322]. DNA methylation is generally associated with transcriptional silencing and plays an essential role in maintaining stem cell pluripotency and controlling other cell- or organ-specific developmental programs [322]. Despite the well-established role of DNA methylation during development, we are only now starting to collect single-base resolution maps of these developmentally regulated DNA methylation changes thanks to improvements in DNA array- and sequencing-based technologies [323]. Such precise DNA methylome maps are important in order to understand how transcriptional networks are controlled during development or how dysregulation of DNA methylation – observed in human diseases like cancer or upon treatment with DNA demethylating agents – may impact cell functions and/or identity [324].

Characterization of mouse models with conditional deletions of genes that encode DNA methyltransferases – enzymes that catalyze the transfer of methyl groups onto DNA – has firmly established that DNA methylation is essential during hematopoietic stem cell differentiation and repopulation [325,326]. Recent reports used enrichment of methylated DNA followed by next-generation DNA sequencing to characterize genome-wide DNA methylation changes in murine hematopoietic stem and progenitor cells, and during erythropoiesis [327,328]. The studies showed a progressive DNA demethylation during erythroid differentiation, and confirmed that groups of CpGs in promoters – so-called CpG islands – tend to be hypomethylated and correlated with active gene expression [327,328].

Although large-scale projects are starting to generate comprehensive DNA methylation datasets [329], they do not capture all stages of human erythroid differentiation and proliferation. In an initial study, investigators monitored genome-wide changes in DNA methylation during human erythropoiesis of bone marrow-derived CD34+ progenitor cells and reported global DNA hypomethylation [330]. We are particularly interested in the epigenetic differences between fetal- and adult-stage erythroblasts that originate, respectively, from the fetal liver and the bone

marrow. The large epigenomic projects (*e.g.* ENCODE, Roadmap Epigenomic, Blueprint) do not profile DNA methylation in these erythroid cells. During this erythroid transition, which occurs around birth in humans, erythroblasts reduce the production of fetal hemoglobin (HbF) and increase the production of adult hemoglobin (HbA) through the transcriptionally regulated fetal-to-adult hemoglobin switch [73]. This gene expression switch is accompanied by progressive DNA hypermethylation of the *HBG2* promoter, which encodes the γ-globin subunit of HbF [188,189]. Understanding the molecular mechanism behind the fetal-to-adult hemoglobin switch is particularly important since re-activating HbF production is the most promising therapy for patients with sickle cell disease and β-thalassemia [73]. *Ex vivo* differentiation protocols now exist to cultivate sufficient number of fetal and adult human erythroblasts to extend the characterization of DNA methylation to the rest of the genome [58].

Here we provide comprehensive DNA methylome maps of human erythroblasts differentiated *ex vivo* from CD34+ progenitor cells purified from the fetal liver or the bone marrow. By analyzing DNA methylation values at 402,819 tested CpGs in fetal and adult erythroblasts, we identified 5,937 differentially methylated CpGs. These differentially methylated regions include the known *β-globin* locus, other genes with known roles in erythropoiesis, as well as several other genes with no previously recognized functions in red blood cell differentiation. We showed that the differentially methylated CpGs cluster within stage-specific erythroid enhancers and are located near binding motifs for transcription factors that regulate hematopoiesis. Finally, we determined that DNA methylation and genetic variation at the *β-globin* locus independently control HbF production in adult-stage cells.

3.4 Methods

3.4.1 Cell culture and differentiation

Primary human fetal and adult CD34+ hematopoietic stem/progenitor cells harvested from 24 anonymous donors (12 of each) were purchased from DV Biologics (http://www.dvbiologics.com/) and Lonza (http://www.lonza.com/), respectively. Primary fetal and adult erythroblasts were generated using two-phase serum-free culture as described previously [58]. Briefly, primary human fetal and adult CD34+ cells are cultured in the expansion medium, composed of StemSpan serum-free expansion medium (SFEM; StemCell Technologies) with 1X CC100 cytokine mix (StemCell Technologies) and 2% penicillin and streptomycin, until day 6. They are then transferred to the differentiation medium, composed of StemSpan SFEM with 2% penicillin and streptomycin, 20 ng/ml stem cell factor, 1 U/ml erythropoietin, 5 ng/ml IL-3, 2 μM dexamethasone, and 1 μM β-estradiol, until cells are collected at day 18. The medium is changed every 3 days. We assessed cell morphology by Wright-Giemsa coloration, and measure cell size and count using the MOXI Z Mini automated cell counter (ORFLO technologies, ID), and hemoglobin production by capillary electrophoresis.

3.4.2 Genomic DNA extraction and methylation assay

We extracted genomic DNA with the Gentra Puregene Cell Kit (Qiagen). DNA was further precipitated with alcohol in order to obtain highly purified DNA. DNA bisulfite conversion was performed using the EZ DNA Methylation Gold Kit (Zymo Research, CA). We used the Infinium HumanMethylation450 BeadChip (Illumina Inc, CA) to measure genome-wide patterns of DNA methylation; the experiment was carried out at the Genome Quebec-McGill Innovation Centre using Illumina's recommended protocol. We assessed data quality with the minfi R package [331] and normalized intensities using the ARRm software, which corrects for probe type, background, dye and position effects [332]. We removed probes that target a genomic sequence annotated to carry genetic variants based on dbSNP version 137 (N=82,694). DNA methylation data have been submitted and are available from the National

Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) repository (accession number [GEO: GSE56491]). To identify single CpG that are differentially methylated between fetal and adult erythroblasts, we converted DNA methylation β-values into M-values [333] and used linear regression in R, correcting for sex effects. To combine results across CpGs located within a functional genomic unit, we used a generalization of Fisher's method that takes into account correlation between nearby CpGs and that is already implemented in the RnBeads software (http://rnbeads.mpi-inf.mpg.de/). We used RefSeq (release 61) coordinates to map CpGs to specific genes. Promoter CpGs are located within 1.5-kilobase upstream of a RefSeq gene and enhancer CpGs map to erythroid enhancers experimentally identified in *ex vivo* differentiated CD34+ cells [334]. In total, we could map 233,894 CpGs to 25,891 gene bodies, 130,854 CpGs to 25,103 promoters and 11,709 CpGs to 4,604 erythroid enhancers. DNA methylation levels at the *HBG2* promoter were measured using the MassArray EpiTYPER platform (Sequenom Inc, CA), which quantitatively analyze fragmented PCR products using MALDI-TOF mass spectrometry. Bisulfite-converted DNA was amplified by PCR as previously described [143].

3.4.3 RNA extraction and gene expression analysis

RNA was extracted using the RNeasy Plus Mini Kit from Qiagen. We determined the quality and quantity of RNA using the RNA 6000 Nano kit on the Bioanalyzer instrument (Agilent). We converted RNA into cDNA with the High Capacity cDNA Reverse Transcription Kit (ABI) and performed quantitative PCR using the Platinum SYBR Green qPCR Mix (Invitrogen). Relative gene expression levels were measured using the $\Delta\Delta C_t$ method and we evaluated statistical significance with *t*-test as implemented in the R software (v. 3.0.0). Primer sequences are listed in **Table S1**. Total RNA for RNA-sequencing was extracted using the miRNEASY kit (Qiagen). Paired-end RNA sequencing was performed on the Illumina Hiseq 2000 platform. Reads were mapped to the genome using Tophat2 (v.2.0.9) and transcript abundances were estimated using Cufflinks (v.2.2.1).

3.4.4 Enrichment analyses

We used binomial tests to measure the enrichment of differentially methylated CpGs in gene bodies, promoters and erythroid enhancers as compared to the rest of the genome surveyed by the Illumina HumanMethylation450 BeadChip. To compare the enrichment of differentially methylated CpGs in erythroid enhancers versus other enhancers present in different cell types, we obtained enhancer coordinates from nine cell lines analysed by the ENCODE Project [216]. We used DAVID to measure the enrichment of gene ontology (GO) terms and biological pathways among genes with a mean β-value difference >0.1 and a combined $P<9x10^{-7}$ in their body or promoter, taking into account the coverage of the Illumina HumanMethylation450 BeadChip [335,336]. To identify transcription factor binding motifs that are enriched at differentially methylated loci, we used the HOMER software and a list of 495 pre-defined transcription factor motifs [337], limiting the search to 200 base pairs on both sides of differentially methylated CpGs. For this analysis, we compared the 5,000 most hypomethylated regions in fetal erythroblasts to the 5,000 most hypomethylated regions in adult erythroblasts, as recommended. HOMER results were stable using different thresholds to select hypomethylated regions.

3.4.5 DNA genotyping and association studies

Genotyping was performed at the Pharmacogenomics Center of the Montreal Heart Institute using the HumanOmniExpress BeadChip array (Illumina). We performed quality-control steps with PLINK [338], removing markers with genotyping success <90% or Hardy-Weinberg $P<1.0x10^{-6}$. All samples had a genotyping success rate >99.8%. We imputed genetic variants with the MaCH/minimac software [339] and reference haplotypes from the 1000 Genomes Project. For association analysis, we only considered genotyped markers or markers with an imputation quality r2_hat >0.6. We tested association between SNP genotypes and methylation levels under an additive model using linear regression as implemented in R. For *BCL11A*, we tested CpGs inside the gene body or the promoter (chr2:60,678,302-60,781,633). For *HBS1L-MYB*, tested CpGs were either inside the gene bodies and promoters of the genes, or in the intergenic region (chr6:135,281,517-135,540,311). For the *β-globin* locus, all CpGs between the promoter of *HBE* and the gene body of *HBB* were included (chr11:5,246,696- 5,527,882).

To test the enrichment of HbF association signals with DNA methylation-implicated erythroid enhancers, we accessed genome-wide association study (GWAS) data of 1,140 African Americans from the Cooperative Study of Sickle Cell Disease (CSSCD). These individuals were genotyped on the Illumina Human610-Quad array, as previously described [340]. SNPs were phased using MaCH (v1.0.16) and imputed with haplotypes from European and African samples generated by the 1000 Genomes Project (phase I) using minimac (v4.4.3). Association P-values with HbF were calculated using mach2qtl (v1.0.8). In total, 6,994,357 SNPs were included in the analysis. We found 63,876 SNPs that overlap with 12,683 erythroid enhancers.

3.5 Results and discussion

3.5.1 *Ex vivo* culture of erythroid progenitor cells

We used a previously established two-phase cell culture protocol to expand and differentiate primary human fetal and adult CD34+ hematopoietic stem/progenitor cells into erythroblasts [58]. For each tissue (fetal liver or bone marrow), we differentiated CD34+ cells from 12 anonymous donors to compare not only DNA methylation changes between fetal and adult erythroblasts, but also between individuals within a developmental stage. We validated the expansion and differentiation cell culture protocol by characterizing cell size, growth and morphology (**Figure S1**). To determine if erythroblasts differentiated *ex vivo* maintain characteristics that are specific to their tissue of origin, we measured expression of genes involved in the fetal-to-adult hemoglobin switch and quantified hemoglobin production by capillary electrophoresis. As expected, erythroblasts derived from bone marrow express significantly more *BCL11A* and *KLF1* than fetal erythroblasts (**Figure S2A-B**). *BCL11A* and *KLF1* are transcriptional repressors of *HBG2* gene expression; *KLF1* also increases *HBB* gene expression, which encodes the β-globin subunit of HbA [58,148,341,342]. Consistently, erythroblasts from the fetal stage produce exclusively HbF (100±0%) whereas differentiated cells from the adult stage produce mostly HbA (84.2±4.8%)(**Figure S2C-E**). Using a targeted assay that measures DNA methylation at seven CpGs within the *HBG2* promoter, we also confirmed

hyper- and hypomethylation of this promoter in adult and fetal erythroblasts, respectively (**Figure S3**)[343]. Taken together, these results confirm that the erythroblasts have maintained their developmental stage specificity.

In the mouse and humans, it is known that global demethylation occurs during erythropoiesis [327,328,330]. For this reason, it is important to confirm that fetal and adult erythroblasts grow and differentiate under similar kinetics. To directly address this concern, we performed a time-course analysis of the expression of genes that code for markers of differentiation and that are localized at the cell surface of erythroblasts: *CD34*, *CD71* (transferrin receptor, *TRFC*), and *CD235a* (glycophorin A, *GYPA*). Recent transcriptomic analyses in *ex vivo*-differentiated human erythroblasts have confirmed that the expression levels of these genes are strongly correlated with the localization of the encoded proteins at the cellular membrane [344,345]. As expected, *CD34* is highly expressed at the beginning of differentiation and decreases thereafter, whereas the expression of *CD71* and *CD235a* increases during differentiation (**Figure S4**). At any given time point, the expression of these three genes was not different between fetal liver- and bone marrow-derived CD34+ progenitors cells, except for *CD235a* at the beginning of differentiation (more expressed in fetal liver cells)(**Figure S4**). These results confirm that fetal and adult erythroblasts grow and differentiate at a similar rate, and therefore that our experiment is adequate to detect differences in DNA methylation that are mostly due to the different developmental stage (fetal liver vs. bone marrow) of these cells.

3.5.2 The DNA methylation landscape in human erythroblasts

We used the Illumina HumanMethylation 450k BeadChip to measure quantitatively DNA methylation across the genome in *ex vivo*-differentiated erythroblasts from 12 fetal liver and 12 bone marrow donors. After quality-control and intensity normalization steps, we obtained DNA methylation β-values (from 0.0 (unmethylated) to 1.0 (fully methylated)) for 402,819 CpGs (**Figure S5**). We only used these CpGs in subsequent analyses. We used DNA methylation data generated by the Roadmap Epigenomics Project (http://www.roadmapepigenomics.org/) using reduced representation bisulfite DNA sequencing (RRBS) from mobilized adult CD34+ cells to validate the DNA methylation levels measured in

erythroblasts with the Illumina HumanMethylation 450k BeadChip. Although not differentiated *ex vivo*, we reasoned that these CD34+ cells would be a good proxy for adult erythroblasts used in our experiment. Across 80,136 CpGs available in both datasets, we observed a strong correlation of DNA methylation values (Pearson's r=0.91, P<2.2x10^{-16}, **Figure S6A**). Importantly, we repeated the same analysis with other RRBS datasets from the Roadmap Epigenomics Project. We observed that DNA methylation values in erythroblasts are more correlated with DNA methylation data from blood-related rather than blood-unrelated cells or tissues (**Figure S6B**). These comparisons confirm the quality and specificity of the DNA methylation data generated with the Illumina 450k array in human erythroblasts.

We analyzed DNA methylation across all CpGs tested using unsupervised clustering methods. The two main clusters accurately distinguish fetal liver- from bone marrow-derived erythroblasts (**Figure 1**). This result confirms our hypothesis that changes in DNA methylation captures developmental differences between fetal and adult erythroblasts. We also noted that the two main stage-specific clusters are sub-divided according to the sex of the sample's donors (**Figure 1**). This sex-specific clustering is mostly dependent on DNA methylation values at CpGs located on the X-chromosome.

3.5.3 Differential DNA methylation between fetal and adult erythroblasts

We sought to identify differentially methylated single CpG sites by comparing methylation β-values for fetal and adult erythroblasts, taking into account potential sex effect. We used two criteria to define differentially methylated CpGs: a difference in β-values of ≥0.2 and a P≤1.25x10^{-7} (significant threshold after Bonferonni correction). Using this definition, we found 5,937 differentially methylated CpGs (**Figure 2A**).

We discuss below two interesting examples of differentially methylated loci detected in human erythroblasts. *GCNT2* encodes the enzymes responsible for the conversion of the blood i- to the I-antigen during the fetal-to-adult transition in erythroblasts. *GCNT2* codes for three isoforms: isoform C is responsible for the I-antigen phenotype in adult erythrocytes [346]. In our cell culture system, we found that cg14322298 in the promoter of isoform C is hypomethylated

in adult erythroblasts ($\Delta\beta_{adult\text{-}fetal}$=-0.47, P=8.0x10^{-16}; **Figure 3A**). We found that *GCNT2-C* expression levels correlate with DNA methylation at cg14322298 in erythroblasts (**Figure 3B**). Isoform B had very low expression in both fetal and adult erythroblasts, whereas isoform A and C explained most of *GCNT2* expression in fetal (72.2%) and adult (75.1%) erythroblasts, respectively (**Figure S7**). Another example of differentially methylated loci include CpGs in the *ARID3A*/*Bright* gene that show a marked differential DNA methylation pattern, being almost completely methylated in fetal-stage cells (*e.g.* cg08894487, $\Delta\beta_{adult\text{-}fetal}$=-0.93, P=3.8x10^{-21})(**Figure 3C**). This gene is required for hematopoietic stem cell development: definitive erythrocyte formation in the fetal liver is impaired in *Bright* knockout mice, but not at the primitive (embryonic) stage [347]. *ARID3A* expression was higher in fetal erythroblasts (**Figure 3D**). Positive correlations between DNA methylation inside gene bodies and gene expression levels have been reported previously [348]. Overall, we observed a global enrichment of positive correlations between DNA methylation at CpGs inside gene bodies and expression levels in erythroblasts. In contrast, there is a marked enrichment of negative correlations between CpGs within promoters or enhancers, and gene expression (**Figure S8**).

3.5.4 Erythroid enhancers are enriched for differentially methylated CpGs

To link changes in DNA methylation to biologically relevant functional units, we combined results from single CpGs that fall within each of 25,891 transcripts defined using RefSeq, 25,103 promoters (1.5-kilobase upstream of RefSeq transcripts) and 4,604 erythroid enhancers [334]. These enhancers, defined using DNase I hypersensitive sites and histone marks, include a set of enhancers that is common to both fetal and adult erythroblasts, as well as developmental lineage-specific enhancers [334]. After Bonferonni correction (significance threshold set at P<9x10^{-7}), we found 77 genes, 116 promoters and 303 erythroid enhancers that are differentially methylated (**Figure 2B-D**).

These analyses highlighted multiple regions of interest. For instance, *C22orf26* was one of the most differentially methylated genes (mean $\Delta\beta_{adult\text{-}fetal}$: -0.60; P=6.2x10^{-18}). Although nothing is known about this gene, it is located just upstream of the microRNA *let7-b* host gene (*MIRLET7BHG*; **Figure 3E**). The microRNA *let-7b* belongs to the *let-7* family, is highly

expressed in adult as compared to fetal erythroblasts, and correlates with *BCL11A* expression and reduced HbF levels [349]. It is possible that DNA methylation at CpGs within or near *C22orf26* regulates the expression of *MIRLET7BHG*. Notably, a nearby enhancer active in adult-stage erythroblasts contain differentially methylated CpGs, which correlates with *MIRLET7BHG* expression (**Figure 3F**). Although the CpGs also correlates with *C22orf26* expression levels (**Figure 3G**), this gene is an order of magnitude less expressed, and other CpGs in its core promoter might also control its expression. Finally, we carried out a pathway-based analysis to identify group of functionally related genes that are differentially methylated. We found multiple pathways enriched for signaling activity, wound healing, oxygen, cytokine production, circulation and cation transport (**Table S2**). Many of these pathways include genes not associated with erythropoiesis previously, and we need to validate if differential DNA methylation translates into biologically important functions.

More generally, we asked whether there was an enrichment of differentially methylated CpGs in genes, promoters and erythroid enhancers when compared to the rest of the genome tested by the Illumina HumanMethylation 450k BeadChip. We observed a strong statistically significant enrichment of differentially methylated CpGs in erythroid enhancers (3.67-fold, P=9.5x10^{-164})(**Figure 2E**). The enrichment was marginally significant in genes (1.03-fold, P=0.04) and we noted a significant depletion of differentially methylated CpGs in promoters (0.86-fold, P=4.3x10^{-10})(**Figure 2E**). The results for promoters and genes are challenging to interpret given the ascertainment bias in the design of the methylation array. However, this technical confounding does not affect the result for the erythroid enhancers since the release of the Illumina HumanMethylation 450k BeadChip pre-dated the publication of the erythroid enhancers [334]. Therefore, differential DNA methylation at erythroid enhancers likely captures transcriptional and developmental differences between fetal and adult erythroblasts.

We then compared enrichment of differentially methylated CpGs in enhancers defined in nine different cell types using data from the ENCODE Project [216]. Although there was enrichment at enhancers from other cell types – perhaps suggesting constitutive regulatory functions – we observed the strongest enrichment for erythroid enhancers defined in differentiated CD34+ cells (significant against enhancers from all other cell types, P<4.5x10^{-17},

Figure 2F)[334]. The enrichment of differentially methylated CpGs within enhancers defined by ENCODE in erythroleukemic K562 cells was similar to the enrichment observed in other, non-hematopoietic cells (**Figure 2F**). This result may highlight a limitation in using K562 cells to study the fine regulatory mechanisms that control erythropoiesis.

Having demonstrated an enrichment of differentially methylated CpGs in erythroid enhancers, we were interested in testing if these genomic regions are more likely to contain genetic variants associated with HbF levels. We analyzed associations between HbF levels measured at baseline and genotypes at 6,994,357 markers in 1,140 adult patients with sickle cell disease [340]. As expected, we observed a strong deviation from the null distribution owing to variants in the *BCL11A*, *HBS1L-MYB* and *β-globin* loci (**Figure 4**). When we considered only the 63,876 SNPs that overlap with erythroid enhancers, we again observed deviation from the null expectation. This departure from the null was even true – although more modest – after excluding enhancers near the known HbF regulators (**Figure 4**). This result directly implicates erythroid enhancer in inter-individual HbF levels variation, and suggests that future genetic experiments to find new HbF regulators should focus on these genomic regions.

3.5.5 Several transcription factor binding motifs are preferentially located near differentially methylated CpGs

Data from the ENCODE Project support a passive role for DNA methylation in controlling gene expression: when a transcription factor is not or less expressed, the CpGs surrounding its consensus binding sites tend to be less accessible to DNase I digestion and to be more methylated [350]. Based on this observation, we reasoned that an analysis of transcription factor binding sites within loci that are differentially methylated between fetal and adult erythroblasts might yield new transcription factors important for erythroid development. We performed this analysis on the 5,000 most hypomethylated CpGs in fetal erythroblasts and compared them to the 5,000 most hypomethylated CpGs in adult erythroblasts. We obtained consistent results when using different thresholds. We included 200 base pairs on each side of the CpGs and considered 495 transcription factor binding motifs defined by the HOMER

software [337]. We identified enrichment for many transcription factors involved in different cellular processes, including many with known roles during erythropoiesis. In **Table 1**, we list the top transcription factors with binding motifs enriched near hypomethylated CpGs.

In fetal erythroblasts, we observed enrichment for the binding sites of SOX6 and GATA1, two key transcription regulators of hematopoiesis [73,351,352]. GATA binding motifs are also enriched at fetal-specific erythroid enhancers [334]. GATA1 and SOX6 are two transcription factors important in fetal and adult erythroblasts, although our results suggest that they may preferentially bind hypomethylated sites in fetal erythroblasts [353]. The motif recognized by NFY was enriched near CpGs hypomethylated in fetal erythroblasts: NFY binds the *HBG2* gene promoters to stimulate chromatin opening [354]. In adult erythroblasts, we detected an enrichment for the motif bound by NF-E2, a transcription factor important for erythroid maturation and *HBB* gene expression [355], as well as RUNX1, an important regulator of mammalian hematopoiesis that acts upstream of NF-E2 [356]. IRF2 was recently established as a transcriptional regulator of erythropoiesis that controls gene expression through adult erythroid enhancers [334]. Binding sites for NRF2, a transcription factor closely related to NF-E2, are also enriched near hypomethylated CpGs in adult erythroblasts. NRF2 is a transcriptional activator of the antioxidant response and its pharmacological induction in K562 cells results in increased HbF production [357]. The most enriched motif in adult erythroblasts belongs to NF1, a family of transcription factors composed of NFIA, NFIB, NFIC and NFIX. Several differentially methylated CpGs are located near *NFIA*, *NIFC* and *NFIX,* suggesting that the expression of these transcription factors, as well as their target genes, may be developmentally regulated by DNA methylation during erythropoiesis. Over-expression of *NFIA* in CD34+ cells leads to increased *HBB* gene expression [358].

3.5.6 DNA methylation and genetic variation control *HBG2* expression

Treatment with DNA demethylating agents like 5-azacytidine induces HbF production in primates and humans [359,360]. The mechanism implies demethylation of the *HBG2* promoter, directly or indirectly through an effect on cellular stress, that results in an increased synthesis of γ-globin chains [188,189]. However, it is unknown whether changes in DNA methylation at loci

unlinked to the *β-globin* cluster on chromosome 11 can also influence the production of HbF in humans. To explore this hypothesis, we tested if common DNA sequence variants associated with HbF production in humans are also associated with changes in DNA methylation levels, that is if they are methylation quantitative trait loci (meQTL). Indeed, recent findings suggest that SNPs associated with complex diseases or traits may exert their phenotypic effect by altering DNA methylation profiles [361,362].

We only analyzed associations between HbF-associated SNPs and DNA methylation levels at the *BCL11A* (chr2:60,678,302-60,781,633), *HBS1L-MYB* (chr6:135,281,517-135,540,311) and *β-globin* (chr11:5,246,696- 5,527,882) loci in adult erythroblasts. We did not include fetal erythroblasts in this analysis because we reasoned that these SNPs affect HbF production in adult erythroid cells. We complemented the 450k DNA methylation data with measures of DNA methylation at CpGs in the *HBG2* promoter obtained using the targeted assay described above (**Figure S3**). We also measured by quantitative PCR the expression of *HBG2* and *HBB* in the same adult cells. In our small sample size (N=12), we did not identify HbF-associated SNPs that are significantly associated with DNA methylation after accounting for the number of tests performed. Focusing on the *β-globin* locus, we noted that increased DNA methylation in the *HBG2* promoter was associated with a decreased *HBG2*/*HBB* gene expression ratio, as expected (**Table 2**). When we included in the prediction model both DNA methylation levels at the *HBG2* promoter and genotypes at rs3759074, both terms were nominally associated in the expected direction with the *HBG2/HBB* expression ratio (**Table 2**). rs3759074 is in linkage disequilibrium with the rs7482144-*Xmn*I variant (r^2=0.74) and falls within the *BCL11A* binding site identified in a functional element important for HbF silencing [363]. In a mouse model, inactivation of *BCL11A* and treatment with demethylating 5-aza-2'-deoxycytidine has a synergistic effect on HbF production [143]. Together with these observations, our results suggest that changes in DNA methylation at the *HBG2* promoter act partly independently from genotypes at the *β-globin* locus to control HbF production.

Table 1. Enrichment of transcription factor binding sites (TFBS) near differentially methylated CpGs.

Motif Name	Consensus	*P*-value	*q*-value	Fold enrichment
TFBS enriched near CpGs hypomethylated in fetal erythroblasts				
SOX2	BCCATTGTTC	$1.0x10^{-13}$	0	1.5
TCF3	ASWTCAAAGG	$1.0x10^{-8}$	0	1.8
REST-NRSF	GGMGCTGTCCATGGTGCTGA	$1.0x10^{-8}$	0	5.5
SOX6	CCATTGTTNY	$1.0x10^{-8}$	0	1.3
GATA1	AGATGKDGAGATAAG	$1.0x10^{-5}$	0.0001	2.4
MAZ	GGGGGGGG	$1.0x10^{-5}$	0.0002	1.1
TCF4	ASATCAAAGGVA	$1.0x10^{-5}$	0.0002	1.4
FOXA1	WAAGTAAACA	$1.0x10^{-5}$	0.0002	1.3
HNF6	NTATYGATCH	$1.0x10^{-4}$	0.0004	1.4
TCFL2	ACWTCAAAGG	$1.0x10^{-4}$	0.0005	1.9
TFBS enriched near CpGs hypomethylated in adult erythroblasts				
NF1	CYTGGCABNSTGCCAR	$1.0x10^{-211}$	0	3.9
IRF2	GAAASYGAAASY	$1.0x10^{-78}$	0	7.3
TLX/NR2E1	CTGGCAGSCTGCCA	$1.0x10^{-76}$	0	2.4
NF1-halfsite	YTGCCAAG	$1.0x10^{-56}$	0	1.4
ISRE	AGTTTCASTTTC	$1.0x10^{-54}$	0	7.5
BACH1	AWWNTGCTGAGTCAT	$1.0x10^{-42}$	0	6.4
RUNX1	AAACCACARM	$1.0x10^{-37}$	0	1.6
RUNX2	NWAACCACADNN	$1.0x10^{-36}$	0	1.7
NRF2	HTGCTGAGTCAT	$1.0x10^{-36}$	0	5.8
c-JUN	GATGASTCATCN	$1.0x10^{-33}$	0	2.3

For these analyses, we used the HOMER software and analyzed TFBS located near CpGs (±200 base pairs) that are hypomethylated in fetal or adult erythroblasts. The top 10 enriched motifs are shown here for each cell type; see Additional file 1: Tables S6-S7 for the complete list of significant TFBS. We calculated q-values using the Benjamini-Hochberg method. We calculated the fold enrichment by comparing the number of hypomethylated CpGs near a given TFBS in fetal and adult erythroblasts. The consensus motif follows the IUPAC nomenclature when more than one base is possible.

Table 2. Genetic and epigenetic control of HBG2 / HBB expression in adult erythroblasts

Model	rs3759074-A		CpG (chr11:5,276,172)	
	BETA (SE)	***P***	**BETA (SE)**	***P***
Model 1: log(*HBG2*/*HBB*) ~ rs3759074	0.98 (0.42)	0.04	-	-
Model 2: log(*HBG2*/*HBB*) ~ CpG	-	-	-6.1 (2.0)	0.01
Model 3: log(*HBG2*/*HBB*) ~ rs3759074 + CpG	5.4 (1.6)	0.008	-0.81 (0.30)	0.02

The A-allele at rs3759074 (chr11:5,257,778) is associated with increased fetal hemoglobin production by genome-wide association studies. In adult erythroblasts, rs3759074-A is associated with increased expression of HBG2/HBB (Model 1). DNA methylation at a CpG located in the HBG2 promoter (chr11:5,276,172) is inversely correlated with HBG2/HBB expression (Model 2). Both genotypes at rs3759074 and DNA methylation at CpG (chr11:5,276,172) are independent predictors of HBG2/HBB expression levels (Model 3). For this analysis, we used adult erythroblasts from 12 donors. Within this dataset, the rs3759074-A allele frequency is 27% and the mean DNA methylation value at the tested CpG is 61 ± 8%. BETA: arbitrary gene expression units; SE: Standard error.

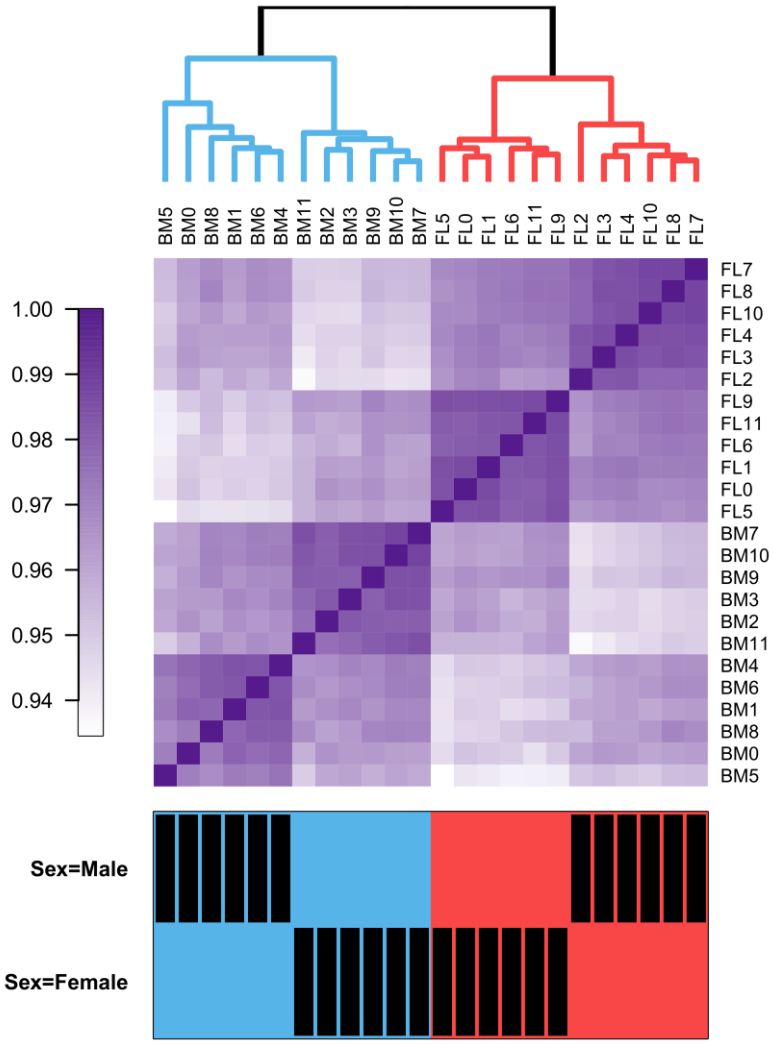


Figure 1. Hierarchical clustering analysis of DNA methylation in erythroblasts.
Unsupervised clustering analysis of DNA methylation at 402,819 CpGs separates precisely fetal liver-derived from bone marrow-derived erythroblasts. Within each developmental stage, cells are sub-divided by the sex of the donors. The top panel represents the dendogram of the different clusters identified. The middle panel summarizes pairwise correlations between all samples. In the bottom panel, black rectangles identify male and female donors. BM: adult erythroblasts; FL: fetal erythroblasts.

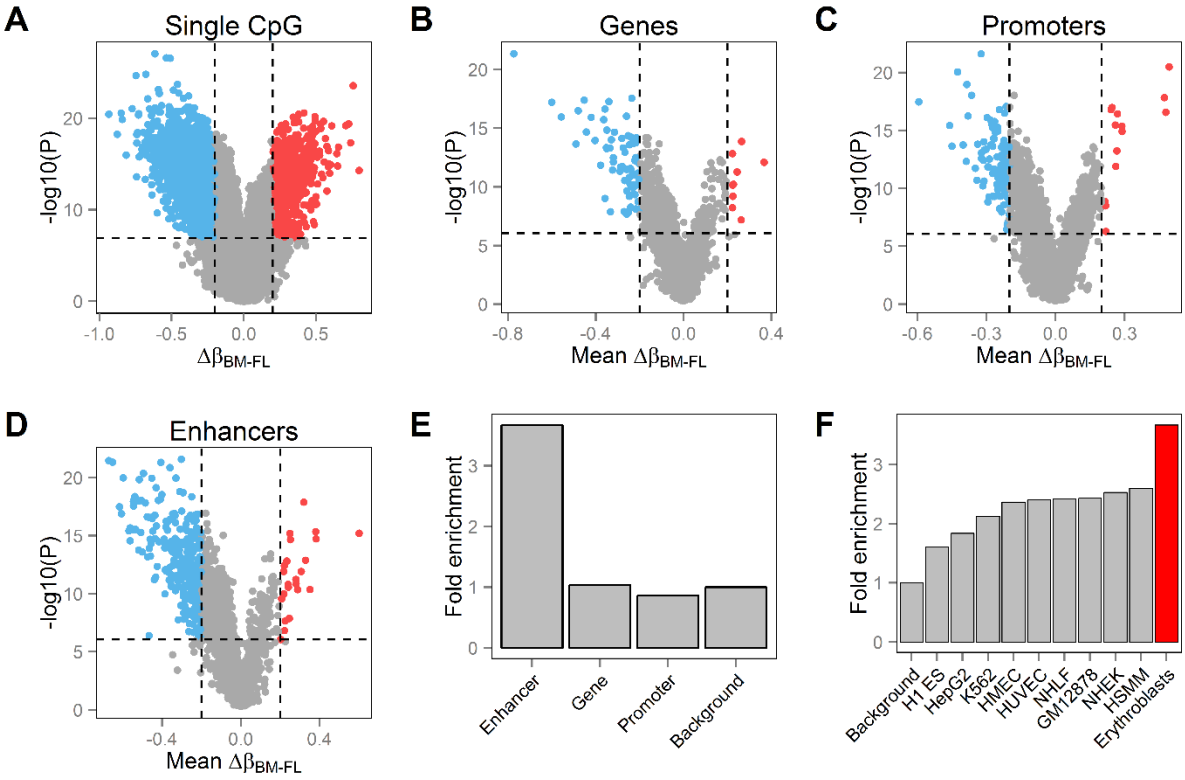


Figure 2. Differential DNA methylation between fetal liver- and bone marrow-derived erythroblasts.

Volcano plots of differentially methylated **(A)** CpGs, **(B)** genes, **(C)** promoters, and **(D)** erythroid enhancers. Differentially methylated single CpGs have a difference in DNA methylation β-values between fetal and adult erythroid cells ≥20% and a $P \leq 1.25 \times 10^{-7}$. Red and blue points correspond to CpGs significantly hypomethylated in fetal- and adult-stage erythroblasts, respectively. For genes, promoters, and erythroid enhancers, we combined P values from CpG sites that fall within each unit using a generalization of Fisher's method to take into account correlation of values at nearby sites. For genes, promoters, and enhancers, we averaged the DNA methylation β-values of the CpGs that fall within the unit (y-axis). We used a Bonferonni-adjusted statistical threshold to define differentially methylated functional regions. **(E)** An enrichment of differentially methylated CpGs was observed in erythroid enhancers when compared to gene or promoter regions also tested on the Illumina HumanMethylation 450 k BeadChip. We used probes that fall outside of these regions to calculate the fold enrichment. **(F)** Erythroid enhancers (red) are enriched in differentially methylated CpGs when compared to enhancers defined in other cell lines using data from the ENCODE Project.

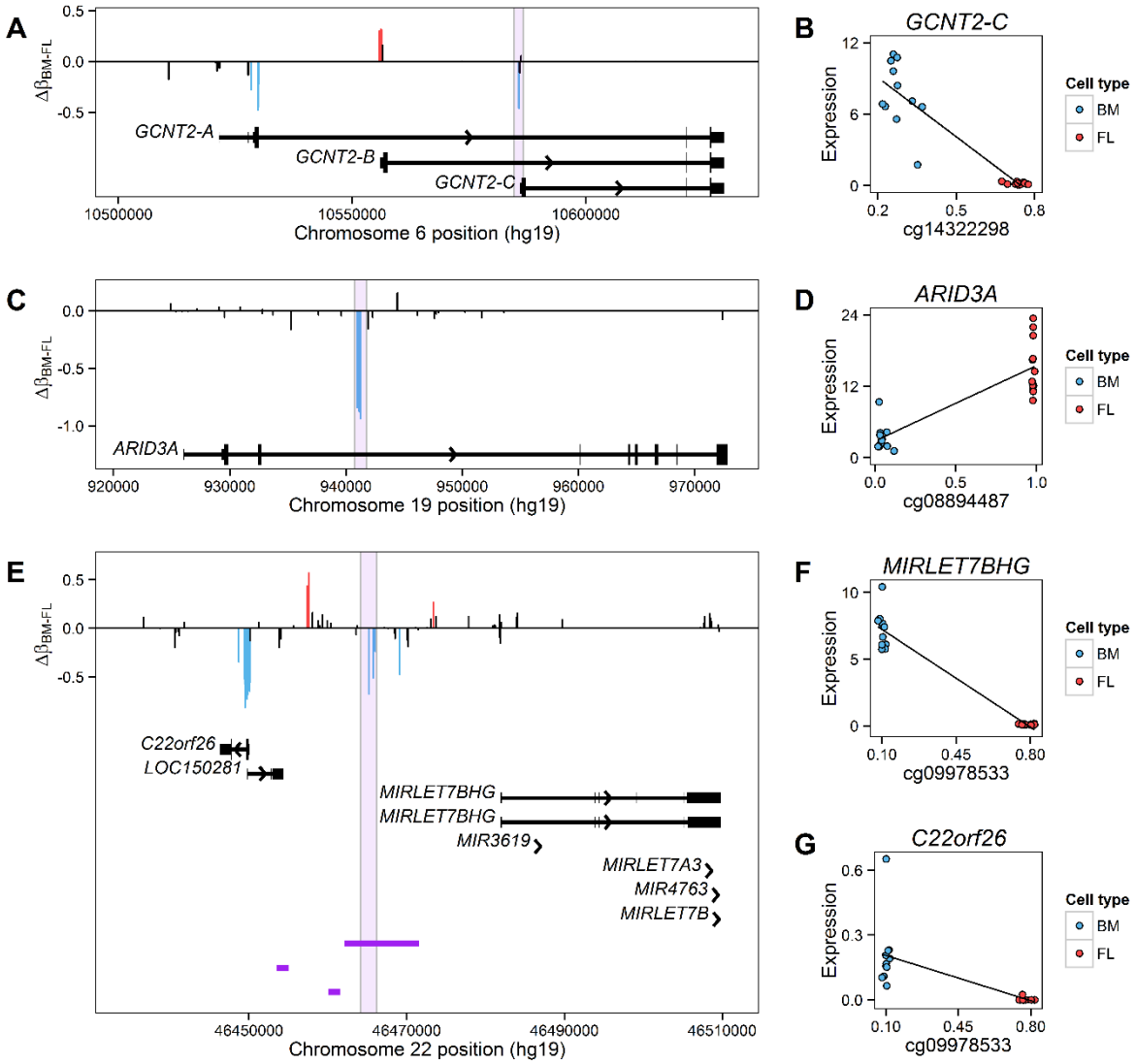


Figure 3. Examples of loci differentially methylated between fetal- and adult stage erythroblasts

Each vertical line represents a targeted CpG. Red and blue lines correspond to CpGs significantly hypomethylated in fetal- and adult-stage erythroblasts, respectively ($P<1.25\times10^{-7}$ and $\Delta\beta\geq0.2$). **(A)** A CpG in the promoter of *GCNT2* isoform C, which is responsible for the conversion of the fetal i-antigen to the adult I-antigen, is hypomethylated in adult-stage erythroblasts. **(B)** *GCNT2-C* expression levels is inversely correlated with DNA methylation at a CpG located in its core promoter (r=-0.97, $P=4.6\times10^{-14}$). **(C)** CpGs in an intron of *ARID3A*, a gene implicated in hematopoiesis, displays hypomethylation in adult-stage erythroblasts. **(D)** DNA methylation inside *ARID3A* (highlighted) positively correlates with its expression levels (r=0.87, $P=9.6\times10^{-8}$). **(E)** Differential DNA methylation near *C22orf26* and the miRNA *let7-b* host gene (*MIRLET7BHG*). A cluster of hypomethylated CpGs in adult erythroblasts overlaps an active erythroid enhancer (purple horizontal bars). DNA methylation inside the enhancer (highlighted) negatively correlates with **(F)** *MIRLET7BHG* ($r=-0.99$, $P=5.6\times10^{-20}$) and **(G)** *C22orf26* expression levels (r=-0.74, $P=3.2\times10^{-5}$).

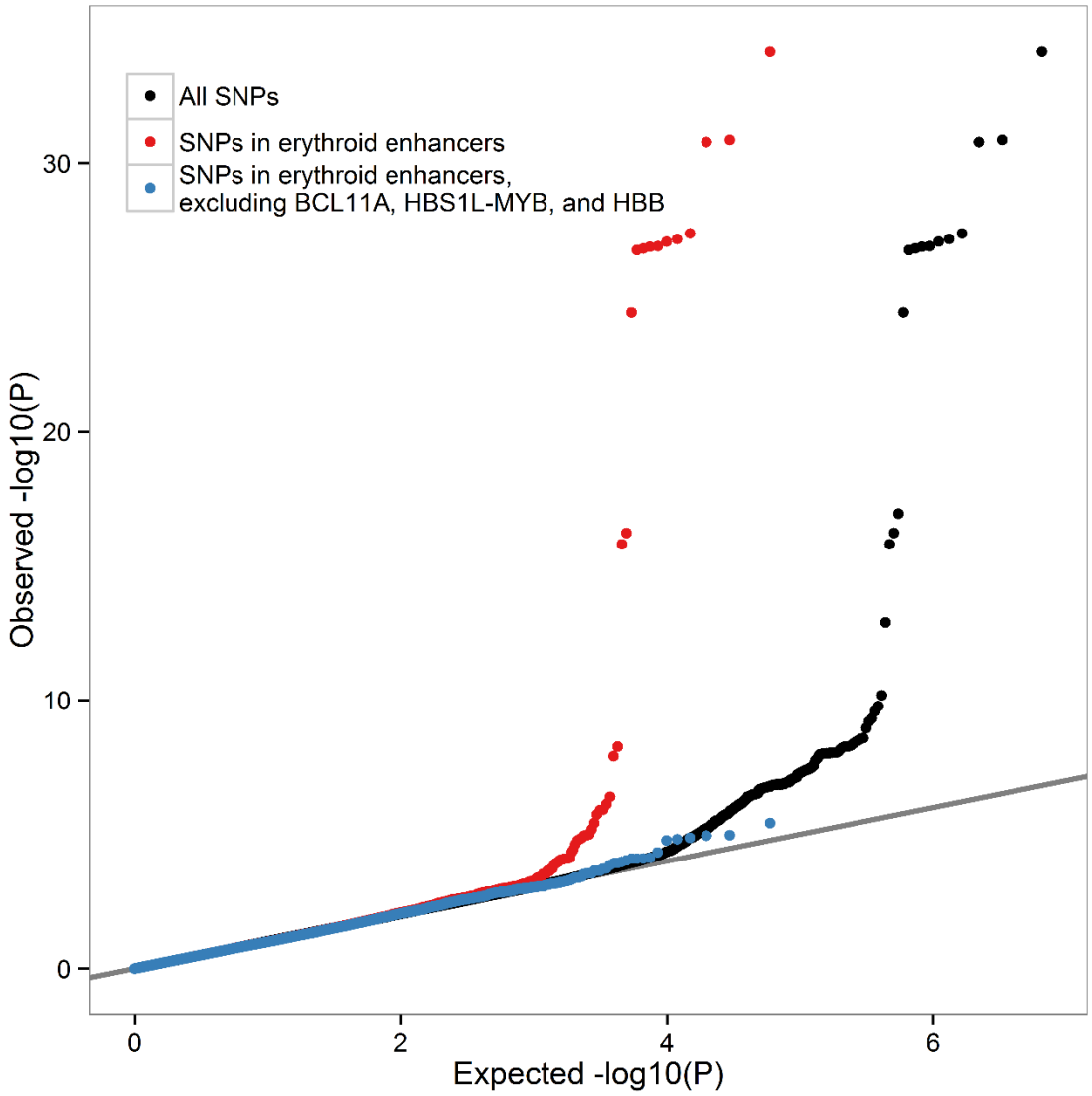


Figure 4. Erythroid enhancers are enriched for SNPs associated with fetal hemoglobin (HbF) levels in patients with sickle cell disease.

Quantile-quantile (QQ) plot of association P-values with HbF levels in 1,140 sickle cell disease patients. The QQ-plot for all 6,994,357 imputed SNPs is shown in black ($\lambda_{GC} = 1.01$). In red are the 63,876 markers that map to erythroid enhancers ($\lambda_{GC} = 1.03$) and in blue are the markers that remain after excluding enhancers near *BCL11A*, *HBS1L-MYB* and the β-globin locus ($\lambda_{GC} = 1.02$).

3.6 Conclusions

We generated comprehensive maps of DNA methylation in human erythroblasts differentiated *ex vivo* from fetal liver or bone marrow CD34+ progenitor cells. At single base pair resolution, we identified 5,937 differentially methylated CpGs that capture many of the transcriptional differences – in terms of transcriptional enhancers and transcription factor binding – between fetal- and adult-stage cells. These analyses also revealed multiple regions that could be of importance for HbF regulation, as indicated by the enrichment of SNPs strongly associated with HbF levels within erythroid enhancers. Most of the differentially methylated CpGs are hypermethylated in fetal erythroblasts. We can explore further these DNA methylation differences to understand what distinguishes fetal from adult erythroblasts during human erythropoiesis.

Our study has some limitations. First, the technology that we used to measure DNA methylation does not distinguish between methylation (5-methylcytosine, 5-mC) and hydroxymethylation (5-hydroxymethylcytosine, 5-hmC). This may be important to functionally explore as 5-mC and 5-hmC have different reported roles in the context of gene regulation [364]. Second, some of the differences observed in DNA methylation levels between fetal and adult erythroblast may be due to slight differences in their growth kinetics. However, we expect this number to be low since our cell morphology and gene expression analyses indicate that these cells are largely undistinguishable. Finally, our analysis of the effect of genetic variation and DNA methylation on globin gene expression is limited by our small sample size. Although our results are consistent with the literature, validation in independent samples is needed to confirm our additive model.

Clinically, one of the most important features of these fetal and adult erythroid cells is their respective production of HbF or HbA. In patients with sickle cell disease or β-thalassemia, increasing HbF production improves disease outcomes. In clinical trials, DNA demethylating agents have shown modest efficacy in increasing HbF production in patients [359,360]. On the other hand, work in a sickle cell mouse model has shown that *BCL11A*-mediated repression and 5-aza-2'-deoxycytidine treatment synergistically control HbF production and improve

hematological parameters [143]. This is consistent with our observation that a SNP within a *BCL11A* binding site in a key regulatory element at the *β-globin* locus and DNA methylation are independent predictors of *HBG2* expression in adult erythroblasts. Together, the mouse and human erythroblast results suggest that a combined strategy to inactivate *BCL11A* and promote *HBG2* demethylation may provide the robust induction of HbF production necessary to treat β-hemoglobinopathy patients.

3.7 Acknowlegdments

We thank Ekat Kritikou and Daniel Bauer for critical reading of this manuscript. We also acknowledge investigators from the Roadmap Epigenomics Project (http://nihroadmap.nih.gov/epigenomics/) for providing access to the RRBS datasets, and thank Dr. Martin H. Steinberg for kindly providing access to the CSSCD GWAS data set (funded by the NHLBI STAMPEED). SL holds fellowships from the "Fonds de Recherche du Québec – Santé" and the "Fondation Pierre Lavoie". This work was supported by grants from the Canadian Institute of Health Research (#123382) and the Canada Research Chair program to GL.

3.8 Supplementary information

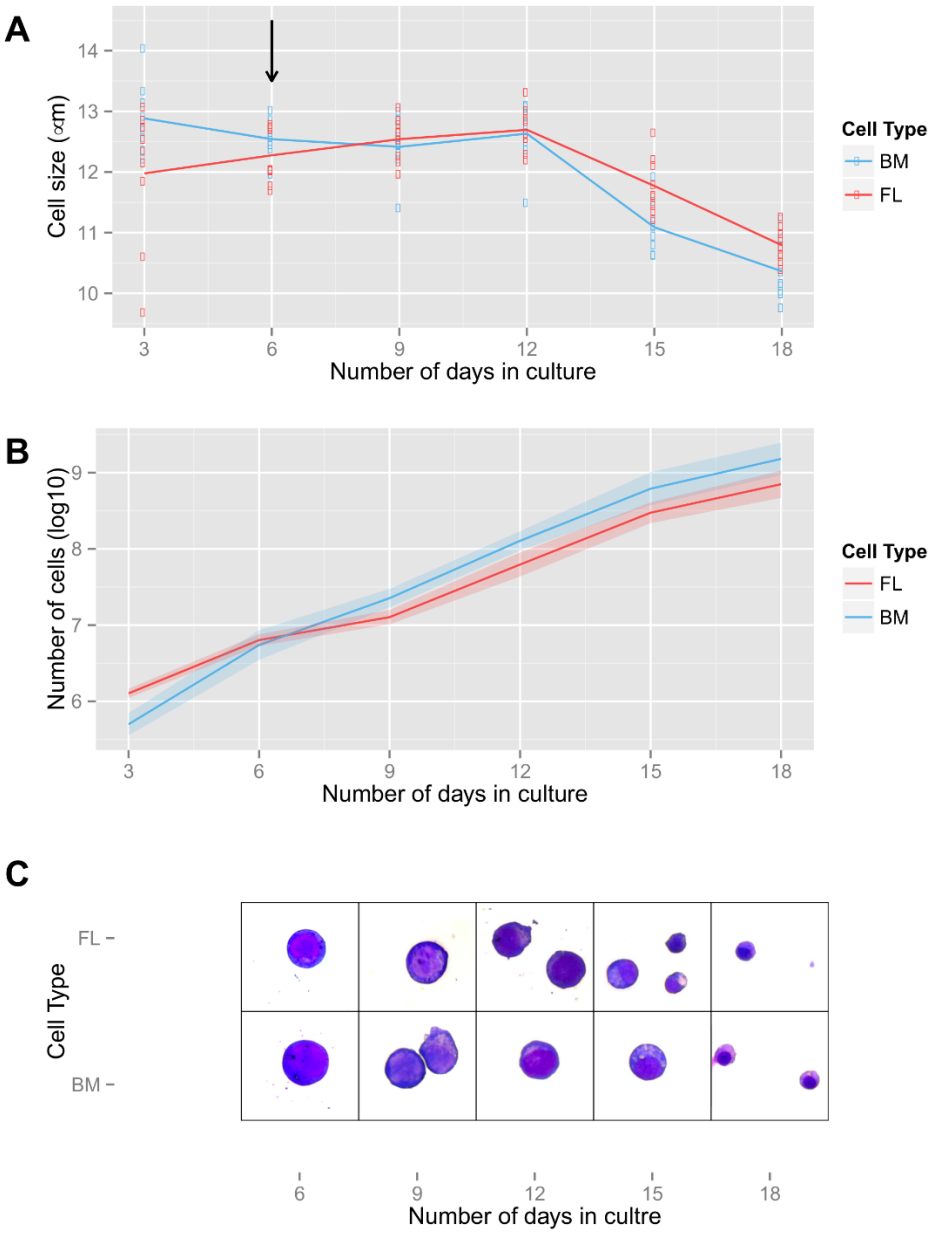


Figure S1. ***Ex vivo*** **differentiation of human CD34+ cells.**
(**A**) Primary human fetal and adult CD34+ cells are cultured in the expansion medium until day 6 (black arrow). They are then transferred to the differentiation medium until cells are collected at day 18. The medium is changed every 3 days. Consistent with differentiation into erythroblasts, cell size decreases during differentiation. (**B**) Differentiating human fetal and adult CD34+ cells have similar growth curves. (**C**) Wright-Giemsa coloration of CD34+ cells during differentiation shows the expected change in morphology (60X): at day 6, cells have myeloid stem cell morphology and at day 18, cells of both types display either orthochromatic or polychromatic erythroblast morphology. FL: CD34+ cells from fetal liver; BM: CD34+ cells from bone marrow (adult).

