An integrated platform to systematically identify causal variants and genes for polygenic human traits.

Damien J. Downes[1], Ron Schwessinger[1,2,*], Stephanie J. Hill[1,*], Lea Nussbaum[1,*], Caroline Scott[1], Matthew E. Gosden[1], Priscila P. Hirschfeld[1], Jelena M. Telenius[1,2], Chris Q. Eijsbouts[1,3,4], Simon J. McGowan[2], Antony J. Cutler[4,5], Jon Kerry[1], Jessica L. Davies[6], Calliope A. Dendrou[4,6], Jamie R.J. Inshaw[5], Martin S.C. Larke[1], A. Marieke Oudelaar[1,2], Yavor Bozhilov[1], Andrew J. King[1], Richard C. Brown[2], Maria C. Suciu[1], James O.J. Davies[1], Philip Hublitz[7], Chris Fisher[1], Ryo Kurita[8], Yukio Nakamura[9], Gerton Lunter[2], Stephen Taylor[2], Veronica J. Buckle[1], John A. Todd[5], Douglas R. Higgs[1], & Jim R. Hughes[1,2,†].

* These authors contributed equally: Ron Schwessinger, Stephanie J. Hill, Lea Nussbaum.

† Corresponding author: jim.hughes@imm.ox.ac.uk

Affiliations:

1 MRC Molecular Haematology Unit, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK

2 MRC WIMM Centre for Computational Biology, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK

3 Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK

4 Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK

5 JDRF/Wellcome Diabetes and Inflammation Laboratory, Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK

6 MRC Human Immunology Unit, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK

7 WIMM Genome Engineering Facility, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK

8 Dept. of Research and Development, Central Blood Institute, Japanese Red Cross Society, Minato-ku, Tokyo, Japan

9 Cell Engineering Division, RIKEN BioResource Research Center, Tsukuba, Ibaraki, Japan

KEY WORDS: GWAS, gene regulation, chromatin conformation, machine learning.

ABSTRACT

Genome-wide association studies (GWAS) have identified over 150,000 links between common genetic variants and human traits or complex diseases. Over 80% of these associations map to polymorphisms in non-coding DNA. Therefore, the challenge is to identify disease-causing variants, the genes they affect, and the cells in which these effects occur. We have developed a platform using ATAC-seq, DNaseI footprints, NG Capture-C and machine learning to address this challenge. Applying this approach to red blood cell traits identifies a significant proportion of known causative variants and their effector genes, which we show can be validated by direct *in vivo* modelling.

INTRODUCTION

Identification of the variation of the genome that determines the risk of common chronic and infectious diseases informs on their primary causes, which leads to preventative or therapeutic approaches and insights. Whilst genome-wide association studies (GWASs) have identified thousands of chromosome regions[1], the identification of the causal genes, variants and cell types remains a major bottleneck. This is due to three major features of the genome and its complex association with disease susceptibility. Trait-associated variants are often tightly associated, through linkage disequilibrium (LD), with tens or hundreds of other variants, mostly single-nucleotide polymorphisms (SNPs), any one or more of which could be causal; the majority (>85%) the variants identified in GWAS lie within the non-coding genome[2]. Although non-coding regions are increasingly well annotated, many variants do not correspond to known regulatory elements, and even when they do, it is rarely known which genes these elements control, and in which cell types. New technical approaches to link variants to the genes they control are rapidly improving but are often limited by their sensitivity and resolution[3–6]; and because so few causal variants have been unequivocally linked to the genes they affect, the mechanisms by which non-coding variants alter gene expression remain unknown in all but a few cases; and, third, the complexity of gene regulation and cell/cell interactions means that knowing when in development, in which cell type, in which activation state, and within which pathway(s) a causal variant exerts its effect is usually impossible to predict. Although significant progress is being made, currently, none of these problems has been adequately solved.

Here, we have developed an integrated platform of experimental and computational methods to prioritise likely causal variants, link them to the genes they regulate, and determine the mechanism by which they alter gene function. To illustrate the approach we have initially focussed on a single haematopoietic lineage: the development of mature red blood cells (RBC), for which all stages of lineage specification and differentiation from a haematopoietic stem cell to a RBC are known, and can be recapitulated *ex vivo* by culture of $CD34^+$ progenitor and stem cells[7–9]. GWASs have identified over 550 chromosome regions associated with changes in the phenotypes of mature RBC[10,11]; within these regions 1,114 index SNPs are in high LD with 30,694 variants, of which, only eight have been claimed as causal regulatory variants through experimental validation[12–16].

We first identify the key cell type(s) throughout erythropoiesis by analysing enrichment of GWAS variants lying within regions of open chromatin. These regions contain the tissue-specific regulatory elements of the genome (promoters, enhancers and boundary elements). We next focus on the ~8% of variants which lie within regulatory elements in the non-coding

genome; with the remaining variants assessed for effects on coding sequences and RNA processing using established programmes[17–19]. The platform addresses the fact that both causal and non-causal variants may lie in open chromatin. Using DNaseI footprinting and a machine learning approach the platform prioritises variants predicted to directly affect the binding of transcription factors or alter chromatin accessibility[20,21]. Having prioritised putative regulatory causal variants, the platform then links the regulatory elements in which they occur to genes using NG Capture-C, the highest resolution chromatin conformation capture (3C) method currently available for targeting numerous loci[22,23]. To validate the predicted molecular changes caused by such GWAS variants we use CRISPR/Cas9 facilitated Homology Dependent Recombination (HDR) to directly model SNP alleles and determine their effects.

Testing our platform against 75 chromosome regions from a previous GWAS of RBC traits[11] we identified putative causal variants at ~80%, their candidate effector genes at ~70%, and three or fewer candidate variants at ~60%. By benchmarking with the eight validated causal variants from previous studies[12–16], and genes at well characterised erythroid loci, we successfully predicted >87% of both causal variants and effector genes. Finally, we used genome editing to directly determine the *in vivo* molecular effects of candidate SNPs in two regions – showing both SNPs to be causal and verifying *JAK2* as a novel RBC trait effector gene. As this platform was developed with methods appropriate for small numbers of cells, and therefore rare cell types, the approach will enable researchers across a wide range of traits or disorders to more readily identify causal variants, the cells in which they exert their effects, their target genes, and the mechanisms by which they alter cell biology, and ultimately, disease risk.

RESULTS

Enrichment of variants influencing RBC traits in highly active erythroid enhancers.

The first stage of an integrated platform for dissecting polygenic traits is the identification of key cell types (Supplementary Fig. 1). Recently, ATAC-seq allowed a comprehensive identification of *cis*-regulatory elements which remain constitutively present or dynamically change throughout haematopoietic lineage specification, differentiation, and maturation[8,24]. To identify regulatory regions in early, intermediate and late erythroid cells we generated ATAC-seq from such cells obtained by *ex vivo* erythroid differentiation of $CD34^+$ stem and progenitor cells[9] from three healthy, non-anaemic individuals (Supplementary Fig. 2). We also examined ATAC-seq profiles from a variety of haematopoietic cells, including erythroid progenitors (Fig. 1a,b); in total identifying 238,918 open-chromatin regions (496-4,136 bp) present at one or more stages of erythropoiesis.

Previously, 1,114 index SNPs, each of which identifies a region of LD, have been associated with specific RBC traits in two extensive GWAS of Asian and European populations[10,11]. These index SNPs are associated, via LD ($r^2 \geq 0.8$), with a total of 30,694 variants. Approximately 8% of these variants, covering ~60% of RBC trait regions, intersected with open chromatin in erythropoietic cells (n=2,590). Intersections were predominantly found in fully committed, intermediate erythroid cells (days 10-13, cumulative binomial distribution, $p=7\times10^{-29}$) rather than in multipotent progenitor cells (Fig. 1a, Supplementary Figs. 3a,b). Enrichment was trait specific as variants associated with immune diseases[25–27] showed minimal enrichment for intersection with erythroid open chromatin but strong enrichment in differentiated lymphocytes (Supplementary Fig. 3d-f) while non-haematological trait variants[28–30] showed no enrichment in either red or white blood cells. For all traits, we saw no enrichment when we analysed ATAC-seq profiles from two non-haematopoietic cell lines.

To further characterise the intersected regulatory elements, we generated ChIP-seq data to distinguish promoters (Histone-3 Lysine-4 trimethylation, H3K4me3), enhancers (H3K4 monomethylation, H3K4me1), boundary elements (CTCF), and “active” sites (H3 Lysine-27 acetylation, H3K27ac) in committed erythroid cells. We applied GenoSTAN[31] to assign a chromatin signature to each element and thereby generated a high-resolution map of open chromatin in erythroid cells with seven functional classes (Fig. 1c, Supplementary Fig. 4). Intersected elements were enriched for enhancers and promoters with high levels of H3K27ac but not those with low levels of H3K27ac, nor ATAC-seq peaks with CTCF enrichment (Fig. 1c). When putative enhancers were ranked for their levels of H3K27ac (Fig. 1d), elements containing RBC variants were significantly enriched amongst the highly activity erythroid enhancers (χ^2: d.f.=3, $p=7\times10^{-54}$).

Cell-specific intersection is consistent with previous studies[32–34], and as shown here, when applied to highly-stratified cell types may help identify the precise cells in which the variant affects function. For example, four variants including the predicted causative SNP rs1175550[13] intersect a *cis*-regulatory element in the Small Integral Membrane Protein 1 (SMIM1; Vel Blood Group) encoding locus. This element is associated with a region of open chromatin which only appears in megakaryocytic-erythroid progenitors (MEPs) and early/intermediate erythroid precursors (Fig. 1b). A meta-analysis of all intersected open chromatin regions showed multiple trajectories of accessibility, including persistent nucleosome depletion, progenitor specific accessibility, and terminal or transient accessibility (Supplementary Figs. 5,6). While overall enrichment of predicted causal variants is strongest in intermediate erythroid cells (day 10-13), RBC traits may also be influenced by variants acting at earlier stages of erythropoiesis.