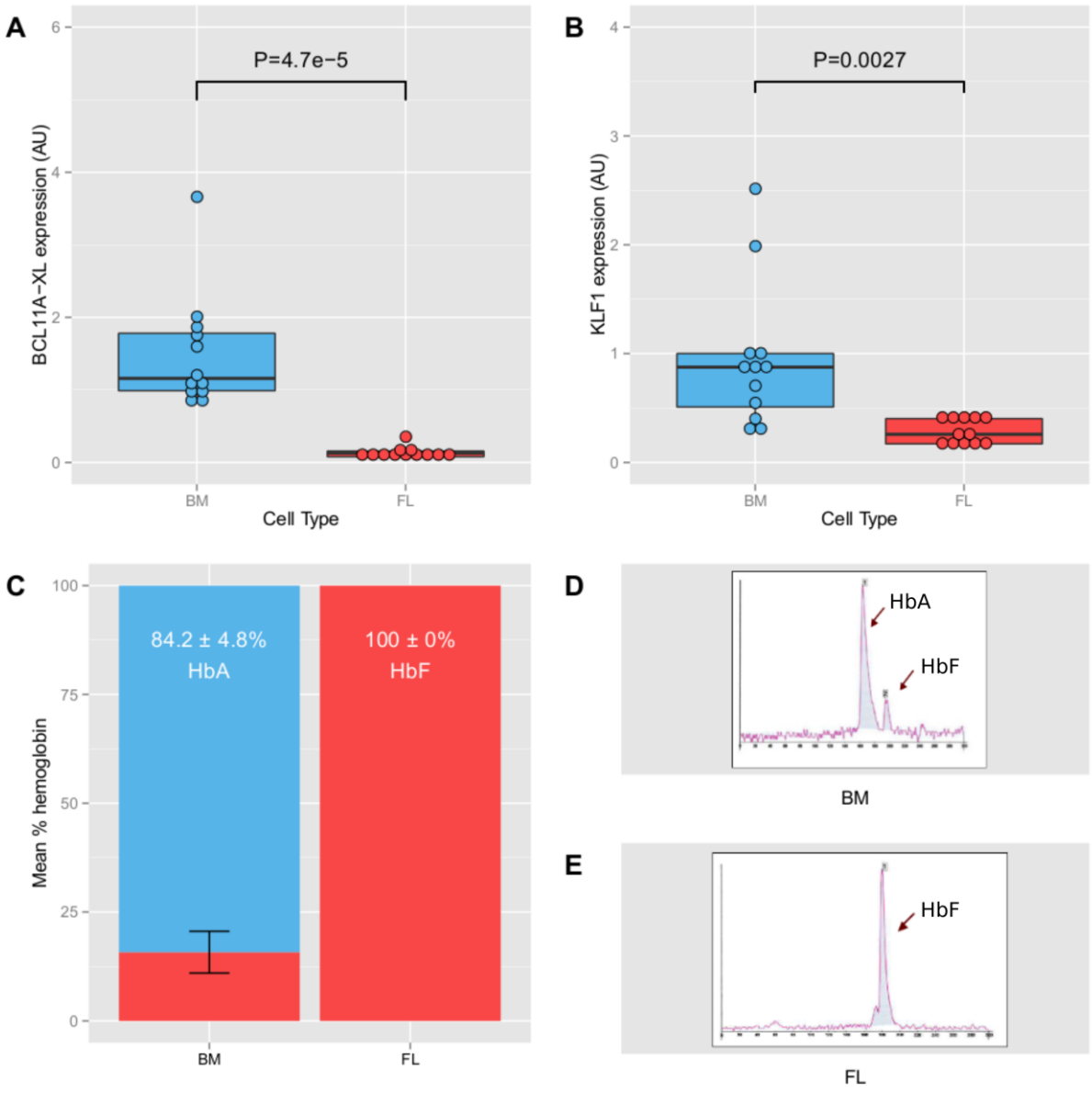


Figure S2. ***Ex vivo*** **differentiated CD34+ cells exhibit gene expression and hemoglobin production that are characteristic of fetal or adult erythroblasts.**

(**A**) Bone marrow-derived erythroblasts express significantly more of the stage-specific γ-globin transcriptomic repressor *BCL11A* as assessed by qPCR. (**B**) Fetal liver-derived erythroblasts cells express significantly less of the *KLF1* gene, which has been show to preferentially activate the *HBB* gene as well as promoting γ-globin repression. For A and B, we calculated *P*-values with Student's *t*-test. (**C**) Erythroblasts from bone marrow produce more adult hemoglobin (HbA), whereas fetal liver-derived erythroblasts produce exclusively fetal hemoglobin (HbF). Error bars are ± SEM. (**D-E**) Representative examples of capillary electrophoresis chromatograms used to quantify hemoglobin in erythroblasts differentiated from bone marrow (BM) or fetal liver (FL) CD34+ cells.

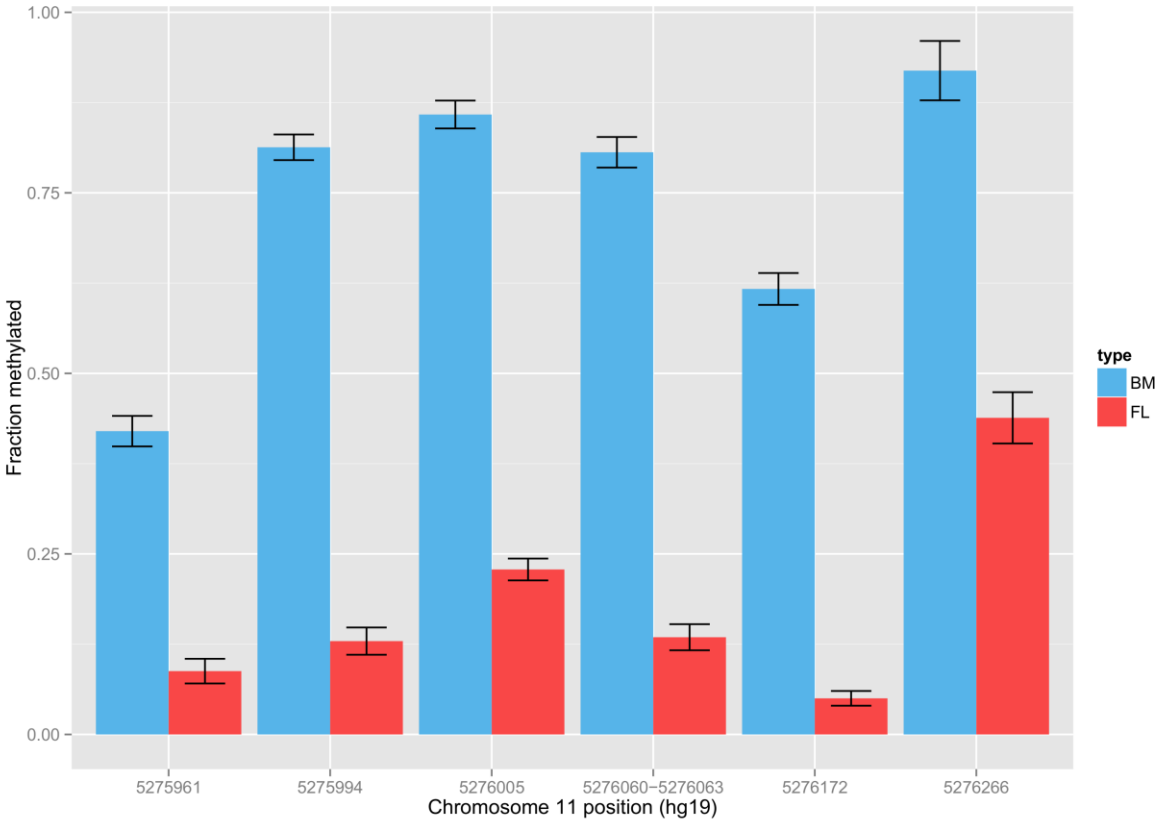


Figure S3. DNA methylation in the *HBG2* (*γ-globin*) promoter
CpGs in the *HBG2* promoter are hypomethylated and hypermethylated in fetal and adult erythroblasts, respectively. Blue: erythroblasts from bone marrow. Red: erythroblasts from fetal liver. Error bars are ± SEM.

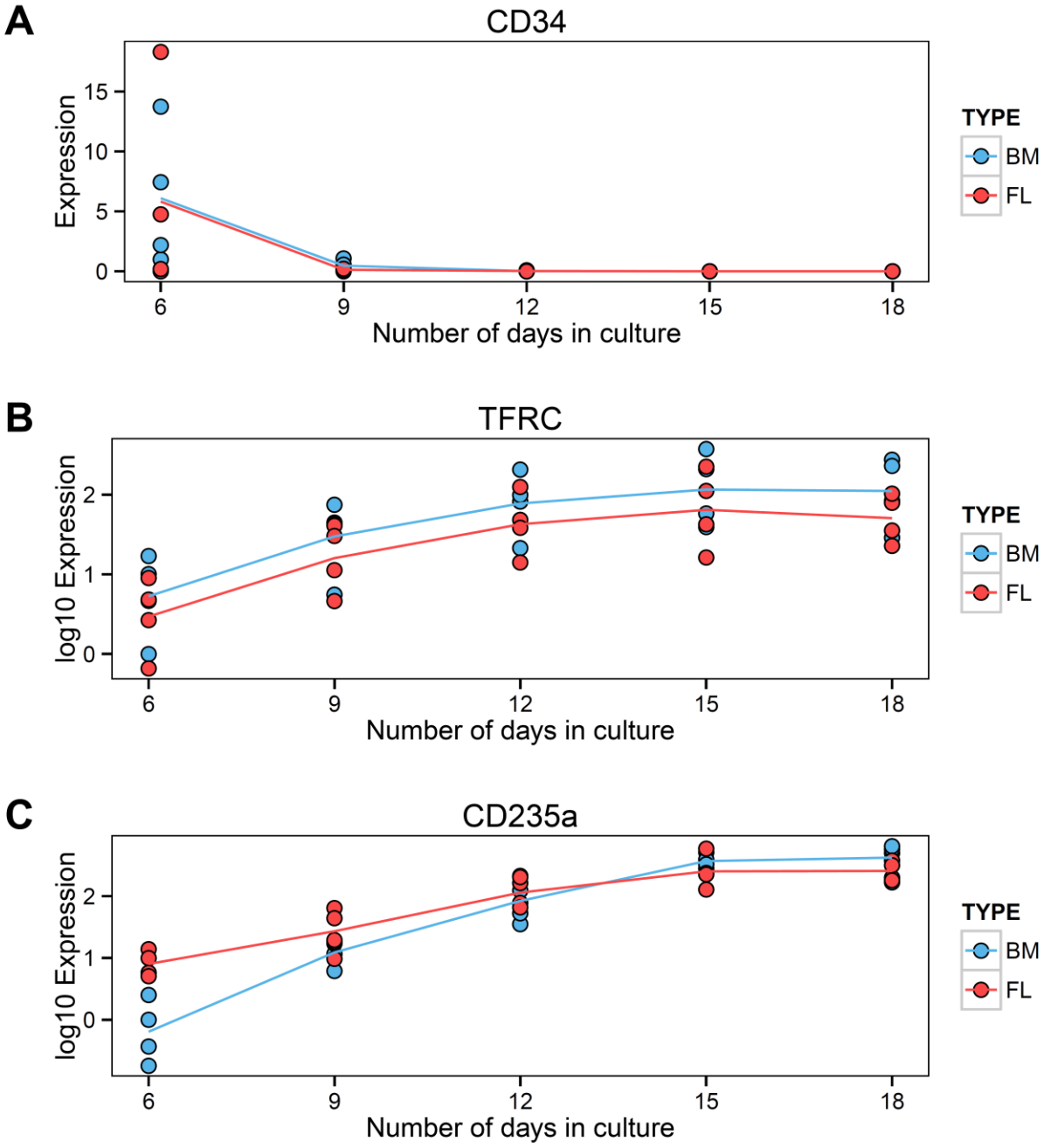


Figure S4. RNA expression of cell surface markers during *ex vivo* differentiation
Relative expression values were obtained by qPCR. (**A**) *CD34* expression decreases during differentiation. (**B**) Expression of *TFRC* (transferrin receptor) increases during differentiation. (**C**) *CD235a* (glycophorin A) expression increases during differentiation. No significant differences were observed between fetal and adult-stage cells for all markers (*t*-test, all $P > 0.15$), expect for *CD235a* at day 6 (*t*-test, $P=0.03$).

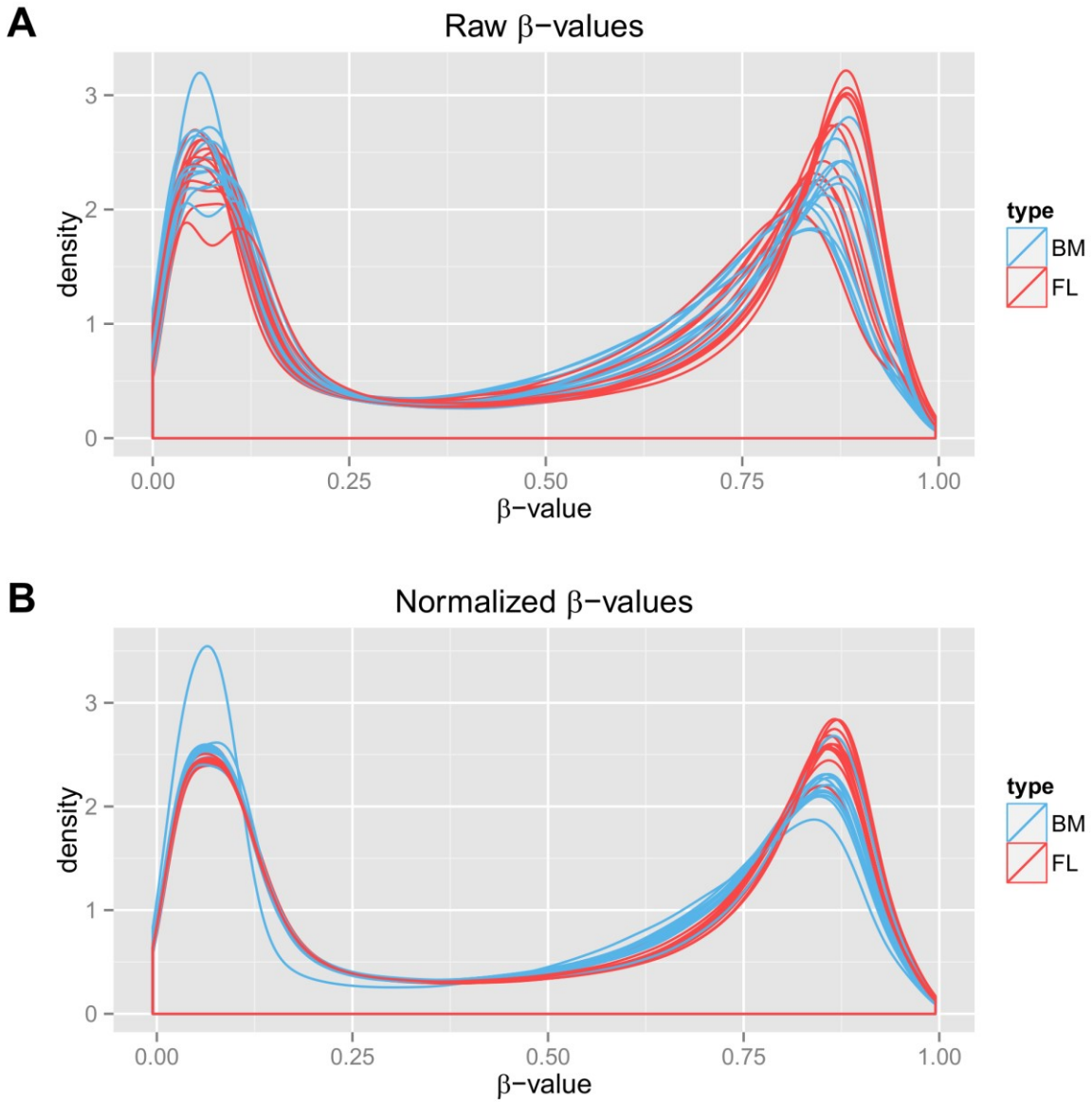


Figure S5. DNA methylation data normalization

DNA methylation β-value distributions across all samples before (**A**) and after (**B**) normalization with ARRm and the removal of probes that contain DNA polymorphisms. Red: fetal liver (FL)-derived erythroblasts; Blue: bone marrow (BM)-derived erythroblasts.

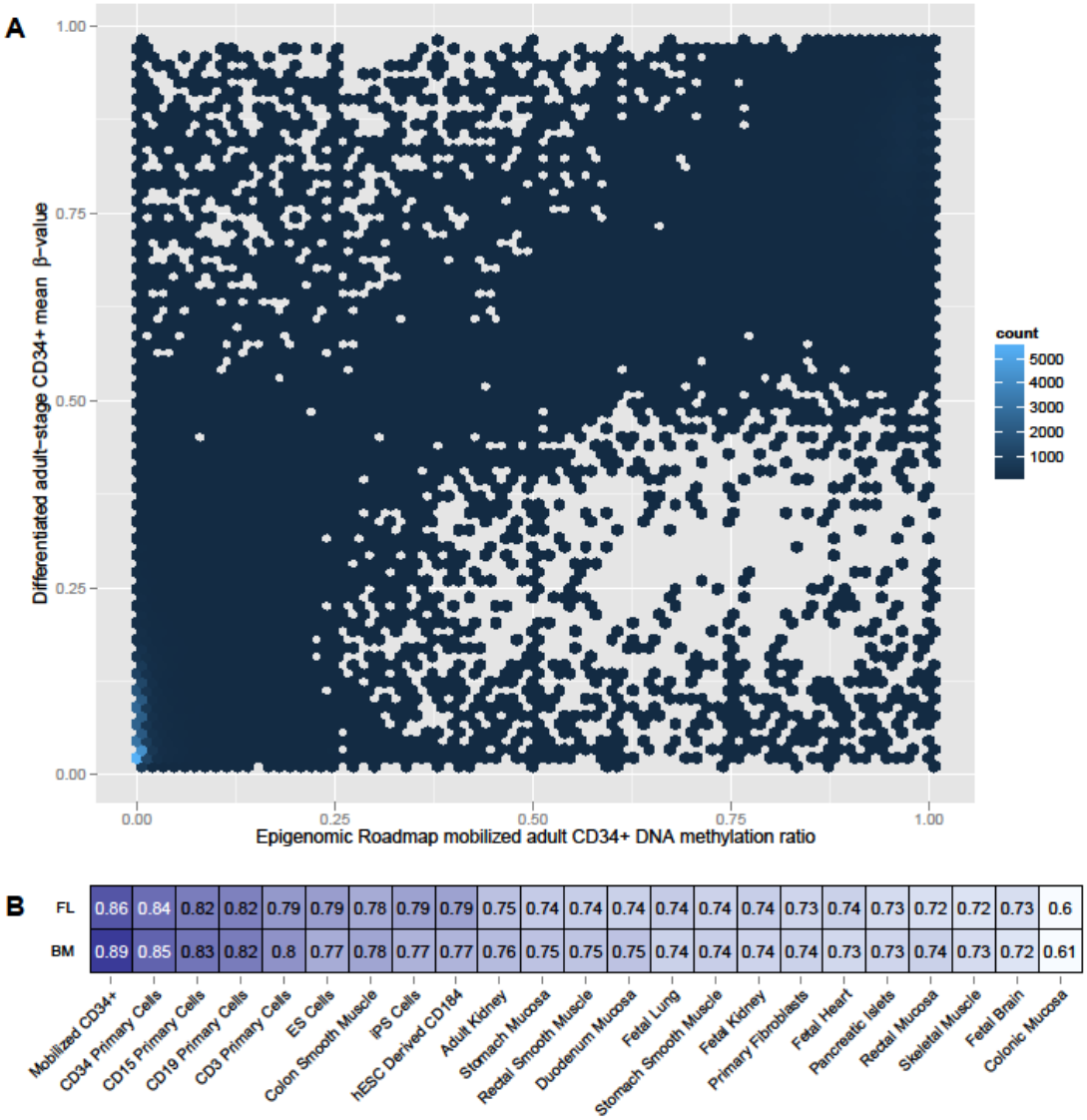


Figure S6. Correlations of DNA methylation between samples of our study and samples from the Roadmap Epigenomics Project.

(A) Correlation between DNA methylation values from our study (adult erythroblasts from 12 donors) and mobilized adult CD34+ cells from the Roadmap Epigenomics Project. For this analysis, we analyzed 80,136 CpGs with DNA methylation values present in both datasets. The correlation is strong (Pearson's r^2=0.83, P<2.2x10^{-16}), thus validating DNA methylation values generated with the Illumina HumanMethylation 450k BeadChip. **(B)** Correlation coefficients (Pearson's r^2) for DNA methylation levels in fetal- (FL) and adult-stage (BM) erythroblasts when compared to reduced representation bisulfite sequencing data from different cell types and tissues from the Roadmap Epigenomics project. We restricted this analysis to CpGs common across all cell types (*i.e.* a subset of the CpGs included in (A)), which explains the difference in r^2 between (A) and (B).

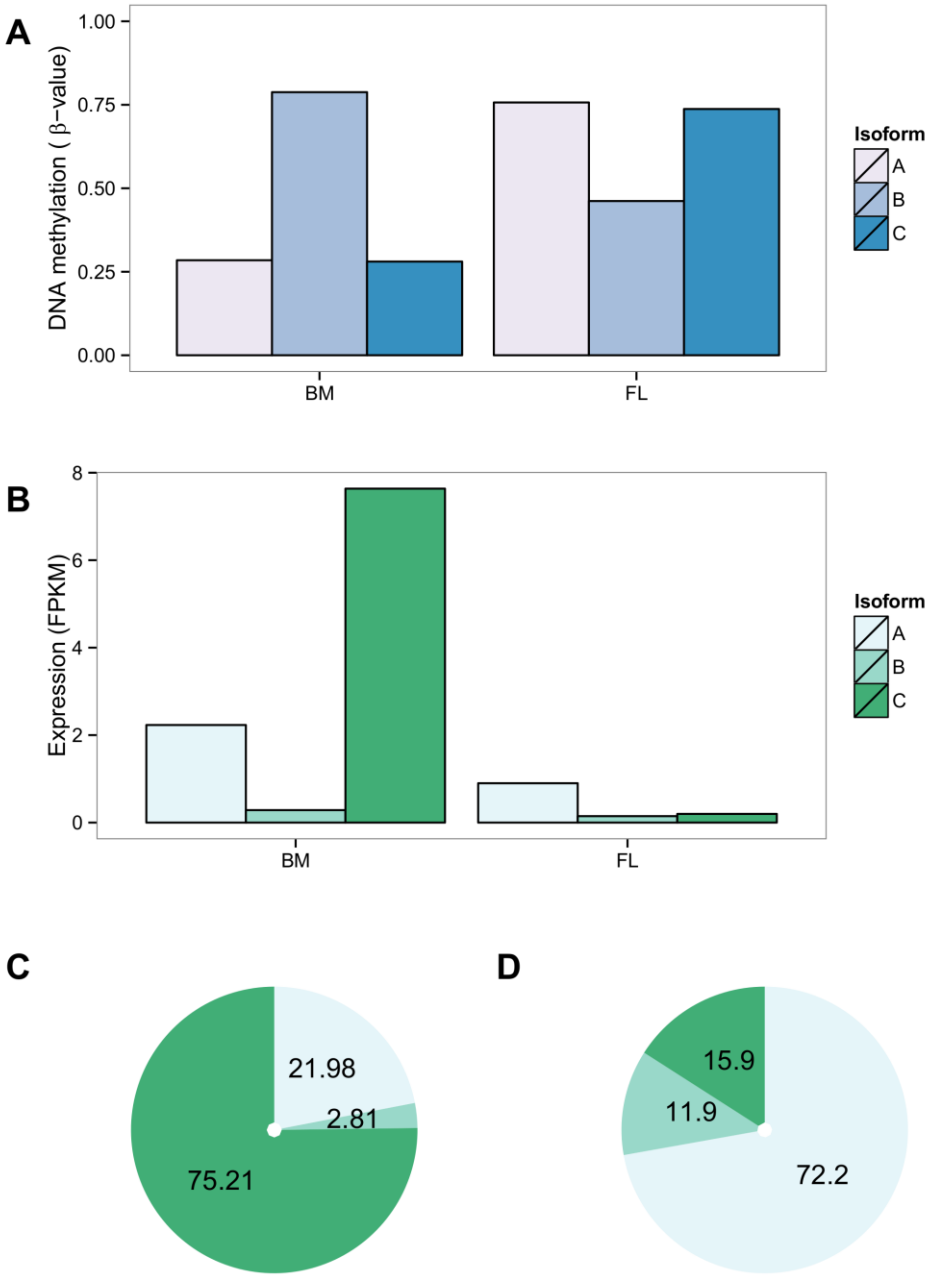


Figure S7. Relationship between DNA methylation and the expression of *GCNT2* isoforms

(**A**) DNA methylation levels at cg22187251 (isoform A exon), cg12695465 (isoform B exon), and cg13322298 (isoform C exon). (**B**) Expression of the different isoforms reported by Cufflinks. C. Relative expression of the GCNT2 isoforms.

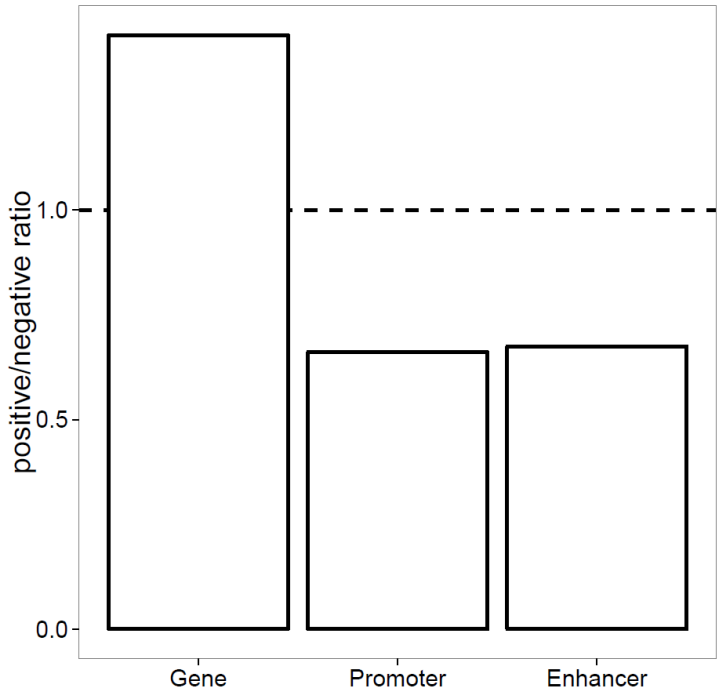


Figure S8. Enrichment of positive and negative correlations of DNA methylation and expression in genes, promoters, and enhancers.

We calculated the correlations between DNA methylation at CpGs in genes, promoters and enhancers with the expression of the corresponding gene. For enhancers, the closest gene was used (maximum distance of 1Mb). We report the ratio of positive and negative correlations (Pearson) with $P<1x10^{-4}$.

Table S1. List of primers used for quantitative PCR analyses.

Target	Sense primer	Antisense primer
BCL11A-XL	TCAAATAGCACTTGACTCTGCCTG	CAGCATTCTTGCAACTTTTCCC
KLF1	GGCAAGAGCTACACCAAGAG	TGTGTTTCCGGTAGTGGC
HBG2	TGGATGATCTCAAGGGCAC	TCAGTGGTATCTGGAGGACA
HBB	CTGAGGAGAAGTCTGCCGTTA	AGCATCAGGAGTGGACAGAT
CD235a	GGCTAAGGTCAGACACTGAC	TGTGCATTGCCACCTCAGTG
CD34	AATGAGGCCACAACAAACATCACA	CTGTCCTTCTTAACCTCCGCACAGC
TFRC	AAAATCCGGTGTAGGCACAG	CCTTTAAATGCAGGGACG
GADPH	GACAGTCAGCCGCATCTTC	GCAACAATATCCACTTTACCAGAG

All primer sequences are 5'-3'.

Table S2. Pathway clusters enriched in differentially methylated genes

Cluster	Number of pathways	Score	Best Pathway	Best P	Corrected P (Benjamini)
Regulation	31	4.38	Signal peptide	$2.7x10^{-9}$	$7.7x10^{-6}$
Healing	12	2.68	Response to wounding	$6.6x10^{-8}$	$9.1x10^{-5}$
Sodium transport	6	2.46	Sodium	$1.0x10^{-3}$	0.046
Oxygen	8	2.46	Dioxygenase	$7.34x10^{-4}$	0.044
Immunity and cytokine production	221	2.38	Immune response	$3.4x10^{-12}$	$9.31x10^{-9}$
Body fluid/ circulation	12	2.00	Regulation of body fluid levels	$1.56x10^{-4}$	0.038

Using the DAVID clustering tool[335,336], we grouped similar pathways with an enrichment $P<0.05$. We report clusters in which at least one pathway was significant (Benjamini-corrected $P<0.05$). Best pathway represent the most enriched significant pathway in the cluster

Chapter 4: RNA-sequencing of *ex vivo* differentiated erythroblasts reveal novel genes and miRNAs associated with fetal and adult erythroid developmental stages

Authors

Samuel Lessard, Mélissa Beaudoin, Daniel E. Bauer, Stuart H. Orkin, Guillaume Lettre

Reference

Lessard S, Beaudoin M, Bauer DE, Orkin SH, Lettre G. *RNA-sequencing of* ex vivo *differentiated erythroblasts reveal novel genes and miRNAs associated with fetal and adult erythroid developmental stages.* [in preparation].

Author's contribution

S.L., D.E.B, S.O., and G.L. conceived and designed the experiments. S.L., M.B. and D.E.B performed the experiments. S.L. and D.E.B. analyzed the data. S.L. and G.L. wrote the manuscript with contributions from all authors.

4.1 Context

Reactivation of fetal hemoglobin (HbF) is still the most promising therapy for the treatment of sickle cell disease. The fetal-to-adult hemoglobin switch is marked by several transcriptional changes. Notably, erythroid cells increase the expression of *BCL11A*, a transcriptional repressor of HbF. Similarly, adult red blood cell (RBC) progenitors highly express several members of the microRNA (miRNAs) let-7 family, which is correlates with reduced expression of the fetal-expressed *LIN28B* and *HBG2* genes. Several studies have highlighted specific genes and miRNAs directly or indirectly modulating HbF levels. However, studies aiming to measure the global transcriptomes of fetal and adult RBC progenitors are limited in term of sample size or depth of information. Here, we measured by RNA-sequencing small and long RNA expression in fetal and adult erythroblasts derived from the fetal liver and bone marrow respectively. The aim of this study is to reveal major transcriptional differences between fetal and adult erythroblasts, and to provide resources for gene prioritization in genome-wide association studies of red blood cell traits.

4.2 Abstract

Sickle cell disease (SCD) and β-thalassemia are two of the most common monogenic disorders in the world. SCD alone affects around 300,000 newborn each year. Red blood cell (RBC) traits are associated with the severity of this disease. In particular, increased fetal hemoglobin (HbF) production in SCD patients is associated with decreased mortality and morbidity. Re-activation of HbF remains the most promising therapy for this disease. We aimed to characterize the transcriptomes of fetal and adult erythroblasts. We use RNA-sequencing to measure gene and micro-RNA (miRNA) expression patterns in erythroblasts derived from fetal liver (FL, N=12) and bone marrow (BM, N=12) hematopoietic progenitor stem cells (HPSC). We identify 3,687 and 4,142 transcripts that are up-regulated in adult and fetal erythroblasts, concurrent with 139 and 263 up-regulated miRNAs. We validate miRNA expression patterns in independent erythroblast samples from the FL (N=3) and BM (N=3), and identify miRNA *let-7* as highly up-regulated in adult cells and enriched for down-regulated targets. Conversely, fetal erythroblasts show stage-specific expression of the 14q32 miRNA cluster. Our results highlight

several genes that may be implicated in erythroid cell-stage specificity, and that may influence HbF production. More globally, our results provide comprehensive datasets that can be used to prioritize genes in genome-wide association studies of RBC traits.

4.3 Introduction

The human fetal-to-adult hemoglobin switch is a major developmental event occurring months following birth. During this transition, erythroblasts decrease the production of fetal hemoglobin (HbF) in favor of adult hemoglobin (HbA).[73,93] The former accounts generally for less than 1% of the total hemoglobin in healthy adults. The β-chain of hemoglobin tetramers are mainly composed of γ- and β-globin in fetal and adult red blood cells (RBC) respectively.[72,73] Sickle cell disease (SCD) and β-thalassemias are genetic disorders resulting from mutations at the β-globin gene (*HBB*) locus. These disorders are among the most common monogenic diseases: SCD alone affects more than 300,000 newborn each year.[73,111,365] The most important modifier of severity for these hemoglobinopathies is HbF: patients that naturally continue to express higher HbF levels have a longer life expectancy and suffer from less complications.[115] Re-activation of HbF remains the most promising therapy for the treatment of these genetic disorders.[73,93]

Loci associated with HbF levels have been identified through candidate-gene DNA sequencing, linkage scans, and more recently genome-wide association studies (GWAS). The association of this trait with variants at the *HBB*, *BCL11A*, and the *HBSIL-MYB* loci have been well replicated.[47,138-141] More recently, whole-genome sequencing in the founder Sardinian population has revealed an association with a rare intronic genetic variant in the *NFIX* gene, although this variant has not been replicated yet.[146] Genetic studies can help prioritize attractive therapeutic targets for HbF production reactivation, but further functional characterisation is needed to understand the role of these variants, and the modulated genes, on the regulation of HbF. For instance, variants act on *BCL11A*, a transcriptional repressor of HbF, by disrupting an erythroid developmental stage-specific enhancer of the gene.[56] On the other hand, it is likely

that variants at the *HBS1L-MYB* locus act by modulating expression of *MYB* resulting in erythroid differentiation defects, and thus might represent a less attractive therapeutic target.[145]

The fetal-to-adult hemoglobin switch is marked by several transcriptomic changes. For instance, the adult erythroid transcriptional program is marked by increased expression of *BCL11A, CA1* and *GCNT2*.[58,346,366] Studies have also implicated microRNAs (miRNAs) in the regulation of HbF production. Specifically, several members of the miRNA *let-7* family are up-regulated in adult erythroblasts.[349,367,368] Overexpression of *LIN28A* and *LIN28B*, which degrades *let-7* miRNAs, leads to increased HbF levels.[349,369,370] miR-96 is up-regulated in adult reticulocytes and overexpression of this miRNA inhibits γ-globin expression.[367] Expression of miR-486 increases during erythropoiesis, and the mature miRNA generated from 3p arm its precursor directly targets *BCL11A* in erythroid cells.[371] Elevated HbF levels in partial trisomy 13 cases have been linked to additional copies of miR-15a and miR-16-1, which themselves affect *MYB* expression levels.[152] Finally, elevated expression of miR-26b, miR-151-3p, miR-148a and miR-494 have been observed in SCD patients treated with the HbF-inducing agent hydroxyurea.[372,373] In particular, overexpression of miR-26b in K562 cells increased γ-globin expression. [374] Nonetheless, studies aiming to quantify the transcriptome of RBC progenitors, and to correlate transcriptomic changes with the fetal-to-adult globin switch in human samples, were limited either in terms of the number of transcripts assayed (using microarrays or qPCR), in terms of sample size, or did not assayed fetal-stage RBC progenitors.[345,367,368,375,376]

Here, we performed RNA and small RNA-sequencing (RNA-seq) of erythroblasts from 12 fetal liver (FL)- and 12 bone marrow (BM)-derived human donors. The transcriptomes of fetal and adult erythroblast captured known developmental stage-specific transcriptional events and displayed substantial differences, both in term of mRNA and miRNA expression. We find that *let-7* miRNAs were predominant in adult erythroblasts. Conversely, the chromosome 14q32 miRNA cluster was substantially up-regulated in fetal erythroblasts, suggesting a novel role of this cluster in the regulation of the fetal transcriptional program. We validated the differential expression of 72 miRNAs by quantitative polymerase chain reaction (qPCR) in an independent

sample of erythroblasts from fetal (N=3) and adult (N=3) donors. Finally, we highlight several miRNA that were enriched for inversely regulated targets.

4.4 Material and methods

4.4.1 Transcript differential expression analysis

We differentiated erythroblasts from hematopoietic stem/progenitor cells (HSPCs) as previously described.[58,377] We purchased human fetal (fetal liver, n=12) and adult (bone marrow, n=12) CD34+ HSPCs from DV Biologics and Lonza, respectively. We also described the RNA extraction and RNA-sequencing protocols elsewhere [377]. Briefly, we performed RNA-sequencing with an Illumina HiSeq2000 sequencer using a stranded cDNA library and a paired-ends 100bp protocol. We mapped reads to the genome (hg19) using Tophat2 (v.2.0.9, with options *--library-type fr-firststrand --microexon-search --coverage-search*).[233] Per gene read counts were obtained using htseq-count (v. 0.6.0, with options *-f bam -r name -s reverse -t exon -i gene_id*) on Ensembl gene sets (release 75). We tested differential expression (DE) of genes using the R package DESeq2 (v.1.12.4).[237] Data is available from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) repository (accession number: GSE90878).

4.4.2 miRNA differential expression analysis

We performed RNA sequencing of small RNAs on an Illumina HiSeq2000 sequencer using a stranded cDNA library with a single-end 50bp protocol. We removed adapter sequences using the fastx-clipper tool (v.0.0.14). We removed reads that mapped to ribosomal DNA using bowtie (v.1.0.0, with options *-n 0 -S –best --chunkmbs 200*). We mapped reads to both mature and precursor miRNA sequences from miRBASE release 21 using mirDeep2 (v.0.0.5).[378] We considered potentially novel miRNA with a mirDeep2 score > 3, which had an estimated 88 ± 2% true positive probability. We tested DE of miRNAs using the R package DESeq2 (v.1.12.4).[237] Data is available from the NCBI GEO repository (accession number: GSE90878).

4.4.3 Enrichment analyses

To identify miRNA that were enriched in differentially expressed targets, we downloaded all validated miRNA targets, as well as the top 35% predicted targets using multiMiR (v.2.1.1) for each miRNA.[379] We only considered miRNA-target pairs that were either validated in at least one resource or that were predicted in at least 3 databases. For each miRNA, we counted how many targets were inversely correlated (FDR<0.05) and were DE between fetal and adult erythroblasts. We calculated the enrichment using hypergeometric tests. In addition, we looked if genes were overrepresented in miRNA targets by assessing how many times each gene was targeted by a differentially expressed miRNA. Since the number of miRNA targeting a specific gene was often small, we permuted miRNA DE labels 10,000 times to assess significance. We counted the number of times a gene was targeted by more DE miRNAs in permutations than what was observed, and derived an empirical *P*-value by calculating the ratio of this value over the total number of permutations. We used the Generally Applicable Gene-set Enrichment (GAGE) (v.2.22.0) method to calculate the enrichment of Kyoto Encyclopedia of Genes and Genomes (KEGG).[237,380,381]

4.4.4 miRNA reverse-transcription quantitative PCR validation

We differentiated erythroblasts from HSPCs from either the fetal liver (N=3) or the bone marrow (N=3) as previously described.[58] We measured expression using a TaqMan Low Density Array covering 768 different miRNAs (ThermoFisher Scientific). Expression levels were derived using the $2^{-\Delta\Delta Ct}$ method. [382] DE was tested using a one-sided *t*-test on the ΔΔCt values for 202 miRNAs that were found DE in the RNA-seq dataset. miRNAs with DE $P<0.05$ in the right direction were considered as replicated.

4.5 Results and discussion

4.5.1 Gene differential expression recapitulates the fetal-to-adult hemoglobin switch

We first measured gene expression levels in erythroblasts differentiated *ex vivo* from HSPCs obtained from 12 fetal liver and 12 bone marrow (adult) human donors, for which we previously describe global DNA methylation differences.[377] Principal component analysis (PCA) of gene expression measures recapitulated cell developmental stage (first dimension, 62% variance explained) and sex (second dimension, 13% variance explained) (**Fig. 1A**). 3,687 and 4,142 transcripts were significantly up-regulated in adult and fetal erythroblasts, respectively (**Fig. 1B**). Of these, 1,129 adult and 2,000 fetal up-regulated genes had a fold change (FC) >=2. As expected, both *HBB* and *HBG2* were highly expressed, and both were differentially expressed (DE) in the expected direction (**Fig. 1B-D**). Although adult erythroblasts expressed *HBG2*, fetal erythroblasts exhibited very low levels of *HBB* compared to adult cells (**Figure 1D**). To a lesser extent, fetal erythroblasts also expressed alpha and beta embryonic globins (*HBZ* and *HBE*; **Fig. 1D**). Genes marking the fetal-to-adult erythroid transition such as *CA1* and *GCNT2*, as well as genes implicated in *HBG2* silencing such as *BCL11A* and *SOX6,* were up-regulated in adult erythroblasts (**Fig. 1B-C**). Conversely, *LIN28B* was highly up-regulated in fetal erythroblasts ($log_2FC=9.4$, $P=2x10^{-254}$). *LIN28B* has been shown to regulate the expression and translation of *IGF2,* which is highly expressed in fetal hepatocytes supporting HSCs. [349,383,384] Although *IGF2* itself was differentially expressed ($log_2FC=1.9$, $P=5x10^{-5}$), IGF2 binding partners *IGF2BP1* ($log_2FC=9.1$, $P<1x10^{-300}$) and *IGF2BP3* ($log_2FC=5.8$, $P<3x10^{-267}$) were much more up-regulated in fetal erythroblasts (**Fig. 1B-C**). Overexpression of *IGF2BP1* in adult RBC progenitors increased HbF levels.[385] Consistent with its role on erythroid cell proliferation, *MYB* was not differentially expressed ($log_2FC=0.08$, $P=0.54$). An intronic variant of *NFIX* was recently associated with HbF levels in the Sardinian population. This gene was up-regulated in adult erythroblasts ($log_2FC=-1.5$, $P=4.5x10^{-31}$). Consistently, CpGs near this gene show differential DNA methylation between fetal and adult erythroblasts, and its binding motif was enriched in differentially methylated regions.[377]

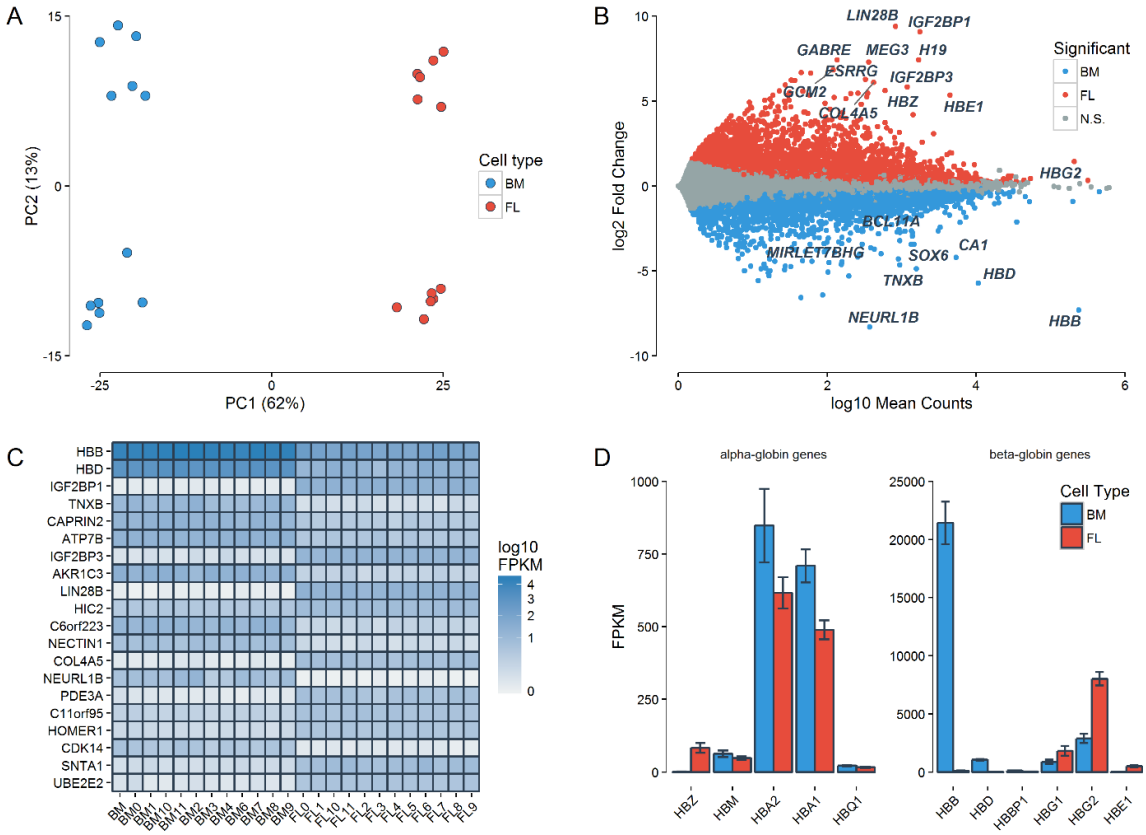


Figure 1. Differential expression (DE) of genes between 12 fetal and 12 adult erythroblast samples.

Fetal and adult erythroblasts were differentiated from CD34+ hematopoietic stem cells (HPSC) collected from the fetal liver (FL) and adult bone marrow (BM), respectively. **(A)** Principal component analysis (PCA) of the 500 top expressed genes. **(B)** Genes DE in fetal (red) and adult (blue) erythroblasts. We tested DE using DESeq2 and genes with a false discovery rate (FDR) <0.05 were considered significantly DE. The y-axis represent the fetal-to-adult expression fold change (FC). The x-axis represent genes mean normalized counts calculated by DESeq2. **(C)** Expression of the top 20 most significant DE genes in order of significance. **(D)** Mean Fragments Per Kilobase of transcript per Million mapped read (FPKM) of genes at the alpha- and beta-globin loci. Error bars represent the standard error of the mean (SEM).

The top pathways associated with adult up-regulated genes included adipocytokine signaling (P=2x10^{-13}), Fanconi anemia pathway (P=5x10^{-10}), and butanoate (P=8x10^{-6}) and fatty acid metabolism (P=8x10^{-5}) (**Table S1**). Mutations in genes causing Fanconi anemia are associated with increased fetal hemoglobin.[386,387] Moreover, short-chain fatty acids like butanoate (butyrate) induce HbF synthesis.[388] Fetal-stage genes were enriched in pathways such as Rap1 signaling (P=6x10^{-15}), cAMP signaling (P=4x10^{-14}), and calcium signaling (P=5x10^{-14}) (**Table S1**). The cAMP pathway has been implicated in sustaining high levels of HbF[389].

Overall, analyses of DE genes between fetal and human adult erythroblasts highlight known (ex. Fanconi anemia, hematopoietic cell lineage, HIF-1 signaling, heparin) and novel (ex. Rap1 signaling, butanoate metabolism) pathways in the developmental progression of human RBC precursors, some of which may directly contribute to the fetal-to-adult beta-globin switch.

Table 1. Top miRNAs enriched for differentially expressed genes

miRNA	Cell type where the miRNA is up-regulated	Total number of targets	Number of DE genes	Ratio of DE expressed genes on the total number of targets	*P*-value	*q*-value
hsa-miR-98-5p	BM	1262	385	0.305	$2.28x10^{-36}$	$7.28x10^{-34}$
hsa-let-7b-5p	BM	1747	470	0.269	$7.96x10^{-28}$	$1.27x10^{-25}$
hsa-let-7g-5p	BM	972	273	0.281	$2.68x10^{-20}$	$2.86x10^{-18}$
hsa-let-7i-5p	BM	1009	279	0.277	$1.09x10^{-19}$	$8.70x10^{-18}$
hsa-let-7a-5p	BM	1218	319	0.262	$1.00x10^{-18}$	$6.42x10^{-17}$
hsa-let-7f-5p	BM	957	260	0.272	$1.06x10^{-17}$	$5.67x10^{-16}$
hsa-let-7d-5p	BM	882	241	0.273	$7.37x10^{-17}$	$3.37x10^{-15}$
hsa-miR-200c-3p	FL	1009	267	0.265	$3.01x10^{-10}$	$1.20x10^{-08}$
hsa-miR-301b	FL	712	188	0.264	$7.48x10^{-09}$	$2.60x10^{-07}$
hsa-miR-495-3p	FL	1292	312	0.241	$8.13x10^{-09}$	$2.60x10^{-07}$
hsa-miR-543	FL	1055	249	0.236	$4.69x10^{-08}$	$1.36x10^{-06}$
hsa-miR-539-5p	FL	1073	260	0.242	$5.75x10^{-08}$	$1.53x10^{-06}$
hsa-miR-381-3p	FL	946	236	0.250	$6.62x10^{-08}$	$1.63x10^{-06}$
hsa-miR-96-5p	BM	910	224	0.246	$1.58x10^{-07}$	$3.60x10^{-06}$
hsa-let-7a-3p	BM	1026	234	0.228	$2.90x10^{-07}$	$6.19x10^{-06}$
hsa-miR-29c-3p	BM	865	210	0.243	$4.17x10^{-07}$	$8.34x10^{-06}$
hsa-miR-323a-3p	FL	451	125	0.277	$1.60x10^{-06}$	$3.00x10^{-05}$
hsa-miR-3163	BM	2652	580	0.219	$1.93x10^{-06}$	$3.43x10^{-05}$
hsa-let-7f-1-3p	BM	1007	222	0.220	$2.38x10^{-06}$	$4.00x10^{-05}$
hsa-miR-182-5p	BM	1050	240	0.229	$3.65x10^{-06}$	$5.82x10^{-05}$
hsa-miR-421	FL	594	141	0.237	$3.82x10^{-06}$	$5.82x10^{-05}$
hsa-let-7b-3p	BM	981	217	0.221	$5.15x10^{-06}$	$7.49x10^{-05}$
hsa-miR-183-5p	BM	697	162	0.232	$3.40x10^{-05}$	$4.73x10^{-04}$
hsa-miR-770-5p	FL	189	60	0.317	$7.77x10^{-05}$	$1.03x10^{-03}$
hsa-miR-301a-3p	FL	975	251	0.257	$8.05x10^{-05}$	$1.03x10^{-03}$
hsa-miR-369-3p	FL	666	153	0.230	$1.14x10^{-04}$	$1.40x10^{-03}$

P-values are calculated using a hypergeometric test. BM: Bone marrow erythroblasts; FL: Fetal liver erythroblasts, DE: Differentially expressed.

4.5.2 The 14q32 miRNA cluster is up-regulated in fetal erythroblasts

We next sequenced small RNAs in all 24 erythroblast samples to measure miRNA expression. Again, principal component of miRNA expression captured cell type (first dimension, 68% variance explained) (**Fig. 2A**). 139 and 263 miRNA were up-regulated in adult and fetal erythroblasts respectively (**Fig. 2B**). miR-486-5p was the most expressed miRNA in both fetal and adult erythroblasts, explaining ~34% of reads mapping to mature miRNAs (**Fig. 2B**). This miRNA is regulated by the erythroid master regulator *GATA1* and its expression increases during erythropoiesis.[390] Members of the *let-7* miRNA family, which are associated with the adult-stage erythroid program, constituted the most up-regulated miRNAs in adult erythroblasts (**Fig 2B-D**).[349,368] When mapping each mature miRNA to the genomic position of their precursor sequences, we found that the large miRNA cluster in the chromosome 14q32 imprinted region was highly up-regulated in fetal erythroblasts (**Fig. 2D**). The most significant miRNA of this locus was miR-381-3p ($\log_2$FC=8.1, P=1.5x10^{-69}). Loss of heterozygosity of this locus in acute lymphoblastic leukemia is associated with up-regulation of *BCL11A*, suggesting that up-regulation of this locus may be implicated in HbF production.[391] Finally, the miRDeep2 software predicted 214 novel miRNA expressed in erythroblasts.[378]

We took advantage of an existing dataset of miRNA measured by qPCR in 3 fetal and 3 adult independent erythroblast samples to validate our miRNA RNA-seq results. Of 202 miRNA differentially expressed in the RNA-seq analysis that were tested in this dataset, 72 showed a consistent DE pattern in the smaller qPCR assay (P<0.05), and 160 were in the expected direction (binomial P=2x10^{-17}). Moreover, there was a significant correlation between fold change of both assays (r=0.73, P=1.1x10^{-34}, **Fig. 3A**). Again, miRNAs consistently up-regulated in adult erythroblasts in both datasets were overrepresented by *let-7* miRNAs (7/14, P=0.01), whereas miRNA up-regulated in fetal erythroblasts were predominantly from the chromosome 14q32 locus (36/58, P=0.02) (**Fig. 3B**). Thus, qPCR expression measures were largely consistent with results obtained by RNA-seq.