Meta-genomic and machine learning approaches effectively prioritise causal variants. As both causal and non-causal variants fall within open chromatin, further assessment of their potential to alter the function of the underlying regulatory elements is required. We applied a combination of meta-genomics and machine learning to further characterise variants found within open chromatin in erythroid cells. Regulatory variants are likely to act by altering the dynamics of transcription factor binding, however only 10-20% of causal SNPs directly alter known transcription factor motifs[35]. This suggests that causal variants may either play an unexplained mechanistic role, or act through uncharacterised transcription factors. Sasquatch uses an unbiased approach to measure the average *in vivo* DNaseI-seq footprint for any given sequence in a specific cell type[20], thus identifying likely transcription factor binding sites for both known and unknown transcription factors, and can therefore be used to evaluate variants in an unbiased manner (Fig. 2a). Using Sasquatch, we found 61.8% of variants in open chromatin in committed erythroid cells were predicted to have at least a weak effect on transcription factor footprints (762/1,233), accounting for variants at ~57% of RBC LD regions (Supplementary Fig. 7). While some of these changes were found in or adjacent to known haematopoietic transcription factors, including SCL/TAL, GATA1, SPI1/PU1, NF-E2, BACH1, and MAFK, footprint changes were also seen for motifs with no known associated transcription factor (Fig. 2, Supplementary Figs. 7,8).

Changes in specific transcription factor binding predicted by Sasquatch were validated by analysis of heterozygous variants using ChIP-seq. Notably, for rs3747093 which falls within an SCL/TAL binding motif, significant allelic imbalance was seen in erythroid SCL/TAL ChIP-seq in three independent datasets (Fig. 2b-c). Similarly, rs77222982 which is directly

adjacent to an AGATAA motif showed allelic imbalance in GATA1 binding (Fig. 2d-e). Often, skew in enrichment was seen across more than one factor in elements affected by a single variant, probably reflecting co-dependency in their binding (Supplementary Fig. 8). Such imbalance is consistent with the alteration of binding predicted by Sasquatch, demonstrating its ability to accurately detect causative variants.

Convoluted neural network based machine learning can predict open chromatin[36,37] and was also used to identify causal variants. We used 936 chromatin-accessibility and epigenetic datasets to train a deep convoluted neural network, deepHaem[21], to predict chromatin accessibility based on DNA sequence across haematological cell-types (Fig. 3a, Supplementary Fig. 9a-d). Using deepHaem it is possible to predict the effect of variants on chromatin accessibility. DeepHaem identified 91 variants in open chromatin with changes greater than 10% of the maximum accessibility score (1.0), with the strongest effects seen in MEP and erythroid populations (Fig. 3b). 45 of the variants predicted to alter chromatin-accessibility had scores greater than 0.1 in erythroid cells. Using ATAC-seq, we identified heterozygous alleles for 15 of these 45 variants in three healthy individuals. Comparison of sequencing from these alleles showed significant bias in ATAC-seq accessibility at 7 of the sites and skew at a further 5 sites (Fig. 3c, Supplementary Fig. 9e) indicating that deepHaem can accurately predict variant-induced changes in chromatin accessibility.

To assess how well the platform performed at identifying causative regulatory variants we used previously characterised RBC trait variants. Currently, no RBC trait variants have been definitively shown to be causative using direct *in vivo* modelling; however, eight regulatory variants have strong support from functional assays[12–16]. The approach established here identified that seven of these eight variants lie in open chromatin in erythroid cells, and therefore had the potential to be regulatory causal variants. Characterisation with Sasquatch or deepHaem further prioritised six of these variants as likely to be causative (Supplementary Table 1). Therefore, the platform accurately prioritises causal variants, and thus identifies variants for functional analysis.

A comprehensive search for all causal variants within an RBC GWAS.

The first major RBC trait GWAS identified 75 index SNPs[11]; the associations identified in that study are likely to represent the most common variants with moderate effect sizes and some rare variants with large effect sizes, therefore we focused specifically on this dataset for in-depth follow-up. A comprehensive GWAS decoding platform must prioritise causal variants by treating all mechanisms as plausible. By examination of the 75 index SNPs, as well as variants in high LD with them ($r^2 \geq 0.8$, 1000 Genomes Project; n=6,420), we identified 486

candidate regulatory variants within 61 of the 75 chromosome regions. In addition to regulatory variants, we considered the possibility for coding and splicing changes across these regions. Putative coding sequence changes were identified using ANNOVAR[17] and then filtered for erythroid expressed genes. This identified 20 variants predicted to alter protein sequence at 14 regions (Supplementary Table 1B). Next, putative alternative-splicing variants were identified using a combination of Splicing Index[18] and a deep learning approach, SpliceAI[19]. Together, these programmes identified 13 putative splice-altering SNPs in 11 erythroid expressed genes across nine regions; however, no variants were highlighted by both algorithms (Supplementary Table 1B). Using these integrated analyses for coding, splicing and regulatory mechanisms we identified candidate causal variants at 63 of the 75 chromosome regions, with 43 of these having three or fewer strong candidates, and the majority of candidates being in tight linkage ($r^2 \geq 0.9$; n=394/515) with their index SNP (Fig. 4, Supplementary Figs. 10,11a).

We next considered why no causal variants were predicted for 12 chromosome regions. Immediate possibilities are that the variant affects gene function in a way that is currently unrecognised, or exerts effects in an untested cell type. Consistently, rs855791, also identified in GWAS for iron status, haemoglobin levels, and erythrocyte volume[38,39], is a missense variant of *TMPRSS6*. *TMPRSS6* is expressed in the liver and encodes Matripase-2, a suppressor of the iron homeostasis master regulator, Hepcidin[40,41]. It is also possible that the causal variant affects mRNA stability. However, there are currently no good predictive software programmes for this. Finally, it may be that causal variants were not identified because the initial GWAS study was not sufficiently powered or used a sub-optimal catalogue of variants; resulting in incompletely resolved genetics. Indeed, index SNPs in unresolved loci were less likely to be replicated in subsequent RBC trait GWAS[10,42] than index SNPs at resolved loci (Supplementary Fig. 11b). Additionally, in a region with multiple unlinked causal variants, incompletely resolved genetics can lead to index SNPs being identified through weak association ($r^2 < 0.8$) with two or more causal variants. Such index SNPs are referred to as tag SNPs[43] (Supplementary Fig. 11a). At the *TMCC2* locus, where rs9660992 is an index SNP[11], moderate linkage (r^2=0.51-0.82) is seen with two independent index SNPs from a subsequent RBC trait GWAS[10]. While rs12137294 and rs1172129 are themselves unlinked (r^2=0.46), each is in tight linkage with several variants in open chromatin (Supplementary Fig. 12); suggesting rs9660992 may be a tag SNP. Therefore, both additional cell types and incomplete genetics can explain unresolved regions.

High resolution 3C mapping accurately identifies effector genes.

The target or effector genes for splicing and coding variants can be directly inferred, but the effector genes of regulatory variants must be identified experimentally. For enhancers to regulate gene expression they often physically interact with target promoters, likely through loop-extrusion and/or phase-separation[44,45]. The close proximity required for regulation can be identified by chromosome conformation capture (3C) to map interactions[46] and this provides a means by which to identify effector genes. NG Capture-C uses biotinylated oligonucleotide probes to target specific loci at high resolution in multiplexed 3C samples[23]; allowing statistical comparison for identification of enhancer-promoter interactions. We designed probes for 214 variant containing *cis*-regulatory elements covering 53/61 chromosome regions with putative regulatory variants; then simultaneously generated 3C interaction data in intermediate erythroid cells, H1-hESCs and HUVECs to link *cis*-regulatory elements with their effector genes. Using a combination of tissue-specificity (DESeq2)[23], Bayesian modelling (Peaky)[47], and promoter proximity (≤5 kb) to call variant-promoter interactions we identified 194 candidate effector genes at 48 of the 53 targeted regions (Fig. 5a-b, Supplementary Table 1). For each targeted region, NG Capture-C identified an average of four genes, which is consistent with the predicted number of gene targets for enhancers[48–50], though whether multiple genes contribute to a GWAS trait at a single locus remains to be determined. Although some methods have indicated that GWAS variants are most frequently found within 20 kb of their target genes[51] we detected interactions up to 992 kb, with a median distance of 83.9 kb (±9.3 kb SE, Fig. 5c). Such long-range interactions were seen at several well characterised erythroid loci including *CITED2* (139 kb), *SLC4A1* (47 kb), *RBM38* (24 kb), *ANK1* (25 kb), *MYB* (85 kb), and *HBA1/2* (63 kb), showing that GWAS variants, as for other enhancer-promoter interactions, may act over large distances (Fig. 5d, Supplementary Figs. 13-18).

The erythroid system and RBC traits have been intensively analysed and characterised; therefore, we were able generate a set of the 24 “most likely” effector genes within the 53 targeted chromosome regions based on prior knowledge of their function (Supplementary Fig. 19a). This set of genes allowed us to benchmark our approach; finding that with NG Capture-C we correctly identified 22 of the 24 most-likely effector genes (Fig. 5a, Supplementary Fig 19b). Of the remaining regions, no genes were identified at one (*miR-181a*), and four incorrect candidates were identified in the region where *TAL1* is almost certainly the effector. With these exceptions, NG Capture-C performs with a high rate of success in identifying effector genes linked to potential causal variants. Three previous attempts with diverse methods to identify effector genes associated with RBC traits have been reported[5,6,11]. These were an annotation-based approach[11], Promoter Capture-HiC[5]

(PC-HiC), and a gene-centric shRNA screen[6]. We directly compared these different approaches with NG Capture-C. There was little consistency between the candidate gene lists from these approaches (Supplementary Fig. 19c), with NG Capture-C the only method to identify *HBA-1/2*, the α-globin encoding genes, which are known to be associated with anaemia and changes in RBC traits[52] (Supplementary Figs. 17,19c). Our approach was also unique in identifying *RPL19*, of interest because ribosomal genes are known to cause Diamond-Blackfan anaemia[53]. Across the 24 benchmark regions, NG Capture-C and the annotation-based approach were the most sensitive methods, respectively identifying 91.7% and 70.8% of the most-likely effector genes correctly (Supplementary Fig. 19 b,d). Overall, the direct comparison of different gene identification methods shows that NG Capture-C is the most successful tool.

Direct modelling of rs9349205 shows reduced expression of its target gene *CCND3*.

The most direct evidence that a particular variant alters gene expression comes from introducing both alleles to an isogenic background and observing an appropriate change in the relevant cell type. Previous studies characterising RBC trait variants have used reporter assays and/or targeted deletions of the regulatory element[13–16]. However, these may not faithfully recapitulate variants effects *in vivo.* Therefore, we used CRISPR/Cas9-facilitated homology directed repair (HDR) to directly model prioritised variants at five GWAS regions in the Human Umbilical Derived Erythroid Progenitor (HUDEP-2) cell line (Supplementary Fig. 20,21); a model of human erythroid differentiation and maturation[54,55]. As previous studies have shown some clonal variation when using such cells[56] it is essential to analyse multiple independently isolated clones. We were able to generate sufficient independent clones for robust analysis of two regions (*CCND3* and *JAK2*).

Using our platform, rs9349205 was identified as tightly linked to the index SNP (rs9349204, r^2=0.841); rs9349205 is the only one of ten linked variants which lies within open chromatin in committed erythroid cells, and shows 3C interaction with *CCND3* (Fig. 6a), which is the most-likely effector gene[16]. rs9349205 also had small effects on both the Sasquatch DNaseI footprint and deepHaem chromatin openness scores (Supplementary Figs. 20,22a). Editing was used to convert rs9349205$^{A/A}$ HUDEP-2 cells to rs9349205$^{G/G}$; a non-erythroid locus was also edited to control for non-specific effects from editing (e.g. spontaneous differentiation). Using ATAC-seq to assess chromatin accessibility, we found that the identified regulatory element in the rs9349205$^{G/G}$ genotype was 54.5% less accessible than in rs9349205$^{A/A}$ cells (Fig. 6b, Supplementary Fig. 22b). The rs9349205$^{G/G}$ clones also showed significantly reduced *CCND3* expression during erythroid differentiation (Fig. 6c). As previously discussed[16], *CCND3* regulates the G2 to S transition during erythropoiesis, and

thus knockout of *CCND3* in mice leads to an increased erythrocyte volume, consistent with linkage to changes in mean cell volume (MCV) detected through GWAS.

rs10758656 causes reductions in chromatin accessibility and *JAK2* expression.

Using NG Capture-C we identified *JAK2*, which encodes Janus Kinase 2, as an effector gene for variants in high LD with the index SNP rs2236496. We confirmed this interaction using NG Capture-C from the *JAK2* promoter (Fig. 7a). Of the 18 linked variants, only two, rs10758656 and rs10739069, intersect open chromatin. Of these two SNPs Sasquatch characterised rs10758656 but not rs10739069 as having the potential to affect transcription factor binding, with the motif strongly matching that of the GATA1 binding motif (Fig. 7b, Supplementary Fig. 23a). DeepHaem also predicted that rs10758656 but not rs10739069 would affect chromatin accessibility. Therefore, HUDEP-2 cells, which are heterozygous A/G for rs10758656, were edited to homozygosity. We generated 16 independent clones homozygous for either A (n=10) or G (n=6). ATAC-seq of these cells during expansion and differentiation showed up to 82% ablation of open chromatin in the rs10758656$^{G/G}$ clones, associated with 86.2% and 58.4% reductions in GATA1 binding and H3K27ac, respectively (Fig. 8a-b, Supplementary Figs. 24b-e). These findings match the prediction of both Sasquatch and deepHaem. Despite being closer to the promoter of *RCL1* than *JAK2*, rs10758656$^{G/G}$ specifically reduced expression of *JAK2* (Fig. 8c, Supplementary Fig. 24f,g), consistent with the specificity of the rs10758656-*JAK2* interaction profile seen in NG Capture-C. JAK2 functions as part of the erythropoietin signalling pathway[57]. Our results demonstrate that *JAK2* is a GWAS effector gene and most likely results in changes to the MCV noted in GWAS through altered signalling responses.

DISCUSSION

Here we have developed and validated a platform to identify causative GWAS variants and link them to the genes whose function they affect. In our platform ATAC-seq analysis allows researchers to identify relevant cell types using the fundamental regulatory elements of the genome: enhancers, promoters and boundary elements. GWAS variants are then assessed *in silico* to predict which variants are likely to alter gene expression or function. Candidate regulatory variants are finally linked to their effector genes using NG Capture-C. Using this method to analyse variants in high LD to RBC trait index SNPs resulted in identification of candidate causal variants and effector genes at a majority of chromosome regions (>70%). Benchmarking also shows that this approach is robust, with 88% of validated causal variants, and 92% of most-likely effector genes identified. Application of this method to fine-mapped GWAS variants is likely to further improve its success. Finally, the functional effects of candidate polymorphisms can then be assessed using allele-specific assays of chromatin accessibility and gene expression.

This platform has been developed and benchmarked using data from purified haematopoietic cells at various stages of commitment, differentiation and maturation along the erythroid pathway to producing RBC. Using haematopoiesis as a model, we show how causal variants can be assigned to the cell types in which they exert their effects and the genes whose expression is perturbed. In principle, this method could be used for any GWAS datasets for which appropriate cell types are available. To ensure that this would apply to rare cell types and a wide range of diseases, we have established a platform that can be effectively applied using as few as 500 cells for ATAC-seq and 20,000 cells for NG Capture-C[24,58]. These data can then be used to improve *in silico* processing and machine learning, meaning that damaging changes can be predicted and prioritised using data from rare cells and those grown under varying conditions of stimulation.

Linking variants to their effector genes using NG Capture-C can easily and reproducibly be applied across a wide range of cell-types at hundreds of specifically targeted sites in either gene- or enhancer- centric designs[22,23]. The ability to compare 3C data from multiple cell types allows tissue-specific and tissue-invariant interactions to be called by a wide range of statistical approaches[23,47,59–61], increasing the throughput of accurate effector gene identification. These candidates can then be validated with functional follow-up, such as screening approaches[6], or as shown here, *in vivo* modelling, to help to explain associated cellular phenotypes.

In addition to elucidating the genes involved at GWAS regions, we have also addressed the multiple molecular mechanisms that may underlie such signals. To date, strong evidence supports a mixture of coding, splicing and regulatory mechanisms[62]. The approach described here identifies enhancer, promoter, RNA processing and coding variants. Despite this we were still unable to identify causative variants at 16% of chromosome regions. This could partly have resulted from the fact that initial variant identification used linkage disequilibrium, which could be improved with either fine mapping or whole genome sequencing[62,63]. Nevertheless, other factors are also likely to contribute. The cell types affected in any complex disease are not necessarily the most obvious candidates. For RBC traits, the most likely affected lineage is erythropoiesis itself. However, other cell types modify erythropoiesis, including those producing growth factors, cytokines, or mediating cell/cell interactions such as macrophages that facilitate enucleation of RBC precursors. Furthermore, causal variants may act in the identified cell type, but only in response to specific environmental cues or signalling. Therefore, platforms such as this must be implemented with a comprehensive appreciation of the systems involved. It is also important to consider that additional untested molecular mechanisms may underlie GWAS signals. Although we found no specific enrichment for variants in CTCF elements, numerous were within CTCF peaks. Recent evidence has shown that disruption of CTCF binding by common variants plays a role in determining the severity of influenza and breast cancer[64,65], thus it likely represents a less common, yet important molecular mechanism. Additionally, modelling of rs10758656 showed near complete loss of open chromatin. It is equally possible that some variants generate, rather than abolish, open chromatin sites. Such a variant has already been described as causing anaemia at the α-globin locus[66]. These sites could only be detected by analysis of individuals with the correct genotype.

Although this integrative platform efficiently identifies variants and genes, it shifts the bottleneck of GWAS follow-up to validation. Using direct *in vivo* modelling we have shown how alleles can alter enhancers and gene expression to different extents. Although we have shown in principle that prioritised variants can be proven to be functionally causative it requires an HDR editable cell type, is labour intensive, and does not work at all loci; this step will require rapid single base editing to enable significant progress. We expect that with the implementation of integrative platforms such as this, and with ongoing advancement of molecular techniques and editing technologies the benefits of GWAS for understanding human physiology and improving health will accelerate.

METHODS

Separation of blood cells: Fresh blood was sourced either as whole blood collected from three healthy donors (two males, one female) using EDTA Vacuettes (Becton Dickson) or 5 ml leukocyte cones (NHS Blood & Transport). Whole cell counts were performed on a Pentra ES60 (Horiba) for donor blood to ensure clinically healthy red blood cell counts (Supplementary Fig. 2). Blood was diluted with PBS and overlaid onto Histopaque-1077 (Sigma) and centrifuged for 30 min at 630 rcf (no brake). Peripheral Blood Mononuclear Cells (PBMCs) were washed in PBS and MACS buffer (PBS, 2 µM EDTA, 0.5% BSA) and stained with Human CD34 Microbead kit (Miltenyi Biotec) following the manufacturer's instruction for 30 minutes (4 ºC) before being passed successively through two LS Columns (Miltenyi Biotec) with three MACs buffer washes. Counting of cells was performed on a Luna FL (Logos) after staining with acridine orange (AO) and propidium iodide (PI). $CD34^+$ cells were either stored in freezing buffer (90% FBS, 10% DMSO) or resuspended in Phase I medium for a three-phase differentiation[9]. $CD34^+$ depleted PBMCs were then sequentially stained and passed over LS or MS columns for selective purification of CD8+, CD14+, CD4+, CD19+ and, NK cell populations using cell-type specific kits (Miltenyi Biotec). For NK cells, non-NK cells were first blocked with a biotin-antibody cocktail before binding to NK microbeads following the manufactures instructions.

Differentiation of $CD34^+$ cells: Cells were differentiated under a three phase *ex vivo* protocol adapted from that used for the BEL-A cell line[7,9,67]. Growth media are listed in Supplementary Table 2. Briefly, for differentiation 0.5-2.5x10^5 cells were resuspended on day 0 in Phase I media at 10^5 cells ml^{-1}. Cell counts were performed on days 3 and 5 with additional Phase I media added to return the concentration to 10^5 cells ml^{-1}. On day 7, cells were counted and pelleted (400 rcf, 5 min, RT) and resuspended in Phase II media at 3x10^5 cells ml^{-1}. Cells were counted on day 9 and diluted to 3x10^5 cells ml^{-1} Phase II media. On day 11, cells were counted and pelleted (400 rcf, 5 min, RT) and resuspended in Phase III media at 10^6 cells ml^{-1}. Cells were counted on days 13 and 15 and diluted to 10^6 cells ml^{-1} in Phase III media. Reproducibility between differentiations was confirmed morphologically with cytospins, immunologically with six FACS cell surface markers and epigenetically with ATAC-seq enhancer staging[24]. For morphological analyses 10^5 cells were resuspended in 200 ml PBS and spun (5 min, 400 rpm) in a Cytospin 4 (ThermoFisher), before staining with modified Wright's Stain on a Hematek (Bayer Health Care), and mounting with DPX (Sigma). Images were taken on an Olympus BX 60 microscope at 10x and 20x magnification. For differentiation and enucleation FACS analyses 10^5 cells were resuspended in FACS buffer (90 % PBS, 10 % FBS) and stained with an erythroid differentiation panel of antibodies (Supplementary Table 4) against CD34, CD36/Fatty acid translocase, CD235a/Glycophorin

A, CD71/Transferrin Receptor, CD233/Band3, CD49d/α-Integrin, and with Hoescht-33258 (ThermoFisher) for live/dead analysis, with Hoescht-33342 (ThermoFisher) for enucleation assays. For immune cell purities, cells were stained with cell-type specific panels of antibodies (Supplementary Fig. 25, Supplementary Table 4). FACS was carried out on an Attune NxT (ThermoFisher), voltages and compensation were set using Ultra Comp eBeads (ThermoFisher) for antibodies, and single stained cells for dyes. Gating was performed using fluorescence minus one (FMO) controls. Analysis was performed using either Attune NxT software (v3.0) or FlowJo (v10.4.2).