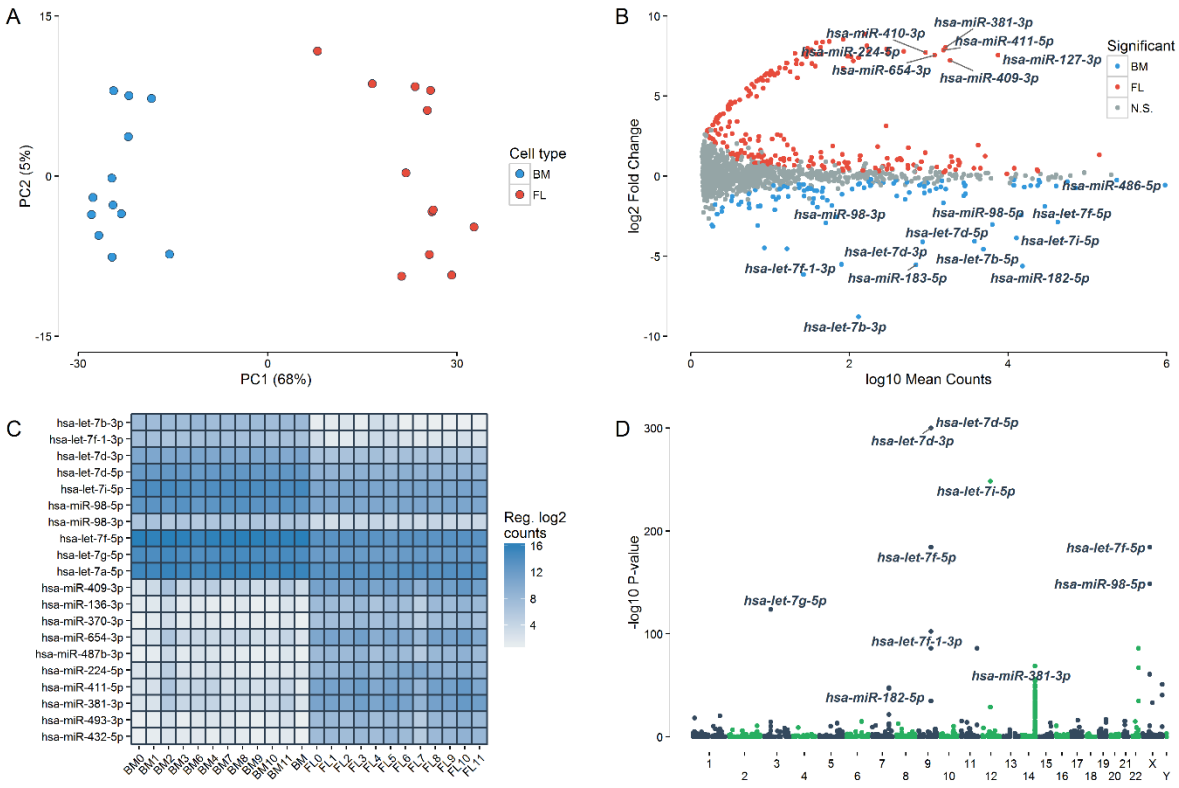


Figure 2. Differential expression (DE) of miRNAs between 12 fetal and 12 adult erythroblast samples.

Fetal and adult erythroblasts were differentiated from CD34+ hematopoietic stem cells (HPSC) collected from the fetal liver (FL) and adult bone marrow (BM) respectively. **(A)** Principal component analysis (PCA) of the 500 top expressed miRNAs. **(B)** miRNAs DE in fetal (red) and adult (blue) erythroblasts. We tested DE using DESeq2 and miRNAs with a false discovery rate (FDR) <0.05 were considered significantly DE. The y-axis represent the fetal-to-adult expression fold change (FC). The x-axis represent miRNAs mean normalized counts calculated by DESeq2. **(C)** Expression of the top 20 most significant DE miRNAs ordered by log2FC. Expression is reported in counts normalized using DESeq2's regularized log2 transformation. **(D)** Manhattan plot of miRNAs DE *P*-values (y-axis) based on the genomic location of their precursor sequences.

4.5.3 Integration of miRNA and mRNA expression data

Because miRNA can modulate gene expression by inhibiting translation and promoting the degradation of their target transcripts, we next asked whether DE target genes were enriched for DE miRNA, but in opposite direction. In other words, if target genes are up-regulated in adult erythroblasts, are their targeting miRNA down-regulated in the same cell type? We linked these miRNAs to their targets using an aggregated list of validated and predicted miRNA-target interactions (**Methods**). We found 18 and 34 adult and fetal up-regulated miRNAs that were

enriched for down-regulated targets (hypergeometric test FDR<0.05; **Fig. 4A-B**, **Table 1**). Of the 18 adult-stage miRNAs, 10 were from the *let-7* miRNA family - an enrichment since only 15 of the 106 adult-stage miRNAs included in this analysis were from the *let-7* miRNA family (Fisher exact test P=0.0003). Conversely, 25 of the 34 fetal-stage miRNAs were part of the 14q32 miRNA cluster (Fisher exact test P=6x10^{-5}). Moreover, 7 of these fetal-stage miRNAs were predicted to target *BCL11A*, including miR-381-3p – the most significantly up-regulated miRNA in fetal erythroblasts (**Fig. 2B-D**). In general, adult-stage miRNAs were more similar in terms of targets, whereas fetal-stage miRNAs were more heterogeneous (**Fig. 4A-B**). In total, 58 down-regulated genes were targeted by 10 or more of the enriched adult-stage miRNAs. These genes included the highly DE gene *IGF2BP1* (Total number of DE miRNAs, n=10), as well as *AGO1* (n=11) and *DICER1* (n=12), both members of the miRNA biogenesis pathway. In contrast, 27 genes were targeted by 10 or more of the 34 enriched fetal-stage miRNAs.

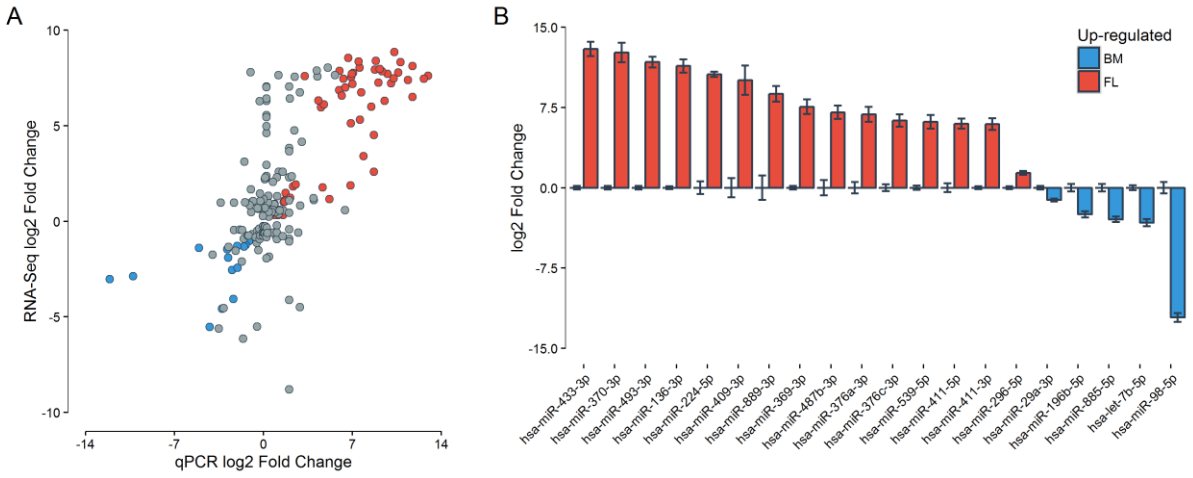


Figure 3. Validation of DE miRNAs by qPCR.
(A) Correlation between miRNA expression fold changes obtained by RNA-sequencing (N_{adult}=12, N_{fetal}=12) and qPCR (N_{adult}=3, N_{fetal}=3). Only miRNA DE in the RNA-sequencing experiment are included. Red and blue dots represent DE genes with P<0.05 in the qPCR experiment that are up-regulated in fetal and adult erythroblasts respectively. **(B)** qPCR log2 fold changes of the 20 most qPCR significant DE miRNAs (all P<5x10^{-3}). Each bar represent the mean log2 fold change for adult (right) and fetal (left) erythroblasts. Error bars represent the standard error of the mean (SEM).

As a complementary analysis, we explored whether some genes were enriched targets of DE miRNAs. We found 27 adult and 7 fetal down-regulated genes that were targeted more often by DE miRNAs than what was expected by chance (FDR<0.05, **Table 2**). The higher

number of adult down-regulated targets was mainly due to being shared targets of *let-7* miRNAs (26 of the 27 targets were targeted by at least one *let-7* miRNA): only *FBN1* was not targeted by *let-7* miRNAs (**Table 2**). The most enriched target in fetal-stage erythroblasts was *TRPS1* ($P<1x10^{-4}$), a transcriptional repressor that specifically targets GATA sequences.[392] *BCL11A* was a potential target of 17 fetal up-regulated miRNAs, but was only nominally enriched (P=0.01, q-value=0.18).

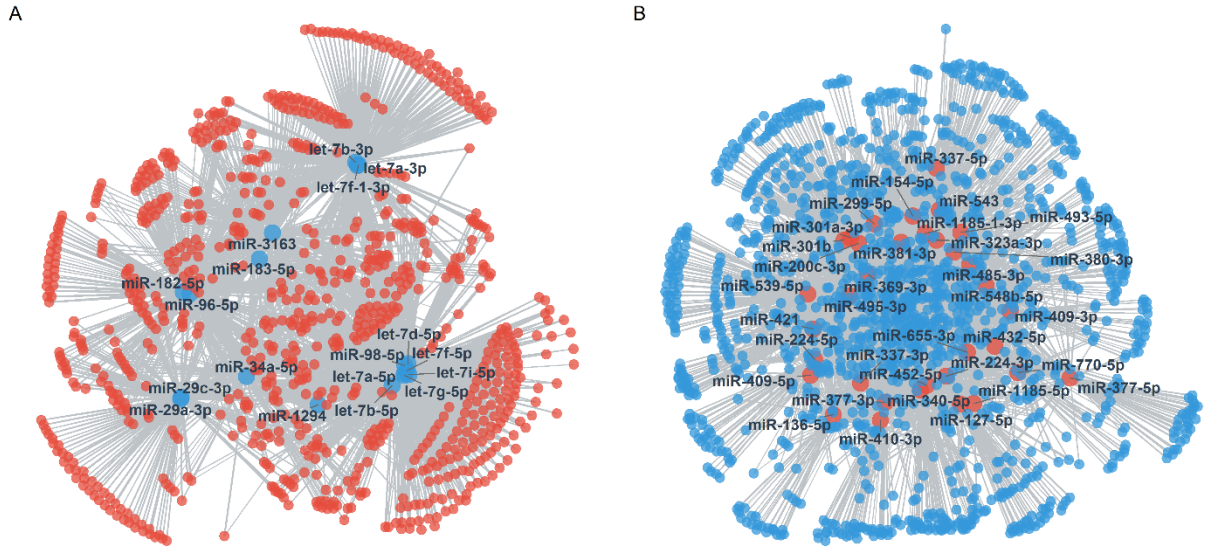


Figure 4. miRNA target networks.
(A) Network of adult up-regulated miRNAs enriched for down-regulated targets. **(B)** Network of fetal up-regulated miRNAs enriched for down-regulated targets. Red and blue dots represent genes or miRNAs up-regulated in fetal and adult erythroblasts respectively.

Table 2. Genes significantly targeted by DE miRNAs

Gene symbol	Annotation	Cell where gene is up-regulated	Total number of miRNAs targeting this gene	Number of DE miRNAs	*P*-value	*q*-value
GYG2	glycogenin 2	FL	14	7	$<1x10^{-4}$	0
PLCB4	phospholipase C beta 4	FL	47	16	$<1x10^{-4}$	0
TRPS1	transcriptional repressor GATA binding 1	BM	137	31	$<1x10^{-4}$	0
DAGLA	diacylglycerol lipase alpha	FL	39	15	$<1x10^{-4}$	0
DZIP1	DAZ interacting zinc finger protein 1	FL	25	11	$<1x10^{-4}$	0
GPAT3	glycerol-3-phosphate acyltransferase 3	FL	13	8	$<1x10^{-4}$	0
PTPRO	protein tyrosine phosphatase	FL	39	15	$<1x10^{-4}$	0
SLAMF6	SLAM family member 6	FL	13	8	$<1x10^{-4}$	0
LIPH	lipase H	FL	13	8	$<1x10^{-4}$	0
COL4A6	collagen type IV alpha 6 chain	FL	26	12	$<1x10^{-4}$	0
GRID2IP	Grid2 interacting protein	FL	13	7	$<1x10^{-4}$	0
HGF	hepatocyte growth factor	BM	28	10	0.0001	$9.65x10^{-3}$
STARD13	StAR related lipid transfer domain containing 13	FL	70	20	0.0001	$9.65x10^{-3}$
CCR7	C-C motif chemokine receptor 7	FL	16	8	0.0002	$1.48x10^{-2}$
STX3	syntaxin 3	FL	23	9	0.0002	$1.48x10^{-2}$
NAT8L	N-acetyltransferase 8 like	FL	34	12	0.0002	$1.48x10^{-2}$
ADRB3	adrenoceptor beta 3	FL	17	8	0.0002	$1.48x10^{-2}$
PHACTR2	phosphatase and actin regulator	FL	87	20	0.0003	$2.09x10^{-2}$
PLD3	phospholipase D family member 3	FL	15	7	0.0004	$2.28x10^{-2}$
C1orf21	chromosome 1 open reading frame 21	FL	87	21	0.0004	$2.28x10^{-2}$
HIF3A	hypoxia inducible factor 3 alpha subunit	FL	29	10	0.0004	$2.28x10^{-2}$
FBN1	fibrillin 1	FL	36	11	0.0004	$2.28x10^{-2}$
MYCN	v-myc avian myelocytomatosis viral oncogene neuroblastoma derived homolog	FL	88	20	0.0005	$2.61x10^{-2}$
JAZF1	JAZF zinc finger 1	BM	108	23	0.0005	$2.61x10^{-2}$
ERRFI1	ERBB receptor feedback inhibitor 1	BM	36	11	0.0006	$2.89x10^{-2}$
SLC46A3	solute carrier family 46 member 3	BM	33	10	0.0006	$2.89x10^{-2}$
ELOVL4	ELOVL fatty acid elongase 4	FL	47	13	0.0007	$3.14x10^{-2}$
NRARP	NOTCH-regulated ankyrin repeat protein	FL	32	10	0.0007	$3.14x10^{-2}$
RIMKLB	ribosomal modification protein rimK like family member B	BM	93	20	0.0009	$3.89x10^{-2}$
HES1	hes family bHLH transcription factor 1	FL	23	8	0.0011	$4.45x10^{-2}$
THRB	thyroid hormone receptor beta	BM	63	15	0.0011	$4.45x10^{-2}$
COL1A1	collagen type I alpha 1 chain	FL	39	11	0.0012	$4.56x10^{-2}$
ASIC1	acid sensing ion channel subunit 1	FL	30	9	0.0012	$4.56x10^{-2}$
ADAMTS 6	ADAM metallopeptidase with thrombospondin type 1 motif 6	FL	71	16	0.0013	$4.79x10^{-2}$

P-values are calculated from 10,000 permutations. BM: Bone marrow erythroblasts; FL: Fetal liver erythroblasts, DE: Differentially expressed.

4.6 Conclusions

In summary, gene expression patterns of fetal and adult RBC progenitors largely capture cell type specificity. We find 3,687 and 4,142 transcripts DE between adult and fetal respectively, of which 1,129 and 2,000 are more than twice as expressed in the respective cell type. DE genes recapitulate known genes associated with cell developmental stages, and are enriched in pathways that have been implicated in HbF production. Concomitantly, we find 139 and 263 miRNAs up-regulated in adult and fetal erythroblasts. Notably, *let-7* miRNAs and their target *IGF2BP1, IGF2BP3,* and *LIN28B* are highly differentially expressed, and have been implicated in HbF regulation.[349,368,385] 14q32 miRNAs show inversed expression pattern. Further studies will be required to define their role on erythroid developmental-stage specificity. We validated 72 of these miRNAs in a smaller dataset. DE miRNAs are enriched for down-regulated targets, suggesting that they may be important contributors to developmental cell-stage specificity. Finally, we highlight several genes that are enriched targets of DE miRNAs.

GWAS have identified hundreds of genes association with RBC indices. As most GWAS, most variants identified fall in non-coding regions, increasing the difficulty of establishing causal links. Transcriptomes characterized in this study provide comprehensive datasets that could aid GWAS loci characterisation. In particular, we identify several genes and miRNAs that show fetal and adult developmental stage-specific expression patterns, and could influence HbF production. Clinically, HbF is the most important modifier of SCD severity and its re-activation remains the most promising therapy.

4.7 Acknowledgments

S.L. holds fellowships from the Canadian Institute of Health Research (CIHR) and the "Fondation Pierre Lavoie". D.E.B. was supported by NIDDK (K08DK093705, R03DK109232), NHLBI (DP2OD022716), Burroughs Wellcome Fund, American Society of Hematology, and the Doris Duke Charitable, Charles H. Hood, and Cooley's Anemia Foundations. G.L. was supported by grants from the CIHR (#123382), the Doris Duke Charitable Foundation, the

National Sciences and Engineering Research Council of Canada (RGPIN-2016-04597), the Montreal Heart Institute Foundation, and the Canada Research Chair Program.

4.8 Supplementary information

Table S1. KEGG pathways enriched in genes up-regulated in fetal or adult erythroblasts.

Pathway	*P*-value	*q*-value	Number of DE genes in pathway	Cell type where pathway is enriched
hsa04670 Leukocyte transendothelial migration	2.83E-19	5.79E-17	101	Fetal
hsa04610 Complement and coagulation cascades	2.02E-18	2.07E-16	70	Fetal
hsa04015 Rap1 signaling pathway	9.11E-17	6.23E-15	198	Fetal
hsa04024 cAMP signaling pathway	7.56E-16	3.88E-14	181	Fetal
hsa04020 Calcium signaling pathway	1.34E-15	5.04E-14	170	Fetal
hsa04974 Protein digestion and absorption	1.47E-15	5.04E-14	86	Fetal
hsa04920 Adipocytokine signaling pathway	2.15E-13	4.42E-11	65	Adult
hsa04140 Regulation of autophagy	5.22E-13	5.35E-11	124	Adult
hsa04510 Focal adhesion	4.10E-12	1.20E-10	194	Fetal
hsa04062 Chemokine signaling pathway	3.28E-11	8.40E-10	157	Fetal
hsa04010 MAPK signaling pathway	2.02E-10	4.27E-09	238	Fetal
hsa04080 Neuroactive ligand-receptor interaction	2.08E-10	4.27E-09	210	Fetal
hsa04261 Adrenergic signaling in cardiomyocytes	2.51E-10	4.68E-09	134	Fetal
hsa04310 Wnt signaling pathway	1.45E-09	2.47E-08	134	Fetal
hsa03460 Fanconi anemia pathway	5.40E-10	3.69E-08	54	Adult
hsa04727 GABAergic synapse	6.47E-09	1.02E-07	77	Fetal
hsa03030 DNA replication	4.75E-09	2.44E-07	36	Adult
hsa04810 Regulation of actin cytoskeleton	1.79E-08	2.62E-07	198	Fetal
hsa04060 Cytokine-cytokine receptor interaction	1.96E-08	2.67E-07	201	Fetal
hsa04114 Oocyte meiosis	1.78E-08	7.30E-07	113	Adult
hsa04664 Fc epsilon RI signaling pathway	6.83E-08	8.76E-07	67	Fetal
hsa04066 HIF-1 signaling pathway	7.95E-08	9.59E-07	96	Fetal
hsa04666 Fc gamma R-mediated phagocytosis	1.02E-07	1.16E-06	88	Fetal
hsa04360 Axon guidance	2.68E-07	2.75E-06	169	Fetal
hsa04380 Osteoclast differentiation	2.58E-07	2.75E-06	120	Fetal
hsa03420 Nucleotide excision repair	4.34E-07	1.48E-05	45	Adult
hsa04611 Platelet activation	1.70E-06	1.66E-05	118	Fetal
hsa03440 Homologous recombination	5.85E-07	1.71E-05	41	Adult
hsa04110 Cell cycle	9.48E-07	2.43E-05	123	Adult

hsa04916 Melanogenesis	3.51E-06	3.27E-05	94	Fetal
hsa04022 cGMP-PKG signaling pathway	3.83E-06	3.41E-05	151	Fetal
hsa04550 Signaling pathways regulating pluripotency of stem cells	5.06E-06	4.33E-05	127	Fetal
hsa04142 Lysosome	5.28E-06	4.33E-05	121	Fetal
hsa00534 Glycosaminoglycan biosynthesis - heparan sulfate / heparin	7.02E-06	5.33E-05	22	Fetal
hsa04961 Endocrine and other factor-regulated calcium reabsorption	6.86E-06	5.33E-05	47	Fetal
hsa04640 Hematopoietic cell lineage	8.85E-06	6.48E-05	86	Fetal
hsa04514 Cell adhesion molecules (CAMs)	1.03E-05	7.25E-05	132	Fetal
hsa04750 Inflammatory mediator regulation of TRP channels	1.07E-05	7.32E-05	93	Fetal
hsa04520 Adherens junction	1.19E-05	7.63E-05	72	Fetal
hsa04971 Gastric acid secretion	1.18E-05	7.63E-05	69	Fetal
hsa04014 Ras signaling pathway	1.28E-05	7.98E-05	212	Fetal
hsa04260 Cardiac muscle contraction	2.69E-05	1.62E-04	70	Fetal
hsa04390 Hippo signaling pathway	2.82E-05	1.65E-04	149	Fetal
hsa04725 Cholinergic synapse	3.34E-05	1.90E-04	106	Fetal
hsa00650 Butanoate metabolism	8.36E-06	1.90E-04	25	Adult
hsa00511 Other glycan degradation	5.18E-05	2.87E-04	17	Fetal
hsa04350 TGF-beta signaling pathway	8.73E-05	4.71E-04	81	Fetal
hsa04918 Thyroid hormone synthesis	9.04E-05	4.75E-04	70	Fetal
hsa04724 Glutamatergic synapse	9.89E-05	4.94E-04	107	Fetal
hsa04977 Vitamin digestion and absorption	9.79E-05	4.94E-04	22	Fetal
hsa00071 Fatty acid degradation	2.94E-05	5.02E-04	39	Adult
hsa00140 Steroid hormone biosynthesis	2.78E-05	5.02E-04	40	Adult
hsa04146 Peroxisome	2.90E-05	5.02E-04	82	Adult
hsa04921 Oxytocin signaling pathway	1.47E-04	7.18E-04	143	Fetal
hsa00982 Drug metabolism - cytochrome P450	1.59E-04	7.57E-04	46	Fetal
hsa04370 VEGF signaling pathway	1.82E-04	8.29E-04	58	Fetal
hsa04978 Mineral absorption	1.82E-04	8.29E-04	48	Fetal
hsa03430 Mismatch repair	5.84E-05	9.06E-04	23	Adult
hsa04141 Protein processing in endoplasmic reticulum	6.19E-05	9.06E-04	163	Adult
hsa01212 Fatty acid metabolism	8.36E-05	1.14E-03	47	Adult

Chapter 5: An erythroid-specific enhancer of *ATP2B4* mediates red blood cell hydration and malaria susceptibility

Authors

Samuel Lessard, Emily N. Stern, Mélissa Beaudoin, Patrick G. Schupp, Falak Sher, Syed Adnan Ali, Sukhpal Prehar, Ryo Kurita, Yukio Nakamura, Esther Baena, Jonathan Ledoux, Delvac Oceandy, Daniel E. Bauer, Guillaume Lettre

Reference

Lessard S, Stern EN, Beaudoin M, Schupp PG, Sher F, Ali SA, Prehar S, Kurita R, Nakamura Y, Baena E, Ledoux J, Oceandy D, Bauer DE, Lettre G. An erythroid-specific enhancer of *ATP2B4* mediates red blood cell hydration and malaria susceptibility. 2017 [Submitted].

Author's contribution

S.L., E.N.S., D.E.B and G.L. conceived and designed the experiments. S.L., E.N.S., M.B., P.G.S., F.S., S.A.A., S.P. and J.L. performed the experiments. S.L., E.N.S, J.L., D.E.B. and G.L. analyzed the data. R.K., Y.N., E.B. and D.O. contributed reagents and materials. S.L., E.N.S., D.E.B and G.L. wrote the manuscript with contributions from all authors.

5.1 Context

Genome-wide association studies have identified >100 loci associated with red blood cell (RBC) traits. Characterization of these genetic variants is challenging as most fall in non-coding regions. These variants may act on the expression of neighboring genes by modulating regulatory elements. Expression quantitative trait loci (eQTL) mapping can be useful in linking non-coding GWAS variants to genes. However, since these variants may act in a cell- or tissue-specific manner, it may be crucial to survey a cell or tissue related to the trait of interest. This study builds on the transcriptional maps generated in Chapter 4. Here, we use eQTL mapping to link genetic variants to gene expression, and relate these to RBC traits with a particular consideration to the *ATP2B4* locus, which has been associated to RBC traits and malaria susceptibility.

5.2 Abstract

Few genotype-phenotype associations identified by genome-wide association studies (GWAS) have been defined mechanistically, precluding thorough assessment of their impact on human health. We conducted an expression quantitative trait loci (eQTL) mapping analysis in erythroblasts and found erythroid-specific eQTLs for *ATP2B4*, the main calcium ATPase of red blood cells (RBC). The same SNPs were previously associated with mean corpuscular hemoglobin concentration (MCHC) and susceptibility to severe malaria infection. We showed that *Atp2b4*$^{-/-}$ mice demonstrate increased MCHC, confirming *ATP2B4* as the causal gene at this GWAS locus. Using CRISPR-Cas9, we fine-mapped the genetic signal to an erythroid-specific enhancer. Erythroid cells with a deletion of the *ATP2B4* enhancer have abnormally high intracellular calcium levels. These results illustrate the importance to combine transcriptome, epigenome, and genome editing approaches in phenotype-relevant cells to characterize non-coding regulatory elements. Our study suggests ATP2B4 as a potential target to modulate RBC hydration in erythroid disorders and malaria infection.

5.3 Introduction

Genome-wide association studies (GWAS) have identified hundreds of loci associated with common human diseases and other clinically relevant traits. Most of these DNA sequence variants map to non-coding regions of the human genome. The functional characterization of genotype-phenotype associations implicating non-coding variants remains a major bottleneck. Some non-coding variants influence phenotypic variation by modulating the activity of cell- or tissue-specific gene regulatory elements[56,300,393]. The statistical enrichment of GWAS-implicated SNPs in regulatory sequences predicted from epigenomic profiling suggests a promising strategy for fine-mapping[394,395]. However, relatively few examples of regulatory mechanisms at individual loci have been described in detail, limiting the ability to design informative high-throughput experiments to characterize causal variants and genes.

Erythropoiesis – the differentiation of hematopoietic stem cells into mature enucleated red blood cells (RBCs) – is an auspicious system to dissect how non-coding genetic variants influence phenotypes. The process is largely cell-autonomous and driven by a small set of master transcription factors. Well-established cell culture protocols exist to monitor proliferation and differentiation. Furthermore, GWAS have already revealed >100 loci associated with the number, size or hemoglobin content of RBCs[13,54]. Fine-mapping these genetic associations with RBC traits promises not only to provide new illustrations of how non-coding variants influence complex phenotypes through effects on gene expression, but will also reveal genes that control RBC biology in health and disease.

5.4 Results

5.4.1 eQTL mapping in erythroblasts identifies novel cell-specific associations with gene expression

We mapped expression quantitative trait loci (eQTLs) in *ex vivo* differentiated human erythroblasts, the nucleated precursors of mature RBCs[377]. To increase statistical power, we

focused the eQTL search on 479 genes that display allelic imbalance (AI) ($P<2x10^{-5}$) (**Methods**). For each of these 479 genes, we tested if nearby SNPs (within 100-kb) were associated with their expression levels and had genotypes consistent with the observed AI effect (**Figure 1A** and **Methods**). We observed a strong enrichment of eQTLs among variants located near AI genes (**Figure 1B**). In total, we identified 6,325 significant eQTLs associated with the expression of 174 different genes at a false discovery rate (FDR) <0.05 (**Figure 1C**). We observed further enrichment of erythroblast eQTLs within erythroid enhancers identified by DNAse I hypersensitive site (DHS) and histone tail modification analyses, ChIP-seq binding sites for the erythroid master transcriptional regulators GATA1 and TAL1, as well as the short binding motifs (12-18 bp) for GATA1 and GATA1::TAL1 (**Figure 1C**). We noted that the co-occuring GATA1::TAL1 motifs showed the greatest inflation among these annotations. Thus, epigenome features prioritize variants that control gene expression in human erythroblasts. Variants associated with RBC traits by GWAS were also over-represented among significant erythroblast eQTLs (**Figure S1**)[13].

We compared our eQTL results to the GTEx dataset[396]. Although GTEx does not include erythroblasts, it is a powerful resource to confirm eQTL effects that are shared across cell types. Of the 5,924 erythroid eQTLs for which results were available in GTEx, 4,502 (76%) were replicated at $P<0.001$ in at least one tissue. On average, human erythroblasts and individual GTEx tissue share 1,755 eQTLs that control the expression of 32 genes (**Figure 1D** and **Figures S2-S4**). We found 63 genes with candidate erythroblast-specific eQTLs. Overall, genes with eQTLs in erythroblasts were enriched for genes implicated in heme biosynthesis ($P<6.6x10^{-7}$) and mouse RBC phenotypes ($P<8.9x10^{-7}$)(**Table S1**).

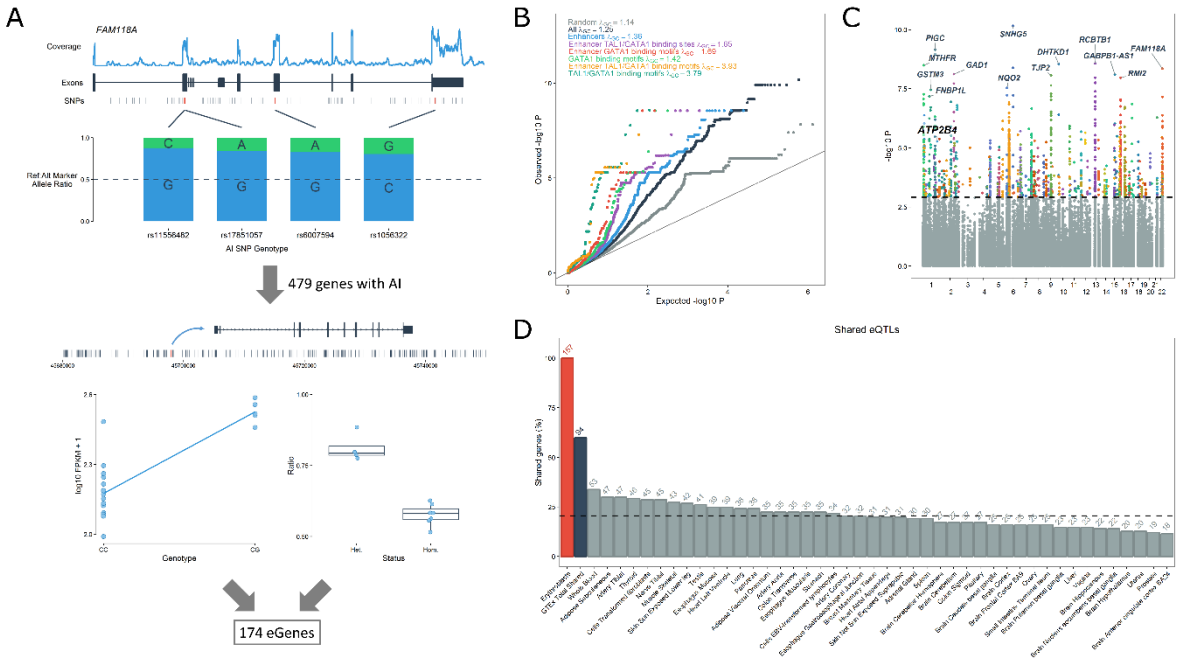


Figure 1. eQTL mapping in erythroblasts.
(**A**) To identify eQTLs in erythroblasts, we first focused on genes that show allelic imbalance (AI) in at least one sample (n=479 AI genes). Then, we tested if SNPs located within 100 kb of these AI genes were associated with their expression level (left panel) and if their genotypes were consistent with the expected AI ratio of "reference allele:alternate allele" (right panel). In this example, we highlight the candidate eQTL variant rs7287869 that is associated with the expression of the AI gene *FAM118A*. (**B**) Quantile-quantile (QQ) plot of eQTL *P*-values for variants located within 100 kb of 479 AI genes in human erythroblasts (black). Given that this analysis is limited to AI genes, we expect to observe a strong inflation of the eQTL test statistics (λ_{GC}=1.25). In comparison, the inflation is reduced (λ_{GC}=1.14) when analyzing variants located near randomly selected 479 non-AI genes (grey). This residual inflation could be explained if some of these genes have real eQTLs in the absence of AI, or if they have AI effects that merely miss statistical significance. We generated subsets of SNPs overlapping erythroid enhancers (blue), GATA1 and TAL1 ChIP-seq peaks inside erythroid enhancers (purple), GATA1 or GATA1-TAL1 binding motifs inside erythroid enhancers (red and yellow, respectively), or all GATA1 or GATA1-TAL1 binding motifs (light and dark green, respectively). These subsets of variants show substantial enrichment (as summarized by the λ_{GC} statistic) when compared to all SNPs (black). (**C**) Manhattan plot of eQTL *P*-values. The dashed line corresponds to FDR *q*-value=0.05. (**D**) Number of genes that share at least one eQTL between erythroblasts and the GTEx tissues (at $P<0.001$). The dashed line corresponds to the mean percentage shared eGenes (mean=20.8%).

5.4.2 *ATP2B4* eQTLs and RBC traits

Of all the genes with erythroblast-specific eQTLs, we were particularly interested by *ATP2B4* (also known as PMCA4) because it encodes the main calcium ATPase of RBCs eQTLs (**Figure 2A** and **Figure S5**). GTEx has identified eQTLs for *ATP2B4*, but these variants are in weak linkage disequilibrium (LD) with the erythroblast-specific eQTLs (r^2<0.09 in the 1000

Genomes Project) and are not associated with *ATP2B4* expression levels in erythroblasts (P>0.05 after correction for multiple testing)(**Figure S6**). We noted that the same SNPs associated with the expression of *ATP2B4* in human erythroblasts had previously been associated with mean corpuscular hemoglobin concentration (MCHC, a measure of RBC hydration) and susceptibility to severe malaria infection by GWAS[61,107,397,398], implicating *ATP2B4* as the likely causal gene for these RBC-related phenotypes.

To test the role of *Atp2b4* in RBC phenotypes, we analyzed blood from mice with a targeted deletion of this gene. *Atp2b4* knockout mice are viable but characterized by male infertility and protection against pathological cardiac hypertrophy[399,400]. We found that MCHC was elevated in these *Atp2b4*$^{-/-}$ mice (**Figure 2B**), consistent with the observation that the allele associated with low *ATP2B4* expression in erythroblasts is associated with higher MCHC in humans[61]. These results corroborate that *Atp2b4* plays a causal role in maintaining MCHC *in vivo*.

To extend the characterization of variants at the *ATP2B4* locus and their effects on RBC phenotypes, we used the first release of the UK Biobank to test the association between *ATP2B4* erythroblast-specific eQTLs and eight RBC traits (**Table S2**). We observed a strong association between the A-allele of rs7551442 and increased MCHC, replicating the signal from previous GWAS (P=2.6x10^{-19}, **Figure 2C-D** and **Figure S7**) and consistent with a recent report[13]. We also detected an association of this allele with decreased red blood cell distribution width (RDW)(P=1.2x10^{-22}) and increased hemoglobin levels (P=2.1x10^{-7})(**Figure 2C-D**). The *ATP2B4* genetic association signals with RBC traits in the UK Biobank were essentially identical to the *ATP2B4* erythroid eQTL association signals in human erythroblasts (**Figure 2C-D**). This concordance supports the hypothesis that the variants act on RBC traits and malaria susceptibility through an effect on the expression of *ATP2B4* in erythroid cells.

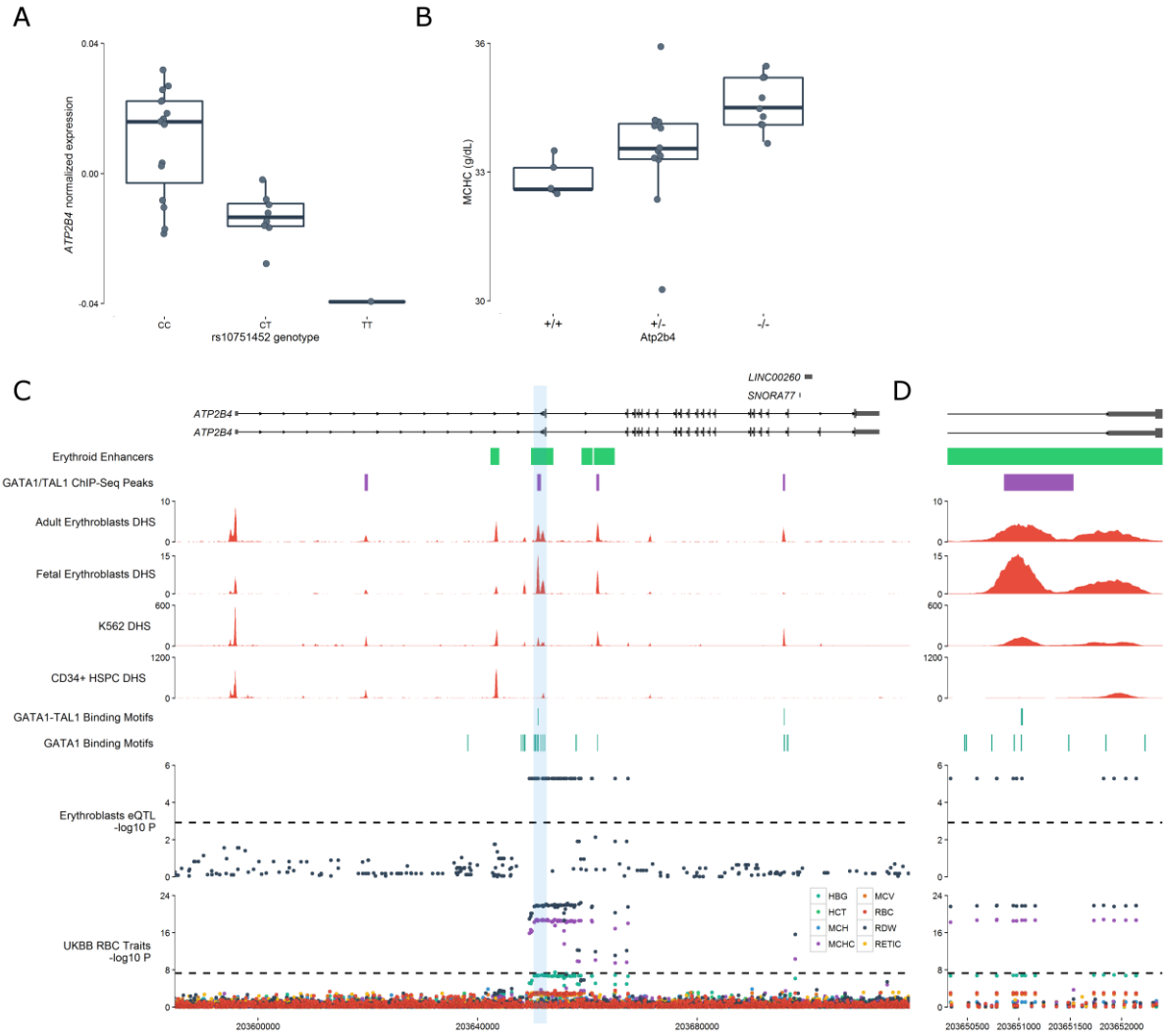


Figure 2. *ATP2B4* eQTLs overlap an erythroid-specific regulatory region and are associated with RBC traits.

(**A**) Association of the rs10751452 C-allele with decreased *ATP2B4* expression in human erythroblasts (P=5.3x10^{6}). Linkage disequilibrium between rs10751452 and the *ATP2B4* sentinel GWAS SNP, rs7551442, is r^2=1.0. Normalized expression corresponds to residuals of $\log_{10}$(FPKM) after correcting for cell developmental stage in a linear regression model. Boxplots represent the median (central line), the first and third quartiles (hinges), and lowest and highest value inside 1.5 times the inter-quartile range from the hinges (whiskers). (**B**) Knockout of *Atp2b4* in mice induces a dose-dependent increase in MCHC (P=0.0027). (**C**) *ATP2B4* eQTLs overlap erythroid enhancers, and cluster around an erythroid-specific DNase 1 hypersensitive site (DHS) bound by GATA1 and TAL1. This site is not present in undifferentiated CD34^{+} hematopoietic stem/progenitor cells. Several GATA1 binding motifs are clustered inside this regulatory region. eQTLs for *ATP2B4* are associated with mean corpuscular hemoglobin concentration (MCHC) and red blood cell distribution width (RDW) in the UK Biobank (UKBB). (**D**) Zoom-in of the erythroid-specific regulatory region (left DHS peak) in intron 1 of *ATP2B4*. We note the near perfect concordance between the UKBB association results and the eQTL results in erythroblasts, as well as the overlap of these SNPs with an erythroid-specific enhancer and GATA1/TAL1 sites.

5.4.3 An erythroid-specific regulatory element is required for ATP2B4 expression

Characterization of DHSs at the *ATP2B4* locus revealed an intronic DHS peak that is present in both adult and fetal erythroblasts (**Figure 2C-D**)[334]. This DHS peak was also present in the K562 erythroleukemic cell line, but absent from 229 other cell types and tissues, including $CD34^+$ hematopoietic stem/progenitor cells (HSPCs)(**Figure S8**)[185,215]. Further annotation of this DHS revealed that it overlaps with an erythroid enhancer chromatin signature[334], several GATA1 binding motifs, and harbors multiple erythroblast-specific eQTLs in strong LD that are associated with *ATP2B4* expression and RBC phenotypes (**Figure 2D** and **Figure 3A**). Supporting the regulatory potential of this DHS in erythroid cells, a recent analysis of ENCODE data found that it shows allele-specific transcription factor binding only in K562 cells[258].

We posited that this region may harbor the causal variant(s) influencing *ATP2B4* expression, RBC traits, and malaria susceptibility. To test this hypothesis, we deleted this region in a human erythroid precursor cell line (HUDEP-2) using the CRISPR-Cas9 system[60]. We introduced two independent pairs of guide RNAs, one of which results in a 927-bp deletion and the other in a 889-bp deletion (**Table S3**). We observed a dose-dependent reduction in *ATP2B4* expression upon enhancer deletion, with biallelic deleted clones displaying only 3% residual *ATP2B4* expression (**Figure 3B**). There was an intermediate phenotype in monoallelic enhancer deleted cells (**Figure 3B**). A clone in which one copy of the enhancer was deleted and the other copy was inverted showed a similar expression pattern to the monoallelic enhancer deleted clones, suggesting that the enhancer can function independent of orientation *in situ* (**Figure 3B**).

To test the requirement of the enhancer element in non-erythroid cells, we generated 293T cells (human embryonic kidney-derived) with the same 927-bp deletions. In contrast to HUDEP-2 cells, no change in *ATP2B4* expression was noted upon enhancer deletion in the 293T cells (**Figure 3C**). These results suggest that the *ATP2B4* intronic element is required for *ATP2B4* expression in erythroid cells but dispensable in non-erythroid cells.

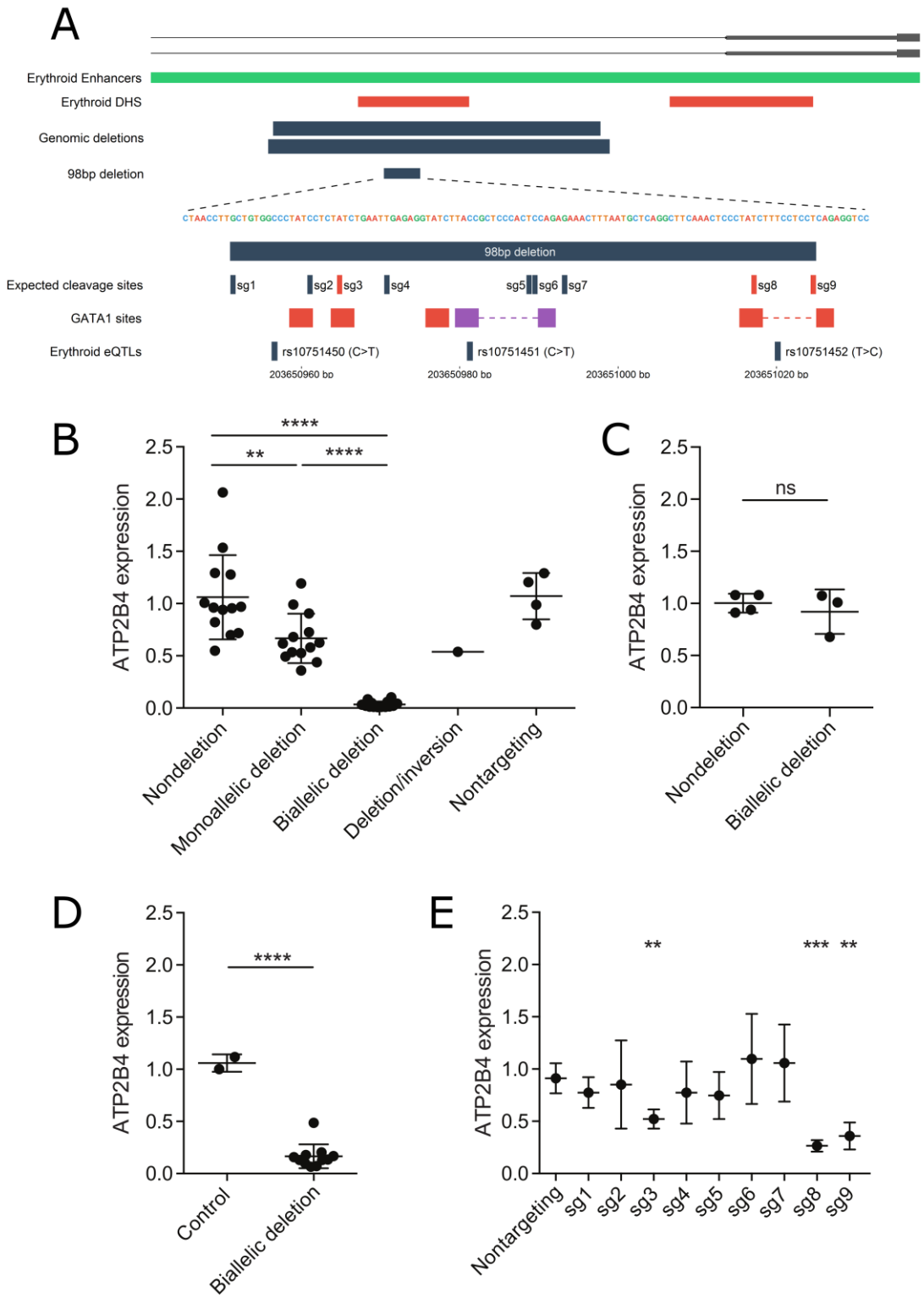


Figure 3. Genome editing demonstrates essential sequences at the trait-associated *ATP2B4* erythroid enhancer.

(**A**) Design of sgRNAs targeting intron 1 of *ATP2B4.* Upper panel shows deletions engineered via CRISPR-Cas9 editing flanking the *ATP2B4* regulatory region. Lower panel shows sgRNA cleavage sites inside the 98 bp core

enhancer. This element overlaps four GATA1 binding motifs and three erythroid eQTLs. The alternative allele of rs10751451 creates a *de novo* GATA1 motif (purple box). Boxes separated by dashed lines correspond to GATA1-TAL1 (half E-box) motifs. (**B**) *ATP2B4* gene expression as measured by RT-qPCR in HUDEP-2 cells with enhancer deletions. Combined analysis of two sgRNA pairs that result in overlapping 927-bp and 889-bp deletions. Non-deletion (n=13) refers to clones exposed to enhancer targeting sgRNA pairs but without a deletion allele, monoallelic deletion (n=13) clones with one allele deleted and the other non-deleted, and biallelic deletion (n=14) clones with both alleles deleted. None of the above clones had an inversion allele detected. One clone was identified with one deletion allele and one inversion allele but no non-deletion allele. Four clones exposed to non-targeting sgRNAs are shown for comparison. Gene expression is normalized to mean of non-deletion clones for the same sgRNA pair. Bars and whiskers show mean and standard deviations. (**C**) *ATP2B4* gene expression in 293T cell clones exposed to enhancer targeting sgRNA pairs but without deletion (n=4) or with biallelic deletion (n=3). ns, non-significant. (**D**) *ATP2B4* gene expression in HUDEP-2 cells with 98-bp core enhancer deletion. Control (n=2) refers to one non-targeting control clone and unedited parental cells as compared to biallelic deletion clones (n=11). (**E**) *ATP2B4* gene expression in HUDEP-2 cells with individual sgRNAs specifying cleavages within the 98-bp core enhancer. Mean and standard deviation are plotted for each of four biological replicates. Gene expression is normalized to unedited cells. All statistical tests performed used Student's *t*-test and two-tailed P-values are reported. **, $P<0.01$; ***, $P<0.001$; ****, $P<0.0001$.

We identified core sequences of the element that included three SNPs (rs10751450, rs10751451, rs10751452) associated with *ATP2B4* expression by eQTL analysis and with RBC traits and malaria susceptibility by GWAS. These core sequences included five GATA1 or composite half-E-box/GATA1 binding motifs, binding sites for the transcription factors GATA1 and TAL1 (**Figure 3A**). We introduced into HUDEP-2 cells a pair of guide RNAs to generate a 98-bp deletion that removed the three SNP positions and these GATA1 motifs (**Figure 3A** and **Table S3**). We observed that clones with bi-allelic deletion of this 98-bp segment had 83% reduction in expression of *ATP2B4* when compared to *ATP2B4* expression in wild-type cells (**Figure 3D**).

To fine-map the gene expression regulatory element, we introduced nine individual guide RNAs to produce small insertions-deletions (indels) at their cleavage sites within the 98-bp enhancer core (**Figure 3A** and **Table S3**). We observed a significant reduction of *ATP2B4* expression with three of the nine guides (**Figure 3E**). These significantly disrupting guide RNAs clustered over two GATA1 binding sites. Overall these results demonstrate a hierarchal requirement for trait-associated sequences at the erythroid-specific enhancer of *ATP2B4*.

Finally, given ATP2B4's role in RBC calcium homeostasis, we measured intracellular calcium concentration by ratiometric imaging in unedited HUDEP-2 cells, as well as cells with a deletion of the *ATP2B4* enhancer element. At baseline or upon stimulation (**Methods**), we found higher intracellular calcium levels in *ATP2B4*-edited HUDEP-2 cells, indicating that cells that do not express *ATP2B4* cannot pump efficiently calcium outside of the cytoplasm (**Figure 4A-B**). In response to endoplasmic reticulum calcium release by caffeine stimulation, *ATP2B4* deficient cells demonstrate exaggerated cytoplasmic calcium accumulation and persistence (**Figure 4C**). These results provides a physiological link between common regulatory SNPs at *ATP2B4*, a gene that encodes a major calcium pump, an ion homeostatic defect in erythroid cells, and human complex phenotypes such as RBC hydration and susceptibility to severe malaria infection.

5.5 Discussion

Few GWAS discoveries have been investigated at the molecular and cellular levels. To explore the genetic architecture of regulatory variants that control RBC traits in humans, we undertook an eQTL search in human in vitro differentiated erythroblasts. Although we identified >4,500 eQTLs that replicated in the GTEx dataset, we acknowledge that some of our findings might be false positive associations; independent replication studies in the same cell type are needed. Furthermore, because we limited our eQTL analyses to genes with AI and variants located within 100 kb of these genes in order to increase statistical power, we would have missed genes without exonic variants (necessary to monitor AI) or that are controlled by long-range regulatory variants. Despite these limitations, we found strong eQTLs for *ATP2B4*, validating our experimental design. Our own functional results and the recent report that the same SNPs are associated with ATP2B4 protein levels in human RBCs[401] strongly argue that this is a true association signal.

Using human erythroblasts, knockout mice, and erythroid cells amenable to genome editing, we undertook the detailed characterization of the *ATP2B4* locus and its roles in RBC biology. This comprehensive approach allowed us (1) to identify the causal regulatory variants within an erythroid-specific enhancer, (2) to confirm *ATP2B4* as the causal gene, and (3) to

highlight calcium homeostasis defect as one possible effector pathway responsible for the association with MCHC and malaria susceptibility (**Figure 4D**). Excess intracellular calcium activates a calcium-activated potassium channel (the Gardos channel), resulting in potassium efflux, RBC volume loss, and elevated MCHC. RBC hydration has been linked with clinical severity in the hemoglobin disorder sickle cell disease[123] and with infectivity by the malaria agent *Plasmodium falciparum*[402]. Supported by our genetic and mechanistic results, the development of therapies that specifically modulate *ATP2B4* activity could have a broad impact on RBC diseases that affect millions of individuals worldwide.