Cell line culture and HUDEP-2 differentiation: Human ESC line H1 (H1-hESC; WiCell) was grown on Matrigel (Corning) coated plates in mTeSR1 medium (StemCell technologies). Cells were harvested as a single cell suspension using Accutase (EDM Millipore); ATAC-seq and fixation were carried out in mTeSR1 medium. Primary neonatal Human Umbilical Vein Endothelial Cells (HUVEC) were sourced from three suppliers to provide genetic diversity (Lonza, Gibco, PromoCell). HUVECs were expanded in endothelial cell growth medium (Sigma) up to five passages following the manufacturer's protocol. Briefly, HUVECs were grown to 60% confluence, washed with HBSS at room temperature and sub-cultured following light trypsination using Trypsin-EDTA (Sigma) at room temperature and terminating the reaction with trypsin inhibitor (Sigma) upon rounding of the cells and gentle release from the flask. HUVECs were fixed in RPMI supplemented with 10 % FBS. Human Umbilical Derived Erythroid Progenitor line 2 cells[54] (HUDEP-2; RIKEN) were maintained at 0.7-1.5x10^6 cells ml^{-1} in HUDEP expansion media (SFEM, 50 ng/ml SCF, 3 IU/ml EPO, 10 µM DEX, 1% L-Glu, 1% Penstrep) and changed into fresh media containing 2x doxycycline (DOX) every two days. For differentiation we used a modified version of the CD34 differentiation protocol. 2-3x10^6 cells were resuspended at 0.3-05x10^6 cells ml^{-1} in HUDEP Phase I media (IMDM, 200 µg/ml Holotransferrin, 10 g/ml Insulin, 3 IU/ml Heparin, 3% Inactivated AB plasma, 2% FBS, 3 IU/ml EPO, 1 ng/ml IL-3, 10 ng/ml SCF, 1x Pen/Strep) with 1x DOX on day 0. On days 1 and 3 cells were counted, pelleted (5 min, 250 rcf, RT) and resuspended to 0.3-0.5x10^6 cells ml^{-1} in fresh HUDEP Phase I media supplemented with 2x DOX. On day 5, cells were counted, pelleted and resuspended to 0.5x10^6 cells ml^{-1} in HUDEP Phase II media (IMDM, 500 µg/ml Holotransferrin, 10 g/ml Insulin, 3 IU/ml Heparin, 3% Inactivated AB plasma, 2% FBS, 3 IU/ml EPO, 1x Pen/Strep) without DOX. On days 7 and 9 cells were counted, pelleted and resuspended to 0.5x10^6 cells ml^{-1} in fresh HUDEP Phase II media. Cytospins and FACS was carried out as for $CD34^+$ differentiation.

HUDEP-2 genome editing: Prior to guide RNA (gRNA) design HUDEP-2 SNPs were genotyped by either Sanger or Next Generation sequencing at MRC WIMM Sequencing

Facility with locus specific primers (Supplementary Table 3). For introduction of SNPs by CRISPR/Cas9 facilitated homology dependent repair (HDR), gRNAs were designed to cut in close proximity to the SNP of interest with the PAM overlapping the SNP where possible, additionally single stranded DNA donors (ssODN; IDT) were offset and thioated to promote integration and reduce degradation. To control for global effects on HUDEP-2 cells caused by CRISPR/Cas9 editing, gRNA and ssODN were designed for a homozygous SNP (rs4508712) with no GWAS associations (www.ebi.ac.uk/gwas), and was not within an erythroid regulatory element or expressed gene. HUDEP-2 genotype specific gRNAs (Merck) were cloned into pX458 plasmid backbone[68] by the Genome Engineering Facility (WIMM, University of Oxford) and purified using Plasmid Midi Kit (Qiagen). pX458 (pSpCas9(BB)-2A-GFP) was a gift from Feng Zhang and is available from Addgene (plasmid #48138). HUDEP-2 cells were then transfected as previously described[55]. Briefly, ~1x10^6 cells were transfected with pairs of 5 µg gRNA plasmid and 4 µg ssODN (Supplementary Table 3) using Amaxa™ Human CD34 Cell Nucleofector™ Kit (Lonza) in the 2B-Nucleofector™ on the U-08 setting. Cells were transferred to 2.5 ml HUDEP expansion media supplemented with 2x DOX and 7.5 µM RAD51-stimulatory compound 1 (RS-1, Sigma). After two days cells were pelleted (5 min, 250 rcf, room temp.) and resuspended in 2.5 ml HUDEP expansion media supplemented with 2x DOX with minimal light exposure. On day 3 cells were single cell sorted on BD FACSAria Fusion flow cytometers (BD Bioscience) into terazaki plates containing 20 µl of expansion media (2x DOX). When colonies reached more than 30 cells they were transferred to a 96-well plate and expanded over two weeks with fresh media and DOX every 2 days until filling two to four wells of a 96-well plate. Half of the cells for each expanded clone were frozen (90% FBS, 10% DMSO) as a stock for recovery post genotyping. For genotyping we followed a 96-well barcoding approach with Next Generation sequencing[69]. Clonally amplified cells were first lysed (50 mM Tris, 1 mM EDTA, 0.5% Tween 20) and the targeted locus was amplified with primers containing a modified m13 adaptor sequence (Supplementary Table 3), the adaptor was then used to for priming with row and column specific primers in a second PCR to barcode each well. Finally, all wells from a single plate were pooled and prepared for sequencing with the NEBNext Ultra II DNA Library Prep kit for Illumina (New England Biolabs). Plates were multiplexed and sequenced on the MiSeq platform (Illumina) using 250 bp paired-end reads (Nano kit, v2 chemistry). Sequences were analysed using platescreen96[69] (v4.0.4, github.com/Hughes-Genome-Group/plateScreen96/releases) to genotype clones. Screening was carried out to exclude clones which appeared homozygous due to microhomology driven large deletions (200-4,000 bp)[70,71] rather than HDR by PCR with locus specific primers (Supplementary Table 3) for rs9349205 and rs4508712. For rs10758656, two

upstream heterozygous SNPs rs7870037 (+129 bp) & rs7855081 (+132 bp) allowed for the exclusion of loss of heterozygosity.

Gene expression analyses: 1-5x10^6 cells were fixed in 1 ml TRI-reagent (Sigma), snap frozen and stored at -80ºC for less than one year. RNA was extracted by addition of 0.1 ml 1-bromo-3-chloropropane, pipette mixing and separation in a Phase Lock gel Heavy tube (5Prime) and then precipitation with 1 µl of GlycoBlue and an equal volume (~500 µl) isopropanol and centrifugation (10 min, 12,000 rcf, 4 ºC). The RNA pellet was washed with 75% ethanol, resuspended in DEPC-treated water, and stored at -80ºC for less than one year. For RT-qPCR RNA was treated with 2U of rDNase I (Invitrogen) and then 1 µg of RNA was used to generate cDNA using SuperScript III First Strand Synthesis SuperMix (Invitrogen) following the manufacturers' instructions. Real-time RT-qPCR was performed on a StepOne Thermocycler (ThermoFisher) using Taqman Universal PCR Master Mix II (Life Tech) and commercially available expression assays (Supplementary Table 5; Life Tech). For RNA-seq total RNA was treated with Turbo DNase (Invitrogen) at 25ºC for 60 min, then RNA was separated using phenol-chloroform isoamylalcohol and a PhaseLock Light-gel tube (5Prime). Treated RNA was precipitated at -80ºC overnight with sodium acetate, glycoblue, and 75% ethanol, before centrifugation (12,000 rcf, 4ºC), 75% ethanol wash and resuspension in DEPC-treated water. Globin and rRNA sequences were depleted from up to 5 µg of treated RNA using Globin-Zero Gold (Illumina), before PolyA selection with NEBNext Poly(A) mRNA Magnetic Isolation module (New England Biolabs), and indexing with NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs) following the manufacturers' instructions. RNA-seq libraries were quantified by qPCR (KAPA) prior to sequencing on the NextSeq platform (Illumina) with 39 bp paired-end reads. Reads were mapped to hg19 (Supplementary Table 6) using STAR[72] (v2.4.2a; --outFilterMultimapNmax 1) and duplicates were filtered using samtools[73] (v1.3; rmdup). For visualisation directional reads were normalised to RPKM using deepTools[74] with no windowing (v2.2.2; bamCoverage --binSize 1 --normalizeUsingRPKM). Uniquely mapped reads were analysed in DESeq2[61] using variance stabilising transformation and exclusion of genes lacking 5 total reads. Violin plots were with the R package generated in ggplot2[75]. Expressed genes were classed as having more than $\log_2$(FPKM) greater than -5.

Chromatin conformation capture and target gene identification: For chromatin conformation, 1-2x10^7, H1-hESC, HUVEC or erythroid cells were crosslinked with 2% formaldehyde which provides optimal *cis/trans* ratios and digestion efficiencies[58]. For each cell type triplicate 3C libraries were prepared using *Dpn*II and standard methods[23] with the following modifications: no douncing was performed, all spins were performed at 300 rcf, and after ligation intact

nuclei were pelleted (15 min, 300 rcf), supernatant was discarded, and nuclei were resuspended in 300 µl Tris-EDTA (TE; Sigma) for phenol chloroform extraction. Digestion efficiency was determined by RT-qPCR with TaqMan and custom oligonucleotides (Supplementary Table 5), and ligation efficiency qualitatively determined by gel electrophoresis. Only 3C libraries with >70% digestion efficiencies were used. 3C libraries were sonicated to 200 bp in a Covaris S220 and indexed with NEB Next Illumina library Prep reagents (NEB). Enrichment for specific viewpoints was performed with 70mer biotinylated oligonucleotides designed using CapSequm[76] (http://apps.molbiol.ox.ac.uk/CaptureC/cgi-bin/CapSequm.cgi). Double capture was performed in multiplexed reactions with pools of oligonucleotides targeting either promoter proximal (within 5 kb of a transcription start site) or promoter distal *DpnII* fragments (Supplementary Table 7) following the described method[23] with each oligonucleotide at a working concentration of 2.9 nM. Captured 3C libraries were sequenced on the NextSeq platform (Illumina) with 150 bp paired-end reads. Reads were mapped and analysed using CCseqBasic5 (github.com/Hughes-Genome-Group/CCseqBasic5) as previously described[77] with the following custom settings (--bowtie2 --globin 2). Briefly, CCseqBasic5 trims adaptor sequences, flashes read pairs, *in silco* digests fragments and uses bowtie2 to map reads before identifying capture and reporter reads. After primary analysis replicates were compared using the comprehensive CaptureCompare software (github.com/Hughes-Genome-Group/CaptureCompare). CaptureCompare normalises *cis* reporter counts per 100,000 *cis* reporters, generates per fragment mean counts for each cell type, calculates difference in mean interactions between cell types, compares differences in raw interaction counts per fragment using DESeq2[45] as previously described[23,78,79], and provides input for peaky interaction calling[47]. Interaction calling using peaky was run with default settings (omega -3.8) and interactions were filtered based upon the Marginal Posterior Probability of Contact (MPPC) within local interaction domains (MPPC > 0.01) or within 1 Mb of the viewpoint (MPPC > 0.1) and assigned to either Refseq transcription start sites (tss) or variants within 500 bp of the interacting fragment. Target genes were first identified as those having a tss within 5kb of an intersecting variant (high proximity), being within 500 bp of a significantly enriched erythroid fragment (FDR <0.05) or with 500 bp of a peaky identified interaction. Candidate genes were subsequently filtered for detectable erythroid expression ($\log_2$(FPKM>-5)). 24 test genes most likely to be effectors were identified based on published functional data (*IKZF1, KIT, TAL1, RBM38, SMIM1, CD164, CCND3, MYB, HBA1, HBA2, BCL11A, JAK2*)[12–14,16,42,80], presence in the Oxford Red Cell Panel for rare inherited anaemia (*KLF1, TFRC, ANK1, HK1, SCL4A1*)[81], containing mutations causing hemochromatosis (*TFR2*)[82], having an erythroid eQTL (*ATP2B4*)[15], and causing altered RBC phenotypes in mouse and zebrafish (*FBXO7, CCNA2, miR-181a, PIEZO1, AKAP10, CITED2*)[83–89].

Chromatin IP, ATAC-seq, and data processing: For ChIP-seq, chromatin was crosslinked with 1% formaldehyde (Sigma) by the addition of 1 ml 10x crosslinking buffer (50 mM HEPES, 1 mM EDTA, 0.5 mM EGTA, 100 mM NaCl, 10% formaldehyde) to 10^7 cells in 9 ml of media and incubation at room temperature for 10 minutes. Crosslinking was quenched with 130 mM glycine, and cells were washed with cold PBS before snap freezing pelleted cells. Fixed material was stored at -80ºC for less than 12 months. Chromatin immunoprecipitation was performed using Agarose ChIP Assay Kit (Merck Millipore). Briefly, 10^7 cells were lysed by incubation on ice with 130 µl lysis buffer for 15 minutes. Lysed cells were transferred to Covaris microtubes and sonicated on the Covaris S220 (Duty cycle: 2%, Intensity: 3, Cycles per burst: 200, Power mode: Frequency sweeping, Duration: 480 sec, Temp.: 6ºC) to generate 200-400 bp fragments. Insoluble material was removed by centrifugation (15,000 rcf, 15 min, 4ºC) and soluble material was diluted to 4 ml with dilution buffer. Immunoprecipitation was performed by incubation of 2 ml diluted chromatin (equivalent to $5x10^6$ input cells) with antibodies for H3K4me1 (3 µg ab195391, lot: GR304893-2; AbCam), H3K4me3 (1 µl 07-473, lot: 2664283; Millipore), H3K27ac (0.3 µg ab4729, lot: GR3205523-1; AbCam), CTCF (10 µl 07-729, lot: 2836929; Millipore) or GATA1 (~7.2 µg ab11852, lot: GR208255-9; AbCam) overnight. Chromatin binding to Protein A/agarose slurry, washes and elution were performed according to the manufacturer's instructions. DNA was purified by phenol-chloroform extraction with PhaseLock tubes (5Prime) and ethanol precipitation with NaOAc, and 2 µl GlycoBlue (Invitrogen). ChIP enrichment was determined by RT-qPCR (Supplementary Table 5) prior to addition of sequencing adaptors using NEBNext Ultra II DNA Library Prep kit for Illumina (New England Biolabs). ATAC-seq was performed as previously described[77,90] using $7x10^5$ cells. ChIP-seq and ATAC-seq libraries were quantified by RT-qPCR with the KAPA Library Quantification Complete Kit (KAPA) prior to sequencing on the NextSeq platform (Illumina) with 39 bp paired-end reads. ATAC-seq, DNaseI-seq and ChIP-seq reads were mapped to the hg19 genome using NGseqBasic[91] (V20; --nextera --blacklistFilter --noWindow) which utilises bowtie. Sequence depth and mapped reads for each sample are provided (Supplementary Table 6). Published GEO repositories[24,66,92–99] were used for ATAC-seq and DNaseI-seq from HSC, CMP, MEP, MPP, Ery (GSE75384), and HUVEC (GSM736575, GSM736533), and ChIP-seq for SCL/TAL (GSE95875, GSE93372, GSE42390, GSE70660, GSE59087, GSE52924), GATA1 (GSE32491, GSE36985, GSE107726, GSE29196), NF-E2 (GSE95875), BACH1 and MAFK (GSE31477), and SPI1/PU1 (GSE70660) were analysed by the same method. For visualisation PCR-duplicate filtered replicates were merged using samtools[73] (v1.3) and converted to bigwigs with minimal smoothing using deepTools[74] (v2.2.2; bamCoverage --binSize 10 --normalizeUsingRPKM --minMappingQuality 30).

644

645 *Imputation and in silico analysis of variants*: The original 75 anaemia index SNPs were
646 imputed with HapMap Phase 2 which is lower resolution than the 1000 Genomes Project
647 Phase 3 dataset[11,100]. Therefore variants in linkage disequilibrium (LD) with index SNPs were
648 identified using the rAggr proxy search online tool (raggr.usc.edu) with default settings
649 ($r^2 \geq 0.8$, distance limit: 500 kb, population panels: All European, All South Asian) for the 1000
650 Genomes Project Phase 3 database[101], which generated 6,420 variants. LD variants for
651 Astle *et al.* (2016) were provided by Lisa Schmunk, Tao Jiang, and Nicole Soranzo
652 (University of Cambridge). Summary statistics for Malaria[102], Multiple Sclerosis[25],
653 Inflammatory Bowel Disease[27], Type 1 Diabetes[26], Type 2 Diabetes[103], Intelligence[28], and
654 Central Corneal Thickness[29] were downloaded from the NHGRI-EBI GWAS Catalog[1]
655 (www.ebi.ac.uk/gwas). For comparison of linkage between index SNPs from van der Harst
656 *et al.* (2012)[11] and Astle *et al.* (2016)[10] we used LDmatrix on the LDlink web tool[104]
657 (http://ldlink.nci.nih.gov/) for European populations. Variants were intersected with peak calls
658 from ATAC-seq or DNaseI-seq for each cell type of interest using bedtools[105]. Enrichment
659 was calculated as the -log(p-value) of a binomial cumulative distribution function b(x; n, p),
660 describing the probability of x or more successes from n Bernoulli trials, with the probability
661 of success for each trial being p. P-values were calculated using the R function pbinom
662 (lower.tail=FALSE) where x was the number of intersecting variants, n was the total number
663 of variants and p was the total number of base-pairs within cell specific peaks divided by the
664 hg19 uniquely mappable base-pairs (2,644,741,479 bp). Variants within the exons and
665 introns of expressed coding genes were tested for predicted damaging effects on coding
666 ANNOVAR[17] or splicing SPIDEX[18] (z-score ≥1.65) and SpliceAI[19] (AI score ≥0.2). Variants
667 within open chromatin were assessed for potential damage to transcription factor binding
668 footprints using Sasquatch[20] (7-mer, WIMM Fibach Erythroid, Exhaustive). Variants within
669 open chromatin were further classified based on their predicted effect on chromatin
670 accessibility using a deep convolutional neural net[21] (deepHaem). Model architecture and
671 data encoding were adapted from DeepSEA[36] with the following modifications. The number
672 of convolutional layers was increased from three to five and batch normalisation was
673 excluded as it did not improve convergence. The network was re-implemented in python
674 using tensorflow (v1.8.0; https://www.tensorflow.org/about/bib). The ENCODE data
675 compendium previously used[36] was supplemented with ATAC-seq and CTCF ChIP-seq data
676 from erythroid differentiations generated for this work, DNaseI-seq[20], and ATAC-seq from
677 sorted progenitor populations[24]. Full model details and architecture are available on GitHub
678 (https://github.com/rschwess/deepHaem).
679

Chromatin segmentation and enhancer based PCA analysis: To ensure identification of all ATAC-seq peaks a combination of the traditional MACS2 approach[106] (v2.0/10 callpeak -B -q 0.01) and digital signal processing with Ritornello[107] (v2.0 default settings) was used. Peak summits from both calls were extended to 500 bp and intersected with bedtools[105] (v2.25.0), and filtered for high ploidy regions in MIG viewer[108] to form peak calls for each cell type (Supplementary Tables 8a-p). Chromatin segmentation was performed using the GenoSTAN[31] hidden Markov model (HMM) which allows a more fine-tuned analysis than ChromHMM[109] as it uses continuous rather than binary signal counts. Segmentation used a peak centric approach, rather than signal across the whole genome, with triplicate H3K4me1, H3K4me3, H3K27ac, and CTCF from day 10 of *ex vivo* CD34 differentiation. Read coverage of each mark was calculated (deepTools v2.4.2) for 1 kb windows over open chromatin peaks (bedtools merge -d 10) to capture histone modifications. The HMM model was trained using Poisson log-normal distributions with 20 initial states. These were manually curated to 8 final states based on similarity of chromatin signature. For Principle Component Analysis (PCA) trajectory plotting combined peak calls from sorted hematopoietic populations covering 176,135 open chromatin regions not within 2 kb of transcription start sites were first used to generate a PCA map of erythroid differentiation from sorted populations of HSC, MPP, CMP, MEP and Erythroid populations[24]. Reads within peaks were normalised (R scale) and the PCA was calculated using the R function prcomp. The read counts from *ex vivo* differentiated cells within the same peak set were then used to calculate sample mapping onto PC1 and PC2, and thus to map differentiation timepoints onto the differentiation trajectory. Heatmaps of intersected peaks were generated with pheatmap[110] (v1.0.8) using z-normalised counts of reads per basepair from all identified peaks. For enhancer activity, peak calls were extended by 250 bp in both directions (bedtools slop) to account for the spreading nature of H3K27ac ChIP-seq signal, enhancers were then ranked based on reads per base pair. To determine the point of inflection between low and high acting enhancers H3K27ac read counts were transformed so that the highest value equalled the number of ranked peaks. The point of inflection where the gradient of the curve became greater than one was used to define low and high enhancer activity. The gradient was calculated based on the local linear gradient of ±200 peaks.