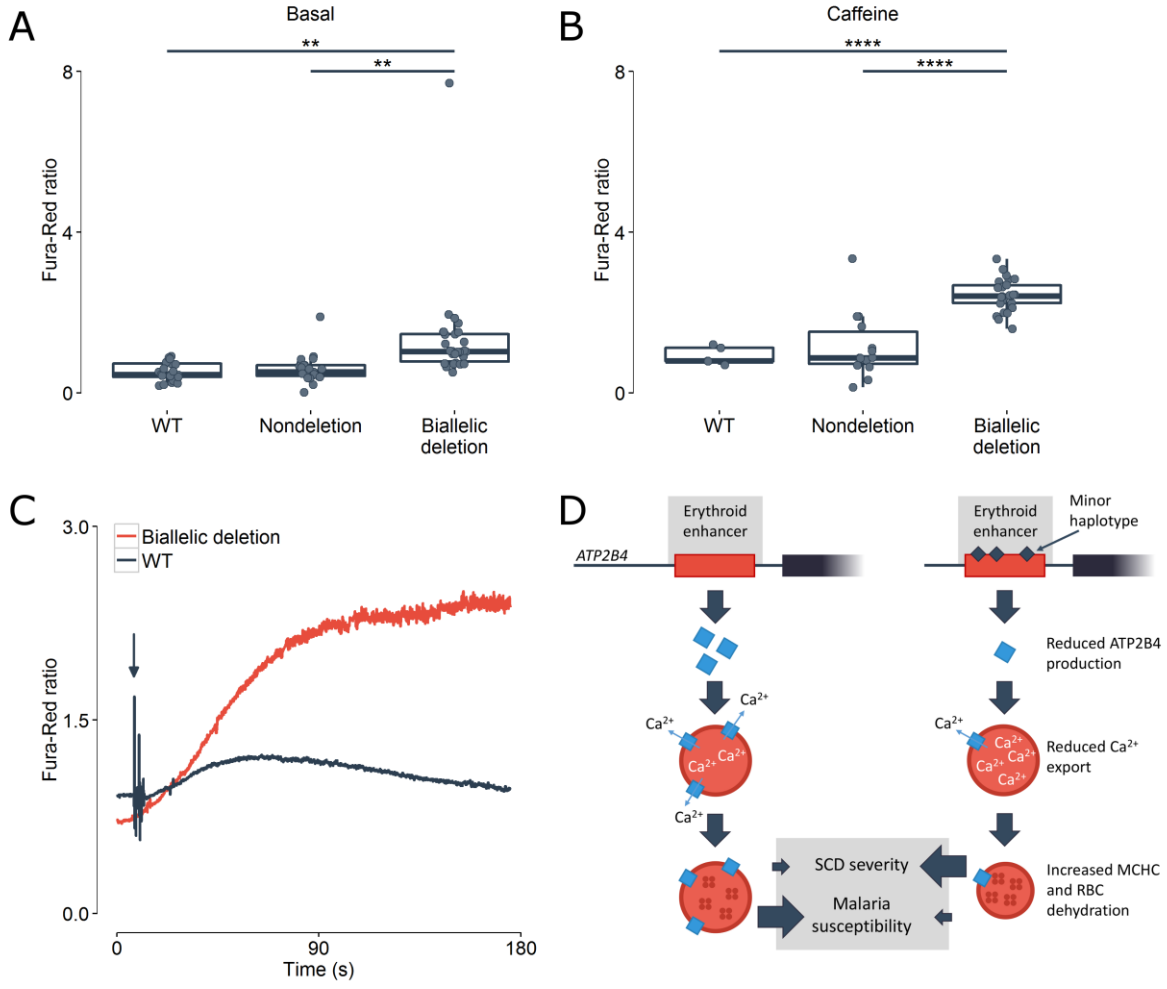


Figure 4. ATP2B4 activity and calcium homeostasis in erythroid cells.
Ratiometric Fura-Red fluorescence (**A**) prior and (**B**) following application of caffeine (10 μM) in HUDEP-2 wild-type (WT), non-deleted (nondeletion), and *ATP2B4* enhancer-deleted cells (biallelic deletion). Both basal levels and caffeine-induced release of calcium are significantly higher in *ATP2B4* enhancer-deleted cells. Basal (n): WT (21), nondeletion (25), and biallelic deletion (30); Caffeine (n): WT (5), nondeletion (14), and biallelic deletion

(27). Statistical analyses using Student's *t*-test. Two-tailed *P*-values are reported. **, $P<0.01$; ****, $P<0.0001$. (**C**) Representative time-course of the fluorescence Fura-Red ratio following exposure to a single bolus of caffeine (black arrow) in a control cell (black, WT) and in a cell with biallelic deletion of the *ATP2B4* enhancer (red; biallelic deletion). Intracellular calcium store release induced by caffeine exposure substantially increased cytoplasmic levels in both cells but elevated calcium concentrations persisted only in cells with the *ATP2B4* enhancer deleted. (**D**) Model of genetic variation at the *ATP2B4* erythroid enhancer influencing red blood cell (RBC) traits and malaria susceptibility. Erythroid cells carrying the minor haplotype at the erythroid-specific enhancer express lower levels of *ATP2B4*. RBCs with reduced *ATP2B4* accumulate cytoplasmic calcium, resulting in dehydration and elevated mean corpuscular hemoglobin concentration (MCHC). Although relatively resistant to infection by the malaria parasite *Plasmodium falciparum*, dehydrated RBCs may increase the severity of erythroid disorders such as sickle cell disease (SCD).

5.6 Methods

5.6.1 Cell culture, RNA-sequencing and DNA genotyping

The cell culture protocol to proliferate and differentiate human CD34+ hematopoietic stem/progenitor cells (HSPCs) into erythroblasts has been described before[58,377]. We purchased human fetal (fetal liver, n=12) and adult (bone marrow, n=12) $CD34^+$ HSCs from DV Biologics and Lonza, respectively. This sample size was selected as it provides 90% power to detect a 3 standard deviations difference in gene expression levels between fetal and adult erythroblasts at $\alpha=1x10^{-5}$. We have also described elsewhere the protocol for RNA extraction and RNA-sequencing[377]. Briefly, we performed RNA-sequencing with an Illumina HiSeq2000 sequencer using stranded cDNA library and a paired-ends 100 bp protocol. We mapped reads to the genome (hg19) using Tophat2 (v.2.0.9, with options *--library-type fr-firststrand --microexon-search --coverage-search*) and estimated transcript abundance with Cufflinks (v.2.2.1, with options *--library-type fr-firststrand --max-bundle-frags 50000000*)[233,235]. Our RNAseq human erythroblast expression dataset is publicly available at NCBI GEO under accession number GSE90878. Genomic DNA extraction, genotyping on the Illumina HumanOmniExpress-12 v1.1 BeadChip array, and quality-control were performed as previously described[377]. We imputed genotypes using the Michigan Imputation Server with the Haplotype Reference Consortium (HRC) panel (v. r1.1)[25].

5.6.2 Allelic imbalance and eQTL mapping

We measured allelic imbalance (AI) at each heterozygous genotype covered by RNA-sequencing in the 24 human erythroblast samples. We only considered SNPs directly genotyped or with high imputation quality (R^2>0.6). We removed duplicated reads using Picard's MarkDuplicates tool (v. 1.96). We counted each read using the samtools (v 1.1) mpileup software and genome build hg19, and kept uniquely mapping reads using the -q 50 argument (mapping quality > 50) and sites with base quality >10. We further restricted the analysis to uniquely mapping sites as per the ENCODE 50-mer mappability track (score=1) and removed sites showing mapping bias in simulations[403]. We excluded sites with less than 30 overlapping reads. For a given heterozygous SNP, we determined the statistical significance of AI, that is the difference between the observed and expected ratio of "reference allele:alternate allele", with a binomial test. To account for read mapping bias, we summed all reads overlapping all heterozygous SNPs in the RNA-sequencing dataset and calculated the expected ratio for each combination of alleles in each sample independently. For SNPs with high sequencing coverage, we down-sampled the number of reads that fell in the top 25th coverage percentile so that the most covered sites do not bias the expected ratio[257]. We used a Bonferonni correction to account for the number of tests performed: the significance threshold for this AI experiment is $\alpha=2x10^{-5}$.

Next, we mapped the regulatory variants responsible for differential gene expression phenotypes. Given our limited sample size, we focused on genes that showed AI, reasoning that the likelihood to find significant expression quantitative trait loci (eQTL) was higher in this subset of genes. We developed a method that combines statistical evidence of AI and eQTL effects. First, we tested by linear regression the association between SNP genotypes (additive model) and gene expression levels (expressed as $\log_{10}$(FPKM+1)), adjusting for cell type (fetal or adult). For these analyses, we only considered SNPs located within 100 kb of the AI genes. Second, we hypothesized that samples that are homozygous for the tested SNP (either reference/reference or alternate/alternate) should not show AI, whereas heterozygote samples should have AI. In other words, the "reference allele:alternate allele" ratio in heterozygote samples should be further from the expected 50:50 ratio than the ratio observed in homozygote

samples. We tested this hypothesis using a one-sided *t*-test. Because the linear regression and *t*-test *P*-values were not correlated, we meta-analyzed these statistics using Fisher's method to obtain a final *P*-value. We used a false discovery rate (FDR) methodology to correct for multiple testing, considering SNPs with a *q*-value <0.05 as significant eQTLs.

5.6.3 Replication of eQTLs in GTEx

We used the GTEx database to replicate the eQTLs that we identified in human erythroblasts[396]. GTEx does not include erythroblasts, but we reasoned that it would still represent a valid source of replication for non-tissue-specific eQTLs. We downloaded from the GTEx portal (version 6) all SNP-gene association results across all available tissues (http://www.gtexportal.org/home/)[396]. We considered as replicated erythroid eQTLs with a *P*<0.001 in any other samples from the GTEx dataset. More stringent thresholds gave consistent results.

5.6.4 eQTL enrichment analyses

We used three sources of information to test the enrichment of erythroid eQTL within specific genomic annotations. First, we obtained the coordinates of erythroid-specific enhancers defined using DNAse hypersensitive sites (DHSs) and histone tail modifications[334]. From the same study, we also obtained genomic coordinates of GATA1 and TAL1 peaks determined by ChIP-seq[334]. Finally, we used the Homer software to identify binding motifs for GATA1 (MA0035.2) and GATA1::TAL1 (co-occurring GATA1 and half-Ebox motifs, MA0140.2) across the human genome, or specifically within erythroid enhancer regions[337,404]. We carried out gene ontology and pathway enrichment analyses using the ToppGene suite (toppgene.cchmc.org)[405].

5.6.5 Red blood cell traits analyses in *Atp2b4*$^{-/-}$ mice

Mice (male-only, n=26) with complete inactivation of *Atp2b4* have been generated and described elsewhere[399,400]. The mice are in mix background 129/sv x C57Bl/6. All mice used were male age between 9-13 weeks. Mice were anaesthetized with isoflurane (2.5%) and blood was collected from the jugular vein by venipuncture. The procedure was performed in accordance with the United Kingdom Animals (Scientific Procedures) Act 1986 and was approved by the University of Manchester Ethics Committee. The samples were measured within 6 hours of collection (RT) in the biological research unit at CRUK Manchester Institute. Evaluation of hematological parameters was carried out in two batches on Sysmex XT-2000iV (Sysmex, Kobe, Japan) automated hematology analyzer using mouse profile. Quality control was carried before running each batch of samples. No randomization was used and experimenters who did the complete blood count analyses were blinded from the animals' *Atp2b4* genotypes.

5.6.6 Replication of the association between *ATP2B4* and RBC phenotypes in the UK Biobank

We tested the association between genotypes at the *ATP2B4* locus (2 Mb) and RBC traits in the July 2015 release of the UK Biobank dataset. We excluded participants with blood cancer, leukemia, lymphoma, bone marrow transplant, congenital or hereditary anemia, HIV, end-stage kidney disease, dialysis, splenectomy, or cirrhosis and those with extreme RBC trait measurements (> 8 standard deviations from the mean). We limited our analysis to participants of British ancestry with imputed genotype data available. In total, we tested the association between eight RBC traits (hemoglobin, hematocrit, RBC count, mean corpuscular volume, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, RBC distribution width, and reticulocyte count) and genotypes (additive model) in 136,727 participants with PLINK1.9 (https://www.cog-genomics.org/plink2). This sample size provides >99% power to replicate the association between *ATP2B4* SNPs and MCHC at α=0.05. After applying exclusion criteria, we corrected the RBC traits for age, sex, recruitment center, and cell counter, and then normalized the residuals using inverse normal transformation. As covariates, we included in the association

tests the ten first principal components calculated using FlashPCA[406]. Genetic analyses were approved by the Montreal Heart Institute Ethics Committee and informed consent was obtained from all participants.

5.6.7 Generating *ATP2B4* deletions in cell lines

HUDEP-2 cells and 293T cells with stable expression of Cas9 were generated by lentiviral transduction (lentiCas9-Blast, Addgene plasmid ID 52962) and blasticidin selection as previously described[59]. Tandem sgRNA lentiviruses were produced based on a modification of lentiGuide-Puro (Addgene plasmid ID 52963) to carry two U6 promoter-guide RNA cassettes per construct as previously described[59]. Tandem sgRNA lentiviruses were transduced into HUDEP-2 cells or 293T cells with stable Cas9 expression. The tandem sgRNA constructs expressed two sgRNAs (**Table S3**) and thus were designed to produce interstitial deletions. Cells were also transduced with a pool of lentiviruses containing ten unique non-targeting sequences (**Table S3**). After transduction, bulk cultures were incubated for 7-10 days with 10 μg mL^{-1} blasticidin and 1 μg mL^{-1} puromycin selection to select for cells with edited alleles. Those bulk cultures transduced with tandem sgRNA lentiviruses were plated clonally at limiting dilution. 96 well plates with greater than 30 clones per plate were excluded to avoid mixed clones. After approximately 14 days of clonal expansion, genomic DNA was extracted using 50 μL QuickExtract DNA Extraction Solution per well (Epicentre). Clones were screened for deletion by conventional PCR, with one PCR reaction using primers internal to a segment to be deleted (non-deletion amplicon) and one gap-PCR reaction using primers across the deletion junction (deletion amplicon) that would produce a characteristic short amplicon in the presence of deletion (**Table S4**). Clones bearing inversion alleles were also identified with one primer outside the segment to be deleted and the other primer inside the segment to be deleted, both in the same orientation with respect to the reference genome, as previously described. PCR was performed using the Qiagen HotStarTaq 2x master mix and the following cycling conditions: 95°C for 15 min; 45 cycles of 95C for 30 s, 60°C for 45 s, 72°C for 1 min, 72°C for 10 min. Biallelic deletion clones were identified based on the presence of a deletion PCR band with absence of a non-deletion PCR band. Inversion clones were also identified as previously

described. Compound deletion-inversion clones had one deleted allele and one inverted allele without the presence of non-deletion alleles. For disruption of individual GATA1 motifs, stable Cas9 expressing HUDEP-2 cells were transduced with lentiviruses carrying individual guide RNAs (lentiGuide-Puro) (**Figure 3A** and **Table S4**). Edited populations of cells were selected with puromycin and blasticidin and RNA was isolated 7-10 days following transduction. Genome editing with indel rates exceeding 75% was confirmed by isolating gDNA from each of these bulk populations of cells, performing a PCR reaction with primers flanking the edited region (**Table S4**), and Sanger sequencing the amplicon with analysis according to a publically available sequence deconvolution algorithm[407].

5.6.8 Reverse transcription-quantitative PCR

RNA was extracted for each selected clone using a kit (Qiagen). 1 ug of RNA per clone was converted to cDNA using the iscript cDNA kit. Quantitative real-time RT-qPCR was subsequently performed using SYBR Select Master Mix (Life). Primers were designed to span exon 5 and exon 6 of the *ATP2B4* gene and were empirically validated for efficiency by serial dilution analysis (**Table S4**). Gene expression was normalized to that of *GAPDH*. All gene expression data reported represents the mean of at least three technical replicates.

5.6.9 Intracellular calcium monitoring

Intracellular calcium levels in HUDEP-2 cell lines were monitored using Fura-Red, a ratiometric fluorescent calcium indicator with a laser-scanning confocal microscope. Cells were first seeded (1 hr) in a coverslip-bottom chamber coated with Cell-Tak (Corning). Cells were then washed with HEPES-PSS and incubated with Fura-Red (10 μM) for 45 min at 37℃. Intracellular calcium levels were recorded on a LSM-Duo confocal microscope (Zeiss) with a 40x objective (Plan APO Oil DIC, 1.3 NA). Single emission fluorescence (LP575) was collected (10 FPS) upon alternate excitation (405 and 489 nm, solid state lasers) on a 512x256 field of view. Stack of interleaved images (16 bits) were then analyzed using FIJI (ImageJ). Upon background subtraction, 20x20 pixels circular regions of interest (ROIs) were manually

positioned on cells and individual ROI's mean fluorescence intensity was measured. Variations in intracellular calcium levels were expressed as the mean ratio of the fluorescence from calcium-bound (excitation wavelength 405 nm) and calcium-free (excitation wavelength 489 nm) Fura-Red of each ROI from 10 consecutives images.

5.7 Acknowledgments

Part of this research has been conducted using the UK Biobank resource under application number 11707. The authors wish to thank Gerald Lilly and Louis R. Villeneuve for assistance, and Seth Alper and Carlo Brugnara for comments. S.L. holds fellowships from the Canadian Institute of Health Research (CIHR) and the "Fondation Pierre Lavoie". E.N.S. was the recipient of an American Society of Hematology HONORS award. S.A.A. holds a Cancer Research UK Clinical Training Award. E.B. is supported by Cancer Research UK start-up grant. D.O. was supported by British Heart Foundation (BHF) Intermediate Fellowship (FS/09/046/28043) and BHF Project Grants (PG/13/12/30017 and PG/16/77/32400). D.E.B. was supported by NIDDK (K08DK093705, R03DK109232), NHLBI (DP2OD022716), Burroughs Wellcome Fund, American Society of Hematology, and the Doris Duke Charitable, Charles H. Hood, and Cooley's Anemia Foundations. G.L. was supported by grants from the CIHR (#123382), the Doris Duke Charitable Foundation, the National Sciences and Engineering Research Council of Canada (RGPIN-2016-04597), the Montreal Heart Institute Foundation, and the Canada Research Chair Program.

5.8 Supplementary information

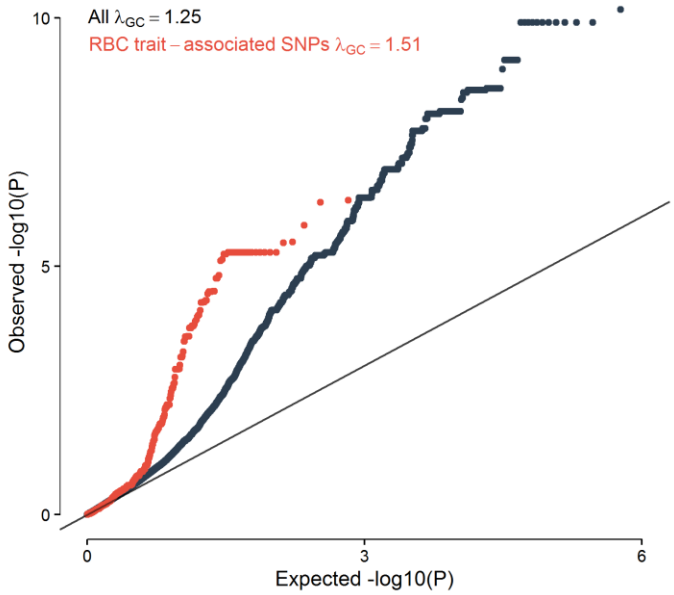


Figure S1. Erythroblasts eQTL are enriched for RBC trait-associated SNPs.
Quantile-quantile (QQ) plot of eQTL *P*-values for variants located within 100 kb of 479 AI genes in human erythroblasts (black, genomic control lambda (λ_{GC})=1.25). We find an enrichment of eQTL *P*-values for SNPs associated with RBC traits (red, λ_{GC}=1.51)[13]. The number of different SNPs included when considering all tested variants and when subsetting on RBC trait-associated SNPs is 258,663 and 178 respectively.

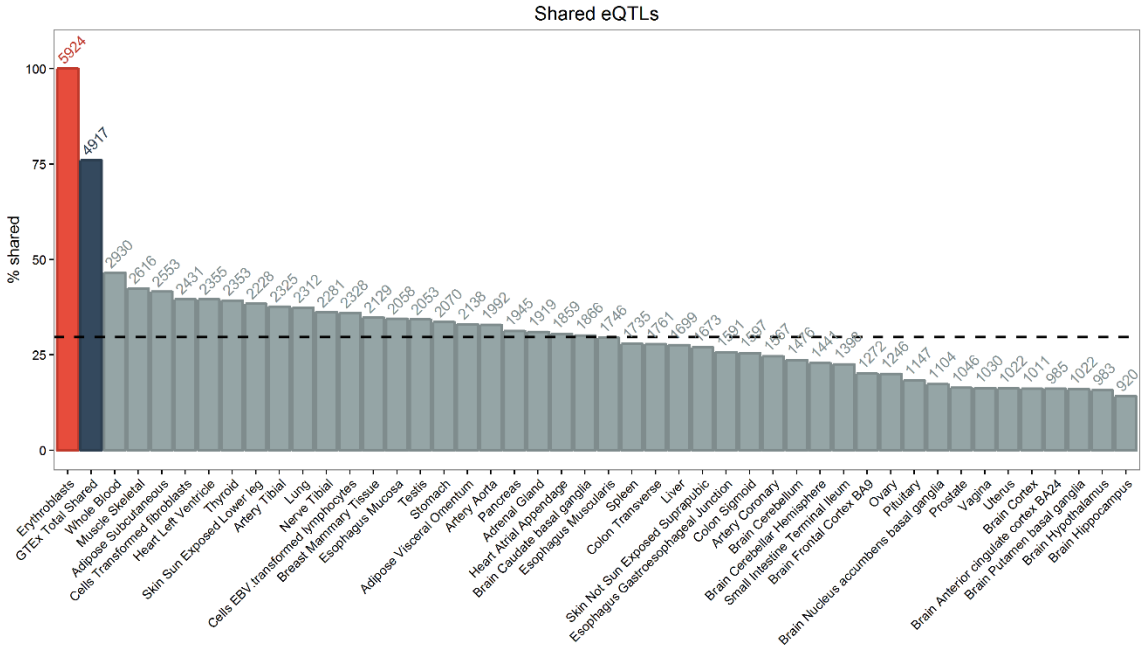


Figure S2. Number of shared eQTLs between erythroblasts and GTEx samples.
The dotted line represents the mean percentage of eQTLs shared between human erythroblasts and any of the GTEx tissues (mean=29.6%).

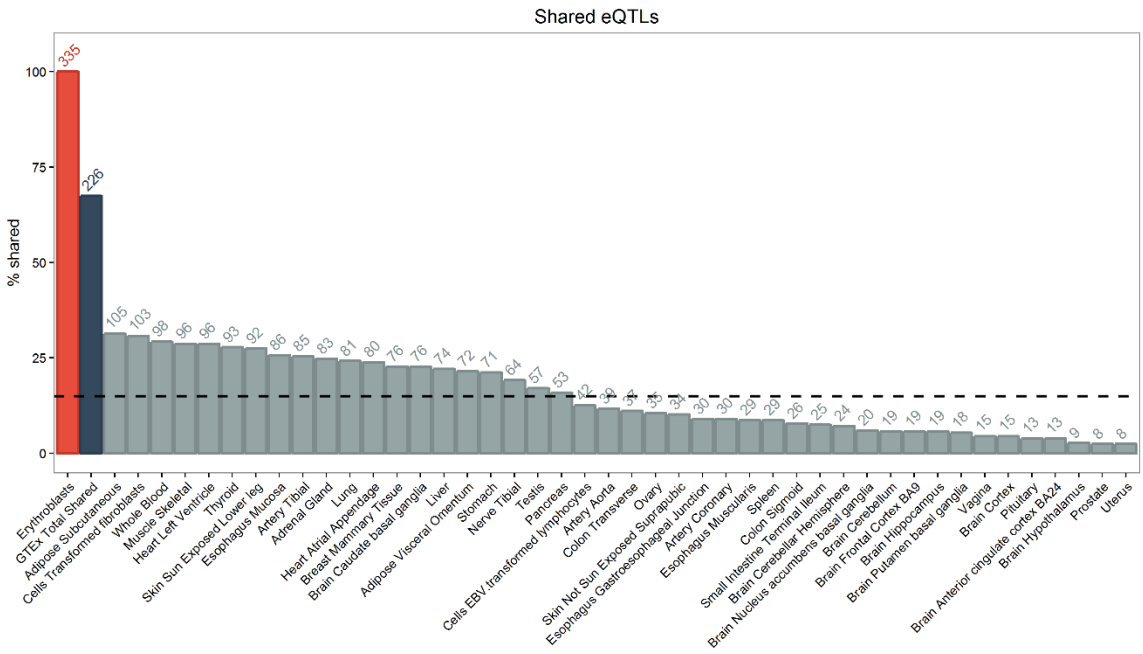


Figure S3. Number of eQTLs in GATA1/TAL1 ChIP-seq peaks shared between erythroblasts and GTEx samples.

The dotted line represents the mean percentage of shared eQTLs that map to GATA1/TAL1 ChIP-seq peaks between human erythroblasts and any of the GTEx tissues (mean=14.9%).

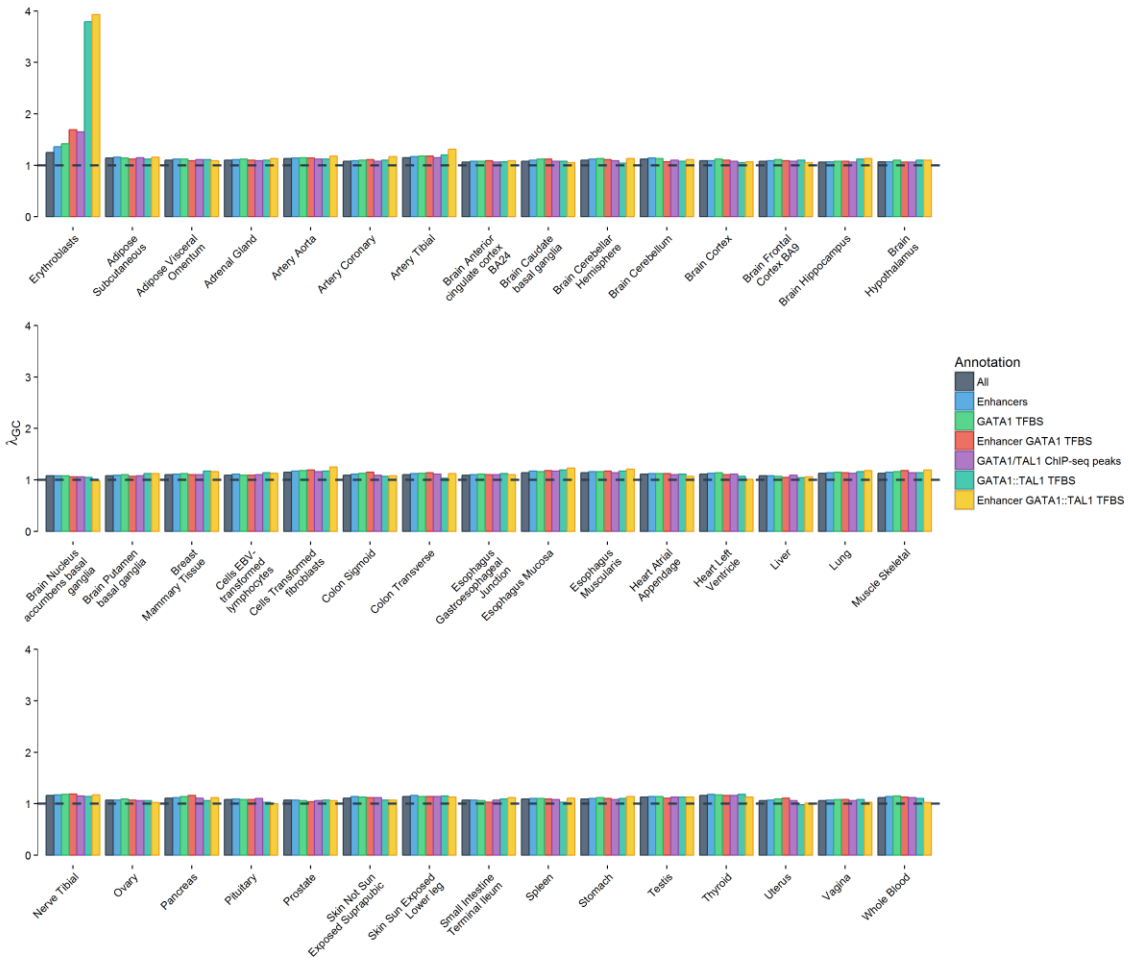


Figure S4. Genomic control lambda (λ_{GC}) of eQTLs subsetted by erythroid regulatory regions.

In this analysis, we compared inflation of eQTL signals observed in human erythroblasts (this study) with the inflation of the same SNPs but in different tissues available in the GTEx resource. We note a marked inflation of eQTL signals for erythroid annotations only in human erythroblasts. Dotted line is λ_{GC}=1.

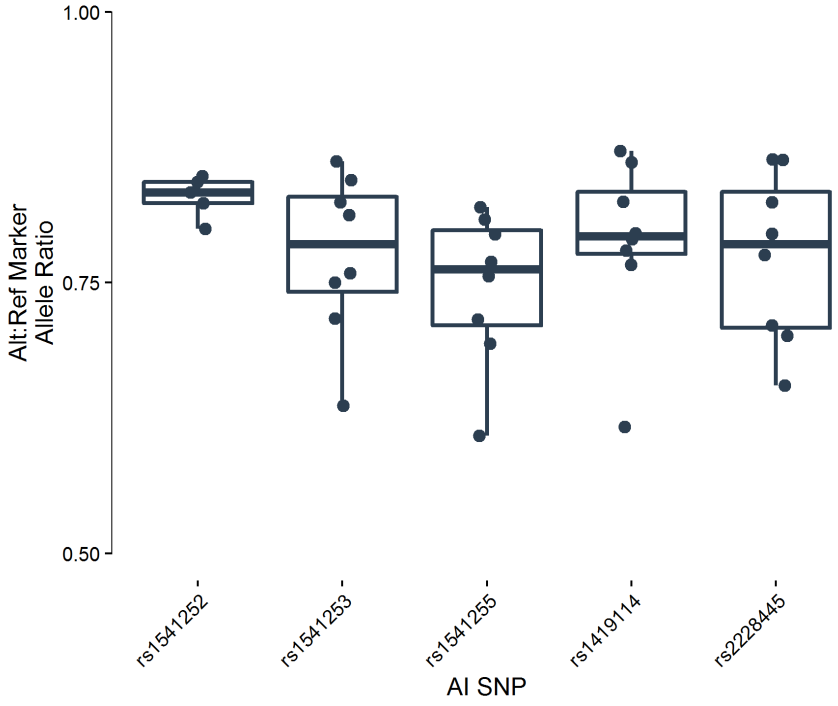


Figure S5. ***ATP2B4*** **allelic imbalance in human erythroblasts.**
Ratio of reads of the alternate allele over the total number of reads overlapping *ATP2B4* exonic SNPs. In the absence of allelic imbalance and read mapping biases, the expected ratio is 0.5. All heterozygous samples at these different exonic SNPs are also heterozygotes for the top *ATP2B4* candidate eQTLs.

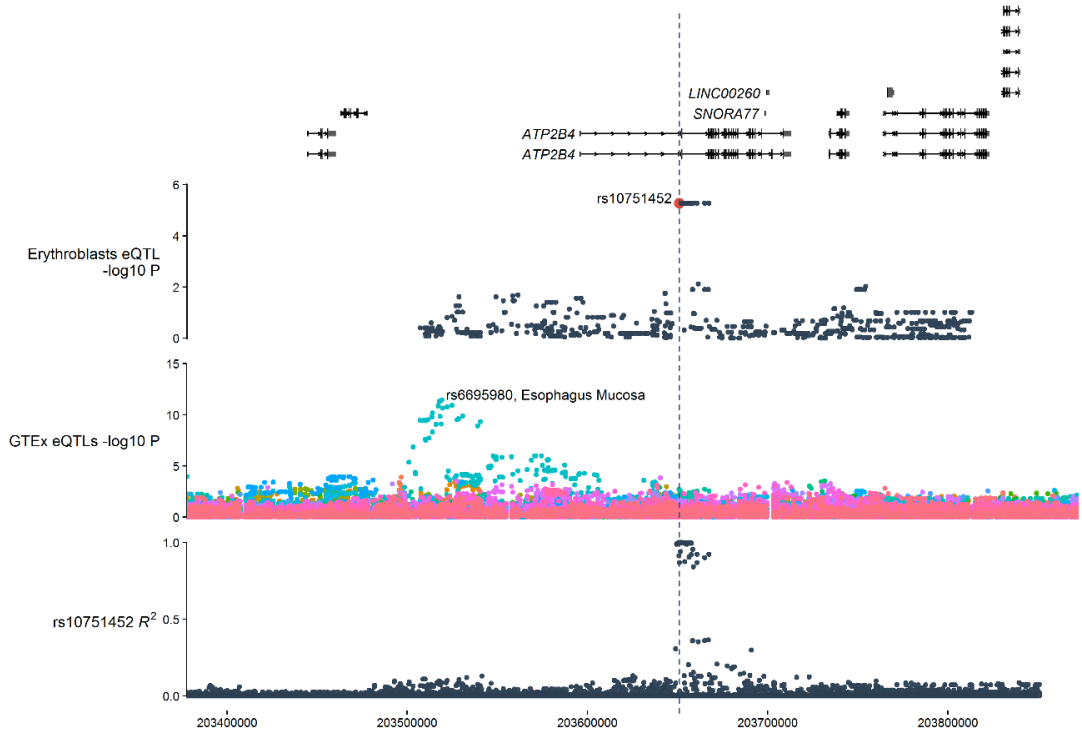


Figure S6. Erythroblasts show an eQTL signal that is independent from *ATP2B4* eQTLs found in GTEx.

The bottom track shows pairwise linkage disequilibrium (LD, r^2) between rs1075152 and surrounding SNPs. There is essentially no LD between the *ATP2B4* erythroblast eQTL (rs1075152) and eQTLs found in esophagus mucosa from the GTEx resource.

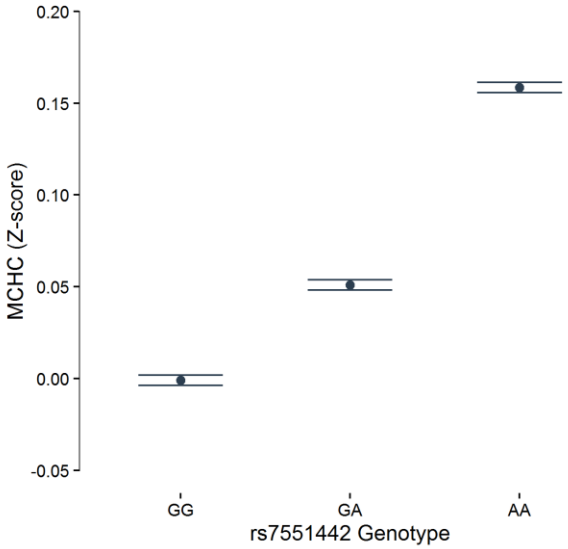


Figure S7. Association of rs7551442 with mean corpuscular hemoglobin concentration (MCHC) in the UK Biobank.

Dots represent the mean normalized MCHC by genotype, while the error-bars represent the standard errors.

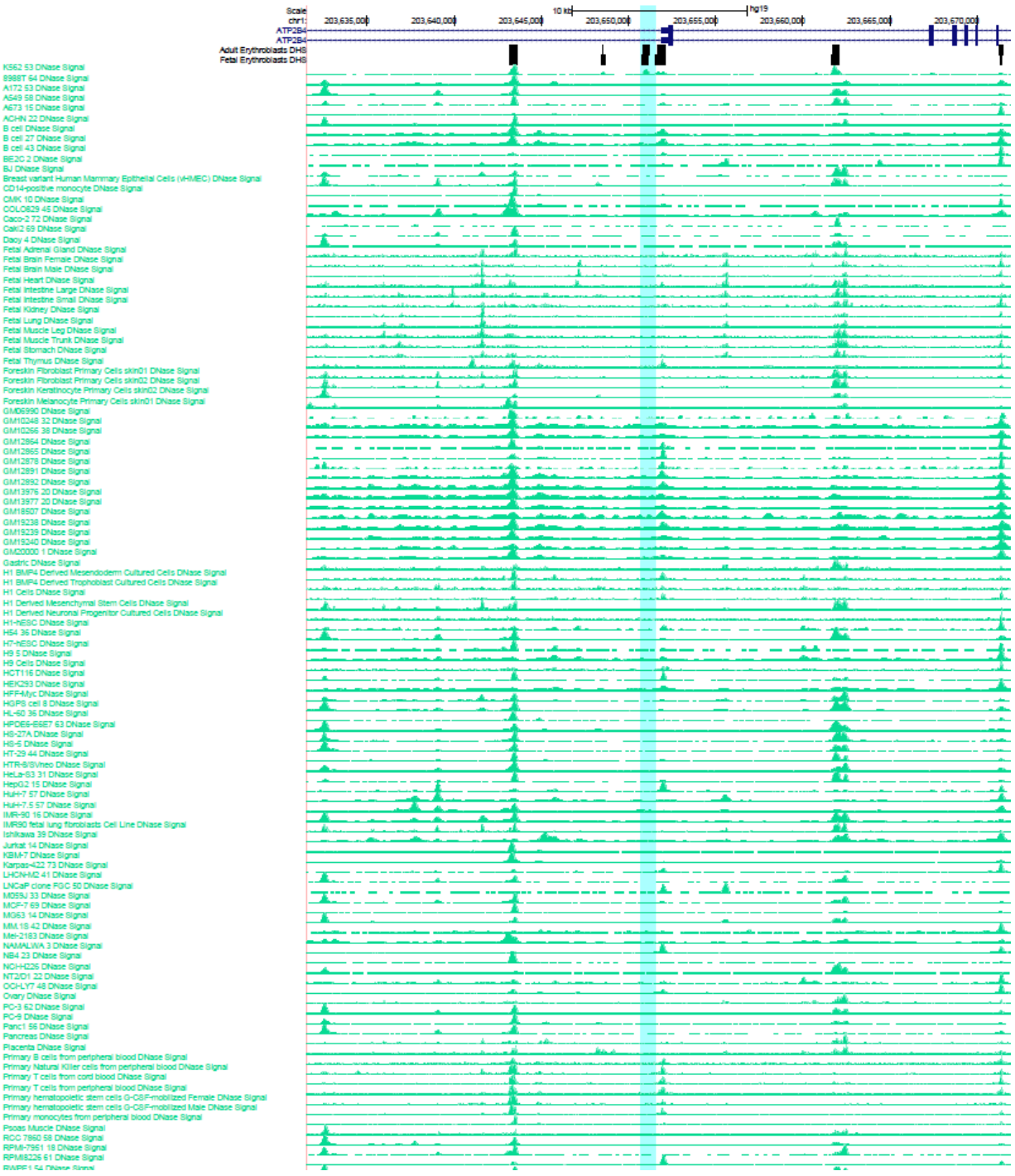


Figure S8. Example of ENCODE and Roadmap Epigenomic Consortia DNAse I hypersensitive sites (DHSs) signals at the *ATP2B4* locus.

Table S1. Top gene ontology terms and mouse phenotypes significantly enriched among erythroblast eGenes.

ID	Name	*P*-value	FDR B&H	Genes from Input	Genes in Annotation
GO: Molecular Function					
GO:0046977	TAP binding	2.18E-05	1.92E-02	3	7
GO:0016404	15-hydroxyprostaglandin dehydrogenase (NAD+) activity	7.52E-05	2.80E-02	2	2
GO:0005344	oxygen transporter activity	2.17E-04	2.80E-02	3	14
GO:0002060	purine nucleobase binding	2.22E-04	2.80E-02	2	3
GO:0016822	hydrolase activity, acting on acid carbon-carbon bonds	2.22E-04	2.80E-02	2	3
GO:0004351	glutamate decarboxylase activity	2.22E-04	2.80E-02	2	3
GO:0016823	hydrolase activity, acting on acid carbon-carbon bonds, in ketonic substances	2.22E-04	2.80E-02	2	3
GO: Biological Processes					
GO:0002486	antigen processing and presentation of endogenous peptide antigen via MHC class I via ER pathway, TAP-independent	6.59E-07	2.42E-03	3	3
GO:0006779	porphyrin-containing compound biosynthetic process	4.78E-06	4.93E-03	5	29
GO:0019885	antigen processing and presentation of endogenous peptide antigen via MHC class I	5.24E-06	4.93E-03	4	14
GO:0033014	tetrapyrrole biosynthetic process	7.94E-06	4.93E-03	5	32
GO:0006335	DNA replication-dependent nucleosome assembly	9.29E-06	4.93E-03	5	33
GO:0034723	DNA replication-dependent nucleosome organization	9.29E-06	4.93E-03	5	33
GO:0002483	antigen processing and presentation of endogenous peptide antigen	9.39E-06	4.93E-03	4	16
GO:0019883	antigen processing and presentation of endogenous antigen	1.96E-05	9.00E-03	4	19
GO:0001916	positive regulation of T cell mediated cytotoxicity	3.62E-05	1.29E-02	4	22
GO:0048534	hematopoietic or lymphoid organ development	4.19E-05	1.29E-02	21	905
GO:0071103	DNA conformation change	4.27E-05	1.29E-02	11	285
GO:0006778	porphyrin-containing compound metabolic process	4.90E-05	1.29E-02	5	46
GO:1901566	organonitrogen compound biosynthetic process	5.27E-05	1.29E-02	28	1445
GO:0002484	antigen processing and presentation of endogenous peptide antigen via MHC class I via ER pathway	5.28E-05	1.29E-02	3	9
GO:0002480	antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-independent	5.28E-05	1.29E-02	3	9
GO:0030097	hematopoiesis	6.08E-05	1.40E-02	20	858
Mouse Phenotypes					
MP:0008956	decreased cellular hemoglobin content	8.88E-07	1.22E-03	4	8
MP:0011177	abnormal erythroblast number	1.48E-06	1.22E-03	6	34
MP:0011188	increased erythrocyte protoporphyrin level	1.58E-06	1.22E-03	4	9
MP:0008954	abnormal cellular hemoglobin content	2.62E-06	1.51E-03	4	10
MP:0011176	abnormal erythroblast morphology	4.02E-06	1.85E-03	6	40
MP:0000611	jaundice	4.93E-06	1.90E-03	5	24
MP:0010375	increased kidney iron level	6.07E-06	2.00E-03	4	12

MP:0002642	anisocytosis	8.16E-06	2.23E-03	6	45
MP:0008742	abnormal kidney iron level	8.70E-06	2.23E-03	4	13
MP:0002812	spherocytosis	1.63E-05	3.14E-03	4	15
MP:0004147	increased porphyrin level	1.63E-05	3.14E-03	4	15
MP:0011989	abnormal porphyrin level	1.63E-05	3.14E-03	4	15
MP:0003657	abnormal erythrocyte osmotic lysis	2.18E-05	3.87E-03	5	32
MP:0010163	hemolysis	2.96E-05	4.88E-03	5	34
MP:0010067	increased red blood cell distribution width	4.14E-05	5.73E-03	7	87

Table S2. Comparison of the association of the rs7551442 A-allele with red blood cell (RBC) traits in the first release of the UK Biobank and effect of the Atp2b4 targeted deletion on RBC phenotypes in mice.

	UK Biobank			*Atp2b4*$^{-/-}$ mice		
Trait	**N**	**Beta (SE)**	***P***	**N**	**Beta (SE)**	***P***
HBG	128,311	0.034 (0.0066)	2.1×10^{-7}	26	0.0557 (0.098)	0.58
HCT	128,311	0.011 (0.0066)	0.098	26	-0.009 (0.003)	2.6×10^{-3}
MCH	128,253	0.012 (0.0065)	0.062	26	-0.095 (0.106)	0.38
MCHC	128,224	0.058 (0.0065)	2.6×10^{-19}	26	0.893 (0.267)	2.7×10^{-3}
RBC	128,312	0.021 (0.0065)	1.4×10^{-3}	26	0.099 (0.056)	0.088
RETIC	128,348	0.022 (0.0066)	0.022	26	-0.295 (0.130)	0.033
RDW	128,231	-0.064 (0.0065)	1.2×10^{-22}	26	0.409 (0.173)	0.026
MCV	128,311	-0.019 (0.0065)	2.8×10^{-3}	26	-1.420 (0.214)	7.1×10^{-7}

In the UK Biobank, effect sizes for the A-allele at rs7551442 (betas and standard errors (SE)) are in standard deviation units. In the mouse, effect sizes for the deletion allele are in the following units: hemoglobin (HGB, g/dL), hematocrit (HCT, ratio), mean corpuscular hemoglobin (MCH, pg), mean corpuscular hemoglobin concentration (MCHC, g/dL), RBC count (RBC, 10^9/L), reticulocyte count (RETIC, %), RBC distribution width (RDW, %), mean corpuscular volume (MCV, fL).

Table S3. Single guide RNA design.

Single-guide RNA Name	Single-guide RNA sequence	Genomic cleavage position (hg19)
ATP2B4 sgL1	GGACTACGTAACCGGGGCTG	chr1:203650624
ATP2B4 sgR1	TAAATGGTTACGACCTCTGA	chr1:203651551
ATP2B4 sgL2	TTAGCACAGCACCAGAACAT	chr1:203650637
ATP2B4 sgR2	CACAGGTCCCTCAAATAGAC	chr1:203651526
ATP2B4 sg1	CTCCCTTGCTAACCTTGCTG	chr1:203650938
ATP2B4 sg2	TCTCAATTCAGATAGAGGAT	chr1:203650951
ATP2B4 sg3	ATACCTCTCAATTCAGATAG	chr1:203650956
ATP2B4 sg4	TATCCTCTATCTGAATTGAG	chr1:203650964
ATP2B4 sg5	CATTAAAGTTTCTCTGGAGT	chr1:203650988
ATP2B4 sg6	AGCATTAAAGTTTCTCTGGAG	chr1:203650989
ATP2B4 sg7	CCTGAGCATTAAAGTTTCTC	chr1:203650994
ATP2B4 sg8	ACCTCTGAGGAGGGAAGATA	chr1:203651026
ATP2B4 sg9	GGACAGAAGGACCTCTGAGG	chr1:203651036
Non-targeting sg1	CACGGAGGCTAAGCGTCGCAA	
Non-targeting sg2	GCGCTTCCGCGGCCCGTTCAA	
Non-targeting sg3	GATCGTTTCCGCTTAACGGCG	
Non-targeting sg4	GTAGGCGCGCCGCTCTCTAC	
Non-targeting sg5	GCCATATCGGGGCGAGACATG	
Non-targeting sg6	GTACTAACGCCGCTCCTACAG	
Non-targeting sg7	GTGAGGATCATGTCGAGCGCC	
Non-targeting sg8	GGGCCCGCATAGGATATCGC	
Non-targeting sg9	GTAGACAACCGCGGAGAATGC	
Non-targeting sg10	GACGGGCGGCTATCGCTGACT	

Table S4. qPCR design.

Forward Primer	Reverse Primer	Description
ATP2B4_enhancer_out_F AGAGGATTGTCAGGAATCGGC	*ATP2B4_enhancer_out_R* TGACTCAAGAGAGGCCCGTT	Deletion amplicon detection
ATP2B4_enhancer_in_F CCCTAGTTAGCATGCGTGAGA	*ATP2B4_enhancer_in_R* CTTTGGCAGTTGGTGACGCA	Non-deletion (internal) amplicon detection
ATP2B4_ enhancer_out_F AGAGGATTGTCAGGAATCGGC	*ATP2B4_enhancer_in_R2* CAGCATTTCTCCCCTCAGAA	Non-inversion amplicon detection
ATP2B4_enhancer_out_F AGAGGATTGTCAGGAATCGGC	*ATP2B4_enhancer_in_F2* CTAACGGGCTTTGGAGGATTT	Inversion amplicon detection
ATP2B4_enh_core_F TTCTGAGGGGAGAAATGCTG	*ATP2B4_ enh_core_R* AAGAGGCTTTTGGGGAGAAG	Deletion/non-deletion amplicon detection
ATP2B4_enh_core_F2 CCCTAGTTAGCATGCGTGAGA	*ATP2B4_enh_core_R2* AAATCCTCCAAAGCCCGTTAG	Amplicon Sanger sequencing
ATP2B4_RT_F1 AAAGACCCCATGTTGCTCTC	*ATP2B4_RT_R1* CCCCTTCGTCATCCTCATTG	RT-qPCR

Chapter 6: Discussion

6.1 Implications

The studies presented in chapters 2 to 5 have several implications on studies of RBC traits. First, I attempted to characterize the effect of rare gene knockouts on anthropometric trait variation. We did not identified significant associations, suggesting the larger and more comprehensive genotyping datasets will be necessary to carry such analyses for RBC traits. I then characterized several DNA methylation and gene expression changes associated with erythroid developmental stages. These results underlined that erythroid enhancers and miRNA can be important contributors to erythroid stage specificity. I found that DNA methylation at the promoter of the γ-globin gene (*HBG2*) and an HbF-associated DNA variant predicted to disrupt a *BCL11A* binding site are independently associated with *HBG2* expression, suggesting that therapies aiming at decreasing *BCL11A* expression and demethylated DNA could have a cooperative effect on HbF reactivation. Moreover, these datasets provide important datasets for future studies of RBC traits, and highlight genes and miRNAs that may be implicated in HbF production such as *let-7* miRNAs and their targets *IGF2BP1*, *IGF2BP3*, and *LIN28B*. Finally, I characterized an erythroid-specific enhancer of *ATP2B4*, which encodes the main calcium ATPase of RBCs, using transcriptomic, epigenomic and genome editing approaches. Genetic variants in this enhancer modulate *ATP2B4* expression, which leads to increased intracellular calcium levels and increased MCHC. Ultimately, this leads to dehydrated RBCs, which are associated with increased SCD severity, but decreased malaria susceptibility.

6.1.1 Characterization of red blood cell trait GWAS loci

The characterization of variants associated with complex trait is challenging given that most fall in non-coding regions.[217] In chapter 2, I investigated if rare predicted human gene KOs contributed to anthropometric trait variation. Determining if the loss of functions of genes affect complex traits can help in the characterization of common non-coding variants in the same loci, as it suggests that these variant regulate these genes. I did not find any significant associations, suggesting that more comprehensive genome sequencing datasets and larger sample sizes will be needed to carry this type of analysis. Alternative approaches, such as prioritizing candidate genes, may also help identify rare KO events associated with complex traits. An alternative to this approach is to use transcriptomic datasets, where the effect of LoF variants can be assessed directly. This strategy was used to characterize rs3211938 (**Annex 3**), a *CD36* nonsense variant associated with RDW.[54] This variant as a frequency of 8.7% in African-Americans, but only 0.01% in individuals of European descent. Of 12 RNA-sequenced fetal erythroblasts, only one sample was heterozygote for this variant. This sample displayed strong allelic imbalance (only ~15% of reads contained the nonsense allele), suggesting nonsense mediated decay. Consistently, this sample showed the lowest expression of *CD36.* This analysis suggests that investigating rare LoF variants in transcriptomes of relevant samples can be useful in the characterization of trait-associated variants.

Molecular characterization of GWAS loci by genome editing can be useful, but require relatively strong hypotheses about the identity of the causal variant(s). Two studies presented in (**Annex 1 and Annex 2**) present approaches to characterize regulatory elements using genome editing. I already presented the characterization of an erythroid-specific enhancer of *BCL11A* in section 1.5.3.[56] My contribution was the genetic fine-mapping of intron 2 of *BCL11A* and conditional analyses of variants associated with HbF levels in SCD patients (**Annex 1**). Overall, the combination of genetic association analyses and gene editing revealed that an enhancer harboring rs1427407 was necessary for stage-specific erythroid expression of *BCL11A.* The second study focused on the intergenic region between *HBS1L* and *MYB* (**Annex 2**).[145] This region has been associated with several RBC traits, including HbF, MCH, MCHC, and RBC count. Genetic fine-mapping of this region in SCD patients was more challenging owing the

extensive LD and limited sample size. A CRISPR/Cas9-based variant-aware saturating mutagenesis approach was used to characterize the loci. Knockdown of *MYB* resulted in a strong differentiation defect in human erythroid cells, which was reflected by increased drop-out of the cells from the screen. This approach revealed 4 DHS with potential regulatory function on *MYB*, although the mutagenesis screen may have been confounded by off-target effects due to repetitive regions, highlighting the challenge of characterizing this type of genetic loci using genome editing.

A final example is the one highlighted in Chapter 5 that is the characterization of the *ATP2B4* erythroid enhancer, associated with MCHC, RDW, hemoglobin levels, and susceptibility to severe malaria. Even though the sample size was small, the eQTL association was very strong. Moreover, the association was erythroid-specific, and would not have been found had we used larger public available datasets such as GTEx.[246] This supports the importance of using trait-related cells when characterizing GWAS loci. This study also offers a mechanism for which the *ATP2B4* variant affects RBC traits. Cells deficient for *ATP2B4* cannot effectively pump calcium outside the cytoplasm. The accumulation of Ca^{2+} inside the cell leads to potassium efflux and volume loss. This in turn increases intracellular hemoglobin concentration (MCHC). Lastly, a recurrent theme of the aforementioned examples is that non-coding variants in erythroid enhancers are important contributors to RBC trait variability, and can act in highly stage- and cell-specific manners.

6.1.2 Modulation of HbF production

DNA demethylating agents can modestly increase HbF production in SCD patients. The global DNA methylation study highlighted several difference between fetal and adult RBC progenitors. Thus, it is likely that DNA demethylating agents have a broad effect on the epigenomes of RBC progenitors. Of note, enhancers showed considerable differences between HbF- and HbA-producing cells, consistent with their role on gene expression regulation. NF1 binding motifs were strongly enriched in differentially methylation regions. Although it has to be replicated, a SNP in the intron of *NFIX*, part of the NF1 transcription factor family, is

associated with HbF levels in Sardinians.[146]. This gene also shows differential expression between fetal and adult erythroblasts. These data suggest that *NFIX* may play a role on developmental stage specificity and HbF regulation, although replication of the association and molecular characterization of the locus will be necessary. HbF-associated SNP inside a *BCL11A* binding site at the *HBB* loci is associated with *HBG2* expression independently of DNA methylation at the promoter of *HBG2.* This is consistent with data of SCD mouse model were repression of *BCL11A* and treatment with 5-aza-2'-deoxycytidine, a DNA demethylating agent, induced a synergistic increased in HbF production.[143] Overall, this suggest that a combination of *BCL11A* inactivation and DNA demethylating agent could provide robust induction of HbF in SCD patients.

Our differential gene expression analyzes revealed that fetal-stage erythroblasts express high levels of *LIN28B*, *IGF2BP1* and *IGF2BP3*. *LIN28B* is a repressor of *let-7* miRNAs, which are highly expressed in adult erythroblasts and inhibit HbF production.[349] *LIN28B* also regulate the insulin-like growth factor 2 gene, *IGF2.* The binding partners of *IGF2, IGF2BP1* and *IGF2BP3* showed substantial expression of fetal erythroblasts and are themselves target of *let-7* miRNAs. Conversely, *IGF2BP1* may interfere with *let-7* miRNA gene regulation by protecting its target genes.[408] Consistently, overexpression of *IGF2BP1* in fetal RBC progenitors increased HbF levels.[385] Taken together, these results suggest that the *LIN28B* and *IGF2BP3* might be important regulators of the fetal-stage program, acting antagonistically to *let-7* miRNAs. These results also support an important role of miRNAs in regulating developmental stage specificity. A potentially important finding of our differential expression study was the up-regulation of the chromosome 14q32 miRNA cluster, which had not been previously linked to HbF regulation. Although the specific role of these miRNA remains to be investigated, down-regulation of a subset of 14q32 miRNAs is associated with *BCL11A* up-regulation, suggesting that these miRNA may participate in HbF regulation by targeting this HbF repressor.[391] To summarize, these results highlight several potential regulators of cell-stage specificity, which could also play a role in HbF production. They also provide comprehensive datasets that can be used to prioritize gene in GWAS.