Ethics: Blood was collected with ethics approval (MREC 03/08/097) and stored according to HTA guidelines (License 12433).

Acknowledgements: This work was carried out as part of the WIGWAM Consortium (Wellcome Investigation of Genome Wide Association Mechanisms) funded by a Wellcome Trust Strategic Award (106130/Z/14/Z). We appreciate the time and support of Kevin Clarke,

Sally-Anne Clarke and Paul Sopp (WIMM Flow Cytometry Facility, Oxford University) during cell sorting. This work was also supported by Medical Research Council (MRC) Core Funding (MC_UU_12009). D.J.D. received funding from the Oxford University Medical Science Internal Fund: Pump Priming (0006152). Wellcome Trust Doctoral Programmes supported R.S. (203728/Z/16/Z), C.Q.E. (203141/Z/16/Z), M.C.S (097309/Z/11/Z), A.M.O. (105281/Z/14/Z), and R.C.B. (203141/Z/16/Z). C.S. was supported by the Congenital Anaemia Network (CAN). A.M.O was supported by the Stevenson Junior Research Fellowship (University College, Oxford). J.O.J.D. is funded by an MRC Clinician Scientist Award (MR/R008108) and received Wellcome Trust Support (098931/Z/12/Z). G.L. is supported by the Wellcome Trust (090532/Z/09/Z) and MRC Strategic Alliance Funding (MC_UU_12025).

Author Contributions: D.J.D., J.K., J.O.J.D., P.H., G.L., J.A.T., S.T., V.J.B., D.R.H., and J.R.H. designed and planned experiments. D.J.D., S.J.H., L.N., C.S., M.E.G., P.P.H., M.C.S., J.L.D., A.J.C., C.A.D., M.S.C.L., A.M.O., Y.B., A.J.K., P.H., and C.F. performed experiments. D.J.D., R.S., J.M.T., C.Q.E., S.J.M., J.R.J.I., and R.C.B. processed and analysed data. R.K., and Y.N. provided essential reagents. D.J.D., J.A.T., D.R.H, J.R.H. wrote the manuscript.

Competing Interest Statement: J.R.H and J.O.J.D. are founders and shareholders of Nucleome Therapeutics.

738 **REFERENCES**

739 1. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide
740 association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
741 2. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide
742 association loci for human diseases and traits. *Proc. Natl. Acad. Sci.* **106**, 9362–9367
743 (2009).
744 3. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts
745 complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
746 4. Baxter, J. S. *et al.* Capture Hi-C identifies putative target genes at 33 breast cancer
747 risk loci. *Nat. Commun.* **9**, (2018).
748 5. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and
749 Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384.e19
750 (2016).
751 6. Nandakumar, S. K. *et al.* Gene-centric functional dissection of human genetic
752 variation uncovers regulators of hematopoiesis. *Elife* **8**, 1–29 (2019).
753 7. Griffiths, R. E. *et al.* Maturing reticulocytes internalize plasma membrane in
754 glycophorin A – containing vesicles that fuse with autophagosomes before exocytosis.
755 *Blood* **119**, 6296–6307 (2012).
756 8. Ludwig, L. S. *et al.* Transcriptional States and Chromatin Accessibility Underlying
757 Human Erythropoiesis. *CellReports* **27**, 3228-3240.e7 (2019).
758 9. Scott, C. *et al.* Modelling erythropoiesis in congenital dyserythropoietic anaemia type I
759 (CDA-I). *bioRxiv* 1–27 (2019). doi:10.1101/744367
760 10. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links
761 to Common Complex Disease. *Cell* **167**, 1415-1429.e19 (2016).
762 11. van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell.
763 *Nature* **492**, 369–75 (2012).
764 12. Stadhouders, R. *et al.* HBS1L-MYB intergenic variants modulate fetal hemoglobin via
765 long-range MYB enhancers. *J. Clin. Invest.* **124**, 1699–1710 (2014).
766 13. Ulirsch, J. C. *et al.* Systematic functional dissection of common genetic variation
767 affecting red blood cell traits. *Cell* **165**, 1530–1545 (2016).
768 14. Raffield, L. M. *et al.* Common α-globin variants modify hematologic and other clinical
769 phenotypes in sickle cell trait and disease. *PLoS Genet.* **14**, 1–21 (2018).
770 15. Lessard, S. *et al.* An erythroid-specific ATP2B4 enhancer mediates red blood cell
771 hydration and malaria susceptibility. *J. Clin. Invest.* **127**, 3065–3074 (2017).
772 16. Sankaran, V. G. *et al.* Cyclin D3 coordinates the cell cycle during differentiation to
773 regulate erythrocyte size and number. *Genes Dev.* **26**, 2075–2087 (2012).
774 17. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic

775 variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
776 18. Xiong, H. Y. *et al.* The human splicing code reveals new insights into the genetic
777 determinants of disease. *Science (80-.).* **347**, 1254806 (2015).
778 19. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning.
779 *Cell* **176**, 535-548.e24 (2019).
780 20. Schwessinger, R. *et al.* Sasquatch : predicting the impact of regulatory SNPs on
781 transcription factor binding from cell- and tissue-specific DNase footprints. *Genome*
782 *Res.* **27**, 1730–1742 (2017).
783 21. Schwessinger, R. *et al.* DeepC : Predicting chromatin interactions using megabase
784 scaled deep neural networks and transfer learning. *bioRxiv* **724005**, (2019).
785 22. Davies, J. O. J., Oudelaar, A. M., Higgs, D. R. & Hughes, J. R. How best to identify
786 chromosomal interactions: A comparison of approaches. *Nat. Methods* **14**, 125–134
787 (2017).
788 23. Davies, J. O. J. *et al.* Multiplexed analysis of chromosome conformation at vastly
789 improved sensitivity. *Nat. Methods* **13**, 74–80 (2016).
790 24. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts
791 human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
792 25. Beecham, A. H. *et al.* Analysis of immune-related loci identifies 48 new susceptibility
793 variants for multiple sclerosis. *Nat. Genet.* **45**, 1353–1360 (2013).
794 26. Onengut-gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and
795 evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat.*
796 *Genet.* **47**, 381–386 (2015).
797 27. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory
798 bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**,
799 979–986 (2015).
800 28. Sniekers, S. *et al.* Genome-wide association meta-analysis of 78,308 individuals
801 identifies new loci and genes influencing human intelligence. *Nat. Genet.* **49**, 1107–
802 1112 (2017).
803 29. Iglesias, A. I. *et al.* Cross-ancestry genome-wide association analysis of corneal
804 thickness strengthens link between complex and Mendelian eye diseases. *Nat.*
805 *Commun.* **9**, 1–11 (2018).
806 30. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic
807 architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
808 31. Zacher, B. *et al.* Accurate promoter and enhancer identification in 127 ENCODE and
809 roadmap epigenomics cell types and tissues by GenoSTAN. *PLoS One* **12**, 1–25
810 (2017).
811 32. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated

812 Variation in Regulatory DNA. *Science (80-.).* **337**, 1190–1195 (2012).
813 33. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell
814 types. *Nature* **473**, 43–49 (2011).
815 34. Iotchkova, V. *et al.* GARFIELD classifies disease-relevant genomic features through
816 integration of functional annotations with association signals. *Nat. Genet.* **51**, (2019).
817 35. Farh, K. K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease
818 variants. *Nature* **518**, 337–343 (2015).
819 36. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep
820 learning–based sequence model. *Nat. Methods* **12**, (2015).
821 37. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset : learning the regulatory code of the
822 accessible genome with deep convolutional neural networks. *Genome Res.* 990–999
823 (2016). doi:10.1101/gr.200535.115
824 38. Benyamin, B. *et al.* Common variants in TMPRSS6 are associated with iron status
825 and erythrocyte volume. *Nat. Genet.* **41**, 1173–1175 (2009).
826 39. Chambers, J. C. *et al.* Genome-wide association study identifies variants in
827 TMPRSS6 associated with hemoglobin levels. *Nat. Genet.* **41**, 1170–1172 (2009).
828 40. Finberg, K. E. *et al.* Mutations in TMPRSS6 cause iron-refractory iron deficiency
829 anemia (IRIDA). *Nat. Genet.* **40**, 569–571 (2008).
830 41. Muckenthaler, M. U., Rivella, S., Hentze, M. W. & Galy, B. A Red Carpet for Iron
831 Metabolism. *Cell* **168**, 344–361 (2016).
832 42. Ulirsch, J. C. *et al.* Interrogation of human hematopoiesis at single-cell and single-
833 variant resolution. *Nat. Genet.* **51**, (2019).
834 43. Asimit, J. L. *et al.* Stochastic search and joint fine-mapping increases accuracy and
835 identifies previously unreported associations in immune-mediated diseases. *Nat.*
836 *Commun.* **10**, (2019).
837 44. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell*
838 *Rep.* **15**, 2038–2049 (2016).
839 45. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A Phase
840 Separation Model for Transcriptional Control. *Cell* **169**, 13–23 (2017).
841 46. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing Chromosome
842 Conformation. *Science (80-.).* **295**, 1306–1311 (2002).
843 47. Eijsbouts, C. Q., Burren, O. S., Newcombe, P. J. & Wallace, C. Fine mapping
844 chromatin contacts in capture Hi-C data. *BMC Genomics* **20**, 1–13 (2019).
845 48. Cao, Q. *et al.* Reconstruction of enhancer – target networks in 935 samples of human
846 primary cells , tissues and cell lines. *Nat. Genet.* **49**, 1428–1436 (2017).
847 49. Fishilevich, S. *et al.* Original article GeneHancer : genome-wide integration of
848 enhancers and target genes in GeneCards. *Database* **2017**, 1–17 (2017).