6.1.3 Sickle cell disease and malaria susceptibility

Reactivation of HbF is still the most promising therapy for SCD.[93] GWAS of HbF revealed genetic variants in *BCL11A, MYB,* and the *HBB* contributing to variation of this trait.[47,138-141] The transcription factor *BCL11A* has essential roles in non-erythroid lineages, such as in neuron development and B-cell lymphopoiesis. Hence, therapeutic targeting of *BCL11A* by small molecules may be challenging. Nonetheless, the characterization of the erythroid stage-restricted enhancer of *BCL11A* may offer an alternative – and potentially very specific – therapeutic strategy of disrupting this enhancer using gene therapy.[93,296] On the other hand, the targeted mutagenesis of *MYB* regulatory elements suggests an indirect effect on HbF production via differentiation defects, and might thus be less amenable to therapies. The significance of the other pathways and genes highlighted with our epigenomic and transcriptomic studies (*let-7*/*LIN28B,* 14q32 miRNAs, *NFIX*) remains to be investigated. The miRNA pathways in particular are implicated in several other processes: *let-7* miRNAs have role in cell cycle, differentiation, proliferation and cancer.[409] Our results suggest that the 14q32 miRNA cluster may be implicated in HbF production, but its specific role remains to be investigated. Finally, the erythroid enhancer of *ATP2B4* increases MCHC and leads to RBC dehydration. This offers a mechanism for which *ATP2B4*-associated variants protect against *Plasmodium falciparum* infection, as RBC dehydration has been linked to infectivity by the parasite.[402] *ATP2B4* variants may also protect by restricting calcium available to the parasite.[401,410,411]. On the other hand, it is possible that the increased dehydration increases the severity of SCD.[123] Thus, therapies that modulate *ATP2B4* could have a beneficial impact on these disorders affecting millions of individuals.

6.1.4 Study limitations

The four studies presented in chapter 2 to 5 have some limitations. In chapter 2, I presented a method to identify potential gene knockouts (KO), and test their association with quantitative traits. A first limitation of this study was the definition of gene knockouts, as I included all variants annotated as stop-loss, stop-gain (nonsense), splice-site, and frameshift indels. Stop-loss are unlikely to be as deleterious as stop-gain variants. On the other hand, I did

not include missense variants, some of which may have a strong phenotypic impact.[317,318] Moreover, I did not account for the transcriptomic context of the variants. Nonsense variants are generally more tolerated at the 3' end of transcripts.[9] Nearby variants can also rescue the impact of some variants, for example, two frameshift indels in close proximity could lead to an in-frame protein. The effect of variants annotated as protein-truncating variants may be visible due to isoforms without the affected exon.[257] These limitations could be addressed using more stringent annotation rules or by validating variants predicted to cause nonsense-mediated decay (NMD) by RNA-sequencing. This study was also limited by the different technologies used to detect LoF variants: We used whole-exome sequencing (WES), low-pass whole-genome sequencing (WGS), and the ExomeChip genotyping array. Low-pass WGS may present a higher rate of false positive for rare variants, and might not be as comprehensive as deep WGS, although we found that the frequencies of LoF variants were similar across datasets. The main advantage of using the ExomeChip was the large sample size (N>100,000), but at the cost of a less comprehensive coverage of LoF variants and no information on the phase of the variants.

The main limitation of the differential expression and differential DNA methylation studies (chapter 3 and chapter 4) was related to the differentiation dynamics of fetal and adult erythroblasts. It is possible that a subset of the difference in expression and DNA methylation between both stages reflect discrepancy in differentiation stages. Although I did not detect significant difference in expression of differentiation markers (CD71, CD235a, and CD34), sorting cells based on surface protein expression of these markers by fluorescence-activated cell sorting (FACS) would have minimized difference in differentiation dynamics. Moreover, I used the Illumina 450K DNA methylation array to assess DNA methylation levels in fetal and adult erythroblasts, which is limited by the number and distribution of CpGs assayed, and does not distinguish between the different steps of DNA methylation (e.g. 5-methylcytosine and 5-hydroxymethylcytosine).

Finally, the erythroblast sample size was a limiting factor when testing the association of genotypes or DNA methylation with expression levels (Chapter 3 and 5). In the eQTL

analysis in particular, the limited sample size reduced the power of detecting SNPs associated with gene expression. Nonetheless, we were able to detect eQTLs for 174 genes, which was aided by the approach of focusing on genes with allelic imbalance. On the other hand, this approach limited the eQTL analysis to 479 genes, and thus might have missed other eQTLs that may have been of interest. Another limiting factor is that I combined fetal and adult erythroblasts to increase sample size. Although I corrected for stage-specific differences statistically, it may have reduced the power to detect certain associations, and stage-specific effects might have been missed. Finally, the disruption of particular GATA motifs in an erythroid-specific regulatory element of *ATP2B4* is essential to its erythroid expression. However, the specific causal SNP was not validated, although the result strongly suggest that it disrupts the aforementioned regulatory element.

6.2 Future studies of RBC traits

Genetic studies of RBC traits are already investigating the effect of rare variants to try and fill the missing heritability. Better variant annotation, larger sample sizes, and more exhaustive imputation panels will certainly help identifying novel rare variant associations. Computational approaches such as gene-based tests or machine learning algorithms may also help find associations not significantly captured by conventional single-variant tests. These approaches may be more challenging to adapt to SCD cohorts where the sample size will remain relatively small, especially compared to large cohorts of hundreds of thousands of patients. Nonetheless, integration of GWAS variants with functional annotations may help prioritize and characterize variants associated with RBC traits.

6.2.1 GWAS of RBC traits

To date, common variants significantly associated with RBC traits explained ~28% of the phenotypic variance.[13] I believe that part of the missing heritability will be explained by rare variants. Astle *et al* already associated 250 low-frequency and rare variants with hematological traits in 170,000 samples, a subset of the 500,000 samples that will be included in the UK

Biobank.[13] Chami et al found 16 new RBC association using the ExomeChip array in a discovery sample of 130,273 individuals.[54] Similarly, Marouli et al. identified 83 new low-frequency height-associated loci using the ExomeChip array in >700,000 samples, increase the heritability of height which together explained an additional 1.7% of the heritability, plus an additional 2.4% when considering novel common loci.[12] This design is however limited to the 240,000 variants tested, and more exhaustive approaches such as WES or WGS are likely to identify additional variants. Considering other types of variants such as copy number variants (CNV) might also detect novel associations. As the Astle *et al* study demonstrates, GWAS can benefit from larger and more comprehensive imputation panels, enabling the inclusion of rare variants.[13,25] Moreover, GWAS were mainly done on individuals of European descent. Studies of complex trait will benefit from association studies in other populations. This is especially well demonstrated with RBC traits, as several variants that are rare in Europeans have increased frequency in African regions endemic for malaria. For example, a SNP near the Duffy antigen/receptor for chemokines (*DARC*) causes the Duffy negative blood groups, which confers resistance to malarial infection. This variant is not present in Europeans, but reaches very high frequency in sub-Saharan African populations (90-100%).[55,412] Isolate population can also display higher frequency of otherwise rare variants. Thus, DNA sequencing studies of isolate populations is another approach to identify rare variants. For example, *BRCA1* and *BRCA2* mutations associated with increased breast and ovarian cancer reaches high frequency in the Ashkenazi Jewish population.[174] WES of Finnish individuals identified that splice variants of *LPA* are associated with decreased lowered plasma lipoprotein(a) and cardiovascular disease protection.[158] Furthermore, population-specific reference panels can increase imputation accuracy in matched target genotyping panels. A study of hemoglobin levels in the Sardinian founder using a Sardinian imputation panel revealed 2 associations not detectable with other reference panels.[146] Finally, alternative strategies, such as those based on machine-learning algorithms, can be used to increase power to detect novel associations, identify potential causal variants, or predict phenotypes based on genotypes.[413,414]

6.2.2 GWAS of HbF

In contrast to RBC traits such as MCHC, MCV, and RBC counts, studies of HbF are more limited in sample size – especially when SCD samples are considered. The largest GWAS of HbF in SCD patients included ~2,000 samples.[93,415] SCD is mainly prevalent in developed countries, and given the technical and financial resources required, it will be challenging to assemble large cohorts in these countries. Even in developed countries such as the United States, the fact that many patients are treated with hydroxyurea – a drug increasing HbF levels - might complicate association studies of this trait. An alternative to SCD cohorts is to study healthy populations. Although β-thalassemia mutations are relatively frequent in the Sardinian population, a study of 6,602 healthy samples from Sardinia captured all known associations with HbF.[146] An alternative to GWAS studies would be WES or WGS studies. Targeted sequencing studies have already implicated rare variants in *KLF1*, *BCL11A,* and *MYB* associated with increased HbF levels.[47,148-150,416,417] Given the small sample size, studies of HbF could benefit from extreme phenotype design studies, where individuals from the extremes of the distribution are selected and submitted to deep sequencing. Studies of rare variants influencing HbF could also benefit from gene-based approaches, as these reduce the number of tests burden (i.e. they typically test the association of ~20,000 genes). However, the recent ExomeChip studies of height and RBC traits, which include hundreds of thousands of samples, gene-based test only identified a few associations, suggesting that they might not have enough power to detect associations in studies with lower sample sizes.[12,54] HbF studies may benefit from loci prioritization using genomic datasets. For example, one could consider only variants in differentially expressed genes between fetal and adult erythroblasts, variants that are erythroid eQTLs, or variants in erythroid enhancers or other regulatory regions. Moreover, it may also be possible to take advantage of the pleiotropic effect of variants. For example, the *HBS1L-MYB* locus is associated with several RBC trait other than HbF such as MCHC. Hence, it could be possible to prioritize variants in SCD cohorts based on their association with other RBC traits in large-scale studies.

6.2.3 Integration with large-scale transcriptomic and epigenomic data

Transcriptomic and epigenomic studies will become increasingly important both for prioritization of genetic variants and characterization of GWAS loci. Epigenomic consortiums are accumulating large sets of data on histone modification, DNA methylation, transcription factor binding sites, gene expression, and chromosome conformation. Datasets, such as the GTEx and BLUEPRINT, are also generating data of association between variants and gene expression (eQTLs and mRNA splicing QTLs) or epigenomic marks (methylation and histone QTLs).[214,246] Whole-blood eQTL studies can be very powerful as a large number of samples can be assayed, and computational methods can help disentangle the contribution of cell types, which may reveal cell-specific eQTLs.[255,418] Single-cell transcriptomics and epigenomics may also be useful for RBC trait studies, especially since it may help resolve differences in differentiation state or cell cycle that can impact population studies.[419,420] Accounting for these differences may also help increase power to detect QTLs.[421] Finally, studies of the 3D structure of the chromatin such as promoter Hi-C may reveal very useful to link enhancer or other regulatory elements to genes, and assign functions to non-coding variants.[203,206] Transcriptomic and epigenetic studies of erythroblasts are still limited, even in large-scale epigenomic projects: the BLUEPRINT dataset contains transcriptomic data for 8 erythroblast samples, only 1 of which is from the fetal liver.[214] Hence, larger sample size will be necessary to conduct exhaustive QTL studies related to RBC traits. In addition to provide candidate causal genes, these studies could be used to prioritize variants in smaller sample sized studies as is the case for HbF.

6.3 Future applications of genome editing

Functional genomic annotations will be useful to prioritize candidate causal genes. However, it will remain necessary to validate these genes through functional experiments. To this effect, genome editing – in particular the CRISPR/Cas9 system – is becoming increasingly useful. This is exemplified by characterization of causal function elements at loci such as *BCL11A, SORT1, FTO,* and *ATP2B4*.[56,300,393] In this section, I will focus on ameliorations of genome editing and their use for genomic characterization.

6.3.1 Modifications and applications of CRISPR/Cas9

One of the limitation of the CRISPR/Cas9 system is its dependency on the 5'-*NGG* PAM at the target loci. One solution is to use orthologs of *Cas9*. For instance, *Staphylococcus aureus Cas9* ortholog recognizes NNGRR(N) PAMs and *Francisella novicida Cpf1* recognizes TTN.[422] Another approach is to engineer modification of *Cas9* to change its PAM specificity. Cas9 VQR and VRER variants recognize NGAN/NGNG and NGCG respectively.[423] A potential problem of using the CRISR/Cas9 system are off-target effects. One approach to address this issue is to use Cas9 nickases, which only induces single-strand breaks. Thus, two nickases targeting the same loci on both sides are necessary to induce DSBs.[424] However, nickases can still induce DSB or point mutation at certain site. Analogous to this approach is to use two dead Cas9 (dCas9), both fused with the FokI nuclease domain to create a dimeric RNA-guided FokI–dCas9 nuclease (RFN). Off-targets are reduced because FokI induces DSBs only when in dimer form, and because two 20bp motifs flanking a 13-18bp target region are necessary.[425] sgRNAs truncated of 2-3bp or sgRNA with an added GG at the 5' end of the sequence have also been shown to reduce off-target events.[424] Two modified Cas9, the enhanced SpCas9 version 1.1 (eSpCas9 1.1) and SpCas9 high-fidelity variant 1 (HF1) have comparable on-target efficiency to Cas9 and, when tested, guides induced very few or no off-target events.[426] Off-target effects can be a potentially large problem for CRISPR/Cas9 experiments. Methods like the genome-wide, unbiased identification of DSBs enabled by sequencing (GUIDE-seq) can be used to assess off-target effects of guides.[427] Several computational tools have been developed in order to predict off-target and on-target effects.[428-432] Future research will be necessary to validate and compare these different methods with experimental data.

Inducible Cas9 can be used to reduce off-target effects or to create conditional knockouts. Examples include a rapamycin-inducible Cas9, photoactivatable Cas9, 4-hydroxytamoxifen-inducible Cas9, and a Cas9 stabilized after treatment with a FKBP12 synthetic ligand Cas9.[433-436] Another approach used an inductile spacer-blocking hairpin (iSBH) structure at the 5′ end of the sgRNA, which can be cleaved by a variety of inducers.[437] Finally, we can take advantage of the specific targeting of Cas9 for other purposes than to induce DSBs. The "dead" Cas9 (dCas9) mentioned above can be created by mutating the two domains

responsible for cleavage. Alone, this dCas9 can repress transcription simply by binding to exons or promoters.[438] Fusing the dCas9 to other proteins can confer additional characteristic to the system: Transcriptional activation can be achieved by fusing repeats of VP16 transactivation domains to dCas9.[439] Other systems, such as the synergistic activation mediator (SAM) system, builds on this by recruiting additional activators.[293,440] Conversely, adding the KRAB domain will transform the Cas9 into a transcription repressor.[292] Targeted epigenomic modifications can be induced by fusing dCas9 with a corresponding protein. For instance, fusing dCas9 to LDS1 or p300 will induce targeted H3K4 demethylation or H3K27 acetylation respectively.[441,442]. dCas9-GFP can be used for chromosomal imaging. [443] Finally, development of CRISPR-based RNA-targeting could be used to track RNAs, induce degradation, alter their localization, or modulate splicing.[444,445]

HDR can be exploited to induce specific mutations at a given loci or to insert genes of interest (knock-in). Conditional KOs can be created inserting flanking LoxP site at a region of interest, which will be targeted by the Cre recombinase. However, compared to the DSB-inducing non-homologous end-joining (NHEJ) pathway, HDR is largely inefficient. Higher efficiency can be achieved by submitting a DNA template with two homologous arms to the targeted site, using single stranded DNA donors, or blocking proteins implicated in NHEJ.[446] Since HDR occurs mainly at phases S and G2 of the cell cycle, using cell cycle synchronizers such as Nocodazole or CCND1 can also be used to promote HDR.[447,448] Future research will focus on increasing the specificity of the CRISPR/Cas9 system, expanding its applications, and increasing the efficiency of HDR.

6.3.2 Large-scale characterization of genes and regulatory elements

With the increasing number of genetic variants associated with complex trait, large scale approaches to characterize the variants will be essential. Massively parallel reporter assays have been used to screen the effect of specific variants on gene expression.[449,450] For example, Ulirsch *et al* tested 2,756 variants association with RBC traits using this approach and identified 32

functional variants representing ~30% of the number of loci tested. [450] The CRISPR/Cas9 system offers an alternative to interrogate regulatory elements. Wakabayashi *et al* used this system to induce deletions at regulatory elements of genes implicated in erythroid disorders.[451] They found that the disruption of regulatory elements could have a drastic impact on gene expression, and that mutations in these site could potentially cause monogenic disorder phenotypes. CRISPR screens can be useful to characterize regulatory elements or genes implicated in various traits. For example, Korkmaz *et al* used CRISPR screens to identify regulatory elements bound by p53 or ERα oncogene-induced senescence and cellular proliferation.[452] CRISPR screen like GeCKO and Avana can be used to characterize the effect of gene KOs.[176,429] Screens of gene activation or repression can also be used to investigate the effect of gene loss or gain of function respectively.[181,292,293] Transcription repressor can be particularly useful to study non-coding RNAs and essential genes. The CRISPR/Cas9 system can also be used to fine-map specific locus, as exemplified by *in situ* mutagenesis of *BCL11A* and *HBS1L-MYB.*[59,145] Other approaches have taken advantage of HDR. For example, Findlay *et al* used a multiplex HDR approach to test the effect of known SNPs in exon 18 of *BRCA1* on its transcript abundance.[453] It will be critical that future CRISPR studies use correct controls such as using different sgRNAs to account for possible off-target effects. This will be especially important for screens or saturating mutagenesis where off-target effects can induce cell death and unanticipated drop-out of sgRNAs from the screen.[145,291,454] Future research will probably aim to expand the number of regulatory elements tested in CRISPR screens. Dissection of regulatory elements will improve our definition of enhancers, as some active regulatory elements do not display the "conventional" signature of enhancers. For instance, a study dissecting regulatory elements of *POU5F1* found that disrupting a class of enhancers had only temporary and reversible effects on gene expression.[454] In short, these studies demonstrate the utility of the CRISPR/Cas9 system for genetic characterization, and the number of applications derived from this system will continue to grow.

6.4 Treatments of RBC disorders

6.4.1 Promising therapies to treat sickle cell disease

The life expectancy of SCD patients in the U.S. is less than 50 years old.[115,455] Hydroxyurea is the only drug approved by the US Food and Drug Administration to treat SCD.[93] This drug increases HbF levels in SCD patients, reducing the number of complications and increasing life expectancy. Although this drug is usually well tolerate and have generally mild side effects, it usually doesn't increase HbF levels enough to completely abrogate SCD complications, especially in adults.[93] The β-globin locus is classic example of chromatin conformation where looping from the locus control region to the γ-globin gene is switched to the β-globin during the fetal-to-adult hemoglobin switch. This is associated with increased repressive epigenetic marks such as DNA methylation at the *HBG2* gene promoter. As such, drugs that target chromatin regulators may present a potential avenue for HbF-activating agents. Decitabine increases HbF in SCD patients and is currently tested in phase 1 clinical trial. Histone deacetylase (HDAC) inhibitors such as Vorinostat are also considered for HbF-reactivation. Vorinostat has shown modest increase in HbF in clinical trials.[456] A caveat of this approach is that chromatin regulators are implicated in the regulation of genes in a multitude of cellular contexts, and thus may show unwarranted non-specific effects and toxicities.

Gene therapies to treat SCD are in development. Therapies aiming to insert the β-globin or γ-globin gene in HPSC of patients using lentiviral vectors have had promising results and are currently in clinical trials.[296] Gene editing might offer an alternative option. A proposed approach would be to isolate HPSC from SCD patients, engineer modifications of genes in these cell, and reinfused them to the patient.[93] A possible avenue would be to correct the SCD mutation in the *HBB* gene, but would require robust HDR.[457] Another approach using the NHEJ pathway is to induce indels in the erythroid-specific enhancer of *BCL11A* to reactivate HbF production at therapeutic levels.[93]

The only cure for SCD is hematopoietic stem cell transplant. This procedure shows over 90% event-free survival when the transplant is received from a HLA-matched donor.[458] However, a major limiting factor of this approach is donor availability, ~14% of U.S. SCD patients have a compatible HLA-matched donor. Unrelated haploidentical donors are a promising alternative, although current efforts have shown high levels of primary graft failure and disease-free survival as low as 38%.[459] HPSC from umbilical cord blood are another alternative. When cord blood is harvested from a related donor, this approach has high event free survival (~90%) and low rates of complications. [460] However, unrelated cord blood grafts show high rates of failure and low event-free survival (~50%).[461] Ongoing efforts at increasing these statistics and reduce complications such as organ rejection, graft-versus-host disease, infertility and infections.

Finally, although the global burden of SCD is expected to rise – especially in developed countries like Nigeria – implementation of health interventions such as antibiotics and vaccination could significantly increase the life expectancy of SCD patients.[111] In the U.S., the healthcare cost imputed to SCD is likely to be over $1.1 Billion annually.[462] In developing countries, the disease has a very large related financial burden for families with children suffering from SCD.[463] Prenatal diagnosis, and education on the genetic inheritance of HbS could reduce the global economic burden of this disease. [111]

6.4.2 Malaria – therapies from human genetics

In 2015, there were more than 212 million malaria cases causing 429 000 related deaths. A major problem that antimalarial therapies are facing is increased drug resistance. Human genetics have already identified several genes implicated in malaria susceptibility. Variants near *DARC* abolishes its expression in erythroid cells, and confers resistance to *P. vivax* infection. [100] Binding of the Duffy antigen to *P. vivax* Duffy-binding protein (PvDBP) is required for parasite infection. Vaccines and small-molecule approach targeting PvDBP are in development.[464] *G6PD* deficiency also confers malaria resistance and is associated with multiple RBC traits. Inhibitors of G6PD could provide an option for the treatment of malaria

infection.[99,465] Individuals that are deficient for the ferrochelatase (*FECH*) gene have erythropoietic protoporphia and are resistant to malarial parasite growth. Thus, FECH inhibitors could present another therapeutic option against malaria.[102] Mutations in the basigin (BSG) gene (Ok blood group antigen) is essential for invasion of the erythrocyte by the parasite.[466] Basigin interacts with *P. falciparum* reticulocyte-binding protein homologue 5 (PfRH5). Acquired antibodies to PfRH5 inhibits parasite growth and is associated with improved outcome to infection.[467] PfRH5-based vaccines are efficacious against malarial infection in *Aotus* monkeys.[468] Alternatively, blocking basigin may present a more efficacious approach.[99] Other variants in *ABO*, *HBB*, *ATP2B4,* and *CD40LG* loci are associated with malaria susceptibility.[107] Understanding how these variants influence malarial infection and RBC functions may help designing therapies for this infectious disease. Therapies that rely on host proteins are not directly influenced by the parasite's genetics and may be harder for parasites to develop drug resistance to, which is an increasing problem for malaria infections.

6.5 Conclusions

In this work, I have used several approaches to try and identify genes implicated in complex traits, and more specifically in RBC traits. First, I have developed a method to identify and test the association between predicted gene KOs and complex traits. I applied this approach on anthropometric traits, but despite the large sample size, it did not reveal significant associations. Comparing DNA methylation states between fetal and adult erythroblasts revealed large genome-wide differences, which were enriched in enhancer regions. These epigenetic changes were also reflected in the large number of transcripts differentially expressed. In particular, *let-7* and chr14q32 miRNAs might be important contributors of cell-type specificity in adult and fetal erythroblasts respectively. Finally, we identify several variants that are eQTL in erythroblasts, and identified a causal regulatory element of *ATP2B4*, for which DNA variants modulate RBC traits and malaria susceptibility.

The lack of association from the KO study suggests that very large sample sizes and comprehensive datasets will be necessary to study the effect of rare LoF variants on complex traits. This kind of analysis might be possible with better imputation panels and large studies such as the UK biobank and UK10K, and could be applied to RBC traits. Comparing fetal and adult erythroblasts identified several candidates that could be implicated in stage-specific RBC functions such as HbF production. Functional annotations will be particularly important to prioritize genes in genome-wide association studies, especially for traits where it is difficult to assemble large cohorts such as for SCD patients.

Characterizing eQTLs – even in modest sample sizes - can be useful to identify causal genes in the context of genetic association studies, as exemplified by the *ATP2B4* example. This study also highlights the importance of using a relevant cell type, as the effect is highly erythroid-specific. It also demonstrates the utility of genome editing in characterizing regulatory elements, for which the number of applications is only growing. Studies characterizing cell-specific regulatory elements as for *ATP2B4* and *BCL11A* opens the possibility of very specific therapies to treat widespread diseases such as malaria and sickle cell disease, especially with the emergence of gene therapies.

References

1. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304-51 (2001).
2. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
3. Yates, A. *et al.* Ensembl 2016. *Nucleic Acids Res* **44**, D710-6 (2016).
4. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
5. Variation in genome-wide mutation rates within and between human families. *Nat Genet* **43**, 712-714 (2011).
6. Francioli, L.C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet* **47**, 822-6 (2015).
7. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-5 (2012).
8. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636-9 (2010).
9. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-8 (2012).
10. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* **43**, D789-98 (2015).
11. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**, 1173-86 (2014).
12. Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186-190 (2017).
13. Astle, W.J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415-1429 e19 (2016).
14. Stranger, B.E., Stahl, E.A. & Raj, T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* **187**, 367-83 (2011).
15. Fisher, R.A. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**, 399-433 (1918).
16. Sturtevant, A.H. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology* **14**, 43-59 (1913).
17. Morgan, T.H. Random Segregation Versus Coupling in Mendelian Inheritance. *Science* **34**, 384 (1911).
18. Kainulainen, K., Pulkkinen, L., Savolainen, A., Kaitila, I. & Peltonen, L. Location on chromosome 15 of the gene defect causing Marfan syndrome. *N Engl J Med* **323**, 935-9 (1990).
19. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**, 95-108 (2005).
20. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**, D896-D901 (2017).
21. Ozaki, K. *et al.* Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat Genet* **32**, 650-4 (2002).

22. Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. Five years of GWAS discovery. *Am J Hum Genet* **90**, 7-24 (2012).
23. International HapMap, C. A haplotype map of the human genome. *Nature* **437**, 1299-320 (2005).
24. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499-511 (2010).
25. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *bioRxiv* (2016).
26. Low-Kam, C. *et al.* Whole-genome sequencing in French Canadians from Quebec. *Hum Genet* **135**, 1213-1221 (2016).
27. Sidore, C. *et al.* Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat Genet* **47**, 1272-81 (2015).
28. Consortium, U.K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).
29. Genome of the Netherlands, C. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46**, 818-25 (2014).
30. Gudbjartsson, D.F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* **47**, 435-44 (2015).
31. Nagasaki, M. *et al.* Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat Commun* **6**, 8018 (2015).
32. Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327-32 (2015).
33. Telenti, A. *et al.* Deep sequencing of 10,000 human genomes. *Proc Natl Acad Sci U S A* **113**, 11901-11906 (2016).
34. Visscher, P.M., Hill, W.G. & Wray, N.R. Heritability in the genomics era--concepts and misconceptions. *Nat Rev Genet* **9**, 255-66 (2008).
35. Reddon, H., Gueant, J.L. & Meyre, D. The importance of gene-environment interactions in human obesity. *Clin Sci (Lond)* **130**, 1571-97 (2016).
36. Nettleton, J.A. *et al.* Gene x dietary pattern interactions in obesity: analysis of up to 68 317 adults of European ancestry. *Hum Mol Genet* **24**, 4728-38 (2015).
37. Mustelin, L., Silventoinen, K., Pietilainen, K., Rissanen, A. & Kaprio, J. Physical activity reduces the influence of genetic effects on BMI and waist circumference: a study in young adult twins. *Int J Obes (Lond)* **33**, 29-36 (2009).
38. Falconer, D.S. & Mackay, T.F.C. *Introduction to Quantitative Genetics*, (Longman, 1996).
39. Silventoinen, K. *et al.* Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Res* **6**, 399-408 (2003).
40. Polderman, T.J. *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat Genet* **47**, 702-9 (2015).
41. Visscher, P.M. *et al.* Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet* **2**, e41 (2006).
42. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-9 (2010).
43. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).

44. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-35 (2015).
45. Robinson, M.R., Wray, N.R. & Visscher, P.M. Explaining additional genetic variation in complex traits. *Trends Genet* **30**, 124-32 (2014).
46. O'Reilly, P.F. *et al.* MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One* **7**, e34861 (2012).
47. Galarneau, G. *et al.* Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat Genet* **42**, 1049-51 (2010).
48. Steinberg, M.H. *Disorders of hemoglobin : genetics, pathophysiology, and clinical management*, xiv, 1268 p. (Cambridge University Press, Cambridge ; New York, 2001).
49. Ulker, P., Sati, L., Celik-Ozenci, C., Meiselman, H.J. & Baskurt, O.K. Mechanical stimulation of nitric oxide synthesizing mechanisms in erythrocytes. *Biorheology* **46**, 121-32 (2009).
50. Jiang, N., Tan, N.S., Ho, B. & Ding, J.L. Respiratory protein-generated reactive oxygen species as an antimicrobial strategy. *Nat Immunol* **8**, 1114-22 (2007).
51. Schnabel, R.B. *et al.* Duffy antigen receptor for chemokines (Darc) polymorphism regulates circulating concentrations of monocyte chemoattractant protein-1 and other inflammatory mediators. *Blood* **115**, 5289-99 (2010).
52. Colin, Y., Le Van Kim, C. & El Nemer, W. Red cell adhesion in human diseases. *Curr Opin Hematol* **21**, 186-92 (2014).
53. Whelihan, M.F. & Mann, K.G. The role of the red cell membrane in thrombin generation. *Thromb Res* **131**, 377-82 (2013).
54. Chami, N. *et al.* Exome Genotyping Identifies Pleiotropic Variants Associated with Red Blood Cell Traits. *Am J Hum Genet* **99**, 8-21 (2016).
55. Chami, N. & Lettre, G. Lessons and Implications from Genome-Wide Association Studies (GWAS) Findings of Blood Cell Phenotypes. *Genes (Basel)* **5**, 51-64 (2014).
56. Bauer, D.E. *et al.* An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science* **342**, 253-7 (2013).
57. Sankaran, V.G. *et al.* Cyclin D3 coordinates the cell cycle during differentiation to regulate erythrocyte size and number. *Genes Dev* **26**, 2075-87 (2012).
58. Sankaran, V.G. *et al.* Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. *Science* **322**, 1839-42 (2008).
59. Canver, M.C. *et al.* BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192-7 (2015).
60. Kurita, R. *et al.* Establishment of immortalized human erythroid progenitor cell lines able to produce enucleated red blood cells. *PLoS One* **8**, e59890 (2013).
61. van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* **492**, 369-75 (2012).
62. Chen, Z. *et al.* Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network. *Hum Mol Genet* **22**, 2529-38 (2013).
63. Shah, R.C., Buchman, A.S., Wilson, R.S., Leurgans, S.E. & Bennett, D.A. Hemoglobin level in older persons and incident Alzheimer disease: prospective cohort analysis. *Neurology* **77**, 219-26 (2011).
64. Sabatine, M.S. *et al.* Association of hemoglobin levels with clinical outcomes in acute coronary syndromes. *Circulation* **111**, 2042-9 (2005).

65. Lo, K.S. *et al.* Genetic association analysis highlights new loci that modulate hematological trait variation in Caucasians and African Americans. *Hum Genet* **129**, 307-17 (2011).
66. Fraser, S.T., Isern, J. & Baron, M.H. Maturation and enucleation of primitive erythroblasts during mouse embryogenesis is accompanied by changes in cell-surface antigen expression. *Blood* **109**, 343-52 (2007).
67. Lin, C.S., Lim, S.K., D'Agati, V. & Costantini, F. Differential effects of an erythropoietin receptor gene disruption on primitive and definitive erythropoiesis. *Genes Dev* **10**, 154-64 (1996).
68. Socolovsky, M. Molecular insights into stress erythropoiesis. *Curr Opin Hematol* **14**, 215-24 (2007).
69. Fischbach, F.T. & Dunning, M.B. *A Manual of Laboratory and Diagnostic Tests*, (Wolters Kluwer Health/Lippincott Williams & Wilkins, 2009).
70. Lokwani, D. *The ABC of CBC: Interpretation of Complete Blood Count and Histograms*, (Jaypee Brothers Medical Publishers, 2013).
71. Steinberg, M.H. *Disorders of hemoglobin : genetics, pathophysiology, and clinical management*, xx, 826 p., 36 p. of plates (Cambridge University Press, New York, 2009).
72. Stamatoyannopoulos, G. Control of globin gene expression during development and erythroid differentiation. *Exp Hematol* **33**, 259-71 (2005).
73. Bauer, D.E., Kamran, S.C. & Orkin, S.H. Reawakening fetal hemoglobin: prospects for new therapies for the beta-globin disorders. *Blood* **120**, 2945-53 (2012).
74. Carter, D., Chakalova, L., Osborne, C.S., Dai, Y.F. & Fraser, P. Long-range chromatin regulatory interactions in vivo. *Nat Genet* **32**, 623-6 (2002).
75. Akashi, K., Traver, D., Miyamoto, T. & Weissman, I.L. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature* **404**, 193-7 (2000).
76. Kondo, M., Weissman, I.L. & Akashi, K. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. *Cell* **91**, 661-72 (1997).
77. Manz, M.G., Miyamoto, T., Akashi, K. & Weissman, I.L. Prospective isolation of human clonogenic common myeloid progenitors. *Proc Natl Acad Sci U S A* **99**, 11872-7 (2002).
78. Hattangadi, S.M., Wong, P., Zhang, L., Flygare, J. & Lodish, H.F. From stem cell to red cell: regulation of erythropoiesis at multiple levels by multiple proteins, RNAs, and chromatin modifications. *Blood* **118**, 6258-68 (2011).
79. Koury, M.J. & Haase, V.H. Anaemia in kidney disease: harnessing hypoxia responses for therapy. *Nat Rev Nephrol* **11**, 394-410 (2015).
80. Hentze, M.W., Muckenthaler, M.U. & Andrews, N.C. Balancing acts: molecular control of mammalian iron metabolism. *Cell* **117**, 285-97 (2004).
81. Ganz, T. & Nemeth, E. Hepcidin and iron homeostasis. *Biochim Biophys Acta* **1823**, 1434-43 (2012).
82. Kautz, L. *et al.* Identification of erythroferrone as an erythroid regulator of iron metabolism. *Nat Genet* **46**, 678-84 (2014).
83. Mozos, I. Mechanisms linking red blood cell disorders and cardiovascular diseases. *Biomed Res Int* **2015**, 682054 (2015).
84. Brugnara, C. Iron deficiency and erythropoiesis: new diagnostic approaches. *Clin Chem* **49**, 1573-8 (2003).

85. Chiabrando, D., Mercurio, S. & Tolosano, E. Heme and erythropoieis: more than a structural role. *Haematologica* **99**, 973-83 (2014).
86. Bergmann, A.K. *et al.* Systematic molecular genetic analysis of congenital sideroblastic anemia: evidence for genetic heterogeneity and identification of novel mutations. *Pediatr Blood Cancer* **54**, 273-8 (2010).
87. Cotter, P.D., Baumann, M. & Bishop, D.F. Enzymatic defect in "X-linked" sideroblastic anemia: molecular evidence for erythroid delta-aminolevulinate synthase deficiency. *Proc Natl Acad Sci U S A* **89**, 4028-32 (1992).
88. Nada, A.M. Red cell distribution width in type 2 diabetic patients. *Diabetes Metab Syndr Obes* **8**, 525-33 (2015).
89. Zalawadiya, S.K. *et al.* Red cell distribution width and mortality in predominantly African-American population with decompensated heart failure. *J Card Fail* **17**, 292-8 (2011).
90. Zalawadiya, S.K., Veeranna, V., Panaich, S.S. & Afonso, L. Red cell distribution width and risk of peripheral artery disease: analysis of National Health and Nutrition Examination Survey 1999-2004. *Vasc Med* **17**, 155-63 (2012).
91. Patel, H.H., Patel, H.R. & Higgins, J.M. Modulation of red blood cell population dynamics is a fundamental homeostatic response to disease. *Am J Hematol* **90**, 422-8 (2015).
92. Patel, K.V. *et al.* Red cell distribution width and mortality in older adults: a meta-analysis. *J Gerontol A Biol Sci Med Sci* **65**, 258-65 (2010).
93. Lettre, G. & Bauer, D.E. Fetal haemoglobin in sickle-cell disease: from genetic epidemiology to new therapeutic strategies. *Lancet* **387**, 2554-64 (2016).
94. Miller, L.H., Ackerman, H.C., Su, X.Z. & Wellems, T.E. Malaria biology and disease pathogenesis: insights for new treatments. *Nat Med* **19**, 156-67 (2013).
95. de Koning-Ward, T.F., Gilson, P.R. & Crabb, B.S. Advances in molecular genetic systems in malaria. *Nat Rev Microbiol* **13**, 373-87 (2015).
96. Trampuz, A., Jereb, M., Muzlovic, I. & Prabhu, R.M. Clinical review: Severe malaria. *Crit Care* **7**, 315-23 (2003).
97. Ockenhouse, C.F. *et al.* Molecular basis of sequestration in severe and uncomplicated Plasmodium falciparum malaria: differential adhesion of infected erythrocytes to CD36 and ICAM-1. *J Infect Dis* **164**, 163-9 (1991).
98. Kwiatkowski, D.P. How malaria has affected the human genome and what human genetics can teach us about malaria. *Am J Hum Genet* **77**, 171-92 (2005).
99. Lelliott, P.M., McMorran, B.J., Foote, S.J. & Burgio, G. The influence of host genetics on erythrocytes and malaria infection: is there therapeutic potential? *Malar J* **14**, 289 (2015).
100. Tournamille, C., Colin, Y., Cartron, J.P. & Le Van Kim, C. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat Genet* **10**, 224-8 (1995).
101. Foo, L.C., Rekhraj, V., Chiang, G.L. & Mak, J.W. Ovalocytosis protects against severe malaria parasitemia in the Malayan aborigines. *Am J Trop Med Hyg* **47**, 271-5 (1992).
102. Smith, C.M. *et al.* Red cells from ferrochelatase-deficient erythropoietic protoporphyria patients are resistant to growth of malarial parasites. *Blood* **125**, 534-41 (2015).

103. Miller, L.H. *et al.* Evidence for differences in erythrocyte surface receptors for the malarial parasites, Plasmodium falciparum and Plasmodium knowlesi. *J Exp Med* **146**, 277-81 (1977).
104. Ruwende, C. & Hill, A. Glucose-6-phosphate dehydrogenase deficiency and malaria. *J Mol Med (Berl)* **76**, 581-8 (1998).
105. Wambua, S. *et al.* The effect of alpha+-thalassaemia on the incidence of malaria and other diseases in children living on the coast of Kenya. *PLoS Med* **3**, e158 (2006).
106. Ackerman, H. *et al.* A comparison of case-control and family-based association methods: the example of sickle-cell and malaria. *Ann Hum Genet* **69**, 559-65 (2005).
107. Malaria Genomic Epidemiology, N. & Malaria Genomic Epidemiology, N. Reappraisal of known malaria resistance loci in a large multicenter study. *Nat Genet* **46**, 1197-204 (2014).
108. Joiner, C.H. Gardos pathway to sickle cell therapies? *Blood* **111**, 3918-9 (2008).
109. Adewoyin, A.S. Management of sickle cell disease: a review for physician education in Nigeria (sub-saharan Africa). *Anemia* **2015**, 791498 (2015).
110. Frenette, P.S. & Atweh, G.F. Sickle cell disease: old discoveries, new concepts, and future promise. *J Clin Invest* **117**, 850-8 (2007).
111. Piel, F.B., Hay, S.I., Gupta, S., Weatherall, D.J. & Williams, T.N. Global burden of sickle cell anaemia in children under five, 2010-2050: modelling based on demographics, excess mortality, and interventions. *PLoS Med* **10**, e1001484 (2013).
112. Grosse, S.D. *et al.* Sickle cell disease in Africa: a neglected cause of early childhood mortality. *Am J Prev Med* **41**, S398-405 (2011).
113. Platt, O.S. *et al.* Pain in sickle cell disease. Rates and risk factors. *N Engl J Med* **325**, 11-6 (1991).
114. Gray, A., Anionwu, E.N., Davies, S.C. & Brozovic, M. Patterns of mortality in sickle cell disease in the United Kingdom. *J Clin Pathol* **44**, 459-63 (1991).
115. Platt, O.S. *et al.* Mortality in sickle cell disease. Life expectancy and risk factors for early death. *N Engl J Med* **330**, 1639-44 (1994).
116. Green, N.S. & Barral, S. Emerging science of hydroxyurea therapy for pediatric sickle cell disease. *Pediatr Res* **75**, 196-204 (2014).
117. Aliyu, Z.Y., Tumblin, A.R. & Kato, G.J. Current therapy of sickle cell disease. *Haematologica* **91**, 7-10 (2006).
118. Okpala, I. & Tawil, A. Management of pain in sickle-cell disease. *J R Soc Med* **95**, 456-8 (2002).
119. Locatelli, F. *et al.* Outcome of patients with hemoglobinopathies given either cord blood or bone marrow transplantation from an HLA-identical sibling. *Blood* **122**, 1072-8 (2013).
120. Nienhuis, A.W. & Nathan, D.G. Pathophysiology and Clinical Manifestations of the beta-Thalassemias. *Cold Spring Harb Perspect Med* **2**, a011726 (2012).
121. Sankaran, V.G., Lettre, G., Orkin, S.H. & Hirschhorn, J.N. Modifier genes in Mendelian disorders: the example of hemoglobin disorders. *Ann N Y Acad Sci* **1214**, 47-56 (2010).
122. Faa, V., Masala, M., Cao, A. & Rosatelli, M.C. Alpha globin gene duplications in beta thalassemia patients with intact beta globin gene. *Blood Cells Mol Dis* **44**, 156-8 (2010).
123. Bartolucci, P. *et al.* Erythrocyte density in sickle cell syndromes is associated with specific clinical manifestations and hemolysis. *Blood* **120**, 3136-41 (2012).

124. Lettre, G. The search for genetic modifiers of disease severity in the beta-hemoglobinopathies. *Cold Spring Harb Perspect Med* **2**(2012).
125. Watson, J. The significance of the paucity of sickle cells in newborn Negro infants. *Am J Med Sci* **215**, 419-23 (1948).
126. Perrine, R.P., Brown, M.J., Clegg, J.B., Weatherall, D.J. & May, A. Benign sickle-cell anaemia. *Lancet* **2**, 1163-7 (1972).
127. Kar, B.C. *et al.* Sickle cell disease in Orissa State, India. *Lancet* **2**, 1198-201 (1986).
128. Castro, O. *et al.* The acute chest syndrome in sickle cell disease: incidence and risk factors. The Cooperative Study of Sickle Cell Disease. *Blood* **84**, 643-9 (1994).
129. Milner, P.F. *et al.* Sickle cell disease as a cause of osteonecrosis of the femoral head. *N Engl J Med* **325**, 1476-81 (1991).
130. Koshy, M. *et al.* Leg ulcers in patients with sickle cell disease. *Blood* **74**, 1403-8 (1989).
131. Stevens, M.C., Padwick, M. & Serjeant, G.R. Observations on the natural history of dactylitis in homozygous sickle cell disease. *Clin Pediatr (Phila)* **20**, 311-7 (1981).
132. Emond, A.M. *et al.* Acute splenic sequestration in homozygous sickle cell disease: natural history and management. *J Pediatr* **107**, 201-6 (1985).
133. Emond, A.M., Holman, R., Hayes, R.J. & Serjeant, G.R. Priapism and impotence in homozygous sickle cell disease. *Arch Intern Med* **140**, 1434-7 (1980).
134. Garner, C. *et al.* Genetic influences on F cells and other hematologic variables: a twin heritability study. *Blood* **95**, 342-6 (2000).
135. Pilia, G. *et al.* Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet* **2**, e132 (2006).
136. Labie, D. *et al.* The -158 site 5' to the G gamma gene and G gamma expression. *Blood* **66**, 1463-5 (1985).
137. Chakalova, L. *et al.* The Corfu deltabeta thalassemia deletion disrupts gamma-globin gene silencing and reveals post-transcriptional regulation of HbF expression. *Blood* **105**, 2154-60 (2005).
138. Menzel, S. *et al.* A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat Genet* **39**, 1197-9 (2007).
139. 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 U S A* **105**, 11869-74 (2008).
140. 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 U S A* **104**, 11346-51 (2007).
141. 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 U S A* **105**, 1620-5 (2008).
142. Sankaran, V.G., Xu, J. & Orkin, S.H. Transcriptional silencing of fetal hemoglobin by BCL11A. *Ann N Y Acad Sci* **1202**, 64-8 (2010).
143. Xu, J. *et al.* Correction of sickle cell disease in adult mice by interference with fetal hemoglobin silencing. *Science* **334**, 993-6 (2011).
144. Jiang, J. *et al.* cMYB is involved in the regulation of fetal hemoglobin production in adults. *Blood* **108**, 1077-83 (2006).
145. Canver, M.C. *et al.* Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci. *Nat Genet* (2017).

146. Danjou, F. *et al.* Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels. *Nat Genet* **47**, 1264-71 (2015).
147. Holmfeldt, P. *et al.* Nfix is a novel regulator of murine hematopoietic stem and progenitor cell survival. *Blood* **122**, 2987-96 (2013).
148. Borg, J. *et al.* Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin. *Nat Genet* **42**, 801-5 (2010).
149. Basak, A. *et al.* BCL11A deletions result in fetal hemoglobin persistence and neurodevelopmental alterations. *J Clin Invest* **125**, 2363-8 (2015).
150. Funnell, A.P. *et al.* 2p15-p16.1 microdeletions encompassing and proximal to BCL11A are associated with elevated HbF in addition to neurologic impairment. *Blood* **126**, 89-93 (2015).
151. Viprakasit, V. *et al.* Mutations in Kruppel-like factor 1 cause transfusion-dependent hemolytic anemia and persistence of embryonic globin gene expression. *Blood* **123**, 1586-95 (2014).
152. 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 U S A* **108**, 1519-24 (2011).
153. Sauna, Z.E. & Kimchi-Sarfaty, C. Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet* **12**, 683-91 (2011).
154. Lim, E.T. *et al.* Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. *Neuron* **77**, 235-42 (2013).
155. Lange, L.A. *et al.* Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am J Hum Genet* **94**, 233-45 (2014).
156. Steinthorsdottir, V. *et al.* Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat Genet* **46**, 294-8 (2014).
157. Gudmundsson, J. *et al.* A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nat Genet* **44**, 1326-9 (2012).
158. Lim, E.T. *et al.* Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet* **10**, e1004494 (2014).
159. Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J.D. & Lin, X. Sequence kernel association tests for the combined effect of rare and common variants. *Am J Hum Genet* **92**, 841-53 (2013).
160. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* **91**, 224-37 (2012).
161. Li, A.H. *et al.* Analysis of loss-of-function variants and 20 risk factor phenotypes in 8,554 individuals identifies loci influencing chronic disease. *Nat Genet* (2015).
162. Do, R. *et al.* Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* **518**, 102-6 (2015).
163. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv* (2015).
164. Alsalem, A.B., Halees, A.S., Anazi, S., Alshamekh, S. & Alkuraya, F.S. Autozygome sequencing expands the horizon of human knockout research and provides novel insights into human phenotypic variation. *PLoS Genet* **9**, e1004030 (2013).
165. Sulem, P. *et al.* Identification of a large set of rare complete human knockouts. *Nat Genet* **47**, 448-52 (2015).

166. Kaiser, V.B. *et al.* Homozygous loss-of-function variants in European cosmopolitan and isolate populations. *Hum Mol Genet* (2015).
167. Jeroncic, A. *et al.* Whole-exome sequencing in an isolated population from the Dalmatian island of Vis. *Eur J Hum Genet* **24**, 1479-87 (2016).
168. Fujikura, K. Multiple loss-of-function variants of taste receptors in modern humans. *Sci Rep* **5**, 12349 (2015).
169. Saleheen, D. *et al.* Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* **544**, 235-239 (2017).
170. Narasimhan, V.M. *et al.* Health and population effects of rare gene knockouts in adult humans with related parents. *Science* **352**, 474-7 (2016).
171. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
172. Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
173. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216-20 (2013).
174. Levy-Lahad, E. *et al.* Founder BRCA1 and BRCA2 mutations in Ashkenazi Jews in Israel: frequency and differential penetrance in ovarian cancer and in breast-ovarian cancer families. *Am J Hum Genet* **60**, 1059-67 (1997).
175. Koike-Yusa, H., Li, Y., Tan, E.P., Velasco-Herrera Mdel, C. & Yusa, K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat Biotechnol* **32**, 267-73 (2014).
176. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84-7 (2014).
177. Zhou, Y. *et al.* High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature* **509**, 487-91 (2014).
178. Hart, T. *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**, 1515-26 (2015).
179. Noorani, I. 140 Genome-wide CRISPR/cas9 Knockout Screens in Human Glioblastoma Identify Genetic Vulnerabilities. *Neurosurgery* **63 Suppl 1**, 157 (2016).
180. Tzelepis, K. *et al.* A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. *Cell Rep* **17**, 1193-1205 (2016).
181. Joung, J. *et al.* Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. *Nat Protoc* **12**, 828-863 (2017).
182. Kamb, A., Harper, S. & Stefansson, K. Human genetics as a foundation for innovative drug development. *Nat Biotechnol* **31**, 975-8 (2013).
183. Dewey, F.E. *et al.* Genetic and Pharmacologic Inactivation of ANGPTL3 and Cardiovascular Disease. *N Engl J Med* (2017).
184. Bernstein, B.E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**, 1045-8 (2010).
185. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
186. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci U S A* **90**, 11995-9 (1993).
187. Jones, P.A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* **13**, 484-92 (2012).

188. Groudine, M., Eisenman, R. & Weintraub, H. Chromatin structure of endogenous retroviral genes and activation by an inhibitor of DNA methylation. *Nature* **292**, 311-7 (1981).
189. Jones, P.A. & Taylor, S.M. Cellular differentiation, cytidine analogs and DNA methylation. *Cell* **20**, 85-93 (1980).
190. Lev Maor, G., Yearim, A. & Ast, G. The alternative role of DNA methylation in splicing regulation. *Trends Genet* **31**, 274-80 (2015).
191. Maunakea, A.K. *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466**, 253-7 (2010).
192. Robertson, K.D. DNA methylation and human disease. *Nat Rev Genet* **6**, 597-610 (2005).
193. Song, C.X. & He, C. Potential functional roles of DNA demethylation intermediates. *Trends Biochem Sci* **38**, 480-4 (2013).
194. Kurdyukov, S. & Bullock, M. DNA Methylation Analysis: Choosing the Right Method. *Biology (Basel)* **5**(2016).
195. Li, B., Carey, M. & Workman, J.L. The role of chromatin during transcription. *Cell* **128**, 707-19 (2007).
196. Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693-705 (2007).
197. Venkatesh, S. & Workman, J.L. Histone exchange, chromatin structure and the regulation of transcription. *Nat Rev Mol Cell Biol* **16**, 178-89 (2015).
198. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215-6 (2012).
199. Zhou, V.W., Goren, A. & Bernstein, B.E. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* **12**, 7-18 (2011).
200. Tsompana, M. & Buck, M.J. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin* **7**, 33 (2014).
201. Schones, D.E. & Zhao, K. Genome-wide approaches to studying chromatin modifications. *Nat Rev Genet* **9**, 179-91 (2008).
202. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109-13 (2012).
203. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* **47**, 598-606 (2015).
204. Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res* **25**, 582-97 (2015).
205. Schoenfelder, S. *et al.* Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet* **42**, 53-61 (2010).
206. Javierre, B.M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384 e19 (2016).
207. Dekker, J., Marti-Renom, M.A. & Mirny, L.A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* **14**, 390-403 (2013).
208. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458-72 (2012).
209. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-80 (2012).

210. Nora, E.P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381-5 (2012).
211. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-80 (2014).
212. Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, 116-20 (2012).
213. Bujold, D. *et al.* The International Human Epigenome Consortium Data Portal. *Cell Syst* **3**, 496-499 e2 (2016).
214. Fernandez, J.M. *et al.* The BLUEPRINT Data Analysis Portal. *Cell Syst* **3**, 491-495 e5 (2016).
215. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).
216. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-9 (2011).
217. Farh, K.K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337-43 (2015).
218. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res* **22**, 1748-59 (2012).
219. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-61 (2014).
220. Hoffman, M.M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **9**, 473-6 (2012).
221. Hoffman, M.M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* **41**, 827-41 (2013).
222. Stricker, S.H., Koferle, A. & Beck, S. From profiles to function in epigenomics. *Nat Rev Genet* **18**, 51-66 (2017).
223. Tak, Y.G. & Farnham, P.J. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics Chromatin* **8**, 57 (2015).
224. Whalen, S., Truty, R.M. & Pollard, K.S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet* **48**, 488-96 (2016).
225. 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).
226. Sheffield, N.C. *et al.* Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res* **23**, 777-88 (2013).
227. He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer-promoter interactome in human cells. *Proc Natl Acad Sci U S A* **111**, E2191-9 (2014).
228. Corradin, O. & Scacheri, P.C. Enhancer variants: evaluating functions in common disease. *Genome Med* **6**, 85 (2014).
229. Ritchie, G.R., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat Methods* **11**, 294-6 (2014).

230. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**, 1413-5 (2008).
231. Oshlack, A., Robinson, M.D. & Young, M.D. From RNA-seq reads to differential expression results. *Genome Biol* **11**, 220 (2010).
232. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat Biotechnol* **26**, 1135-45 (2008).
233. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).
234. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol* **17**, 13 (2016).
235. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-5 (2010).
236. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31**, 46-53 (2013).
237. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
238. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-40 (2010).
239. Jonas, S. & Izaurralde, E. Towards a molecular understanding of microRNA-mediated gene silencing. *Nat Rev Genet* **16**, 421-33 (2015).
240. Ha, M. & Kim, V.N. Regulation of microRNA biogenesis. *Nat Rev Mol Cell Biol* **15**, 509-24 (2014).
241. Mackowiak, S.D. Identification of novel and known miRNAs in deep-sequencing data with miRDeep2. *Curr Protoc Bioinformatics* **Chapter 12**, Unit 12 10 (2011).
242. Nicolae, D.L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* **6**, e1000888 (2010).
243. Zhong, H. *et al.* Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet* **6**, e1000932 (2010).
244. Ramasamy, A. *et al.* Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat Neurosci* **17**, 1418-28 (2014).
245. Zhang, X. *et al.* Genetic associations with expression for genes implicated in GWAS studies for atherosclerotic cardiovascular disease and blood phenotypes. *Hum Mol Genet* **23**, 782-95 (2014).
246. Consortium, G.T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-60 (2015).
247. Zhernakova, D.V. *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nat Genet* **49**, 139-145 (2017).
248. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).
249. Pickrell, J.K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet* **48**, 709-17 (2016).
250. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414 e24 (2016).