50. Hait, T. A., Amar, D., Shamir, R. & Elkon, R. FOCS : a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer – promoter map. *Genome Biol.* **19**, 1–14 (2018).
51. Kumasaka, N., Knights, A. J. & Gaffney, D. J. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nat. Genet.* **51**, (2019).
52. Kan, Y. W. *et al.* Molecular basis of hemoglobin-H disease in the Mediterranean population. *Blood* **54**, 1434 LP – 1438 (1979).
53. Tolosano, E. & Chiabrando, D. Diamond Blackfan anemia at the crossroad between ribosome biogenesis and heme metabolism. *Adv. Hematol.* **2010**, (2010).
54. Kurita, R. *et al.* Establishment of Immortalized Human Erythroid Progenitor Cell Lines Able to Produce Enucleated Red Blood Cells. *PLoS One* **8**, (2013).
55. Moir-Meyer, G. *et al.* Robust CRISPR/Cas9 Genome Editing of the HUDEP-2 Erythroid Precursor Line Using Plasmids and Single-Stranded Oligonucleotide Donors. *Methods Protoc.* **1**, 28 (2018).
56. Vinjamur, D. S. & Bauer, D. E. Growing and Genetically Manipulating Human Umbilical Cord Blood-Derived Erythroid Progenitor (HUDEP) Cell Lines. in *Erythropoiesis: Methods and Protocols* (ed. Lloyd, J. A.) 275–284 (Springer New York, 2018). doi:10.1007/978-1-4939-7428-3_17
57. Bittorf, T., Jaster, R., Lüdtke, B., Kamper, B. & Brock, J. Requirement for JAK2 in erythropoietin-induced signalling pathways. *Cell. Signal.* **9**, 85–89 (1997).
58. Oudelaar, A. M., Davies, J. O. J., Downes, D. J., Higgs, D. R. & Hughes, J. R. Robust detection of chromosomal interactions from small numbers of cells using low-input Capture-C. *Nucleic Acids Res.* **45**, (2017).
59. Geeven, G., Teunissen, H., Laat, W. De & Wit, E. De. peakC : a flexible , non-parametric peak calling package for 4C and Capture-C data. **46**, (2018).
60. Klein, F. A. *et al.* FourCSeq: Analysis of 4C sequencing data. *Bioinformatics* **31**, 3085–3091 (2015).
61. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).
62. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
63. Spain, S. L. & Barrett, J. C. Strategies for fi ne-mapping complex traits. *Human* **24**, 111–119 (2015).
64. Liu, Y. *et al.* Identification of breast cancer associated variants that modulate transcription factor binding. *PLoS Genet.* 1–21 (2017).
65. Allen, E. K. *et al.* SNP-mediated disruption of CTCF binding at the IFITM3 promoter is

associated with risk of severe influenza in humans. *Nat. Med.* **23**, 975–983 (2017).

66. Gobbi, M. De *et al.* A Regulatory SNP Causes a Human Transcriptional Promoter. 1215–1218 (2006).

67. Trakarnsanga, K. *et al.* An immortalized adult human erythroid line facilitates sustainable and scalable generation of functional red cells. *Nat. Commun.* **8**, 1–7 (2017).

68. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).

69. Nussbaum, L. *et al.* High-Throughput Genotyping of CRISPR/Cas Edited Cells in 96-Well Plates. *Methods Protoc.* **1**, 29 (2018).

70. Kosicki, M., Tomberg, K. & Bradley, A. Repair of CRISPR–Cas9-induced double-stranded breaks leads to large deletions and complex rearrangements. *Nat. Biotechnol.* **36**, 765–771 (2018).

71. Owens, D. D. G. *et al.* Microhomologies are prevalent at Cas9-induced larger deletions. *Nucleic Acids Res.* **47**, 7402–7417 (2019).

72. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

73. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

74. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).

75. Wickham, H. ggplot2: elegant graphics for data analysis. *J. Stat. Softw.* **77**, 3–5 (2017).

76. Hughes, J. R. *et al.* Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.* **46**, 205–12 (2014).

77. Hay, D. *et al.* Genetic dissection of the α-globin super-enhancer in vivo. *Nat. Genet.* 1–12 (2016). doi:10.1038/ng.3605

78. Hanssen, L. L. P. *et al.* Tissue-specific CTCF – cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. *Nat. Cell Biol.* **19**, 952–961 (2017).

79. Simon, C. S. *et al.* Functional characterisation of cis -regulatory elements governing dynamic Eomes expression in the early mouse embryo. *Development* **144**, 1249–1260 (2017).

80. Sankaran, V. G. *et al.* Human Fetal Hemoglobin Expression Is Regulated by the Developmental Stage-Specific Repressor BCL11A. *Science (80-.).* **322**, 1839–1842 (2008).

81. Roy, N. B. A. *et al.* A novel 33-Gene targeted resequencing panel provides accurate,

clinical-grade diagnosis and improves patient management for rare inherited anaemias. *Br. J. Haematol.* **175**, 318–330 (2016).

82. Camaschella, C. *et al.* The gene TFR2 is mutated in a new type of haemochromatosis mapping to 7q22. *Nat. Genet.* **25**, 14–15 (2000).
83. Ludwig, L. S. *et al.* Genome-wide association study follow-up identifies cyclin A2 as a regulator of the transition through cytokinesis during terminal erythropoiesis. *Am. J. Hematol.* **90**, 386–391 (2015).
84. Jayapal, S. R. *et al.* Cyclin A2 regulates erythrocyte morphology and numbers. *Cell Cycle* **15**, 3070–3081 (2016).
85. Bielczyk-Maczyńska, E. *et al.* A Loss of Function Screen of Identified Genome-Wide Association Study Loci Reveals New Genes Controlling Hematopoiesis. *PLoS Genet.* **10**, (2014).
86. Chen, F. *et al.* High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. *Nat. Methods* **8**, 753–757 (2011).
87. Randle, S. J., Nelson, D. E., Patel, S. P. & Laman, H. Defective erythropoiesis in a mouse model of reduced Fbxo7 expression due to decreased p27 expression. *J. Pathol.* **237**, 263–272 (2015).
88. Cahalan, S. M. *et al.* Piezo1 links mechanical forces to red blood cell volume. *Elife* **4**, 1–12 (2015).
89. Figueroa, A. A. *et al.* miR-181a regulates erythroid enucleation via the regulation of Xpo7 expression. *Haematologica* **103**, e341–e344 (2018).
90. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–8 (2013).
91. Telenius, J. M. & Hughes, J. R. NGseqBasic - a single-command UNIX tool for ATAC-seq, DNaseI-seq, Cut-and-Run, and ChIP-seq data mapping, high-resolution visualisation, and quality control. *bioRxiv* 393413 (2018). doi:10.1101/393413
92. Canver, M. C. *et al.* Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci. *Nat. Genet.* **49**, 625–634 (2017).
93. ENCODE. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
94. Huang, J. *et al.* Dynamic Control of Enhancer Repertoires Drives Lineage and Stage-Specific Transcription during Hematopoiesis. *Dev. Cell* **36**, 9–23 (2016).
95. Xu, J. *et al.* Developmental control of Polycomb subunit composition by GATA factors mediates a switch to non-canonical functions Jian. *Mol. Cell* **57**, 304–316 (2015).

960 96. Pinello, L., Xu, J., Orkin, S. H. & Yuan, G. Analysis of chromatin-state plasticity
961 identifies cell-type – specific regulators of H3K27me3 patterns. *Proc. Natl. Acad. Sci.*
962 344-E353 (2014). doi:10.1073/pnas.1322570111
963 97. Kang, Y. *et al.* Autophagy Driven by a Master Regulator of Hematopoiesis. *Mol. Cell.*
964 *Biol.* 226–239 (2012). doi:10.1128/MCB.06166-11
965 98. Trompouki, E. *et al.* Lineage Regulators Direct BMP and Wnt Pathways to Cell-
966 Specific Programs during Differentiation and Regeneration. *Cell* **147**, 577–589 (2011).
967 99. Mciver, S. C. *et al.* The exosome complex establishes a barricade to erythroid
968 maturation. *Blood* **124**, 2285–2298 (2019).
969 100. Vries, P. S. De *et al.* Comparison of HapMap and 1000 Genomes Reference Panels
970 in a Large-Scale Genome- Wide Association Study. *PLoS One* 1–22 (2017).
971 doi:10.1371/journal.pone.0167742
972 101. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74
973 (2015).
974 102. Ravenhall, M. *et al.* Novel genetic polymorphisms associated with severe malaria and
975 under selective pressure in North-eastern Tanzania. *PLoS Genet.* **14**, e1007172
976 (2018).
977 103. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47
978 (2016).
979 104. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring
980 population-specific haplotype structure and linking correlated alleles of possible
981 functional variants. *Bioinformatics* **31**, 3555–3557 (2015).
982 105. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing
983 genomic features. *Bioinformatics* **26**, 841–842 (2010).
984 106. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using
985 MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).
986 107. Stanton, K. P., Jin, J., Lederman, R. R., Weissman, S. M. & Kluger, Y. Ritornello: high
987 fidelity control-free chromatin immunoprecipitation peak calling. *Nucleic Acids Res.*
988 **45**, (2017).
989 108. McGowan, S. J., Hughes, J. R., Han, Z. P. & Taylor, S. MIG: Multi-Image Genome
990 viewer. *Bioinformatics* **29**, 2477–2478 (2013).
991 109. Ernst, J. & Kellis, M. ChromHMM: Automating chromatin-state discovery and
992 characterization. *Nat. Methods* **9**, 215–216 (2012).
993 110. Kolde, R. Pheatmap: pretty heatmaps. (2012).
994