251. Fairfax, B.P. *et al.* Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat Genet* **44**, 502-10 (2012).
252. Dimas, A.S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325**, 1246-50 (2009).
253. Nica, A.C. *et al.* The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet* **7**, e1002003 (2011).
254. Barreiro, L.B. *et al.* Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proc Natl Acad Sci U S A* **109**, 1204-9 (2012).
255. Fu, J. *et al.* Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet* **8**, e1002431 (2012).
256. Montgomery, S.B., Lappalainen, T., Gutierrez-Arcelus, M. & Dermitzakis, E.T. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet* **7**, e1002144 (2011).
257. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-11 (2013).
258. Cavalli, M. *et al.* Allele-specific transcription factor binding to common and rare variants associated with disease and gene expression. *Hum Genet* **135**, 485-97 (2016).
259. McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science* **342**, 747-9 (2013).
260. Maurano, M.T. *et al.* Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat Genet* **47**, 1393-401 (2015).
261. Adoue, V. *et al.* Allelic expression mapping across cellular lineages to establish impact of non-coding SNPs. *Mol Syst Biol* **10**, 754 (2014).
262. Kumasaka, N., Knights, A.J. & Gaffney, D.J. Corrigendum: Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat Genet* **48**, 473 (2016).
263. Ran, F.A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**, 2281-308 (2013).
264. Scherer, S. & Davis, R.W. Replacement of chromosome segments with altered DNA sequences constructed in vitro. *Proc Natl Acad Sci U S A* **76**, 4951-5 (1979).
265. Urnov, F.D., Rebar, E.J., Holmes, M.C., Zhang, H.S. & Gregory, P.D. Genome editing with engineered zinc finger nucleases. *Nat Rev Genet* **11**, 636-46 (2010).
266. Lieber, M.R. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu Rev Biochem* **79**, 181-211 (2010).
267. Brandsma, I. & Gent, D.C. Pathway choice in DNA double strand break repair: observations of a balancing act. *Genome Integr* **3**, 9 (2012).
268. Doudna, J.A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096 (2014).
269. Chevalier, B.S. *et al.* Design, activity, and structure of a highly specific artificial endonuclease. *Mol Cell* **10**, 895-905 (2002).
270. Li, L., Wu, L.P. & Chandrasegaran, S. Functional domains in Fok I restriction endonuclease. *Proc Natl Acad Sci U S A* **89**, 4275-9 (1992).
271. Kim, H. & Kim, J.S. A guide to genome engineering with programmable nucleases. *Nat Rev Genet* **15**, 321-34 (2014).
272. Holkers, M. *et al.* Differential integrity of TALE nuclease genes following adenoviral and lentiviral vector gene transfer into human cells. *Nucleic Acids Res* **41**, e63 (2013).

273. Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *J Bacteriol* **169**, 5429-33 (1987).
274. Mojica, F.J., Diez-Villasenor, C., Soria, E. & Juez, G. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol Microbiol* **36**, 244-6 (2000).
275. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S.D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551-61 (2005).
276. Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **60**, 174-82 (2005).
277. Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653-63 (2005).
278. Tang, T.H. *et al.* Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus. *Proc Natl Acad Sci U S A* **99**, 7536-41 (2002).
279. Makarova, K.S., Grishin, N.V., Shabalina, S.A., Wolf, Y.I. & Koonin, E.V. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct* **1**, 7 (2006).
280. Haft, D.H., Selengut, J., Mongodin, E.F. & Nelson, K.E. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol* **1**, e60 (2005).
281. Jansen, R., Embden, J.D., Gaastra, W. & Schouls, L.M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43**, 1565-75 (2002).
282. van der Oost, J., Westra, E.R., Jackson, R.N. & Wiedenheft, B. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nat Rev Microbiol* **12**, 479-92 (2014).
283. Makarova, K.S. *et al.* Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* **9**, 467-77 (2011).
284. Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602-7 (2011).
285. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-21 (2012).
286. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc Natl Acad Sci U S A* **109**, E2579-86 (2012).
287. Sander, J.D. & Joung, J.K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat Biotechnol* **32**, 347-55 (2014).
288. Cho, S.W., Kim, S., Kim, J.M. & Kim, J.S. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat Biotechnol* **31**, 230-2 (2013).
289. 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**, 21312-24 (2014).
290. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096-101 (2015).

291. Aguirre, A.J. *et al.* Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. *Cancer Discov* **6**, 914-29 (2016).
292. Gilbert, L.A. *et al.* Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**, 647-61 (2014).
293. Konermann, S. *et al.* Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**, 583-8 (2015).
294. Fulco, C.P. *et al.* Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769-773 (2016).
295. Choi, P.S. & Meyerson, M. Targeted genomic rearrangements using CRISPR/Cas technology. *Nat Commun* **5**, 3728 (2014).
296. Hoban, M.D., Orkin, S.H. & Bauer, D.E. Genetic treatment of a molecular disorder: gene therapy approaches to sickle cell disease. *Blood* **127**, 839-48 (2016).
297. Tabebordbar, M. *et al.* In vivo gene editing in dystrophic mouse muscle and muscle stem cells. *Science* **351**, 407-11 (2016).
298. Nelson, C.E. *et al.* In vivo genome editing improves muscle function in a mouse model of Duchenne muscular dystrophy. *Science* **351**, 403-7 (2016).
299. Long, C. *et al.* Postnatal genome editing partially restores dystrophin expression in a mouse model of muscular dystrophy. *Science* **351**, 400-3 (2016).
300. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med* **373**, 895-907 (2015).
301. Hardison, R.C. & Blobel, G.A. Genetics. GWAS to therapy by genome edits? *Science* **342**, 206-7 (2013).
302. Alkuraya, F.S. Human knockout research: new horizons and opportunities. *Trends Genet* **31**, 108-115 (2015).
303. Antonarakis, S.E. & Beckmann, J.S. Mendelian disorders deserve more attention. *Nat Rev Genet* **7**, 277-82 (2006).
304. Amselem, S. *et al.* Laron dwarfism and mutations of the growth hormone-receptor gene. *N Engl J Med* **321**, 989-95 (1989).
305. Clement, K. *et al.* A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction. *Nature* **392**, 398-401 (1998).
306. Jackson, R.S. *et al.* Small-intestinal dysfunction accompanies the complex endocrinopathy of human proprotein convertase 1 deficiency. *J Clin Invest* **112**, 1550-60 (2003).
307. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
308. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-9 (2012).
309. Gorlov, I.P., Gorlova, O.Y., Sunyaev, S.R., Spitz, M.R. & Amos, C.I. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet* **82**, 100-12 (2008).
310. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-97 (2007).
311. Hogan, M.C. *et al.* PKHDL1, a homolog of the autosomal recessive polycystic kidney disease gene, encodes a receptor with inducible T lymphocyte expression. *Hum Mol Genet* **12**, 685-98 (2003).

312. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).
313. Cramer, S.D., Ferree, P.M., Lin, K., Milliner, D.S. & Holmes, R.P. The gene encoding hydroxypyruvate reductase (GRHPR) is mutated in patients with primary hyperoxaluria type II. *Hum Mol Genet* **8**, 2063-9 (1999).
314. Nissel, R. *et al.* Body growth after combined liver-kidney transplantation in children with primary hyperoxaluria type 1. *Transplantation* **82**, 48-54 (2006).
315. Dent, C.E. & Stamp, T.C. Treatment of primary hyperoxaluria. *Arch Dis Child* **45**, 735-45 (1970).
316. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. & Goldstein, D.B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* **9**, e1003709 (2013).
317. Majithia, A.R. *et al.* Rare variants in PPARG with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. *Proc Natl Acad Sci U S A* **111**, 13127-32 (2014).
318. Thormaehlen, A.S. *et al.* Systematic cell-based phenotyping of missense alleles empowers rare variant association studies: a case for LDLR and myocardial infarction. *PLoS Genet* **11**, e1004855 (2015).
319. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum Mutat* **34**, E2393-402 (2013).
320. Lettre, G., Lange, C. & Hirschhorn, J.N. Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet Epidemiol* **31**, 358-62 (2007).
321. Eppig, J.T. *et al.* The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Res* **43**, D726-36 (2015).
322. Holliday, R., Ho, T. & Paulin, R. Gene silencing in mammalian cells. in *Epigenetic mechanisms of gene regulation* (eds. Russo, V.E.A., Martienssen, R.A. & Riggs, A.D.) 47-60 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, USA, 1996).
323. Bock, C. Analysing and interpreting DNA methylation data. *Nat Rev Genet* **13**, 705-19 (2012).
324. Heyn, H. & Esteller, M. DNA methylation profiling in the clinic: applications and challenges. *Nat Rev Genet* **13**, 679-92 (2012).
325. Trowbridge, J.J., Snow, J.W., Kim, J. & Orkin, S.H. DNA methyltransferase 1 is essential for and uniquely regulates hematopoietic stem and progenitor cells. *Cell Stem Cell* **5**, 442-9 (2009).
326. Challen, G.A. *et al.* Dnmt3a is essential for hematopoietic stem cell differentiation. *Nat Genet* **44**, 23-31 (2012).
327. Hogart, A. *et al.* Genome-wide DNA methylation profiles in hematopoietic stem and progenitor cells reveal overrepresentation of ETS transcription factor binding sites. *Genome Res* **22**, 1407-18 (2012).
328. Shearstone, J.R. *et al.* Global DNA demethylation during mouse erythropoiesis in vivo. *Science* **334**, 799-802 (2011).
329. Ziller, M.J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477-81 (2013).

330. Yu, Y. *et al.* High resolution methylome analysis reveals widespread functional hypomethylation during adult human erythropoiesis. *J Biol Chem* **288**, 8805-14 (2013).
331. Maksimovic, J., Gordon, L. & Oshlack, A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol* **13**, R44 (2012).
332. Fortin, J.P., Greenwood, C.M.T. & Labbe, A. ARRmNormalization: Adaptive Robust Regression normalization for Illumina methylation data. in *R package* 1.0.0 edn (2013).
333. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010).
334. Xu, J. *et al.* Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Dev Cell* **23**, 796-811 (2012).
335. Huang da, W., Sherman, B.T. & Lempicki, R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1-13 (2009).
336. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).
337. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-89 (2010).
338. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
339. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu Rev Genomics Hum Genet* **10**, 387-406 (2009).
340. 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**, 1815-22 (2010).
341. Sankaran, V.G. *et al.* Developmental and species-divergent globin switching are driven by BCL11A. *Nature* **460**, 1093-7 (2009).
342. Zhou, D., Liu, K., Sun, C.W., Pawlik, K.M. & Townes, T.M. KLF1 regulates BCL11A expression and gamma- to beta-globin gene switching. *Nat Genet* **42**, 742-4 (2010).
343. Mabaera, R. *et al.* Developmental- and differentiation-specific patterns of human gamma- and beta-globin promoter DNA methylation. *Blood* **110**, 1343-52 (2007).
344. Shi, L. *et al.* Developmental transcriptome analysis of human erythropoiesis. *Hum Mol Genet* (2014).
345. An, X. *et al.* Global transcriptome analyses of human and murine terminal erythroid differentiation. *Blood* **123**, 3466-77 (2014).
346. Inaba, N. *et al.* A novel I-branching beta-1,6-N-acetylglucosaminyltransferase involved in human blood group I antigen expression. *Blood* **101**, 2870-6 (2003).
347. Webb, C.F. *et al.* The ARID family transcription factor bright is required for both hematopoietic stem cell and B lineage development. *Mol Cell Biol* **31**, 1041-53 (2011).
348. Kulis, M. *et al.* Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat Genet* **44**, 1236-42 (2012).
349. Lee, Y.T. *et al.* LIN28B-mediated expression of fetal hemoglobin and production of fetal-like erythrocytes from adult human erythroblasts ex vivo. *Blood* **122**, 1034-41 (2013).

350. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82 (2012).
351. Xu, J. *et al.* Transcriptional silencing of {gamma}-globin by BCL11A involves long-range interactions and cooperation with SOX6. *Genes Dev* **24**, 783-98 (2010).
352. Orkin, S.H. & Zon, L.I. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**, 631-44 (2008).
353. Palis, J. Primitive and definitive erythropoiesis in mammals. *Front Physiol* **5**, 3 (2014).
354. Duan, Z., Stamatoyannopoulos, G. & Li, Q. Role of NF-Y in in vivo regulation of the gamma-globin gene. *Mol Cell Biol* **21**, 3083-95 (2001).
355. Zhou, Z. *et al.* USF and NF-E2 cooperate to regulate the recruitment and activity of RNA polymerase II in the beta-globin gene locus. *J Biol Chem* **285**, 15894-905 (2010).
356. Ichikawa, M. *et al.* AML1/Runx1 negatively regulates quiescent hematopoietic stem cells in adult hematopoiesis. *J Immunol* **180**, 4402-8 (2008).
357. Macari, E.R. & Lowrey, C.H. Induction of human fetal hemoglobin via the NRF2 antioxidant response signaling pathway. *Blood* **117**, 5987-97 (2011).
358. Starnes, L.M. *et al.* NFI-A directs the fate of hematopoietic progenitors to the erythroid or granulocytic lineage and controls beta-globin and G-CSF receptor expression. *Blood* **114**, 1753-63 (2009).
359. DeSimone, J., Heller, P., Hall, L. & Zwiers, D. 5-Azacytidine stimulates fetal hemoglobin synthesis in anemic baboons. *Proc Natl Acad Sci U S A* **79**, 4428-31 (1982).
360. Ley, T.J. *et al.* 5-azacytidine selectively increases gamma-globin synthesis in a patient with beta+ thalassemia. *N Engl J Med* **307**, 1469-75 (1982).
361. Gamazon, E.R. *et al.* Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. *Mol Psychiatry* **18**, 340-6 (2013).
362. Grundberg, E. *et al.* Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am J Hum Genet* **93**, 876-90 (2013).
363. Sankaran, V.G. *et al.* A functional element necessary for fetal hemoglobin silencing. *N Engl J Med* **365**, 807-14 (2011).
364. Guibert, S. & Weber, M. Functions of DNA methylation and hydroxymethylation in mammalian development. *Curr Top Dev Biol* **104**, 47-83 (2013).
365. Modell, B. & Darlison, M. Global epidemiology of haemoglobin disorders and derived service indicators. *Bull World Health Organ* **86**, 480-7 (2008).
366. Brady, H.J. *et al.* Expression of the human carbonic anhydrase I gene is activated late in fetal erythroid development and regulated by stage-specific trans-acting factors. *Br J Haematol* **76**, 135-42 (1990).
367. Azzouzi, I. *et al.* MicroRNA-96 directly inhibits gamma-globin expression in human erythropoiesis. *PLoS One* **6**, e22838 (2011).
368. Noh, S.J. *et al.* Let-7 microRNAs are developmentally regulated in circulating human erythroid cells. *J Transl Med* **7**, 98 (2009).
369. Piskounova, E. *et al.* Determinants of microRNA processing inhibition by the developmentally regulated RNA-binding protein Lin28. *J Biol Chem* **283**, 21310-4 (2008).
370. Heo, I. *et al.* Lin28 mediates the terminal uridylation of let-7 precursor MicroRNA. *Mol Cell* **32**, 276-84 (2008).

371. Lulli, V. *et al.* MicroRNA-486-3p regulates gamma-globin expression in human erythroid cells by directly modulating BCL11A. *PLoS One* **8**, e60436 (2013).
372. Saki, N. *et al.* MicroRNA Expression in beta-Thalassemia and Sickle Cell Disease: A Role in The Induction of Fetal Hemoglobin. *Cell J* **17**, 583-92 (2016).
373. Walker, A.L. *et al.* Epigenetic and molecular profiles of erythroid cells after hydroxyurea treatment in sickle cell anemia. *Blood* **118**, 5664-70 (2011).
374. Alijani, S. *et al.* Evaluation of the Effect of miR-26b Up-Regulation on HbF Expression in Erythroleukemic K-562 Cell Line. *Avicenna J Med Biotechnol* **6**, 53-6 (2014).
375. Goh, S.H. *et al.* The human reticulocyte transcriptome. *Physiol Genomics* **30**, 172-8 (2007).
376. Doss, J.F. *et al.* A comprehensive joint analysis of the long and short RNA transcriptomes of human erythrocytes. *BMC Genomics* **16**, 952 (2015).
377. Lessard, S., Beaudoin, M., Benkirane, K. & Lettre, G. Comparison of DNA methylation profiles in human fetal and adult red blood cell progenitors. *Genome Med* **7**, 1 (2015).
378. Friedlander, M.R., Mackowiak, S.D., Li, N., Chen, W. & Rajewsky, N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* **40**, 37-52 (2012).
379. Ru, Y. *et al.* The multiMiR R package and database: integration of microRNA-target interactions along with their disease and drug associations. *Nucleic Acids Res* **42**, e133 (2014).
380. Luo, W., Friedman, M.S., Shedden, K., Hankenson, K.D. & Woolf, P.J. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* **10**, 161 (2009).
381. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).
382. Livak, K.J. & Schmittgen, T.D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402-8 (2001).
383. Polesskaya, A. *et al.* Lin-28 binds IGF-2 mRNA and participates in skeletal myogenesis by increasing translation efficiency. *Genes Dev* **21**, 1125-38 (2007).
384. Chou, S. & Lodish, H.F. Fetal liver hepatic progenitors are supportive stromal cells for hematopoietic stem cells. *Proc Natl Acad Sci U S A* **107**, 7799-804 (2010).
385. Tumburu, L. *et al.* IGF2BP1 Reverses Hemoglobin Switching in Adult Erythroblasts. *Blood* **126**, 639-639 (2015).
386. Krauss, J.S., Hahn, D.A., Jonah, M.H. & Trinh, M. Estimation of highly increased concentrations of fetal hemoglobin in Fanconi's anemia. *Clin Chem* **31**, 1737-8 (1985).
387. Alter, B.P. *et al.* Genetic regulation of fetal haemoglobin in inherited bone marrow failure syndromes. *Br J Haematol* **162**, 542-6 (2013).
388. Bhatia, H. *et al.* Short-chain fatty acid-mediated effects on erythropoiesis in primary definitive erythroid cells. *Blood* **113**, 6440-8 (2009).
389. Ikuta, T., Kuroyanagi, Y., Odo, N. & Liu, S. A common signaling pathway is activated in erythroid cells expressing high levels of fetal hemoglobin: a potential role for cAMP-elevating agents in beta-globin disorders. *J Blood Med* **4**, 149-59 (2013).
390. Shaham, L. *et al.* MicroRNA-486-5p is an erythroid oncomiR of the myeloid leukemias of Down syndrome. *Blood* **125**, 1292-301 (2015).
391. Agueli, C. *et al.* 14q32/miRNA clusters loss of heterozygosity in acute lymphoblastic leukemia is associated with up-regulation of BCL11a. *Am J Hematol* **85**, 575-8 (2010).

392. Malik, T.H. *et al.* Transcriptional repression and developmental functions of the atypical vertebrate GATA protein TRPS1. *EMBO J* **20**, 1715-25 (2001).
393. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714-9 (2010).
394. Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-5 (2012).
395. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-47 (2013).
396. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).
397. Timmann, C. *et al.* Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature* **489**, 443-6 (2012).
398. Li, J. *et al.* GWAS of blood cell traits identifies novel associated loci and epistatic interactions in Caucasian and African-American children. *Hum Mol Genet* **22**, 1457-64 (2013).
399. Schuh, K. *et al.* Plasma membrane Ca2+ ATPase 4 is required for sperm motility and male fertility. *J Biol Chem* **279**, 28220-6 (2004).
400. Mohamed, T.M. *et al.* The plasma membrane calcium ATPase 4 signalling in cardiac fibroblasts mediates cardiomyocyte hypertrophy. *Nat Commun* **7**, 11074 (2016).
401. Zambo, B. *et al.* Decreased calcium pump expression in human erythrocytes is connected to a minor haplotype in the ATP2B4 gene. *Cell Calcium* (2017).
402. Tiffert, T. *et al.* The hydration state of human red blood cells and their susceptibility to invasion by Plasmodium falciparum. *Blood* **105**, 4853-60 (2005).
403. Panousis, N.I., Gutierrez-Arcelus, M., Dermitzakis, E.T. & Lappalainen, T. Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. *Genome Biol* **15**, 467 (2014).
404. Mathelier, A. *et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **44**, D110-5 (2016).
405. Chen, J., Bardes, E.E., Aronow, B.J. & Jegga, A.G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* **37**, W305-11 (2009).
406. Abraham, G. & Inouye, M. Fast principal component analysis of large-scale genome-wide data. *PLoS One* **9**, e93766 (2014).
407. Brinkman, E.K., Chen, T., Amendola, M. & van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res* **42**, e168 (2014).
408. Busch, B. *et al.* The oncogenic triangle of HMGA2, LIN28B and IGF2BP1 antagonizes tumor-suppressive actions of the let-7 family. *Nucleic Acids Res* **44**, 3845-64 (2016).
409. Roush, S. & Slack, F.J. The let-7 family of microRNAs. *Trends Cell Biol* **18**, 505-16 (2008).
410. Gazarini, M.L., Thomas, A.P., Pozzan, T. & Garcia, C.R. Calcium signaling in a low calcium environment: how the intracellular malaria parasite solves the problem. *J Cell Biol* **161**, 103-10 (2003).
411. Tiffert, T., Staines, H.M., Ellory, J.C. & Lew, V.L. Functional state of the plasma membrane Ca2+ pump in Plasmodium falciparum-infected human red blood cells. *J Physiol* **525 Pt 1**, 125-34 (2000).

412. Howes, R.E. *et al.* The global distribution of the Duffy blood group. *Nat Commun* **2**, 266 (2011).
413. Mieth, B. *et al.* Combining Multiple Hypothesis Testing with Machine Learning Increases the Statistical Power of Genome-wide Association Studies. *Sci Rep* **6**, 36671 (2016).
414. Okser, S. *et al.* Regularized machine learning in the genetic prediction of complex traits. *PLoS Genet* **10**, e1004754 (2014).
415. 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**, 1961-2 (2012).
416. Liu, D. *et al.* KLF1 mutations are relatively more common in a thalassemia endemic region and ameliorate the severity of beta-thalassemia. *Blood* **124**, 803-11 (2014).
417. Huang, J. *et al.* Compound heterozygosity for KLF1 mutations is associated with microcytic hypochromic anemia and increased fetal hemoglobin. *Eur J Hum Genet* **23**, 1341-8 (2015).
418. Westra, H.J. *et al.* Cell Specific eQTL Analysis without Sorting Cells. *PLoS Genet* **11**, e1005223 (2015).
419. Psaila, B. *et al.* Single-cell profiling of human megakaryocyte-erythroid progenitors identifies distinct megakaryocyte and erythroid differentiation pathways. *Genome Biol* **17**, 83 (2016).
420. Liu, S. & Trapnell, C. Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res* **5**(2016).
421. Wills, Q.F. *et al.* Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat Biotechnol* **31**, 748-52 (2013).
422. Ding, Y., Li, H., Chen, L.-L. & Xie, K. Recent Advances in Genome Editing Using CRISPR/Cas9. *Frontiers in Plant Science* **7**(2016).
423. Kleinstiver, B.P. *et al.* Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481-5 (2015).
424. Tsai, S.Q. & Joung, J.K. Defining and improving the genome-wide specificities of CRISPR-Cas9 nucleases. *Nat Rev Genet* **17**, 300-12 (2016).
425. Tsai, S.Q. *et al.* Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat Biotechnol* **32**, 569-76 (2014).
426. Kleinstiver, B.P. *et al.* High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490-495 (2016).
427. Tsai, S.Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol* **33**, 187-97 (2015).
428. Haeussler, M. *et al.* Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol* **17**, 148 (2016).
429. Doench, J.G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol* **34**, 184-91 (2016).
430. Sanjana, N.E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat Methods* **11**, 783-4 (2014).
431. Chari, R., Mali, P., Moosburner, M. & Church, G.M. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat Methods* **12**, 823-6 (2015).
432. Bae, S., Kweon, J., Kim, H.S. & Kim, J.S. Microhomology-based choice of Cas9 nuclease target sites. *Nat Methods* **11**, 705-6 (2014).

433. Zetsche, B., Volz, S.E. & Zhang, F. A split-Cas9 architecture for inducible genome editing and transcription modulation. *Nat Biotechnol* **33**, 139-42 (2015).
434. Nihongaki, Y., Kawano, F., Nakajima, T. & Sato, M. Photoactivatable CRISPR-Cas9 for optogenetic genome editing. *Nat Biotechnol* **33**, 755-60 (2015).
435. Davis, K.M., Pattanayak, V., Thompson, D.B., Zuris, J.A. & Liu, D.R. Small molecule-triggered Cas9 protein with improved genome-editing specificity. *Nat Chem Biol* **11**, 316-8 (2015).
436. Senturk, S. *et al.* Rapid and tunable method to temporally control gene editing based on conditional Cas9 stabilization. *Nat Commun* **8**, 14370 (2017).
437. Ferry, Q.R., Lyutova, R. & Fulga, T.A. Rational design of inducible CRISPR guide RNAs for de novo assembly of transcriptional programs. *Nat Commun* **8**, 14633 (2017).
438. Qi, L.S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173-83 (2013).
439. Cheng, A.W. *et al.* Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. *Cell Res* **23**, 1163-71 (2013).
440. Vora, S., Tuttle, M., Cheng, J. & Church, G. Next stop for the CRISPR revolution: RNA-guided epigenetic regulators. *FEBS J* **283**, 3181-93 (2016).
441. Kearns, N.A. *et al.* Functional annotation of native enhancers with a Cas9-histone demethylase fusion. *Nat Methods* **12**, 401-3 (2015).
442. Hilton, I.B. *et al.* Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat Biotechnol* **33**, 510-7 (2015).
443. Ma, H. *et al.* Multicolor CRISPR labeling of chromosomal loci in human cells. *Proc Natl Acad Sci U S A* **112**, 3002-7 (2015).
444. Abudayyeh, O.O. *et al.* C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **353**, aaf5573 (2016).
445. O'Connell, M.R. *et al.* Programmable RNA recognition and cleavage by CRISPR/Cas9. *Nature* **516**, 263-6 (2014).
446. Chu, V.T. *et al.* Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nat Biotechnol* **33**, 543-8 (2015).
447. Richardson, C.D., Ray, G.J., DeWitt, M.A., Curie, G.L. & Corn, J.E. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nat Biotechnol* **34**, 339-44 (2016).
448. Zhang, J.P. *et al.* Efficient precise knockin with a double cut HDR donor after CRISPR/Cas9-mediated double-stranded DNA cleavage. *Genome Biol* **18**, 35 (2017).
449. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**, 271-7 (2012).
450. Ulirsch, J.C. *et al.* Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* **165**, 1530-45 (2016).
451. Wakabayashi, A. *et al.* Insight into GATA1 transcriptional activity through interrogation of cis elements disrupted in human erythroid disorders. *Proceedings of the National Academy of Sciences* **113**, 4434-4439 (2016).
452. Korkmaz, G. *et al.* Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat Biotechnol* **34**, 192-8 (2016).
453. 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**, 120-3 (2014).

454. 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**, 900-13 (2016).
455. Lanzkron, S., Carroll, C.P. & Haywood, C., Jr. Mortality rates and age at death from sickle cell disease: U.S., 1979-2005. *Public Health Rep* **128**, 110-6 (2013).
456. Okam, M.M. *et al.* Phase 1/2 trial of vorinostat in patients with sickle cell disease who have not benefitted from hydroxyurea. *Blood* **125**, 3668-9 (2015).
457. DeWitt, M.A. *et al.* Selection-free genome editing of the sickle mutation in human adult hematopoietic stem/progenitor cells. *Sci Transl Med* **8**, 360ra134 (2016).
458. Bhatia, M. & Sheth, S. Hematopoietic stem cell transplantation in sickle cell disease: patient selection and special considerations. *J Blood Med* **6**, 229-38 (2015).
459. Dallas, M.H. *et al.* Long-term outcome and evaluation of organ function in pediatric patients undergoing haploidentical and matched related hematopoietic cell transplantation for sickle cell disease. *Biol Blood Marrow Transplant* **19**, 820-30 (2013).
460. Locatelli, F. *et al.* Related umbilical cord blood transplantation in patients with thalassemia and sickle cell disease. *Blood* **101**, 2137-43 (2003).
461. Radhakrishnan, K. *et al.* Busulfan, fludarabine, and alemtuzumab conditioning and unrelated cord blood transplantation in children with sickle cell disease. *Biol Blood Marrow Transplant* **19**, 676-7 (2013).
462. Kauf, T.L., Coates, T.D., Huazhi, L., Mody-Patel, N. & Hartzema, A.G. The cost of health care for children and adults with sickle cell disease. *Am J Hematol* **84**, 323-7 (2009).
463. Ngolet, L.O. *et al.* Sickle-Cell Disease Healthcare Cost in Africa: Experience of the Congo. *Anemia* **2016**, 2046535 (2016).
464. Sampath, S. *et al.* Glycan masking of Plasmodium vivax Duffy Binding Protein for probing protein binding function and vaccine development. *PLoS Pathog* **9**, e1003420 (2013).
465. Preuss, J. *et al.* Identification and characterization of novel human glucose-6-phosphate dehydrogenase inhibitors. *J Biomol Screen* **18**, 286-97 (2013).
466. Crosnier, C. *et al.* Basigin is a receptor essential for erythrocyte invasion by Plasmodium falciparum. *Nature* **480**, 534-7 (2011).
467. Tran, T.M. *et al.* Naturally acquired antibodies specific for Plasmodium falciparum reticulocyte-binding protein homologue 5 inhibit parasite growth and predict protection from malaria. *J Infect Dis* **209**, 789-98 (2014).
468. Douglas, A.D. *et al.* A PfRH5-based vaccine is efficacious against heterologous strain blood-stage Plasmodium falciparum infection in aotus monkeys. *Cell Host Microbe* **17**, 130-9 (2015).
469. Fugger, L., McVean, G. & Bell, J.I. Genomewide association studies and common disease--realizing clinical utility. *N Engl J Med* **367**, 2370-1 (2012).
470. Paul, D.S. *et al.* Maps of open chromatin guide the functional follow-up of genome-wide association signals: application to hematological traits. *PLoS Genet* **7**, e1002139 (2011).
471. Nuinoon, M. *et al.* A genome-wide association identified the common genetic variants influence disease severity in beta0-thalassemia/hemoglobin E. *Hum Genet* **127**, 303-14 (2010).
472. Bhatnagar, P. *et al.* Genome-wide association study identifies genetic variants influencing F-cell levels in sickle-cell patients. *J Hum Genet* **56**, 316-23 (2011).

473. Xu, J. *et al.* Corepressor-dependent silencing of fetal hemoglobin expression by BCL11A. *Proc Natl Acad Sci U S A* **110**, 6518-23 (2013).
474. Liu, P. *et al.* Bcl11a is essential for normal lymphoid development. *Nat Immunol* **4**, 525-32 (2003).
475. Yu, Y. *et al.* Bcl11a is essential for lymphoid development and negatively regulates p53. *J Exp Med* **209**, 2467-83 (2012).
476. Farber, M.D., Koshy, M. & Kinney, T.R. Cooperative Study of Sickle Cell Disease: Demographic and socioeconomic characteristics of patients and families with sickle cell disease. *J Chronic Dis* **38**, 495-505 (1985).
477. Kassouf, M.T. *et al.* Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome Res* **20**, 1064-83 (2010).
478. Soler, E. *et al.* The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes Dev* **24**, 277-89 (2010).
479. Turner, D.J. & Hurles, M.E. High-throughput haplotype determination over long distances by haplotype fusion PCR and ligation haplotyping. *Nat Protoc* **4**, 1771-83 (2009).
480. Tyson, J. & Armour, J.A. Determination of haplotypes at structurally complex regions using emulsion haplotype fusion PCR. *BMC Genomics* **13**, 693 (2012).
481. Leid, M. *et al.* CTIP1 and CTIP2 are differentially expressed during mouse embryogenesis. *Gene Expr Patterns* **4**, 733-9 (2004).
482. Kowalczyk, M.S. *et al.* Intragenic enhancers act as alternative promoters. *Mol Cell* **45**, 447-58 (2012).
483. Lee, H.J., Kim, E. & Kim, J.S. Targeted chromosomal deletions in human cells using zinc finger nucleases. *Genome Res* **20**, 81-9 (2010).
484. John, A. *et al.* Bcl11a is required for neuronal morphogenesis and sensory circuit formation in dorsal spinal cord development. *Development* **139**, 1831-41 (2012).
485. Patwardhan, R.P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**, 265-70 (2012).
486. Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J. Human genetic variation and its contribution to complex traits. *Nat Rev Genet* **10**, 241-51 (2009).
487. Koehler, A.N. A complex task? Direct modulation of transcription factors with small molecules. *Curr Opin Chem Biol* **14**, 331-40 (2010).
488. Joung, J.K. & Sander, J.D. TALENs: a widely applicable technology for targeted genome editing. *Nat Rev Mol Cell Biol* **14**, 49-55 (2013).
489. van den Akker, E., Satchwell, T.J., Pellegrin, S., Daniels, G. & Toye, A.M. The majority of the in vitro erythroid expansion potential resides in CD34(-) cells, outweighing the contribution of CD34(+) cells and significantly increasing the erythroblast yield from peripheral blood samples. *Haematologica* **95**, 1594-8 (2010).
490. Bredemeyer, A.L. *et al.* ATM stabilizes DNA double-strand-break complexes during V(D)J recombination. *Nature* **442**, 466-70 (2006).
491. Wu, M.C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93 (2011).
492. He, A., Kong, S.W., Ma, Q. & Pu, W.T. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proc Natl Acad Sci U S A* **108**, 5632-7 (2011).

493. McDevitt, M.A., Fujiwara, Y., Shivdasani, R.A. & Orkin, S.H. An upstream, DNase I hypersensitive region of the hematopoietic-expressed transcription factor GATA-1 gene confers developmental specificity in transgenic mice. *Proc Natl Acad Sci U S A* **94**, 7976-81 (1997).
494. Cermak, T. *et al.* Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res* **39**, e82 (2011).
495. Miller, J.C. *et al.* A TALE nuclease architecture for efficient genome editing. *Nat Biotechnol* **29**, 143-8 (2011).
496. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-23 (2013).
497. Ran, F.A. *et al.* Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380-9 (2013).
498. Hsu, P.D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* **31**, 827-32 (2013).
499. 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**, 4935-45 (2011).
500. Stadhouders, R. *et al.* HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers. *J Clin Invest* **124**, 1699-710 (2014).
501. Mtatiro, S.N. *et al.* Genome wide association study of fetal hemoglobin in sickle cell anemia in Tanzania. *PLoS One* **9**, e111464 (2014).
502. Ganesh, S.K. *et al.* Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nat Genet* **41**, 1191-8 (2009).
503. Soranzo, N. *et al.* A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat Genet* **41**, 1182-90 (2009).
504. Kamatani, Y. *et al.* Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat Genet* **42**, 210-5 (2010).
505. 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**, 101-5 (2013).
506. Esvelt, K.M. *et al.* Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat Methods* **10**, 1116-21 (2013).
507. Mali, P., Esvelt, K.M. & Church, G.M. Cas9 as a versatile tool for engineering biology. *Nat Methods* **10**, 957-63 (2013).
508. Ran, F.A. *et al.* In vivo genome editing using Staphylococcus aureus Cas9. *Nature* **520**, 186-91 (2015).
509. Kleinstiver, B.P. *et al.* Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9 by modifying PAM recognition. *Nat Biotechnol* **33**, 1293-1298 (2015).
510. Zetsche, B. *et al.* Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**, 759-71 (2015).
511. Rajagopal, N. *et al.* High-throughput mapping of regulatory DNA. *Nat Biotechnol* **34**, 167-74 (2016).
512. Pinello, L. *et al.* Analyzing CRISPR genome-editing experiments with CRISPResso. *Nat Biotechnol* **34**, 695-7 (2016).
513. 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).

514. Doench, J.G. *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol* **32**, 1262-7 (2014).
515. Shalem, O., Sanjana, N.E. & Zhang, F. High-throughput functional genomics using CRISPR-Cas9. *Nat Rev Genet* **16**, 299-311 (2015).
516. Chen, S. *et al.* Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *Cell* **160**, 1246-60 (2015).
517. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823-6 (2013).
518. Pinello, L., Xu, J., Orkin, S.H. & Yuan, G.C. Analysis of chromatin-state plasticity identifies cell-type-specific regulators of H3K27me3 patterns. *Proc Natl Acad Sci U S A* **111**, E344-53 (2014).
519. Whyte, W.A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-19 (2013).
520. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-8 (2011).
521. Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* **42**, D142-7 (2014).
522. Giarratana, M.C. *et al.* Proof of principle for transfusion of in vitro-generated red blood cells. *Blood* **118**, 5071-9 (2011).
523. Huang, Y.L. *et al.* Prognostic value of red blood cell distribution width for patients with heart failure: a systematic review and meta-analysis of cohort studies. *PLoS One* **9**, e104861 (2014).
524. Whitfield, J.B. & Martin, N.G. Genetic and environmental influences on the size and number of cells in the blood. *Genet Epidemiol* **2**, 133-44 (1985).
525. Evans, D.M., Frazer, I.H. & Martin, N.G. Genetic and environmental causes of variation in basal levels of blood cells. *Twin Res* **2**, 250-7 (1999).
526. Lin, J.P. *et al.* Evidence for linkage of red blood cell size and count: genome-wide scans in the Framingham Heart Study. *Am J Hematol* **82**, 605-10 (2007).
527. Auer, P.L. *et al.* Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. *Nat Genet* **46**, 629-34 (2014).
528. Eicher, J.D. *et al.* Platelet-related variants identified by exome chip meta-analysis in 157,293 individuals. (Submitted).
529. Tajuddin, S.M. *et al.* Large-scale exome-wide association analysis identifies loci for white blood cell traits and pleiotropy with immune-mediated diseases. (Submitted).
530. Grove, M.L. *et al.* Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium. *PLoS One* **8**, e68095 (2013).
531. Wells, Q.S. *et al.* Whole exome sequencing identifies a causal RBM20 mutation in a large pedigree with familial dilated cardiomyopathy. *Circ Cardiovasc Genet* **6**, 317-26 (2013).
532. Winkler, T.W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* **9**, 1192-212 (2014).
533. Limongelli, G., Elliott, P., Charron, P., Mogensen, J. & McKeown, P.P. Approaching genetic testing in cardiomyopathies. *ESC Council for Cardiology Practice* (2012).
534. Olson, T.M., Michels, V.V., Thibodeau, S.N., Tai, Y.S. & Keating, M.T. Actin mutations in dilated cardiomyopathy, a heritable form of heart failure. *Science* **280**, 750-2 (1998).

535. Liu, D.J. *et al.* Meta-analysis of gene-level tests for rare variant association. *Nat Genet* **46**, 200-4 (2014).
536. Price, A.L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* **86**, 832-8 (2010).
537. Zhang, X. *et al.* Synthesis of 53 tissue and cell line expression QTL datasets reveals master eQTLs. *BMC Genomics* **15**, 532 (2014).
538. Astner, I. *et al.* Crystal structure of 5-aminolevulinate synthase, the first enzyme of heme biosynthesis, and its link to XLSA in humans. *EMBO J* **24**, 3166-77 (2005).
539. Halpain, S. & Dehmelt, L. The MAP1 family of microtubule-associated proteins. *Genome Biol* **7**, 224 (2006).
540. Takei, Y., Kikkawa, Y.S., Atapour, N., Hensch, T.K. & Hirokawa, N. Defects in Synaptic Plasticity, Reduced NMDA-Receptor Transport, and Instability of Postsynaptic Density Proteins in Mice Lacking Microtubule-Associated Protein 1A. *J Neurosci* **35**, 15539-54 (2015).
541. Pankratz, N. *et al.* Meta-analysis of rare and common exome chips variants identifies S1PR4 and other novel genes influencing blood cell traits. *Nature Genetics* (In press).
542. Mukherjee, S. Multivariate analysis of whole exome sequence data identifies rare variants with pleiotropic effects on obesity-related metabolic traits in 31,000 participants of the Regeneron Genetics Center – Geisinger MyCode collaborative project – DiscovEHR. *American Society of Human Genetics (ASHG) Conference 2015*.
543. Dehghan, A. *et al.* Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels. *Circulation* **123**, 731-8 (2011).
544. Global Lipids Genetics, C. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274-83 (2013).
545. Taylor, K.C. *et al.* A gene-centric association scan for Coagulation Factor VII levels in European and African Americans: the Candidate Gene Association Resource (CARe) Consortium. *Hum Mol Genet* **20**, 3525-34 (2011).
546. de Vries, P.S. *et al.* A meta-analysis of 120 246 individuals identifies 18 new loci for fibrinogen concentration. *Hum Mol Genet* **25**, 358-70 (2016).
547. Kooner, J.S. *et al.* Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat Genet* **43**, 984-9 (2011).
548. AMP-T2D_Program, T2D-GENES_Consortium, SIGMA_T2D_Consortium, type2diabetesgenetics.org & 2016_March_17. http://www.type2diabetesgenetics.org/variantInfo/variantInfo/rs1800961.
549. Pradel, L.C., Vanhille, L. & Spicuglia, S. [The European Blueprint project: towards a full epigenome characterization of the immune system]. *Med Sci (Paris)* **31**, 236-8 (2015).
550. Aulchenko, Y.S. *et al.* Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat Genet* **41**, 47-55 (2009).
551. Kullo, I.J. *et al.* Complement receptor 1 gene variants are associated with erythrocyte sedimentation rate. *Am J Hum Genet* **89**, 131-8 (2011).
552. Dobbeling, U. The effects of cyclosporin A on V(D)J recombination activity. *Scand J Immunol* **45**, 494-8 (1997).

553. Zarza, R. *et al.* Molecular characterization of the PK-LR gene in pyruvate kinase deficient Spanish patients. Red Cell Pathology Group of the Spanish Society of Haematology (AEHH). *Br J Haematol* **103**, 377-82 (1998).
554. Valentini, G. *et al.* Structure and function of human erythrocyte pyruvate kinase. Molecular basis of nonspherocytic hemolytic anemia. *J Biol Chem* **277**, 23807-14 (2002).
555. Van Sligtenhorst, I. *et al.* Cardiomyopathy in alpha-kinase 3 (ALPK3)-deficient mice. *Vet Pathol* **49**, 131-41 (2012).
556. Peloso, G.M. *et al.* Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am J Hum Genet* **94**, 223-32 (2014).
557. Gu, Y. *et al.* The t(4;11) chromosome translocation of human acute leukemias fuses the ALL-1 gene, related to Drosophila trithorax, to the AF-4 gene. *Cell* **71**, 701-8 (1992).
558. Guastadisegni, M.C. *et al.* CBFA2T2 and C20orf112: two novel fusion partners of RUNX1 in acute myeloid leukemia. *Leukemia* **24**, 1516-9 (2010).
559. Serbanovic-Canic, J. *et al.* Silencing of RhoA nucleotide exchange factor, ARHGEF3, reveals its unexpected role in iron uptake. *Blood* **118**, 4967-76 (2011).
560. Okumura, N., Tsuji, K. & Nakahata, T. Changes in cell surface antigen expressions during proliferation and differentiation of human erythroid progenitors. *Blood* **80**, 642-50 (1992).
561. Kiefer, C.R. & Snyder, L.M. Oxidation and erythrocyte senescence. *Curr Opin Hematol* **7**, 113-6 (2000).
562. Nicholson, A.C., Han, J., Febbraio, M., Silversterin, R.L. & Hajjar, D.P. Role of CD36, the macrophage class B scavenger receptor, in atherosclerosis. *Ann N Y Acad Sci* **947**, 224-8 (2001).
563. Auer, P.L. *et al.* Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. *Am J Hum Genet* **91**, 794-808 (2012).
564. Elbers, C.C. *et al.* Gene-centric meta-analysis of lipid traits in African, East Asian and Hispanic populations. *PLoS One* **7**, e50198 (2012).
565. Ayodo, G. *et al.* Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. *Am J Hum Genet* **81**, 234-42 (2007).
566. Aitman, T.J. *et al.* Malaria susceptibility and CD36 mutation. *Nature* **405**, 1015-6 (2000).
567. Bhatia, G. *et al.* Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. *Am J Hum Genet* **89**, 368-81 (2011).

Annex 1: An erythroid enhancer of *BCL11A* subject to genetic variation determines fetal hemoglobin level

Authors

Daniel E. Bauer, Sophia C. Kamran, Samuel Lessard, Jian Xu, Yuko Fujiwara, Carrie Lin, Zhen Shao, Matthew C. Canver, Elenoe C. Smith, Luca Pinello, Peter J. Sabo, Jeff Vierstra, Richard A. Voit, Guo-Cheng Yuan, Matthew H. Porteus, John A. Stamatoyannopoulos, Guillaume Lettre, and Stuart H. Orkin

Reference

Bauer, D.E. *et al.* An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science* **342**, 253-7 (2013).

Abstract

Genome-wide association studies (GWAS) have ascertained numerous trait-associated common genetic variants, frequently localized to regulatory DNA. We find that common genetic variation at *BCL11A* associated with fetal hemoglobin (HbF) level lies in noncoding sequences decorated by an erythroid enhancer chromatin signature. Fine-mapping uncovers a motif-disrupting common variant associated with reduced transcription factor binding, modestly diminished BCL11A expression and elevated HbF. The surrounding sequences function *in vivo* as a developmental stage-specific lineage-restricted enhancer. Genome engineering reveals the enhancer is required in erythroid but not B-lymphoid cells for BCL11A expression. These findings illustrate how GWAS may expose functional variants of modest impact within causal elements essential for appropriate gene expression. We propose the GWAS-marked *BCL11A* enhancer represents an attractive target for therapeutic genome engineering for the β-hemoglobinopathies.

Main text

Genome-wide association studies (GWAS) have identified numerous common single nucleotide polymorphisms (SNPs) associated with human traits and diseases. However advancing from genetic association to causal biologic process has been challenging.[469] Recent genome-scale chromatin mapping studies have highlighted the enrichment of GWAS variants in regulatory DNA elements, suggesting many causal variants may affect gene regulation.[61,185,216,394,470] GWAS of HbF level have identified trait-associated variants at *BCL11A* (see supplementary online text).[138,141,340,471,472] The transcriptional repressor BCL11A has been validated as a direct regulator of HbF level.[58,143,341,351,417,473] Although constitutive BCL11A deficiency results in embryonic lethality and impaired lymphocyte development[474,475], erythroid-specific deficiency of BCL11A counteracts developmental silencing of embryonic and fetal globin genes and rescues the hematologic and pathologic features of sickle cell disease (SCD) in mouse models.[143]

To further understand how common genetic variation impacts *BCL11A*, HbF level and β-globin disorder severity, we compared the distribution of the HbF-associated SNPs at *BCL11A* with DNase I sensitivity, an indicator of chromatin state suggestive of regulatory potential. In primary human erythroblasts, three peaks of DNase I hypersensitivity were observed in intron-2, adjacent to and overlying the HbF-associated variants (Fig. 1A). We term these DNase I hypersensitive sites (DHSs) +62, +58 and +55 based on distance in kb from the transcription start site (TSS) of *BCL11A*. Brain and B-lymphocytes, two tissues that express high levels, and T-lymphocytes, which do not express BCL11A, showed unique patterns of DNase I sensitivity at the *BCL11A* locus, with a paucity of hypersensitivity overlying the trait-associated SNPs (Figs. 1A).

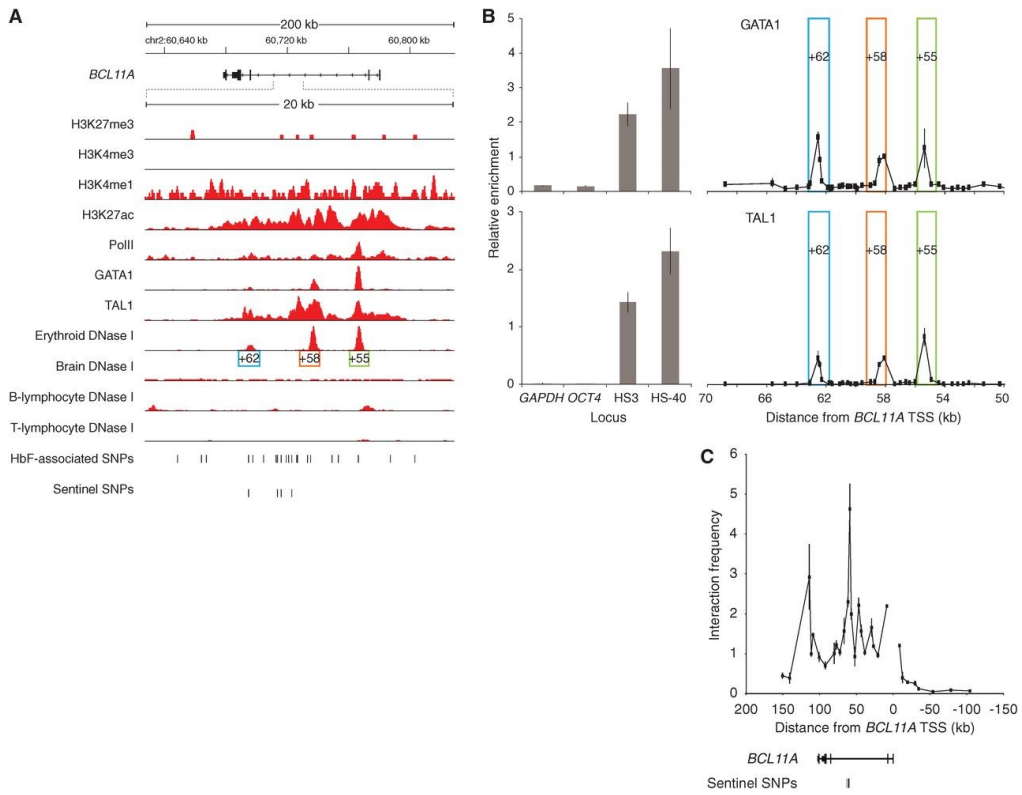


Figure 1. Chromatin state and TF occupancy at *BCL11A*

(**A**) ChIP-seq from human erythroblasts with indicated antibodies. DNase I cleavage densities are from indicated human tissues. Three erythroid DHSs termed +62, +58, and +55 are based on distance in kilobases from *BCL11A* TSS. *BCL11A* transcription is from right to left. (**B**) ChIP–quantitative PCR from human erythroblasts at *BCL11A* intron-2. DHSs +62, +58, and +55 are boxed. Enrichment at negative (*GAPDH* and *OCT4*) and positive control (β-globin LCR HS3 and α-globin HS-40) loci are displayed. (**C**) Chromosome conformation capture in human erythroblasts using *BCL11A* promoter as anchor. Error bars indicate SD.

ChIP-seq demonstrated histone modifications with an enhancer signature overlying the trait-associated SNPs at *BCL11A* intron-2, including the presence of H3K4me1 and H3K27ac, and absence of H3K4me3 and H3K27me3 marks (Figs. 1A). The major erythroid transcription factors (TFs) GATA1 and TAL1 also occupy this enhancer region. ChIP-qPCR confirmed three discrete peaks of GATA1 and TAL1 binding within *BCL11A* intron-2, each falling within an erythroid DHS (Fig. 1B). A common feature of distal regulatory elements is long-range interaction with cognate promoters. We evaluated the interactions between the *BCL11A* promoter and fragments across 250 kb of the *BCL11A* locus using a chromosome conformation capture assay. The greatest promoter interaction was observed within the region of intron-2 containing the trait-associated SNPs (Fig. 1C).

We hypothesized that the causal trait-associated SNPs could function by modulating critical cis-regulatory elements. Therefore we performed extensive genotyping of SNPs within the three erythroid DHSs +62, +58 and +55 in 1,263 DNA samples from the Cooperative Study of SCD (CSSCD).[476] 1,178 individuals and 38 SNPs were used for association testing. Analysis of common variants (MAF > 1%) revealed that rs1427407 in DHS +62 had the strongest association to HbF level ($P = 7.23 \times 10^{-50}$; Figs. 2A). We identified associations to HbF level within the three DHSs that remained following conditioning on rs1427407 (Figs. 2A), consistent with the hypothesis that multiple functional SNPs within the composite enhancer act combinatorically to influence BCL11A regulation. The most significant residual association was for rs7606173 in DHS +55 ($P = 9.66 \times 10^{-11}$).

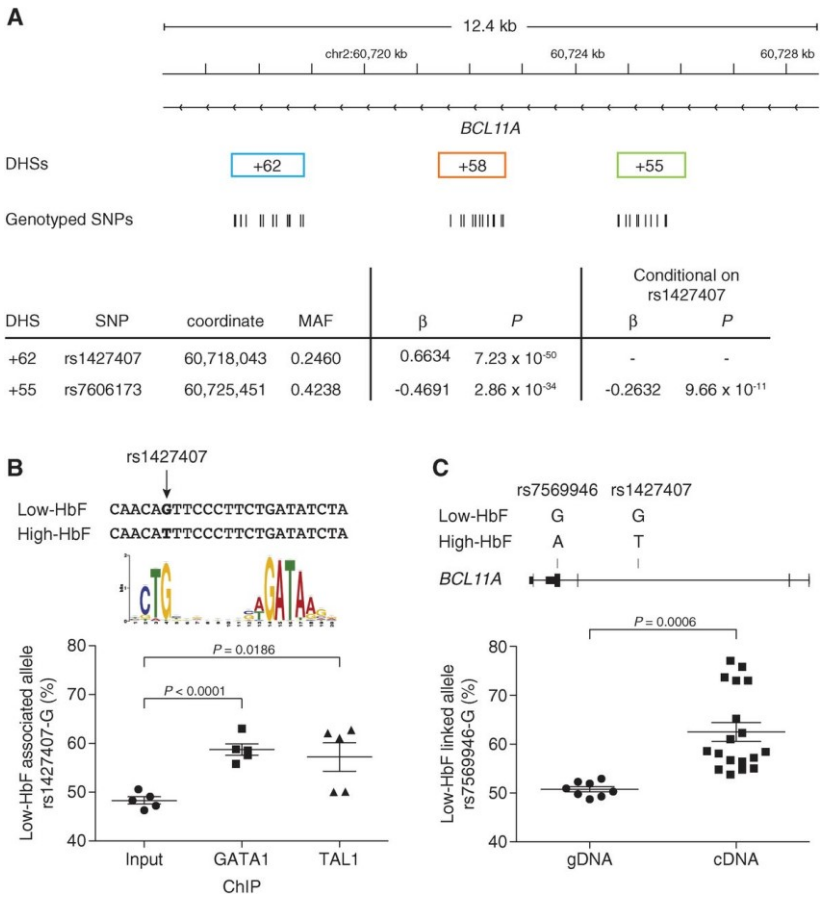


Figure 2. Regulatory variants at *BCL11A*.

(**A**) Genotype data obtained in 1178 individuals from CSSCD for 38 variants within *BCL11A* +62, +58, or +55 DHSs. Shown are most highly significant associations to HbF level among common (MAF > 1%) SNPs ($n = 10$ variants) before (rs1427407) or after (rs7606173) conditional analysis on rs1427407. SNP coordinates are chromosome 2, build hg19. (**B**) Chromatin from erythroblasts of individuals heterozygous for rs1427407, immunoprecipitated by GATA1 or TAL1 and pyrosequenced to quantify the relative abundance of the rs1427407-

G allele. Composite half E-box–GATA motif previously identified is shown[477]. (**C**) gDNA and cDNA from erythroblasts of individuals heterozygous for rs1427407, rs7606173, and rs7569946. Haplotyping demonstrated rs7569946-G, rs1427407-G, and rs7606173-C on the same chromosome in each. Pyrosequencing was performed to quantify the relative abundance of the rs7569946-G allele.

The SNP rs1427407 falls within a peak of GATA1 and TAL1 binding (Figs. 1A). The minor T-allele disrupts the G-nucleotide of a sequence element resembling a half E-box/GATA composite motif [CTG(n_9)GATA], a consensus sequence enriched for chromatin bound by GATA1 and TAL1 complexes in erythroid cells.[477,478] We identified five primary erythroblast samples from individuals heterozygous for the major G-allele and minor T-allele at rs1427407 and subjected these samples to ChIP followed by pyrosequencing. As anticipated, we observed an even balance of alleles in the input DNA. However, we detected more frequent binding to the G-allele compared to the T-allele in both the GATA1 and TAL1 immunoprecipitated chromatin samples (Fig. 2B).

As the common synonymous SNP rs7569946 lies within exon-4 of *BCL11A*, it can be used to discriminate expression of alleles. We identified three primary erythroblast samples doubly heterozygous for the rs1427407–rs7606173 haplotype and rs7569946. For each sample, we determined by molecular haplotyping that the major rs7569946 G-allele was in phase with the low-HbF associated rs1427407–rs7606173 G–C haplotype[479,480] Pyrosequencing revealed that whereas the alleles were balanced in genomic DNA (gDNA), significant imbalance was observed in complementary DNA (cDNA) with 1.7-fold increased expression of the low-HbF linked G-allele of rs7569946 (Fig. 2C).

To understand the context within which these apparent regulatory trait-associated SNPs play their role, we explored the function of the harboring composite element. We cloned a 12.4 kb (+52.0–64.4 kb from TSS) human gDNA fragment containing the three erythroid DHSs to assay enhancer potential in a murine transgenic *lacZ* reporter assay. Endogenous BCL11A shows abundant expression throughout the developing central nervous system with much lower expression observed in the fetal liver.[481] In contrast, we observed in the transgenic embryos reporter gene expression largely confined to the fetal liver, the site of definitive erythropoiesis, with weaker expression noted in the central nervous system (Fig. 3A).

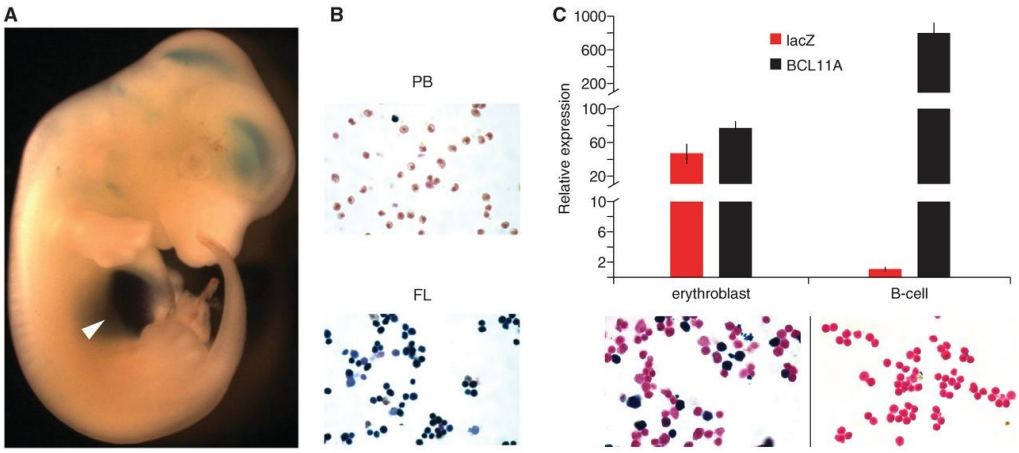


Figure 3. The GWAS-marked BCL11A enhancer is sufficient for adult-stage erythroid expression.

(**A**) A 12.4-kb fragment of *BCL11A* intron-2 (+52.0 to 64.4 kb from TSS) was cloned to a *lacZ* reporter construct. Shown is a transient transgenic mouse embryo from 12.5 dpc X-gal stained. Arrowhead indicates liver. (**B**) Cell suspensions isolated from peripheral blood (PB) and fetal liver (FL) of stable transgenic embryos at 12.5 dpc X-gal stained. (**C**) Sorted erythroblasts and B-lymphocytes from young adult stable transgenic mice subject to X-gal staining or RNA isolation followed by quantitative reverse transcription (RT)–PCR. Gene expression was normalized to glyceraldehyde-3-phosphate dehydrogenase and expressed relative to T-lymphocytes. Error bars indicate SD.

A characteristic feature of globin gene and *BCL11A* expression is developmental regulation. In stable transgenic *BCL11A* +52.0–64.4 reporter lines at 12.5 dpc, circulating primitive erythrocytes failed to stain for X-gal whereas definitive erythroblasts in fetal liver robustly stained positive (Fig. 3B). Endogenous BCL11A was expressed at 10.4-fold higher levels in B-lymphocytes as compared to erythroblasts. LacZ expression was restricted to erythroblasts and not observed in B-lymphocytes (Fig. 3C). These results indicate that the GWAS-marked *BCL11A* intron-2 regulatory sequences are sufficient to specify developmentally-restricted, erythroid-specific gene expression.

We aimed to disrupt the enhancer to investigate its requirement for BCL11A expression. Since there are no suitable adult-stage human erythroid cell lines, we turned to the mouse erythroleukemia (MEL) cell line. We observed an orthologous enhancer signature at intron-2 of mouse *Bcl11a* indicated by sequence homology, erythroid-specific DNase I hypersensitivity, characteristic histone marks and GATA1/TAL1 occupancy.[478,482] Sequence-specific nucleases can produce small chromosomal deletions via NHEJ-mediated repair.[483] We engineered TALENs to introduce double-strand breaks to flank the orthologous 10 kb *Bcl11a* intron-2

sequences carrying the erythroid enhancer chromatin signature. Three unique clones were isolated that had undergone biallelic excision of the intronic segment. BCL11A transcript was profoundly reduced in the absence of the orthologous erythroid composite enhancer (Fig. 4A). BCL11A protein expression was not detectable in the enhancer-deleted clones (Fig. 4B). In the absence of the BCL11A enhancer, embryonic globin gene derepression was pronounced, with the ratio of embryonic εy to adult β1/2 globin increased by a mean of 364-fold.

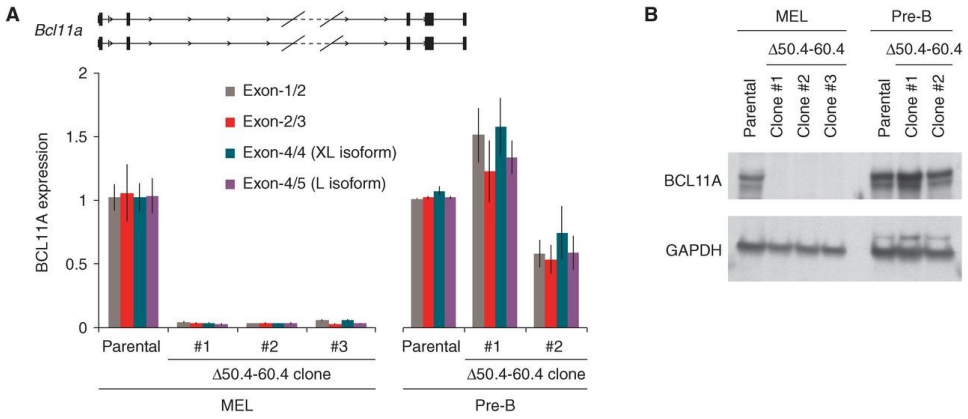


Figure 4. The GWAS-marked *BCL11A* enhancer is necessary for erythroid but dispensable for nonerythroid expression.

(**A**) Three MEL and two pre-B lymphocyte clones with biallelic deletion of the orthologous *Bcl11a* erythroid enhancer (Δ50.4 to 60.4) subject to quantitative RT-PCR. (**B**) Immunoblot of Δ50.4 to 60.4 MEL and pre-B lymphocyte clones.

To examine potential lineage-restriction of the requirement for the +50.4–60.4 kb intronic sequences for BCL11A expression, we evaluated their loss in a non-erythroid context. The same strategy of introduction of two pairs of TALENs to obtain clones with NHEJ-mediated deletion was employed in a pre-B lymphocyte cell line. In contrast to the erythroid cells, BCL11A expression was retained in the Δ50.4–60.4 kb enhancer deleted pre-B cell clones at both the RNA and protein levels (Figs. 4A and 4B). These results indicate the orthologous erythroid enhancer sequences are essential for erythroid gene expression but not required in B-lymphoid cells for integrity of transcription from the *Bcl11a* locus.

The prior identification of BCL11A as a critical repressor of HbF levels has raised new hope for mechanism-based therapeutic approaches to the β-hemoglobinopathies.[73] However, the

paradox that genetic variation at *BCL11A* is common, well-tolerated and disease-protective despite the critical roles of BCL11A in neurogenesis and lymphopoiesis[474,475,484] has remained unresolved. Here we demonstrate that the HbF-associated variants localize to an erythroid enhancer of *BCL11A*. By allele-specific analyses, we show that genetic variation within this enhancer is associated with modest impact on TF binding, BCL11A expression and HbF level. Relatively small effect sizes associated with individual variants may not be surprising given that most single nucleotide substitutions, even within critical motifs, result in only modest loss of enhancer activity.[449,485] In contrast, loss of the *BCL11A* enhancer results in the absence of BCL11A expression in the erythroid lineage. Most trait-associated SNPs identified by GWAS are noncoding and have small effect size.[469,486] The impact of GWAS-identified SNPs on biological processes is often uncertain. Our findings underscore how a modest influence engendered by an individual noncoding variant neither predicts nor precludes a profound contribution of an underlying regulatory element.

Challenges to inhibiting BCL11A for mechanism-based reactivation of HbF include the supposedly "undruggable" nature of transcription factors[487] and its important non-erythroid functions.[475,484] With recent developments in their efficiency and precision, sequence-specific nucleases can be designed to exquisitely target genomic sequences of interest.[265,488] We propose the GWAS-identified enhancer of *BCL11A* as a particularly promising therapeutic target for genome engineering in the β-hemoglobinopathies. Disruption of this enhancer would impair BCL11A expression in erythroid precursors with resultant HbF derepression, while sparing BCL11A expression in non-erythroid lineages. Rational intervention might mimic common protective genetic variation.

Material and methods

Cell culture Human

CD34+ cells from mobilized peripheral blood of healthy donors were obtained from Centers of Excellence in Molecular Hematology at Yale University, New Haven, Connecticut and Fred Hutchinson Cancer Research Center, Seattle, Washington. The cells were subject to

ex vivo erythroid maturation with a two-phase serum-free liquid culture protocol as previously described.[334] Peripheral blood mononuclear cells (PBMCs) were obtained from healthy donors from Boston Children's Hospital. Erythroid differentiation from PBMCs was performed as previously described.[489] Mouse erythroleukemia (MEL) cells and 293T cells were cultured as previously described.[490] Stably v-Abl transformed pre-B lymphocyte murine cells (derived as described[490]) were cultured in RPMI plus 2% penicillin-streptomycin, 15% FCS, 2% HEPES, 1% nonessential amino acids, 1% sodium pyruvate, 1% L-glutamine and 100 μM β-mercaptoethanol.

ChIP and DNase I sensitivity

Chromatin immunoprecipitation and massively parallel sequencing were performed as described.[334] The following antibodies were used: H3K27me3 (Millipore, 07-449), H3K4me3 (Millipore, 04-745), H3K4me1 (Abcam, ab8895), H3K27ac (Abcam, ab4729), RNA Polymerase II (PolII, Santa Cruz, sc-899), GATA1 (Abcam, ab11852) and TAL1 (Santa Cruz, sc-12984). DNase I cleavage density performed and analyzed as previously described.[350] For ChIP-qPCR, relative enrichment was determined by comparing amplification of ChIP material to 1% input chromatin by the ΔCt method. Loci previously reported to be occupied and non-occupied by GATA1 and TAL1 were used as positive and negative controls respectively.[334]

Chromosome conformation capture (3C)

3C assay was performed as previously described[334] except as below. Nuclei from formaldehyde cross-linked primary human erythroid precursors were digested with HindIII prior to ligation and reversal of cross-links. Quantitative real-time PCR was performed using iQ SYBR Green Supermix (Bio-Rad, 170-8880). A fragment containing the *BCL11A* promoter was used as the anchor region. To correct for amplification efficiency of different primers, a control template was prepared by digesting and ligating an equimolar mixture of two bacterial artificial chromosomes (BACs) comprising the complete human *BCL11A* locus (RP11-606L8 and RP11-139C22) and one the human β- globin cluster (CTD-3055E11). An interaction between fragments in HS1/HS2 and HS3 of the human β-globin locus control region (LCR) served as a

positive control. Interaction frequency was expressed as amplification relative to the known LCR interaction, normalized to the BAC control template.

Fine-mapping *BCL11A* locus

Markers (all coordinates hg19) were selected from within the three *BCL11A* intron-2 DHSs +62 (chr2:60,717,492-60,718,860), +58 (chr2:60,721,411-60,722,674) and +55 (chr2:60,724,802-60,726,084). 21 markers were identified from the 1000 Genomes Project database using the North European (CEU), Nigerian (YRI) and African-American (ASW) reference populations. 38 additional variants were present in 3 dbSNP135. We sequenced by Sanger chemistry the three DHS intervals in the DNA of 52 and 36 sickle cell disease (SCD) patients from the CSSCD cohort with high (> 8%) and low (< 2%) HbF levels, respectively. From this sequencing effort, seven novel sequence variants were identified. Because most markers cluster in small genomic intervals, it was not possible to design genotyping assays for some of them. Of 66 non-redundant variants identified in the three DHSs, genotyping assays for 40 markers were performed in 1,263 participants from the CSSCD, an African-American SCD cohort for which genomic DNA (gDNA) is available and HbF levels are known.[476] Markers were genotyped using the Sequenom iPLEX platform. Individuals and DNA sequence variants with a genotyping success rate < 90% were excluded. Overall genotype concordance estimated from triplicates was 100%. SNPs passing quality control (QC; n = 38) are shown schematically in Figs. 2A below the three DHSs. A substantial fraction of the genotyped SNPs are rare in the reference populations so not surprisingly monomorphic in the CSSCD (n = 18). After QC, 1,178 individuals and 20 polymorphic SNPs remained for the analysis. HbF levels were modeled as previously described.[47,139] Association and conditional analyses of single variants (MAF > 1%) were performed with PLINK[338] using linear regression under an additive genetic model. Analysis of common variants (MAF > 1%) revealed that rs1427407 in DHS +62 had the strongest association to HbF level (P=7.23x10^{-50}; Figs. 2A). Conditional analysis demonstrated that after conditioning on rs1427407 and rs7606173, no more SNPs were significant. Adjusting for principal components (PCs) on 855 individuals for whom genome-wide genotyping data was available to account for admixture and other confounders yielded similar results.

For rare and low-frequency variants (MAF < 5%), we performed set-based analyses using each of the three DHSs +62, +58 and +55 as the testing unit. For these analyses, we used the sequence kernel association test (SKAT-O) program[491] with default parameters. We selected the 5% threshold for MAF in order to maximize statistical power given our limited sample size, but note that markers with a MAF between 1% and 5% were also analyzed in the single variant analyses presented above. This variant overlap is accounted for using conditional analyses with the common variants independently associated with HbF levels. Two sets were found to be statistically significant, namely DHS +62 and DHS +55, but after conditioning on rs1427407 and rs7606173, results were no longer statistically significant, suggesting weak LD between the rare/low-frequency variants and the common SNPs. We did not find evidence that rare and low-frequency sequence variants within the *BCL11A* DHSs influence HbF levels in SCD subjects, despite Sanger re-sequencing these DHSs in 88 subjects with extreme HbF phenotype.

The rs1427407–rs7606173 haplotype frequencies in CSSCD are: T–G 24.5%, T– C 0.085%, G–C 42.3%, G–G 33.1%. The mean HbF level is 4.05% (SD 3.10) in 213 rs1427407–rs7606173 G–C individuals, 7.08% (SD 4.50) in 254 rs1427407–rs7606173 T–G/G–T heterozygotes and 11.21% (SD 4.37) in 60 rs1427407–rs7606173 T–G individuals. For comparisons of HbF levels between genotypes, the *P*-values were determined by one-tailed student *t*-tests.

Molecular haplotyping

For two heterozygous SNPs on the same chromosome, there are two possible phases: A–B/a–b (model 1) and A–b/a–B (model 2). For SNPs within the 12-kb *BCL11A* intron-2 fragment +52.0-64.4 kb, phase was determined by cloning PCR products and determining co-distribution of SNP alleles. To determine phase of rs7569946 and rs1427407 alleles (separated by 30.1 kb on chromosome 2), emulsion fusion PCR was performed as previously described[479,480] with minor modification. Fusion PCR brings two regions of interest, from separate parts of the same chromosome, together into a single product. By carrying out the reaction in emulsion with aqueous microdroplets surrounded by oil, the preponderance of amplicons are derived from a single template molecule. Genomic DNA from individuals known to be doubly heterozygous for rs7569946 and rs1427407 served as template in the following 100 µl reaction (with final

concentrations listed): KOD Hot Start DNA Polymerase (14 U, Novagen, 71086), KOD buffer (1X), MgSO4 (1.5 mM), dNTPs (0.2 mM each), rs7569946-F and rs1427407-R primers (1 µM each), rs7569946-R primer (30 nM), rs7569946-R-revcomp-rs1427407–F bridging inner primer (30 nM), gDNA (200 ng). The 100 µl aqueous reaction was added dropwise with stirring to 200 µl oil phase to create an emulsion. Two 125 µl aliquots of emulsion were amplified under the following conditions: 95 degrees 2 minutes; 45 cycles of 95 degrees 20 seconds, 60 degrees 10 seconds, 70 degrees 30 seconds; 70 degrees 2 minutes. Hexane extracted fusion PCR product was subject to nested PCR in 25 µl as follows: KOD Hot Start DNA Polymerase (0.5 U), KOD buffer (1X), MgSO4 (1.5 mM), dNTPs (0.2 mM each), rs7569946-nested-F and rs1427407-nested-R primers (300 nM each), extracted fusion PCR product (75 nl); 95 degrees 2 minutes; 35 cycles of 95 degrees 20 seconds, 60 degrees 10 seconds, 70 degrees 30 seconds; 70 degrees 2 minutes. The nested product was confirmed by agarose gel electrophoresis to constitute a single band of expected size. The purified product was cloned with the Zero Blunt PCR Cloning kit (Life Technologies, K2700-20). The Sanger sequencing of fusion amplicons enumerated clones of 4 possible sequences: A–B, a–b, A–b and a–B. The likelihood of each phase was calculated based on a multinomial distribution assumption. The likelihood ratio for the two configurations was calculated as a measure for the statistical significance of the data fitting haplotype model 1 (as compared to model 2). A ratio approaching infinity suggests model 1, a ratio of 1 suggests equipoise and a ratio approaching zero suggests model 2.

Pyrosequencing

Healthy CD34+ cell donors were screened to identify five donors heterozygous for rs1427407. These CD34+ cells were subject to ex vivo erythroid differentiation. Chromatin was isolated and ChIP performed with GATA1 and TAL1 antibodies. Input chromatin as compared to GATA1 or TAL1 precipitated material was subject to pyrosequencing to determine allelic balance of rs1427407. Healthy CD34+ donors were screened to identify three donors heterozygous for the rs1427407–rs7606173 G–C/T–G haplotype. These CD34+ cells were subject to ex vivo erythroid differentiation. Complementary DNA (cDNA) and gDNA were subject to pyrosequencing to determine allelic balance of rs7569946.

PCR conditions as follows: 2X HotStarTaq master mix (Qiagen, 203443), MgCl2 (final concentration 3 mM), template DNA (0.1-1 ng) and SNP-specific forward and reverse-biotinylated primers (200 nM each). PCR cycling conditions were: 94ºC 15 min; 5 45 cycles of 94ºC 30 s; 60ºC 30 s; 72ºC 30 s; 72ºC 5 min. One primer of each pair was biotinylated. The PCR product strand containing the biotinylated primer was bound to streptavidin beads and combined with a specific sequencing primer. The primed single stranded DNA was sequenced and genotype analyzed using the Pyrosequencing PSQ96 HS System (Qiagen Pyrosequencing) following the manufacturer's instructions.

Transgenic mice

The enhancer reporter construct pWHERE-Dest was obtained from Dr. William Pu. Modified from pWHERE (Invivogen, pwhere) as previously described[492], the construct has murine H19 insulators flanking a CpG-free lacZ variant driven by a minimal Hsp68 minimal promoter with a Gateway destination cassette at the upstream MCS. Enhancer fragments were amplified from mouse gDNA, recombined into pDONR221 vector (Invitrogen, 12536-017) by BP clonase (Invitrogen, 11789020) and recombined into pWHERE-Dest vector with LR clonase (Invitrogen, 11791020). Plasmids were digested with PacI to remove vector backbone. The lacZ enhancer reporter fragments were purified by gel electroelution and then concentrated using Wizard DNA Clean-Up System (Promega, A7280). Transgenic mice were generated by pronuclear injection to FVB fertilized eggs. Approximately 10 ng/µl of DNA solution was used for series of injections. CD-1 females were used as recipients for injected embryos. 10.5 to 14.5 dpc embryos were dissected from surrogate mothers with whole-mount and tissue X-gal staining performed as previously described.[493] X-gal stained cytospins were counterstained with Nuclear Fast Red (Vector Laboratories, H-3403). Tails used for PCR genotyping. Animal procedures were approved by the Children's Hospital Instititutional Animal Care and Use Committee.

Human erythroid precursor enhancer assay

Genomic DNA fragments containing putative enhancer elements were cloned into pLVXPuro (Clontech, 632164) upstream of a minimal TK promoter and GFP reporter gene as described[334]. 293T cells were transfected with FuGene 6 reagent (Promega, E2691) according to manufacturer's protocol. The media was changed after 24 hours to SFEM medium supplied

with 2% penicillin-streptomycin, and after 36 hours, supernatant was collected and filtered. $CD34^+$ cell-derived erythroid cultures were transduced with lentivirus on expansion days 4 and 5 by spin-infection as previously described.[334] Cells were resuspended in erythroid differentiation media 24 hours after the second infection. Selection with puromycin 1 μg/ml commenced 48 hours after infection. Transduced cells were analyzed after five days in differentiation media by flow cytometry for GFP mean fluorescence intensity.

Flow cytometry

Live cells were gated by exclusion of 7-aminoactinomycin D (7-AAD, BD Pharmingen, 559925). Bone marrow (for erythroblast) and spleen (for lymphocyte) suspensions were isolated from young adult transgenic mice. Following hypotonic lysis of mature red blood cells, live cells (7-AAD-) sorted based on staining with CD71-biotin (BD, 557416), streptavidin-APC (BD, 554067), Ter-119-PE (BD, 553673), CD19-APC (BD, 550992) or CD3-PE (BD, 100308). CD71+ Ter119+ , CD19+ and CD3+ sorted populations used for cytospin and RNA isolation.

TALEN-mediated chromosomal deletion

Transcription activator-like effector nucleases (TALENs) were designed to generate cleavages at mouse *Bcl11a* intron-2 at sites +50.4 kb (termed 5' site) and +60.4 kb (3' site) relative to the TSS. The TALENs recognize the following sequences: CTTAAGGCAAGAATCACT (5' left), CCATGCCTTTCCCCCCCT (5' right), GAGTTAAAATCAGAAATCT (3' left), CTGACTAATTGATCAT (3' right). TALENs were synthesized with Golden Gate cloning[494] using the NN RVD to recognize G. The synthesized DNA binding domains were cloned into pcDNA3.1 (Invitrogen, V790-20) with the FokI nuclease domain, Δ152 N-terminal domain and +63 C-terminal domain previously described.[495] 2.5 μ g of each of the four TALEN plasmids with 0.5 μg pmaxGFP (Lonza) were delivered to 2 x 106 MEL or pre-B cells by electroporation per manufacturer's protocol (Lonza, VCA-1005). GFP-positive cells were sorted by flow cytometry after 48 hours. Cells seeded by limiting dilution in 96-well plates to isolate individual clones. Clones screened by PCR of gDNA to detect the amplification of a short product from upstream of the 5' site and downstream of the 3' site indicating deletion of the intervening segment. Monoallelic deleted clones were subject to a second round of TALEN-mediated deletion to obtain biallelic deleted clones. Clones with

biallelic deletion were identified by detecting absence of amplification from within the deleted fragment. Deletion frequency was approximately one in 50 alleles. Deletion was validated with Southern blotting. Genomic DNA was digested with BmtI; a 561-bp probe (amplified from gDNA upstream of the 5' site) hybridizes to a 3.6 kb fragment from the wild-type allele and a 8.9 kb fragment from the Δ50.4-60.4 deleted allele.

RT-qPCR and immunoblotting

RNA isolation with RNeasy columns (Qiagen, 74106), reverse transcription with iScript cDNA synthesis kit (Bio-Rad, 170-8890), qPCR with iQ SYBR Green Supermix (BioRad, 170-8880) and immunoblotting performed as described (39). For the mouse β- globin cluster genes, a common primer pair recognizes the adult β-globins β2 and β1 while independent primers recognize the embryonic β-globins εy and βH1. The following antibodies were used for immunoblotting: BCL11A (Abcam, ab19487), GAPDH (Santa Cruz, sc-25778).

Acknowledgments

Thanks to A. Woo, A. Cantor, M. Kowalczyk, S. Burns, J. Wright, J. Snow, J. Trowbridge and members of the Orkin laboratory, particularly C. Peng, P. Das, G. Guo, M. Kerenyi, and E. Baena, for discussions. C. Guo and F. Alt provided the pre-B cell line, A. He and W. Pu the pWHERE lacZ reporter construct, C. Currie and M. Nguyen technical assistance, D. Bates and T. Kutyavin expertise with sequence analysis, R. Sandstrom help with data management, G. Losyev and J. Daley aid with flow cytometry and J. Desimini graphical assistance. L. Yan at EpigenDx (Hopkinton, Massachusetts) conducted the custom pyrosequencing reactions. This work was funded by grants from the Doris Duke Charitable Foundation (#2009089) and Canadian Institute of Health Research (#123382) to G.L.; Amon Carter Foundation, Hyundai Hope on Wheels, NIH, Lucille Packard Foundation to M.H.P.; NIH grants U54HG004594 and U54HG007010 to J.A.S.; and NIH R01HL032259, P01HL032262, and P30DK049216 (Center of Excellence in Molecular Hematology) to S.H.O. D.E.B. is supported by NIDDK Career Development Award K08DK093705. A patent application related to this work was filed by Boston Children's Hospital, and D.E.B., J.X., and S.H.O. are inventors.

Annex 2: Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci

Authors

Matthew C Canver, Samuel Lessard, Luca Pinello, Yuxuan Wu, Yann Ilboudo, Emily N Stern, Austen J Needleman, Frédéric Galactéros, Carlo Brugnara, Abdullah Kutlar, Colin McKenzie, Marvin Reid, Diane D Chen, Partha Pratim Das, Mitchel A Cole, Jing Zeng, Ryo Kurita, Yukio Nakamura, Guo-Cheng Yuan, Guillaume Lettre, Daniel E Bauer & Stuart H Orkin

Author contribution

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.

Reference

Canver, M.C. *et al.* Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci. *Nature Genetics* (2017).

Abstract

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.

Introduction

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[394]. 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[56,59]. However, targeted deletion has limited throughput, efficiency, and resolution[289]. 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 extrachromosomal 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[56,59,225].

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[59]. 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[59,263,289,496-498]. 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.

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[47,139-141,499-501]. 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[61,62,502-505]. 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[500]. 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).

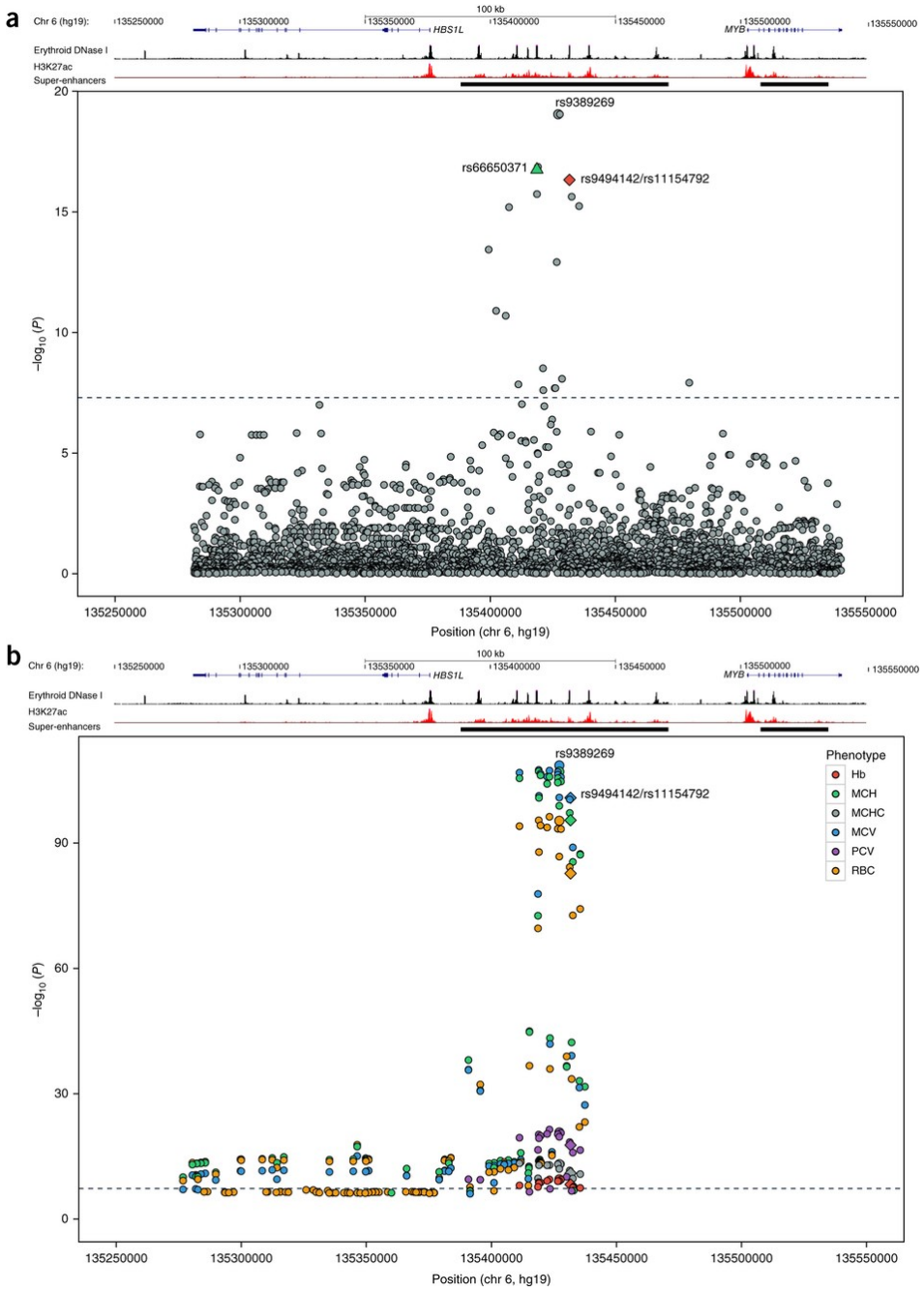


Figure 1. Trait associations of the *HBS1L-MYB* intergenic region.
(**a**) Meta-analysis of HbF-associated SNPs from SCD cohorts (n = 2,222). rs66650371 (green triangle) and rs9494142/rs11154792 (red diamond) have previously been implicated as possible functional SNPs affecting *MYB* expression[500]. 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 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)[61]. 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.

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[61] (Fig. 1b). Owing to extensive LD and limited sample size, conditional analysis of HbF-associated SNPs could not confidently pinpoint a specific set of causal variants. 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)[500]. 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[56] (Fig. 1). The trait-associated SNPs from both meta-analyses are concentrated in an 83-kb intergenic super-enhancer (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). 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.

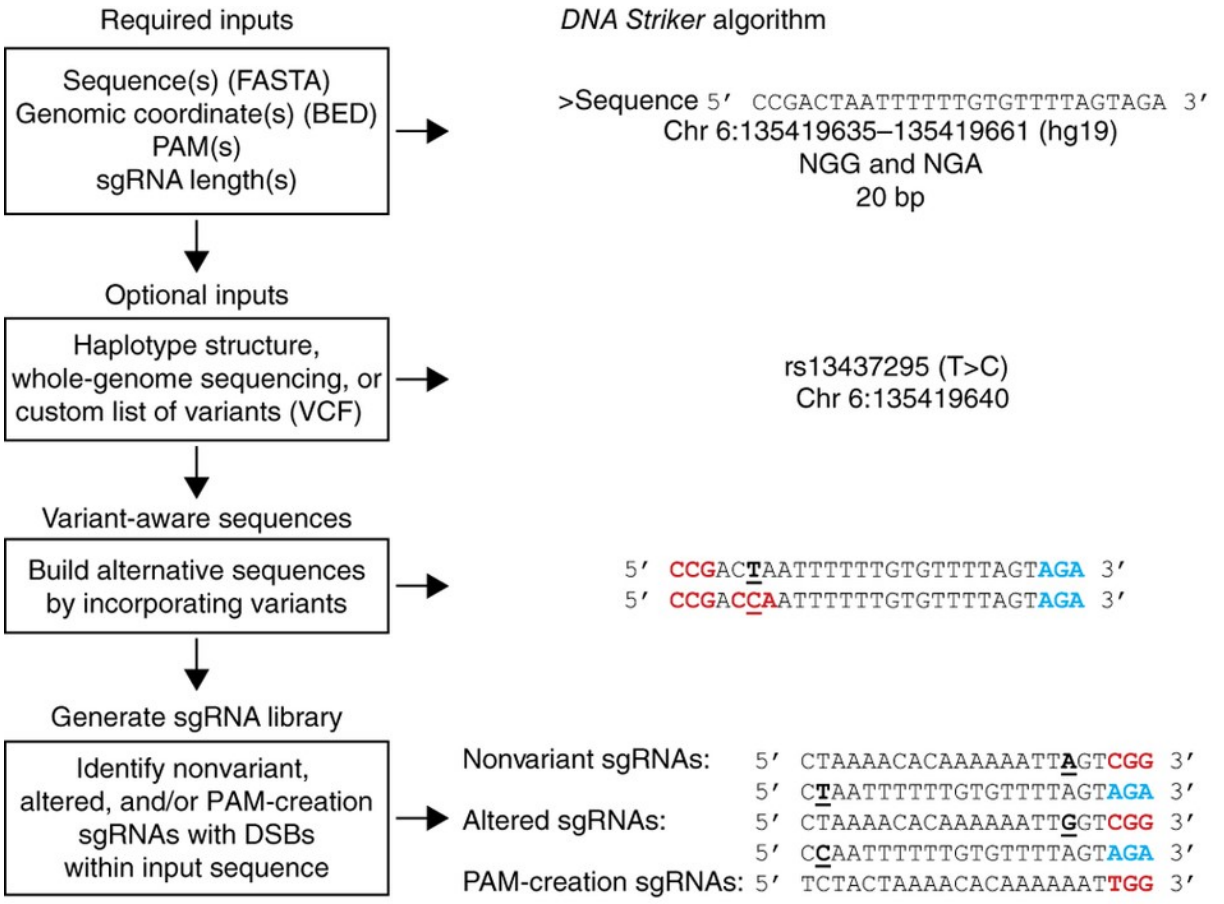


Figure 2. DNA Striker algorithm

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.

Saturating-mutagenesis-library design

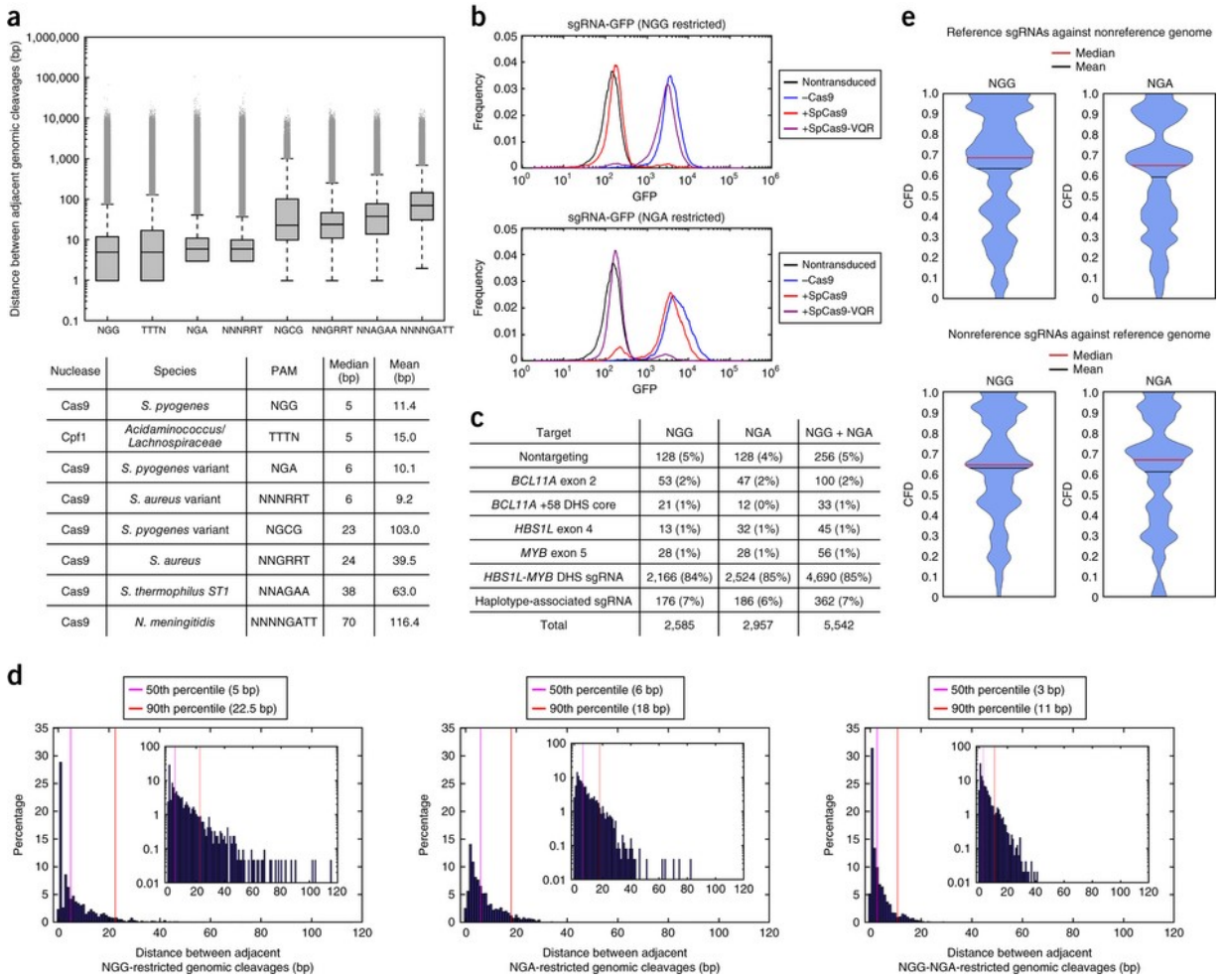
CRISPR-associated nucleases with unique PAM-recognition sequences have been reported for genome editing[263,423,496,506-510]. The frequency of each PAM varies throughout the genome (Fig. 3a). Given the sequence dependence of PAM availability, feature-specific variation in cleavage density for each nuclease was observed in DHSs, enhancers, and repressed regions as well as genes.

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.

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)[423], 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[59], *HBS1L* exon 4, and *MYB* exon 5 (Fig. 3c). 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. 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.



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.

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). Both NGG- and NGA-restricted sgRNA libraries were synthesized and successfully batch-cloned into lentiviral constructs.

Cutting-frequency determination (CFD) has previously been used to assess the activity of imperfect-match sgRNAs[429]. We used CFD 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[59,60]. 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[59]. 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[59] (Fig. 4b). *BCL11A* +58 DHS–targeted sgRNAs were not underrepresented, whereas *BCL11A* exon 2 sgRNAs showed modest dropout, in agreement with previous findings[59](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. 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. 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[152].

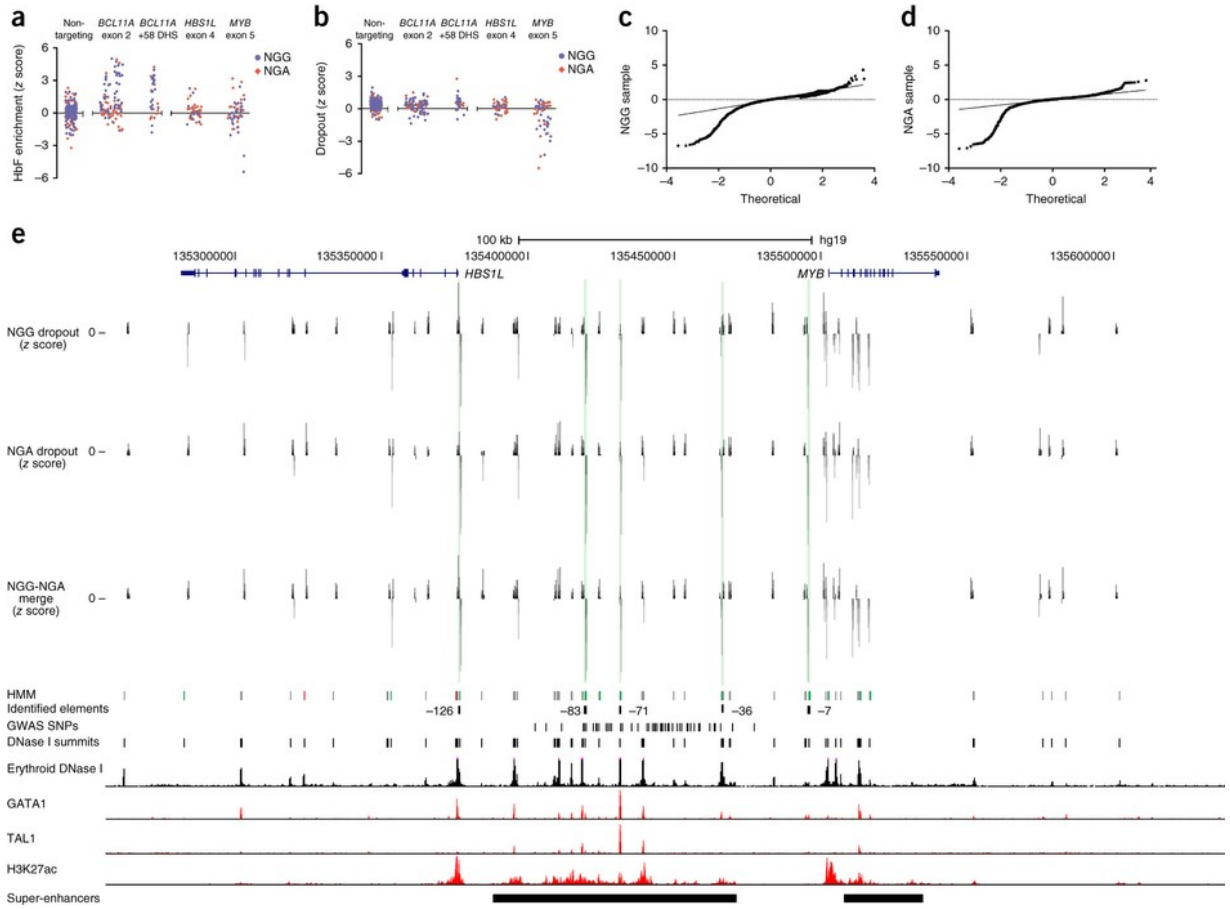


Figure 4. Mapping NGG- and NGA-restricted sgRNA dropout scores to genomic cleavage position identifies putative functional elements

(**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).

Introduction of a 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. The same sgRNA targeting *MYB* also demonstrated a cell-proliferation defect in $CD34^+$ HSPC-derived human erythroblasts. 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. Furthermore, targeting *MYB* led to a decrease in MYB expression. 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[59]. 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).

Notably, the utilization of SpCas9 and SpCas9-VQR species together enhanced resolution at these DHSs by decreasing the gaps between adjacent genomic cleavages. In addition, a higher sgRNA density enhanced the reliability of functional sequence detection by HMM analysis. Notably, the −83 and −71 DHSs fell within an annotated super-enhancer region, and each of these five DHSs colocalized with GATA1 and/or GATA1–TAL1 binding (Fig. 4e). 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[500]. 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). rs9389268, which is highly associated with erythroid traits, is located within the −83 DHS (Fig. 5). 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). 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[511]. The top-scoring sgRNAs at the −71 element specified a cleavage ~200 bp from the peak of DNase I sensitivity and GATA1–TAL1 binding[511] (Fig. 4e). The highly trait-associated SNP within the −71 DHS that disrupts a GATA1 motif, rs9494142, also known as rs11154792 (ref.[499]; 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). rs66650371 is a 3-bp indel that disrupts a 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).

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[291,454]. 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). shRNA-mediated knockdown of *MYB* expression demonstrated that loss of MYB decreases cellular fitness in HUDEP-2 and $CD34^+$ HSPC-derived erythroblasts. 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, Fig. 6a,b). However, increased genomic matches for a given sgRNA have also been predicted to decrease cellular fitness[291,454]. 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.

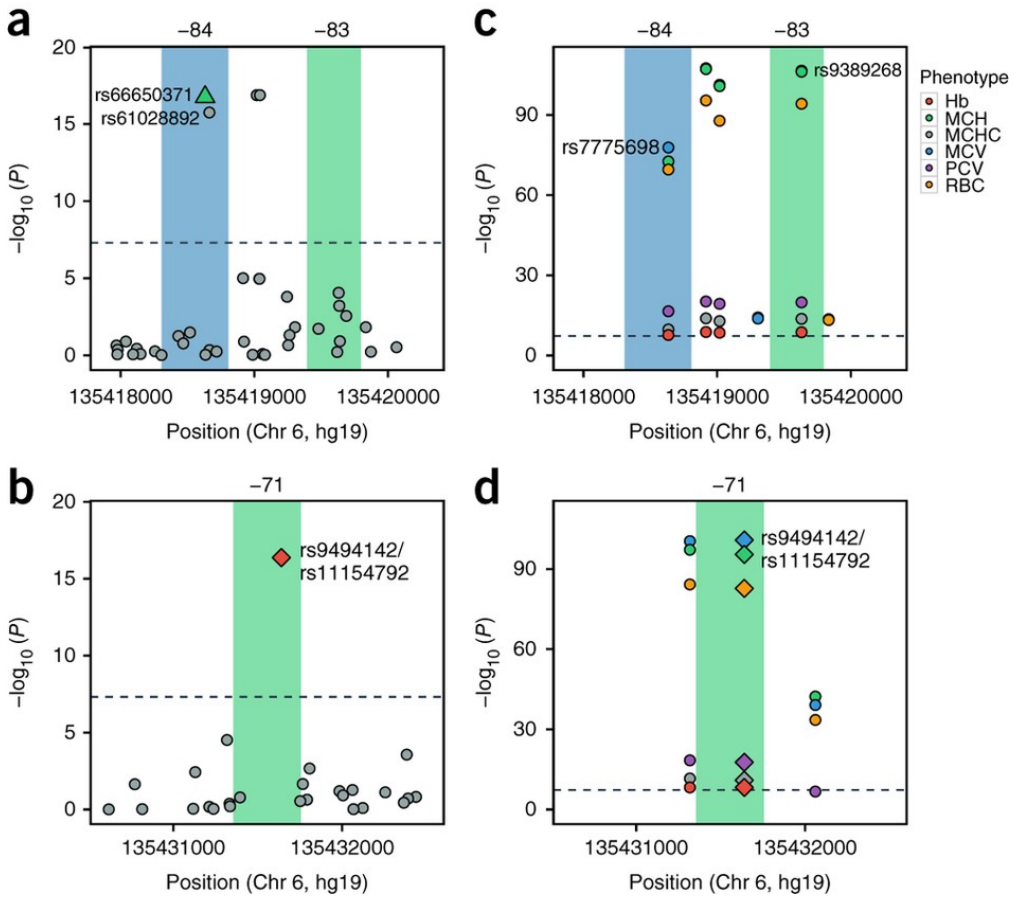


Figure 5. Trait-associated SNPs mark essential enhancer elements

(**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[500]. −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[500]. −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[500]. −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 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[430]. 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 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. 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. We used sgRNA 1910, which had the maximal off-target score (indicative of lower off-target potential) in the −36 region. 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, 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[500,512]. We used an NGG-restricted guide targeting the −71 DHS (sgRNA 1582) that produced a DSB directly adjacent to the rs9494142/rs11154792/GATA1 motif. Targeting this motif in $CD34^+$ HSPC-derived erythroblasts resulted in successful mutagenesis but did not alter cellular proliferation or *MYB* expression.

An NGA-restricted guide targeting the −84 DHS (sgRNA 1500) was used with a DSB position 1 bp from the implicated TAL1-binding motif[500]. Targeting this motif in $CD34^+$ HSPC-derived erythroblasts resulted in moderate levels of editing but did not alter cellular proliferation. *MYB* expression trended toward a decrease; however, this effect did not reach statistical significance. Notably, sgRNA 1500 resulted in a predominance of indels sparing the adjacent TAL1- and GATA1-binding motifs. 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); this sgRNA demonstrated significant dropout in the saturating-mutagenesis screen. 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. 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). All four of these DHSs contain repetitive sequence. 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.

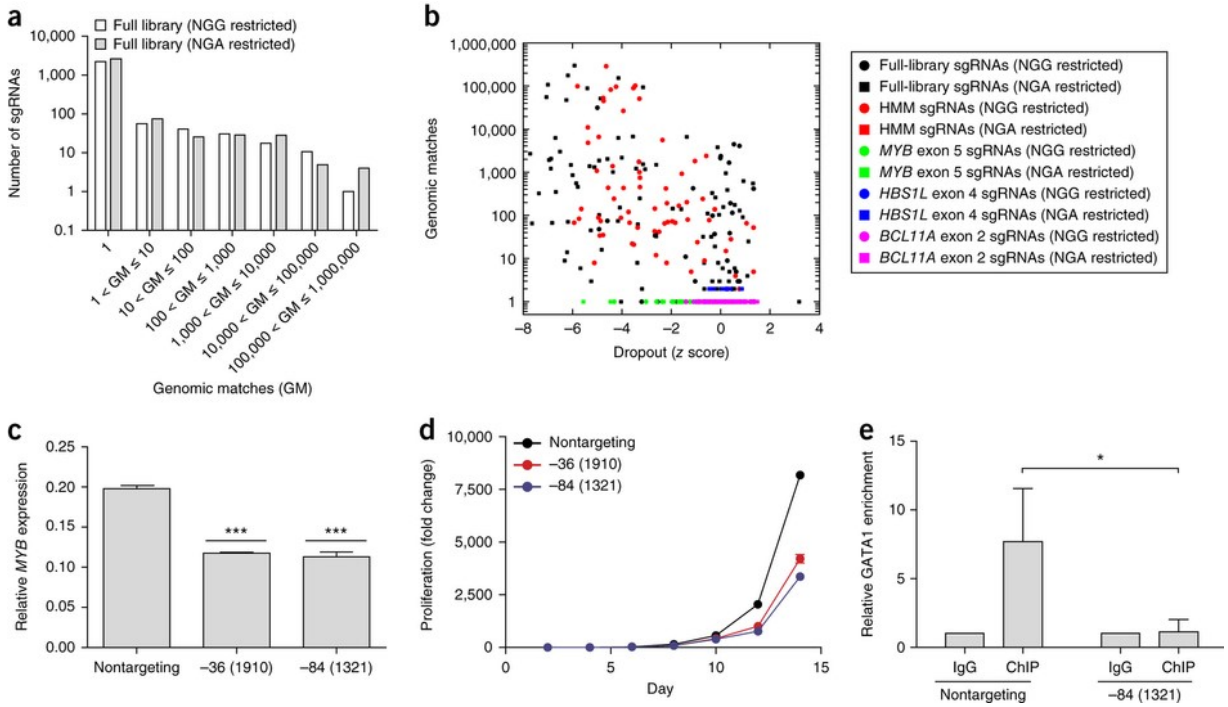


Figure 6. The *HBS1L-MYB* intergenic region contains highly repetitive genomic sequences

(**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; ***$P < 0.0001$.

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. Similarly, a cellular proliferation defect was observed in the CD34$^+$ HSPC-derived erythroblasts. Targeting *MYB* coding sequence had an intermediate phenotype. Targeting *HBS1L* and *BCL11A* coding sequence, −84 DHS (1329), -49 DHS (5430), and −71 DHS (1582) had no effect on cellular proliferation. 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. *HBS1L* expression levels were unchanged. A moderate differentiation block was also observed after targeting of the −126, −83, −71, and −7 DHSs.

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 a 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[453]; 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. [500]). Our approach allowed for high-resolution functional mapping of all DHSs in a ~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[513]. 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.

Methods

HUDEP-2 cell culture.

HUDEP-2 cells were used as previously described[59,60] 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*.[423] (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[514]. 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).

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.*[423]. 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.*[423].:

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

where *n* is the number of potential off-target 'hits', and $S_{hit}(h_i)$ 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[56]. 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[59]. 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[59], *HBS1L* exon 4, and *MYB* exon 5 (Fig. 3c). 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)[4]. 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%. 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[59,430,515,516] 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[496,510,517]. 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[496,506,508,510,517]. 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).

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.*[429], 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[59]. 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 nontargeting 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[59,185]. 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[185]. Repressed regions were used from previously published data[518].

Super-enhancer analysis.

The ROSE algorithm was used to perform super-enhancer analysis[519].

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[56].

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})[520]. The most recent version of the JASPAR database with hg19 sequences was used for the analysis[521].

Hidden Markov model (HMM) analysis.

HMM analysis to identify repressive, active, and neutral sequences was performed as previously described[59].

Red-blood-cell trait meta-analysis.

Red-blood-cell-associated SNPs were taken from a previously published meta-analysis[61]. 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[59,176]. 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[59,176]. 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[512].

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: *MYB* shRNA 1 (TRCN0000295917), *MYB* shRNA 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 (1:1,000; 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[59,522]. 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.

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.

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).

Annex 3: Exome genotyping identifies pleiotropic variants associated with red blood cell traits

Authors

Nathalie Chami, Ming-Huei Chen, Andrew J. Slater, John D. Eicher, Evangelos Evangelou, Salman M. Tajuddin, Latisha Love-Gregory, Tim Kacprowski, Ursula M. Schick, Akihiro Nomura, Ayush Giri, Samuel Lessard, Jennifer A. Brody, Claudia Schurmann, Nathan Pankratz, Lisa R. Yanek, Ani Manichaikul, Raha Pazoki, Evelin Mihailov, W. David Hill, Laura M. Raffield, Amber Burt, Traci M. Bartz, Diane M. Becker, Lewis C. Becker, Eric Boerwinkle, Jette Bork-Jensen, Erwin P. Bottinger, Michelle L. O'Donoghue, David R. Crosslin, Simon de Denus, Marie-Pierre Dubé, Paul Elliott, Gunnar Engström, Michele K. Evans, James S. Floyd, Myriam Fornage, He Gao, Andreas Greinacher, Vilmundur Gudnason, Torben Hansen, Tamara B. Harris, Caroline Hayward, Jussi Hernesniemi, Heather M. Highland, Joel N. Hirschhorn, Albert Hofman, Marguerite R. Irvin, Mika Kähönen, Ethan Lange, Lenore J. Launer, Terho Lehtimäki, Jin Li, David C.M. Liewald, Allan Linneberg, Yongmei Liu, Yingchang Lu, Leo-Pekka Lyytikäinen, Reedik Mägi, Rasika A. Mathias, Olle Melander, Andres Metspalu, Nina Mononen, Mike A. Nalls, Deborah A. Nickerson, Kjell Nikus, Chris J. O'Donnell, Marju Orho-Melander, Oluf Pedersen, Astrid Petersmann, Linda Polfus, Bruce M. Psaty, Olli T. Raitakari, Emma Raitoharju, Melissa Richard, Kenneth M. Rice, Fernando Rivadeneira, Jerome I. Rotter, Frank Schmidt, Albert Vernon Smith, John M. Starr, Kent D. Taylor, Alexander Teumer, Betina H. Thuesen, Eric S. Torstenson, Russell P. Tracy, Ioanna Tzoulaki, Neil A. Zakai, Caterina Vacchi-Suzzi, Cornelia M. van Duijn, Frank J.A. van Rooij, Mary Cushman, Ian J. Deary, Digna R. Velez Edwards, Anne-Claire Vergnaud, Lars Wallentin, Dawn M. Waterworth, Harvey D. White, James G. Wilson, Alan B. Zonderman, Sekar Kathiresan, Niels Grarup, Tõnu Esko, Ruth J.F. Loos, Leslie A. Lange, Nauder Faraday, Nada A. Abumrad, Todd L. Edwards, Santhi K. Ganesh, Paul L. Auer, Andrew D. Johnson, Alexander P. Reiner, Guillaume Lettre

Reference

Chami, N. *et al.* Exome Genotyping Identifies Pleiotropic Variants Associated with Red Blood Cell Traits. *Am J Hum Genet* **99**, 8-21 (2016).

Abstract

Red blood cell (RBC) traits are important heritable clinical biomarkers and modifiers of disease severity. To identify novel coding genetic variants associated with these traits, we conducted meta-analyses of seven RBC phenotypes in 130,273 multi-ethnic individuals from

studies genotyped on an exome array. Following conditional analyses and replication in 27,480 independent individuals, we identified 16 new RBC variants. We found low-frequency missense variants in *MAP1A* (rs55707100, minor allele frequency (MAF=3.3%, $P=2x10^{-10}$ for hemoglobin (HGB)) and *HNF4A* (rs1800961, MAF=2.4%, $P<3x10^{-8}$ for hematocrit (HCT) and HGB). In African Americans, we identified a nonsense variant in *CD36* associated with higher RBC distribution width (rs3211938, MAF=8.7%, $P=7x10^{-11}$), and showed that it is associated with lower *CD36* expression and strong allelic imbalance in *ex vivo* differentiated human erythroblasts. We also identified a rare missense variant in *ALAS2* (rs201062903, MAF=0.2%) associated with lower mean corpuscular volume and mean corpuscular hemoglobin ($P<8x10^{-9}$). Mendelian mutations in *ALAS2* are a cause of sideroblastic anemia and erythropoietic protoporphyria. Gene-based testing highlighted three rare missense variants in *PKLR*, a gene mutated in Mendelian non-spherocytic hemolytic anemia, associated with HGB and HCT (SKAT $P<8x10^{-7}$). The novel rare, low-frequency, and common RBC variants showed pleiotropy, being also associated with platelet, white blood cell, and lipid traits. Our association results and functional annotation suggest the involvement of new genes in human erythropoiesis. We also confirm that rare and low-frequency variants play a role in the architecture of complex human traits, although their phenotypic effect is generally smaller than originally anticipated.

Introduction

One in four cells in the human body is a mature enucleated red blood cell (RBC), also called an erythrocyte. RBC mean lifespan in adults is 100-120 days, requiring constant renewal. To that end, we produce on average 2.4 million RBCs per second in the bone marrow. This massive, yet well-orchestrated cell proliferation process is necessary to accommodate RBCs' main function: to transport oxygen from the lungs to the peripheral organs, and carbon dioxide from the organs to the lungs. Hemoglobin (HGB), the metalloprotein that constitutes by far the most abundant biomolecule found in mature RBC, is responsible for oxygen transport. In addition to their critical role in the circulatory system, RBCs also have secondary, often less appreciated, functions. Within blood vessels, they respond to shear stress and produce the vasodilator nitric oxide to regulate vascular tonus[49]. RBCs participate in antimicrobial strategies

to fight hemolytic pathogens[50] or in the inflammatory response, acting as a reservoir for multiple chemokines[51]. Furthermore, the direct involvement of RBC in adhering to the vascular endothelium or supporting thrombin generation may help to promote blood coagulation or thrombosis[52,53].

Given the paramount importance of RBCs in physiology, it is not surprising that monitoring their features is common practice in medicine to assess the overall health of patients. An excessive number of circulating RBCs (erythrocytosis (MIM 133100)) may suggest a primary bone marrow disease, a myeloproliferative neoplasm such as polycythemia vera (MIM 263300), or chronic hypoxemia due to congenital heart defects. Low HGB concentration and hematocrit (HCT) levels (anemia) may indicate inherited HGB or RBC structural gene mutations, malnutrition, or kidney diseases. By considering the volume (mean corpuscular volume (MCV)), hemoglobin content (mean corpuscular hemoglobin (MCH) and mean corpuscular hemoglobin concentration (MCHC)), or the distribution width (RDW) of RBCs, a physician can distinguish between the different causes of anemia (*e.g.* microcytic/hypochromic due to iron deficiency[84]). In addition, epidemiological studies have correlated high RDW values with a worse prognosis in heart failure patients[523]. RDW is also an independent predictor of overall mortality in healthy individuals, as well as a predictor of mortality in patients with various conditions such as cardiovascular diseases, obesity, malignancies, and chronic kidney disease[88-92].

RBC count and indices vary among individuals, and 40-90% of this phenotypic variation is heritable[135,524-526]. Identifying the genes and biological pathways that contribute to this inter-individual variation in RBC traits could highlight modifiers of severity and/or therapeutic options for several hematological diseases. Already, large-scale genome-wide association studies (GWAS) have found dozens of single nucleotide polymorphisms (SNPs) associated with one or more of these RBC traits[61,62]. However, owing to their design, GWAS are largely insensitive to rare (minor allele frequency (MAF) <1%) and low-frequency (1%≤MAF<5%) genetic variants. Using an exome array, we previously performed an association study for HGB

and HCT in 31,340 European-ancestry individuals, and identified rare coding or splice site variants in the erythropoietin and β-globin genes[527]. Within the framework of the Blood-Cell Consortium (BCX)[528,529], we now report a larger genotyping-based exome survey of seven RBC traits conducted in up to 130,273 individuals, including 23,896 participants of non-European ancestry. By performing this experiment, our initial goals were to expand the list of rare and low-frequency coding or splice site variants associated with RBC traits, and to explore whether the exome array can complement the GWAS approach to fine-map RBC causal genes.

Subjects and methods

Study participants

The Blood-Cell Consortium (BCX) aims to identify novel common and rare variants associated with blood-cell traits using an exome array. BCX is comprised of over 134,021 participants from 24 discovery cohorts and five ancestries: European, African American, Hispanic, East Asian, and South Asian. BCX is interested in the genetics of all main hematological measures and is divided into three main working groups: RBC, white blood cell (WBC)[529], and platelet (PLT)[528]. For the RBC working group, we analyzed seven traits available in up to 130,273 individuals: RBC count ($x10^{12}$/L), HGB (g/dL), HCT (%), MCV (fL), MCH (pg), MCHC (g/dL), and RDW (%). The BCX procedures were in accordance with the institutional and national ethical standards of the responsible committees and proper informed consent was obtained.

Genotyping and quality-control steps

Participants from the different studies were genotyped on one of the following exome chip genotyping arrays: Illumina ExomeChip v1.0, Illumina ExomeChip v1.1_A, Illumina ExomeChip-12 v1.1, Affymetrix Axiom Biobank Plus GSKBB1, Illumina HumanOmniExpressExome Chip. Genotypes were then called either 1) with the Illumina GenomeStudio GENCALL and subsequently recalled using zCALL; or 2) by the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium Exome Chip effort[530]. The same quality-control steps were followed by each participating study. We

excluded variants with low genotyping success rate (<95%, except for WHI that used a cutoff <90%). Samples with call rate <95% (except for SOLID-TIMI 52 and STABILITY that used 94.5% and 93.5% cutoffs, respectively) after joint or zCALL calling and with outlying heterozygosity rate were also excluded. Other exclusions were deviation from Hardy-Weinberg equilibrium ($P<1x10^{-6}$) and gender mismatch. We performed principal component analysis (PCA) or multidimensional scaling (MDS) and excluded sample outliers from the resulting plots through visual inspection, using populations from the 1000 Genomes Project to anchor these analyses. Keeping only autosomal and X-chromosome variants for the analysis, we aligned all variants on the forward strand and created a uniform list of reference alleles using the --force alleles command in PLINK[338]. Finally, an indexed variant call format file (VCF) was created by each study and checked for allele alignment and any allele or strand flips using the checkVCF package[531]. Prior to performing meta-analyses of the association results provided by each participating study, we ran the EasyQC protocol[532] to check for population allele frequency deviations and proper trait transformation in each cohort.

Phenotype modeling and association analyses

When possible, we excluded individuals with blood cancer, leukemia, lymphoma, bone marrow transplant, congenital or hereditary anemia, HIV, end-stage kidney disease, dialysis, splenectomy, and cirrhosis, and those with extreme measurements of RBC traits. We also excluded individuals on erythropoietin treatment or chemotherapy. Additionally, we excluded pregnant women and individuals with acute medical illness at the time the complete blood count (CBC) was done. For the seven RBC traits, within each study, we adjusted for age, age-squared, gender, the first 10 principal components, and, where applicable, other study-specific covariates such as study center using a linear regression model. Within each study, we then applied inverse normal transformation on the residuals and tested the phenotypes for association with the ExomeChip variants using either RVtests (version 20140416)[533] or RAREMETALWORKER.0.4.9[534].

Discovery meta-analyses

Score files generated by RVtests or RAREMETALWORKER from each participating study were used to carry out meta-analyses of the single variant association results using

RareMETALS v.5.9[535]. All analyses were performed separately in each of European (EA) and African-American (AA) ancestries. In the multi-ancestry meta-analyses, we combined individuals of European, African-American, Hispanic, East-Asian, and South-Asian ancestries (All). We included variants with allele frequency difference between the highest and lowest MAF <0.3 for European and African-American ancestries, and <0.6 for the combined ancestry meta-analyses. For the gene-based analyses, we used score files and variance-covariance matrices from the study-specific association results, and applied the sequence kernel association test (SKAT)[491] and variable threshold (VT) algorithms[536] in RareMETALS considering only missense, nonsense and splice site variants with a MAF <1%. Gene-based analyses were also stratified by ancestry. Significance thresholds were determined using Bonferroni correction assuming ~250,000 independent variants ($P<2x10^{-7}$ for the single variant analyses) and ~17,000 genes tested on the ExomeChip ($P<3x10^{-6}$ for the gene-based tests).

Conditional analysis and replication

In order to identify independent signals, we performed conditional analyses. In each round of conditional analysis, we conditioned on the most significant single variant in a 1 Mb window. These conditional analyses were performed at the meta-analysis level using RareMETALS. We repeated this step until there were no new signals identified in each region, defined as $P<2x10^{-7}$. We then checked for linkage disequilibrium (LD) within the list of variants that was retained from the conditional analyses. For variants that were in moderate-to-strong LD ($r^2 \geq 0.3$), we kept the most significant. We attempted replication of the final list of independent variants in eight additional studies that contributed a total of 27,480 individuals (N=21,473 for EA and N=6,007 for AA). The division of discovery and replication samples was dictated by timing as we collected all groups we were aware of for initial discovery then found others who could only participate much later and hence were used for replication. These studies followed similar analytical procedures and steps as those followed by the discovery analysis (see above). A joint meta-analysis of the discovery and the replication results was carried out using a fixed-effects model and inverse-variance weighting as implemented in METAL[312]. We considered as replicated markers those with a nominal $P_{replication}<0.05$ and an effect on phenotype in the same direction as in the discovery results.

Allelic imbalance and expression of *CD36*

We checked for allelic imbalance (AI) of the rs3211938 variant in *CD36* (MIM 173510) as well as the expression of the gene in 12 samples of fetal liver erythroblasts obtained from anonymous donors. Details on the protocol including RNA extraction and sequencing can be found elsewhere[377]. We calculated the difference in the ratio of reads of the reference allele (T) and the alternate allele (G) of rs3211938. Briefly, reads overlapping rs3211938 were counted using samtools (v 1.1) mpileup software using genome build hg19. We kept uniquely mapping reads using -q 50 argument (mapping quality > 50) and sites with base quality >10. Statistical significance of the difference in the ratio of reads between the reference allele and the alternate allele was assessed using a binomial test. For each sample, we summed all reads overlapping all heterozygous SNPs and calculated the expected ratio within each SNP allele combination. Reads that fall in the top 25th coverage percentile were down-sampled so that the highest covered sites do not bias the expected ratio[257]. For rs3211938, the expected T:G ratio was 0.507.

Expression quantitative trait loci (eQTL) analysis

We cross-referenced our list of RBC novel variants with over 100 separate expression quantitative trait loci (eQTL) published datasets. Datasets were collected through publications, publically available sources, or private collaborations. A general overview of a subset of >50 eQTL studies has been published[537], with specific citations for >100 datasets included in the current query followed here. A complete list of tissues and studies used can be found in the **Supplemental Data**. We considered SNPs that are themselves expression SNPs (eSNP) when they meet a $P<0.0001$ threshold or when they are in LD ($r^2>0.3$) with the best eSNP ($P<0.0001$).

Results

Single-variant meta-analyses

We meta-analyzed ExomeChip results for seven RBC-related phenotypes (RBC count, HCT, HGB, MCH, MCHC, MCV, and RDW) available in up to 130,273 individuals from 24 studies and five ancestries. Across these different phenotypes, a total of 226 variants reached exome-wide significance ($P<2x10^{-7}$) in the combined ancestry analyses (**Figure 1**). Given that

some of these RBC traits are correlated, these associated variants highlight 71 different loci (defined using a 1 Mb interval). Overall, we observed only modest inflation of the test statistics (λ_{GC}=1.03-1.05), consistent with little confounding due to technical artifacts, population stratification, or cryptic relatedness.

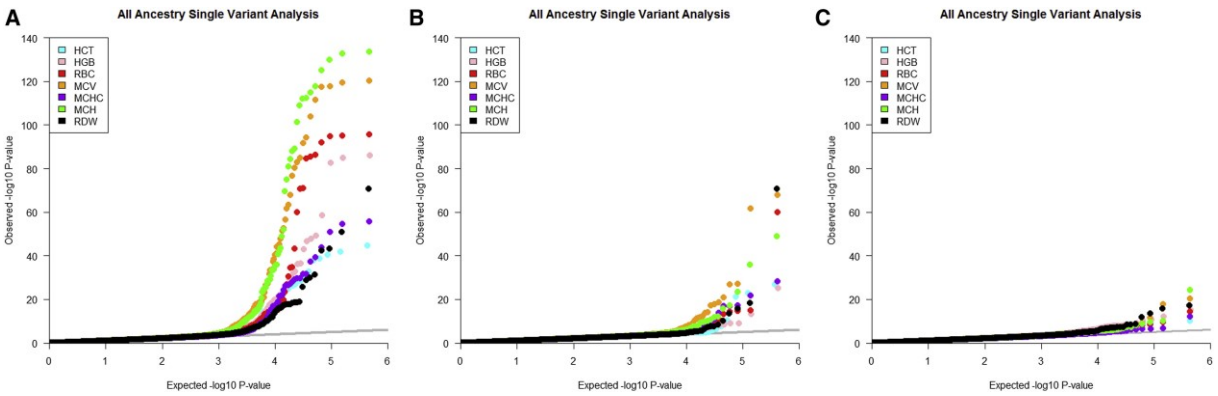


Figure 1. Quantile-Quantile (QQ) plots of single variant association results in the all ancestry meta-analyses for the seven red blood cell (RBC) traits analyzed.
(A) Distribution of the single variant results for all variants tested on the exome array. (**B**) Only markers with a minor allele frequency <5% are shown here. (**C**) Variants outside of known RBC GWAS regions. Variants that are within 1 Mb from a previously published RBC GWAS locus were excluded for this QQ plot. HCT, hematocrit; HGB, hemoglobin; RBC, red blood cell count; MCV, mean corpuscular volume; MCHC, mean corpuscular hemoglobin concentration; MCH, mean corpuscular hemoglobin; RDW, red blood cell distribution width.

In order to identify independent variants, we performed conditional analyses at the meta-analysis level adjusting for the effect of the most significant variant in a 1 Mb region in a stepwise manner (**Subjects and Methods**). Following this analysis, we obtained a list of 126 independent variants associated with at least one RBC trait at $P<2\times10^{-7}$. Selecting only variants that lie more than 1 Mb away from a known GWAS locus resulted in 23 independent variants located within 20 novel RBC loci (**Table 1**). We attempted to replicate these 126 variants in eight independent cohorts totaling 27,480 participants. Overall, we observed a strong replication, with 94 of the 126 variants showing consistent direction of effect between the discovery and replication analyses (binomial $P=3\times10^{-8}$). Of the 23 novel RBC variants, we replicated 16 at nominal $P<0.05$ for at least one RBC trait (binomial $P=3\times10^{-16}$, **Table 1**). Out of these 16 novel and replicated RBC variants, there are five missense variants, including two variants with MAF<5% in *MAP1A* (MIM 600178) and *HNF4A* (MIM 600281), and one

nonsense variant in *CD36* (**Table 1**). Among the remaining nine novel and replicated RBC variants, there are five intronic, one synonymous, one 5'UTR, and one intergenic markers (**Table 1**).

Prioritization of candidate genes and genetic variants

Our single-variant analyses in European-ancestry samples identified one rare missense variant in *ALAS2* (MIM 301300) associated with MCV and MCH (rs201062903, p.Pro507Leu, MAF=0.2%) (**Table 1**). The association with this variant did not replicate, potentially because of limited statistical power (the replication sample size for this rare marker was 5,044; see also **Discussion**). *ALAS2* encodes 5-aminolevulinate synthase 2, the rate-controlling enzyme of erythroid heme synthesis. Additionally, rare mutations in *ALAS2* cause X-linked sideroblastic anemia (MIM 300751) and erythropoietic protoporphyria (MIM 300752). Thus, despite the lack of replication, *ALAS2* remains an excellent candidate gene to modulate RBC traits. The *ALAS2* p.Pro507Leu variant, which is not reported in the ClinVar database, maps between two amino acids (p.Tyr506 and p.Thr508) that are important for catalytic activity and known to be mutated in cases of sideroblastic anemia[538].

Two low-frequency missense variants identified in our analyses implicate *MAP1A* and *HNF4A* in RBC biology (**Table 1**). *MAP1A* encodes microtubule-associated protein 1A, a gene highly expressed in the nervous system and mostly studied in the context of neuronal diseases, although it is expressed in many additional tissues, including hematopoietic cells[539]. Deletion of *MAP1A* in the mouse causes defects in synaptic plasticity[540]. This observation is interesting given that inactivation of *ANK1* (MIM 612641), another gene that codes for a cytoskeleton protein expressed in neurons and RBC, is associated with neurological dysfunction in the mouse and spherocytosis and hemolytic anemia in humans (MIM 182900). Our meta-analyses confirmed two known independent *ANK1* variants associated with MCHC: an intronic SNP (rs4737009, MAF=19.8%, P=1.3x10^{-8}) and a low-frequency missense variant (rs34664882, p.Ala1462Val, MAF=2.9%, P=1.7x10^{-16}) [61,541].

Table 1. Association results of variants in novel loci associated with red blood cell (RBC) traits.

	Marker Info					Discovery				Replication				Combined	
Trait	**Position**	**A1/A2**	**SNP**	**Annotation**	**Gene**	**N**	**AF (A2)**	**Beta (se)**	**P-value**	**N**	**AF (A2)**	**Beta (se)**	**P-value**	**Beta (se)**	**P-value**
RDW-EA	1:25768937	A/G	**rs10903129**	intron	*TMEM57-RHD*	45573	0.544	0.037(0.007)	1.19E-07	18475	0.560	0.023(0.011)	0.0373	0.033(0.006)	2.41E-08
RDW-All	1:25768937	A/G	**rs10903129**	intron	*TMEM57-RHD*	56194	0.568	0.034(0.006)	9.58E-08	24474	0.600	0.021(0.01)	0.0252	0.03(0.005)	1.32E-08
HCT-All	1:155162067	C/T	**rs4072037**	synonymous	*MUC1*	109875	0.554	0.025(0.005)	5.82E-08	25006	0.563	0.038(0.009)	5.96E-05	0.027(0.004)	3.47E-11
HGB-All	2: 27741237	T/C	rs780094	intron	*GCKR*	130,273	0.626	0.024 (0.004)	7.14E-08	3162	0.626	-0.012(0.026)	0.6410	0.023(0.044)	1.68E-07
RBC-All	2:219509618	C/A	**rs2230115**	missense	*ZNF142*	74488	0.509	0.033(0.006)	9.74E-09	27442	0.477	0.024(0.01)	0.0167	0.031(0.005)	7.11E-10
HCT-All	3:56771251	A/C	**rs3772219**	missense	*ARHGEF3*	109875	0.338	-0.028(0.005)	2.38E-09	25006	0.366	-0.021(0.01)	0.0292	-0.027(0.004)	2.56E-10
HGB-All	3:56771251	A/C	**rs3772219**	missense	*ARHGEF3*	130273	0.336	-0.026(0.004)	3.76E-09	27749	0.367	-0.02(0.009)	0.0331	-0.025(0.004)	4.33E-10
HCT-EA	4:88008782	G/A	rs236985	intron	*AFF1*	87444	0.394	0.032(0.005)	3.89E-10	19968	0.405	0.02(0.011)	0.0626	0.03(0.005)	1.14E-10
RBC-EA	4:88008782	G/A	**rs236985**	intron	*AFF1*	60231	0.393	0.034(0.006)	3.50E-08	21435	0.405	0.023(0.011)	0.0273	0.031(0.005)	4.22E-09
HGB-EA	4:88030261	G/T	**rs442177**	intron	*AFF1*	106377	0.595	-0.034(0.005)	3.97E-13	21743	0.586	-0.029(0.01)	0.0052	-0.033(0.004)	8.23E-15
RDW-EA	5:127371588	A/G	**rs10063647**	intron	*LINC01184 - SLC12A2*	45573	0.463	-0.05(0.007)	1.72E-13	18475	0.480	-0.033(0.011)	0.0018	-0.045(0.006)	2.88E-15
RDW-All	5:127371588	A/G	**rs10063647**	intron	*LINC01184 - SLC12A2*	56194	0.506	-0.044(0.006)	2.11E-12	24474	0.545	-0.03(0.01)	0.0014	-0.04(0.005)	2.37E-14
RDW-EA	5:127522543	C/T	**rs10089**	utr_5p	*LINC01184 - SLC12A2*	45573	0.21	0.051(0.008)	8.45E-10	16692	0.215	0.058(0.014)	2.71E-05	0.053(0.007)	1.15E-13
RDW-All	5:127522543	C/T	**rs10089**	utr_5p	*LINC01184 - SLC12A2*	56194	0.207	0.044(0.008)	4.08E-09	22691	0.208	0.045(0.012)	0.0001	0.044(0.006)	2.73E-12
HGB-All	6: 7247344	C/A	**rs35742417**	missense	*RREB1*	130273	0.174	0.030 (0.005)	1.17E-08	4074	0.207	0.065 (0.028)	0.0190	0.032 (0.005)	1.50E-09
RDW-AA	7:80300449	T/G	**rs3211938**	nonsense	*CD36*	6666	0.087	0.174(0.031)	2.36E-08	5999	0.086	0.139(0.032)	1.83E-05	0.161(0.025)	7.09E-11
RDW-All	7:80300449	T/G	**rs3211938**	nonsense	*CD36*	55510	0.012	0.171(0.029)	5.29E-09	22691	0.023	0.139(0.032)	1.61E-05	0.157(0.022)	5.12E-13
RDW-EA	8:126490972	A/T	**rs2954029**	intergenic	*TRIB1*	45573	0.46	0.036(0.007)	1.53E-07	16692	0.466	0.026(0.011)	0.0210	0.034(0.006)	1.29E-08
RDW-All	8:126490972	A/T	**rs2954029**	intergenic	*TRIB1*	56194	0.439	0.032(0.006)	1.83E-07	22691	0.432	0.021(0.01)	0.0298	0.029(0.005)	2.54E-08
MCH-All	10:105659826	T/C	rs2487999	missense	*OBFC1*	66318	0.869	0.047(0.009)	4.12E-08	26749	0.861	0.025(0.013)	0.0601	0.041(0.007)	1.75E-08
MCH-AA	11:92722761	G/A	rs1447352	intergenic	*MTNR1B*	8273	0.557	0.089(0.016)	1.85E-08	5038	0.562	-0.022(0.02)	0.2713	0.07(0.014)	1.08E-06
HGB-EA	15:43820717	C/T	**rs55707100**	missense	*MAP1A*	106377	0.033	-0.071(0.013)	1.65E-08	21743	0.0223	-0.099(0.033)	0.0028	-0.075(0.012)	2.31E-10
MCV-AA	16:1551082	A/G	**rs2667662**	intron	*TELO2*	10849	0.725	-0.099(0.015)	1.79E-10	5034	0.724	-0.093(0.022)	3.02E-05	-0.098(0.014)	7.32E-12
MCV-AA	16:2812939	C/A	**rs2240140**	missense	*SRRM2*	8525	0.118	0.134(0.025)	7.08E-08	6002	0.124	0.106(0.027)	0.0001	0.128(0.022)	5.24E-09
HCT-EA	17:59017025	T/C	rs8080784	intron	*BCAS3-TBX2*	79344	0.158	-0.039(0.007)	2.62E-08	19968	0.148	0.011(0.014)	0.4349	-0.029(0.006)	3.39E-06
HGB-EA	17:59483766	C/T	rs8068318	intron	*BCAS3-TBX2*	106377	0.722	-0.026(0.005)	1.53E-07	21743	0.730	-0.021(0.011)	0.0565	-0.025(0.005)	2.55E-08

MCV-EA	20:31140165	C/T	**rs4911241**	intron	*NOL4L*	61462	0.241	-0.04(0.007)	1.25E-08	21714	0.252	-0.025(0.012)	0.0302	-0.036(0.006)	2.01E-09
RDW-EA	20:31140165	C/T	**rs4911241**	intron	*NOL4L*	45573	0.242	0.043(0.008)	5.79E-08	18475	0.240	0.049(0.012)	7.44E-05	0.045(0.007)	2.01E-11
RDW-All	20:31140165	C/T	**rs4911241**	intron	*NOL4L*	56194	0.235	0.038(0.007)	1.56E-07	24474	0.222	0.044(0.011)	6.10E-05	0.04(0.006)	4.60E-11
HCT-EA	20:43042364	C/T	**rs1800961**	missense	*HNF4A*	79344	0.024	0.083(0.015)	1.44E-08	19968	0.033	0.082(0.028)	0.0037	0.083(0.013)	1.91E-10
HGB-EA	20:43042364	C/T	**rs1800961**	missense	*HNF4A*	98277	0.032	0.073(0.013)	2.53E-08	21743	0.032	0.062(0.027)	0.0232	0.071(0.012)	1.93E-09
HCT-All	20:43042364	C/T	**rs1800961**	missense	*HNF4A*	100313	0.022	0.077(0.014)	2.31E-08	25006	0.027	0.091(0.028)	0.0010	0.08(0.012)	9.88E-11
HGB-All	22:44324727	C/G	rs738409	missense	*PNPLA3*	130273	0.223	0.028(0.005)	2.24E-08	4074	0.218	0.053(0.027)	0.0504	0.029(0.005)	4.81E-09
MCH-EA	X:55039960	G/A	rs201062903	missense	*ALAS2*	52758	0.002	-0.324(0.053)	7.32E-10	5855	0.001	-0.291(0.235)	0.215	-0.323(0.052)	5.81E-10
MCH-All	X:55039960	G/A	rs201062903	missense	*ALAS2*	65067	0.002	-0.322(0.051)	3.36E-10	10893	0.001	-0.276 (0.224)	0.218	-0.321(0.051)	2.68E-10
MCV-EA	X:55039960	G/A	rs201062903	missense	*ALAS2*	60211	0.002	-0.285 (0.049)	7.11E-09	5044	0.001	-0.178 (0.248)	0.472	-0.282 (0.049)	6.11E-09

Variants in novel loci with $P<2x10^{-7}$ and that were retained after conditional analyses are presented here. All variants are >1Mb apart from a known GWAS signal for RBC traits. Chromosome positions are given on human genome build hg19. Allele frequency and effect size are given for the alternate (A2) allele. Replication was carried out in six cohorts for EA and two cohorts for AA and was performed in RareMetals; meta-analyses of the discovery and replication cohorts are presented under "Combined" and were carried out in METAL. Variants that replicated with a nominal $P<0.05$ are highlighted in bold. EA, European American; AA, African American; All, combined ancestry (EA + AA + Asians + Hispanics); A1, reference allele; A2, alternate allele; N, sample size; AF, allele frequency; se, standard error; HCT, hematocrit; HGB, hemoglobin; RBC, red blood cell count; MCV, mean corpuscular volume; MCHC, mean corpuscular hemoglobin concentration; MCH, mean corpuscular hemoglobin; RDW, red blood cell distribution width

In the accompanying BCX PLT article[528], we report that the same *MAP1A* rs55707100 allele (p.Pro2349Leu) associated here with decreased HGB concentration is also associated with increased PLT count. Furthermore, recent studies have identified associations between rs55707100 and HDL-cholesterol and triglycerides levels[542]. Adding to the complexity, the GTEx dataset indicates that rs55707100 is an expression quantitative trait locus (eQTL) for *ADAL* (P_{eQTL}=9x10^{-11}) but not for *MAP1A*[246]. *ADAL* is a poorly characterized adenosine deaminase-like protein that is highly expressed in human erythroblasts. However, the eQTL association between rs55707100 and *ADAL* could simply reflect "LD shadowing" from nearby markers that are much stronger eQTL variants for *ADAL*. Indeed, rs3742971 (a common variant located in *ADAL*'s 5'UTR) is in partial LD with rs55707100 (r^2=0.18 in European populations from the 1000 Genomes Project) and strongly associated with *ADAL* expression levels (P_{eQTL}=6x10^{-49}).

The second low-frequency missense variant associated with HGB and HCT maps within the coding sequence of the transcription factor HNF4A (**Table 1**). This marker, rs1800961 (p.Thr117Ile), has previously been associated with HDL- and total cholesterol, C-reactive protein, fibrinogen, and coagulation factor VII levels[543-546]. Mutations in *HNF4A* cause maturity-onset diabetes of the young (MODY, MIM 125851) and a common intronic SNP in *HNF4A* (rs4812829) has been associated with type 2 diabetes (MIM 125853) risk[547]. The missense rs1800961 associated with HGB and HCT is only in weak LD with rs4812829 (r^2=0.021 in European-ancestry populations from the 1000 Genomes Project). Querying recently released ExomeChip data, we found that rs1800961 is also associated with T2D risk in ~82,000 participants (P=9.5x10^{-7}, odds ratio=1.16)[548]. *HNF4A* is expressed in the kidney and could influence HGB and HCT through the regulation of erythropoietin production[396]. It is also abundantly expressed in the liver, where it could indirectly affect HGB and HCT levels through an effect on blood lipid levels (see **Discussion**). *HNF4A* is detectable at low levels in erythroblasts, and the BLUEPRINT Project has found that some *HNF4A* isoforms may be more highly expressed in this cell type[549].

In AA, we identified a nonsense variant (rs3211938, p.Tyr325Ter, MAF=8.7%, P=7.1x10^{-11}) in *CD36* associated with RDW. This variant displays a wide variation in allele frequency between AA and EA (MAF_{EA}=0.01%). The association is slightly improved in the ancestry-combined meta-analysis (P=5.1x10^{-13}) because there is also evidence of association in Hispanics (MAF=1.9%, P=0.022) (**Table 1**). We examined a dataset of *ex vivo* differentiated human erythroblasts to determine if this *CD36* nonsense variant shows allelic imbalance (AI)[377]. All samples were homozygous at rs3211938 for the reference allele with the exception of one heterozygous sample (FL11). FL11 had the lowest level of *CD36* expression among the 12 samples tested, and demonstrated strong AI where we observe 705 sequence reads for the reference allele (T) *vs*. 126 for the alternate allele (G) (P=4.9x10^{-95}, **Figure 2**). To confirm this finding in independent samples, we queried the GTEx dataset, which has compiled RNA-sequencing and genotype information from multiple human tissues[246]. GTEx does not include data for human erythroblasts. However, it detected a strong eQTL effect of rs3211938 on *CD36* expression in whole blood (P_{eQTL}=1.1x10^{-15}), with carriers of the G-allele expressing less *CD36*. Furthermore, GTEx reported evidence for moderate AI in multiple tissues for *CD36*-rs3211938, with the G-allele being under-represented among sequence reads. These results strongly support our observations in human erythroblasts.

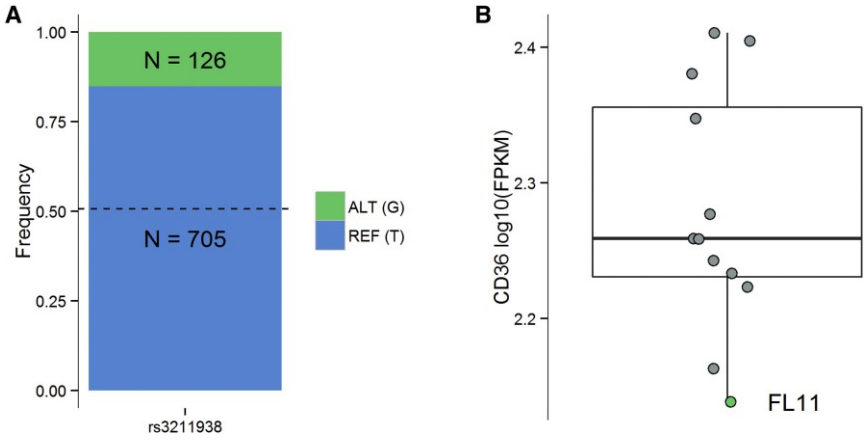


Figure 2. *CD36* expression in human erythroblasts.
(A) In a dataset of 12 human fetal liver erythroblasts, all samples were homozygous at rs3211938 for the reference T-allele with the exception of one heterozygous sample (FL11). FL11 demonstrated strong allelic imbalance: we observed 705 reads for the reference allele (T) and 126 reads for the alternate allele (G)(binomial P=4.9x10^{-95}). (B) FL11 (in green) shows the lowest *CD36* expression level when compared to the other 11 samples. FPKM, fragments per kilobase of transcript per million mapped reads.

eQTL analysis

To prioritize additional causal genes at RBC loci that contain non-coding variants, we cross-referenced our list of novel variants with over 100 published eQTL datasets (**Subjects and Methods**). Overall, 12 variants were significant eQTLs in a wide variety of tissues. The most interesting eQTL finding is the association between rs10903129, a common marker associated with RDW in our analyses and located within an intron of *TMEM57* (MIM 610301), and the expression of *RHD* (MIM 111680) in whole blood. *RHD* is located 112 kb downstream of *TMEM57* and encodes the D antigen of the clinically significant Rhesus (Rh) blood group. rs10903129 has also been associated with total cholesterol levels and erythrocyte sedimentation rate (ESR)[550,551]. The association with ESR is particularly intriguing given that it is considered a non-specific indicator of inflammation. As described above, RDW is also abnormal in chronic diseases, such as atherosclerosis and diabetes, which have an important inflammation component.

Gene-based association testing

Despite our large sample size, statistical power remains limited for rare variants of weak-to-moderate phenotypic effect. To try to capture these genetic factors, we performed gene-based testing by aggregating coding and splice site variants with MAF <1% within each gene (**Subjects and Methods**). The SKAT analyses identified two genes: *ALAS2* associated with MCH, and *PKLR* (MIM 609712) associated with HGB and HCT (**Table 2**). The *ALAS2* signal was driven by a single rare missense variant (rs201062903) and was described above. *PKLR* encodes the erythrocyte pyruvate kinase (PK) that catalyzes the last step of glycolysis. PK deficiency, usually caused by recessive mutations, is one of the main causes of non-spherocytic hemolytic anemia (MIM 266200). In fact, one of the variants identified in our meta-analysis (rs116100695, p.Arg486Trp, MAF=0.3%, $beta_{HGB}$= -0.242 g/dl, $P_{HGB}=1.2x10^{-5}$) is a frequent cause of PK deficiency in Italian and Spanish patients[552,553]. This variant was confirmed in the replication cohorts ($P_{replication}$=0.039). Two additional *PKLR* rare missense variants contribute to the gene-based association statistic with HGB and HCT: rs61755431 (p.Arg569Gln, MAF=0.2%, $beta_{HGB}$=-0.179 g/dl , P_{HGB}=0.006) and rs8177988 (p.Val506Ile, MAF=0.6%,

$beta_{HGB}$=+0.116 g/dl , P_{HGB}=0.003). It is noteworthy that the p.Val506Ile substitution is associated with increased HGB concentration given that this amino acid maps to a PKLR structural domain necessary for protein interaction[554]. This heterogeneity of effect among the *PKLR* missense variants may explain why SKAT's result is more significant than VT's for this gene (**Table 2**). A third gene, *ALPK3*, was identified only in the VT analysis for association with MCHC (**Table 2**). *ALPK3* encodes a kinase previously implicated in cardiomyocyte differentiation[555]. We could not test for replication the association between *ALPK3* and MCHC given the rarity of its coding alleles.

Table 2. Gene-based association results

Trait	Gene	N	Number of variants analyzed	VT *P*-value	SKAT P-value	Top variant	Top-variant MAF	Top-variant P-value
HGB-EA	*PKLR*	106,377	15	$1.92x10^{-5}$	7.02x10-	rs116100695	0.003	$1.17x10^{-5}$
HGB-All	*PKLR*	130,273	15	0.00016	$6.57x10^{-7}$	rs116100695	0.003	$1.94x10^{-5}$
HCT-All	*PKLR*	109,875	15	$3.96x10^{-5}$	$7.95x10^{-7}$	rs116100695	0.003	$2.49x10^{-5}$
MCH-EA	*ALAS2*	54,009	11	$4.78x10^{-6}$	$5.79x10^{-7}$	rs201062903	0.002	$7.32x10^{-10}$
MCHC-All	*ALPK3*	84,841	28	$1.95x10^{-6}$	0.793	rs202037221	$3.0x10^{-5}$	0.0005

Gene-based results of the VT and SKAT algorithms for genes associated with RBC traits at $P<3x10^{-6}$. We analyzed non-synonymous coding (nonsense, missense) and splice site variants with a minor allele frequency (MAF) <1%. EA, European American; All, combined ancestry (EA + AA + Asians + Hispanics); N, sample size; HCT, hematocrit; HGB, hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCH, mean corpuscular hemoglobin.

RBC variants and pleiotropic effects

Besides the overlap within the RBC traits themselves, we identified seven novel RBC variants associated with other blood-cell type traits, or with lipid levels (**Figure 3** and **Table 3**). To assess whether the genetic associations with RBC traits are independent of lipid levels, we performed additional analyses in a subset of BCX participants from 3 of our studies (FHS, MHIBB, and FHS) ranging from ~10,000 to 23,000 individuals. We repeated the association analyses for five RBC loci (*TMEM57-RHD* rs10903129, *AFF1* rs442177, *TRIB1* rs2954029,

MAP1A rs55707100, and *HNF4A* rs1800961) additionally adjusting for the respective lipid trait, and combined the results across the 3 studies using fixed-effect meta-analysis. There was little or no change in the effect size or P-values associated with the five RBC trait loci upon adjustment for the corresponding lipid trait, suggesting that the RBC and lipid associations are independent of one another and thus represent true "pleiotropic" genetic effects.

Table 3. Overlap of red blood cell (RBC) markers with other blood cell traits and/or lipids

SNP	Position	A1/A2	AF (A2)	Annotation	Gene	Trait	Beta	*P*-value
rs10903129	1:25768937	A/G	0.568	intron	*TMEM57-RHD*	RDW	0.037	1.19×10^{-7}
						TC[550]	0.061	5.40×10^{-10}
						PLT	-0.021	7.06×10^{-6}
rs3772219	3:56771251	A/C	0.338	missense	*ARHGEF3*	HCT*	-0.028	2.38×10^{-9}
						HGB*	-0.026	3.76×10^{-9}
						PLT	0.031	5.93×10^{-10}
						HGB	-0.034	3.97×10^{-13}
rs442177	4:88030261	G/T	0.595	intron	*AFF1*	TG[544]	-0.031	1.00×10^{-18}
						BASO	-0.030	1.99×10^{-5}
rs2954029	8:126490972	A/T	0.439	intergenic	*TRIB1*	RDW	0.036	1.53×10^{-7}
						TG[544]	-0.076	1.0×10^{-107}
rs55707100	15:43820717	C/T	0.033	missense	*MAP1A*	HGB	-0.071	1.65×10^{-8}
						PLT	0.095	7.03×10^{-14}
						TG[556]	0.090	1.40×10^{-17}
						MCV	-0.040	1.25×10^{-8}
rs4911241	20:31140165	C/T	0.241	intron	*NOL4L*	RDW	0.043	5.79×10^{-8}
						BASO	-0.051	1.35×10^{-10}
						MONO	-0.033	3.57×10^{-5}
rs1800961	20:43042364	C/T	0.032	missense	*HNF4A*	HCT	0.083	1.44×10^{-8}
						HGB	0.073	2.53×10^{-8}
						HDL[544]	-0.127	2.00×10^{-34}

Shown here are significant novel variants from the RBC traits association analyses that overlap with other blood-cell traits or with lipids. Results for the white blood cell and platelet traits are from the Blood-Cell Consortium, and results for lipids are from the published literature. Results are presented for European-ancestry individuals, except

in the presence of an asterisk (*), which stands for result from "All" ancestry. The allele frequency and direction of the effect (beta) is given for the A2 allele. A1, reference allele; A2, alternate allele; AF, allele frequency; HCT, hematocrit; HGB, hemoglobin; MCV, mean corpuscular volume; RDW, red blood cell distribution width; TC, total cholesterol; PLT, platelet; TG, triglycerides; WBC, white blood cells; BASO, basophils; MONO, monocytes; HDL, HDL-cholesterol.

A correlated response to, or role in inflammation might explain why some of the RBC variants are also associated with WBC, PLT, or lipid traits. Another plausible explanation for the concomitant association of several markers with RBC, WBC, and PLT phenotypes could be a more general effect of these genes on the proliferation or differentiation of hematopoietic progenitor cells. This is most likely the case for *JAK2* (MIM 147796) and *SH2B3* (MIM 605093), two key regulators of hematopoietic cells (**Figure 3**). In this category, we also observed two novel findings, *AFF1* (MIM 159557) and *NOL4L*, which are associated with RBC and WBC phenotypes, and have been previously implicated in leukemia[557,558]. Additionally, we identified a novel missense variant in *ARHGEF3* (MIM 612115) associated with HGB and HCT. Other than its association with PLT traits, *ARHGEF3* plays a role in the regulation of iron uptake and erythroid cell maturation[559].

Discussion

We present multi-ethnic meta-analyses of seven RBC traits using ExomeChip results of 130,273 individuals. Our statistical thresholds to declare significance at the discovery stage ($P<2\text{x}10^{-7}$ in the single-variant analyses) was adjusted for the approximate number of variants genotyped on the ExomeChip (Bonferroni correction for 250,000 variants), but we decided not to adjust it for the seven RBC phenotypes tested because of the high correlation between some of these traits. Instead, we relied on independent replication to distinguish true from likely false positive associations. Despite the limited size of our replication set (27,480 individuals), it was encouraging to detect a strong replication of direction of effect for known and novel RBC variants, suggesting a low false discovery rate. In total, we identified 23 novel variants associated with RBC traits in the single-variant analyses, and a collection of three rare missense variants in *PKLR* associated with HGB and HCT in the gene-based analyses. Out of the 23 novel RBC variants, 16 were replicated at $P<0.05$ in the independent samples (**Table 1**). To inform

our replication criteria, we conducted a power analysis using a sample size of 20,000 and considering multiple combinations of allele frequencies and effect sizes. Based on allele frequency and effect size, one of our most difficult to replicate variant was rs1800961 (MAF=0.022, Beta=0.028). However, we still had approximately 56% power to detect this association in the replication stage.

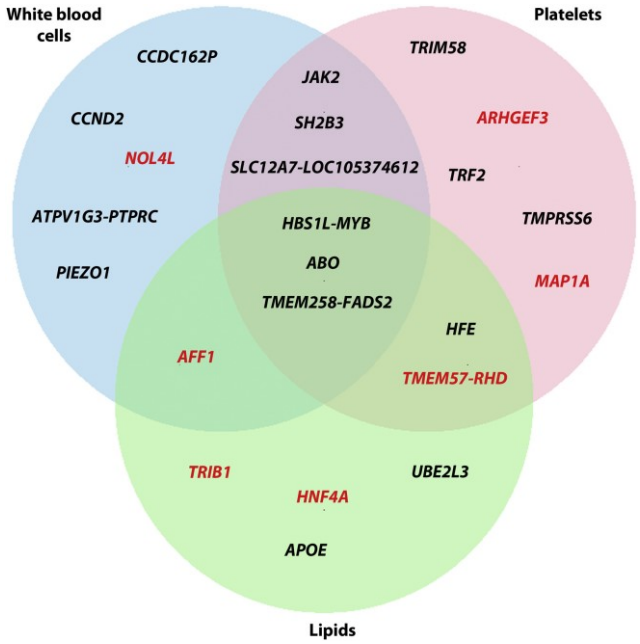


Figure 3. Venn diagram summarizing pleiotropic effects for genetic variants associated with red blood cell (RBC) traits.

We only considered variants if their association P-values with white blood cell (WBC) traits, platelet (PLT) traits, or with lipid levels was $P<1x10^{-4}$. Results for WBC and PLT are from the accompanying Blood-Cell Consortium articles. Results for lipids have previously been published (**Table 3**). Genes highlighted in red are novel RBC trait findings.

We identified a nonsense variant in *CD36* associated with RDW in African Americans. CD36 is a type B scavenger receptor located on the surface of many cell types, including endothelial cells, platelets, monocytes, and erythrocytes. CD36 is a marker of erythroid progenitor differentiation[560] and may also be involved in macrophage-mediated clearance of red blood cells[561]. Furthermore, CD36 plays a role in many biological pathways such as lipid metabolism/transport and atherosclerosis, hemostasis, and inflammation[562]. The nonsense *CD36* variant identified in our RDW meta-analysis (rs3211938, **Table 1**) has previously been

associated with platelet count, HDL-cholesterol, and C-reactive protein levels in African Americans[563,564], and malaria resistance in Africans[565,566]. The *CD36* locus shows a signature of natural selection in African-ancestry populations[567] and the MAF of rs3211938 varies widely between continents: in the 1000 Genomes Project, the minor allele is absent from European populations but reaches frequency of 24-29% in some African populations[4]. To characterize the molecular mechanism by which rs3211938 may impact RDW, we identified one heterozygous sample among a collection of *ex vivo* differentiated human erythroblasts[377]. In erythroblasts from this donor, we noted a strong allelic imbalance (**Figure 2**). Importantly, this result was confirmed in independent samples from the GTex dataset. At the molecular level, this *CD36* expression phenotype could be explained by nonsense-mediated mRNA decay, or the regulatory effect of non-coding genetic variants in LD with rs3211938.

We observed that many new RBC variants are pleiotropic, being often associated with more than one RBC indices as well as with WBC, PLT, and lipid traits (**Figure 3**). These shared effects could imply that the underlying causal genes at these RBC loci generally controlled blood cell proliferation or modulate inflammatory responses. An additional explanation for the link between RBC traits and lipid variants may be the cholesterol content of RBC membranes. As mentioned earlier, RBC corresponds to a large fraction (~25%) of the cells found in the human body. Genetic variation that modulates RBC count or volume could impact circulating lipid levels. In support of this hypothesis, it has been observed that a thalassemia allele is strongly associated with cholesterol levels in the Sardinian population[27]. In total, we found ten loci associated with lipid levels and RBC indices, including four novel RBC variants (*AFF1*, *TMEM57-RHD*, *TRIB1*, *HNF4A*) (**Figure 3**).

In summary, our multi-ethnic meta-analyses have expanded the genetic knowledge of erythrocyte biology and identified new common, low-frequency, and rare RBC variants. Many of the new RBC variants are pleiotropic, affecting other complex traits such as WBC, PLT, and blood lipid levels. While our report demonstrates the utility of the ExomeChip for genetic discovery, it also highlights the challenge to attribute gene causality based only on association results. This is particularly evident for loci with common variants, for which coding and non-coding markers are often statistically equivalent. For instance, we found no examples of RBC

coding variants that entirely explain RBC GWAS signals among the seven loci that had both a sentinel GWAS variant and ExomeChip coding markers. Although increasing sample sizes will continue to yield additional RBC loci, it has become incredibly clear that only a combination of well-powered genetic studies, transcriptomic and epigenomic surveys, and functional experiments (*e.g.* using genome editing) will ultimately pinpoint causal variants and genes that control RBC phenotypes.

Acknowledgments

We thank all participants, staff, and study coordinating centers. We also thank Raymond Doty and Jan Abkowitz for discussion of the *ALAS2* finding. We would like to thank Liling Warren for contributions to the genetic analysis of the SOLID-TIMI-52 and GSK-STABILITY datasets. YFS acknowledges the expert technical assistance in the statistical analyses by Ville Aalto and Irina Lisinen. EGCUT thanks co-workers at the Estonian Biobank, especially Mr V. Soo, Mr S. Smith and Dr L. Milani. Airwave thanks Louisa Cavaliero who assisted in data collection and management as well as Peter McFarlane and the Glasgow CARE, Patricia Munroe at Queen Mary University of London, Joanna Sarnecka and Ania Zawodniak at Northwick Park for their contributions to the study. This work was supported by the Fonds de Recherche du Québec–Santé (FRQS, scholarship to NC), the Canadian Institute of Health Research (Banting-CIHR, scholarship to SL and operating grant MOP#123382 to GL), and the Canada Research Chair program (to GL). PLA was supported by NHLBI R21 HL121422-02. NAA is funded by NIH DK060022. AN was supported by the Yoshida Scholarship Foundation. SK was supported by a Research Scholar award from the Massachusetts General Hospital (MGH), the Howard Goodman Fellowship from MGH, the Donovan Family Foundation, R01HL107816, and a grant from Fondation Leducq. The authors declare no conflicts of interest.