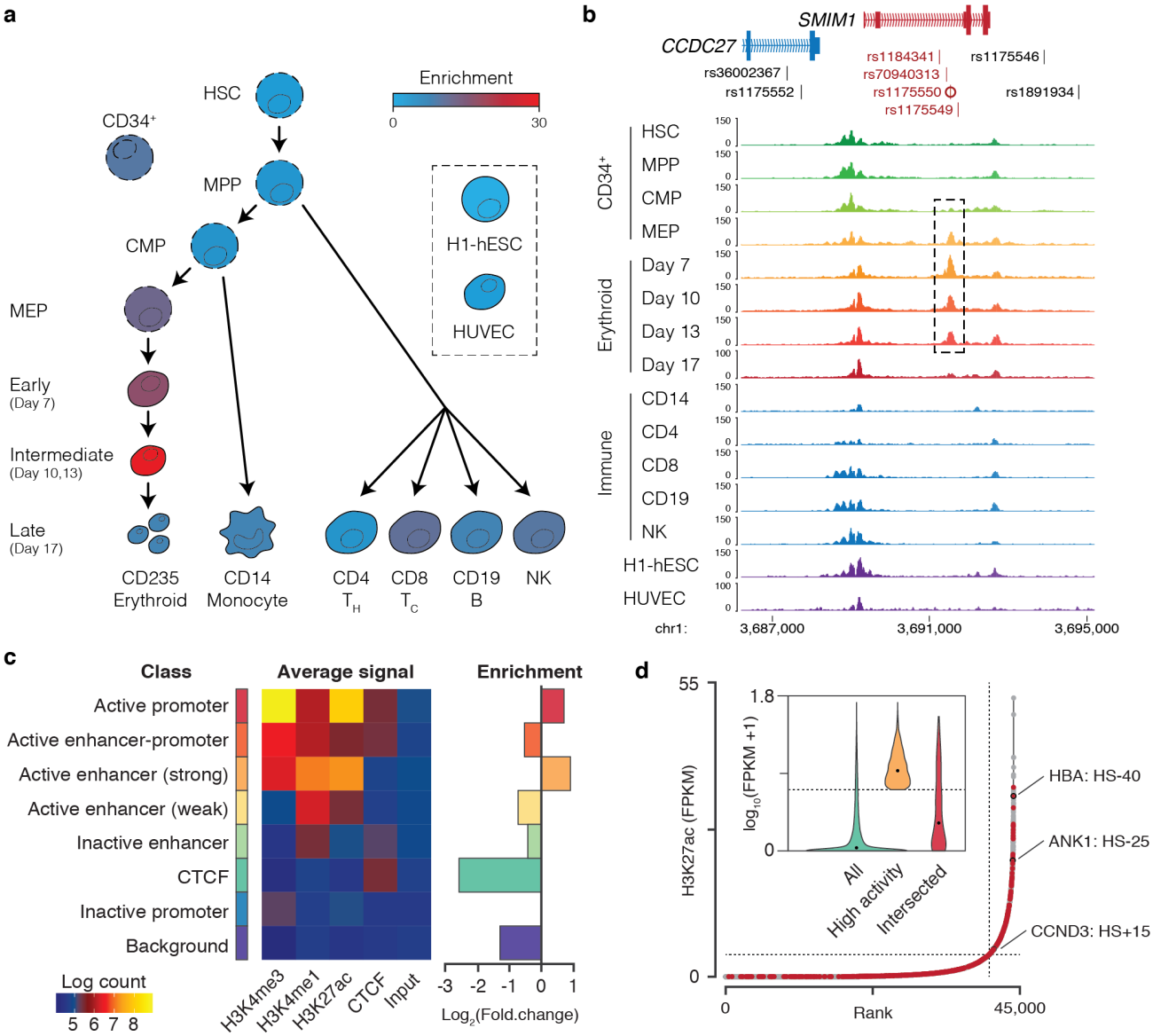

Fig. 1 I Variants associated with RBC traits lie within highly active enhancers. a, Schematic of selected cells from human haematopoiesis showing enrichment (-log(p) of a cumulative Binomial Distribution) for RBC associated variants within open chromatin regions of haematopoietic stem cells (HSC), multi-potent progenitors (MPP), common myeloid progenitors (CMP), megakaryocyte-erythroid progenitors (MEP), early, intermediate, and late erythroid cells from *in vitro* culture, CD14 monocytes, CD4 helper and CD8 cytotoxic T-cells, CD19+ B-cells, natural killer cells (NK), human embryonic stem cells (H1-hESC) and human umbilical vein endothelial cells (HUVEC). **b,** ATAC-seq tracks showing location of open chromatin intersecting variants (red) at the SMIM1 locus. Intersected peaks are highlighted with a dashed box. The index SNP rs1175550 is marked (circle). **c,** GenoSTAN classification and average signal of open chromatin based upon epigenetic marks with the enrichment/depletion in representation of each class amongst elements containing variants. Note, no intersection with inactive promoters was detected so was excluded from enrichment analysis **d,** Open chromatin regions distal (≥2kb) to annotated transcription start sites were ranked by level of H3K27ac ChIP-seq signal (FPKM), with highly active enhancers defined as those above the point of inflection of the curve (marked with a dashed line). Open chromatin regions containing RBC variants (dots coloured red) are enriched for highly active enhancer elements. Hypersensitive sites (HS) near important erythroid genes are shown. A violin plot of H3K27ac levels on all distal regions, highly active distal regions, and variant containing distal regions is inset; the median level is marked (black dot).

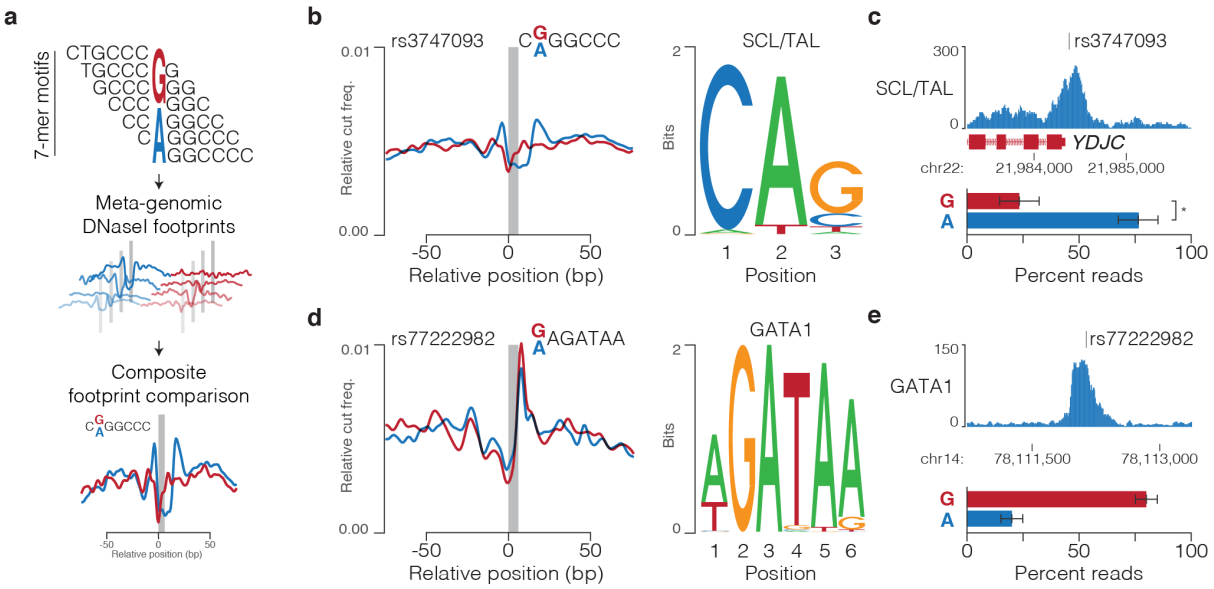

Fig. 2 | Sasquatch provides an unbiased prediction of variant effect. a, Sasquatch analyses *in vivo* generated DNaseI footprints over 7-mer motifs within open chromatin regions to generate meta-genomic footprints. Comparison of Relative cut frequency for each profile is used to generate predictive footprint-change scores. **b,** rs3747093 is within a 7-mer motif (grey bar) which is predicted to alter the DNaseI footprint of SCL/TAL based on presence of the SCL/TAL binding motif. **c,** SCL/TAL ChIP-seq shows allelic skew over rs3747093 as shown by percent of reads containing either allele (*P=0.0468, Ratio paired t-test, n =3). **d,** rs77222982 is within a 7-mer motif (grey bar) which is predicted to alter the DNaseI footprint of GATA1 based on presence of the GATA1 binding motif. **e,** GATA1 ChIP-seq shows allelic skew over rs77222982 as shown by percent of reads containing either allele (n=2). Error bars depict standard error of the mean.

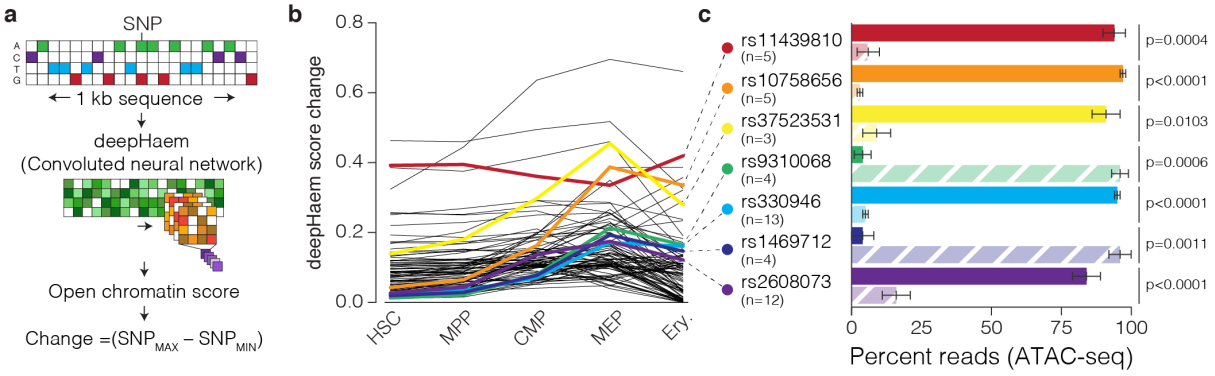

Fig. 3 | Deep Learning predicts variant driven changes in chromatin accessibility. a, deepHaem, a deep convoluted neural network, calculates a chromatin openness score using 1 kb of DNA sequence which can be used to compare variant alleles. **b**, Comparison of alleles for all RBC trait variants in open chromatin (n=2,662) identifies variants with a predicted to change deepHaem openness scores by more than 0.1, or 10% of the maximum openness score (n=91). **c,** Mean percentage of day 10 and day 13 erythroid ATAC-seq reads on either the reference (dark bar) or variant (light dashed bar) allele from heterozygous individuals with a minimum of 5 reads. Error bars depict the standard error of the mean with the number of independent replicates from either multiple donors and/or multiple differentiations shown in parentheses. p-values shown are for a ratio paired t-test.

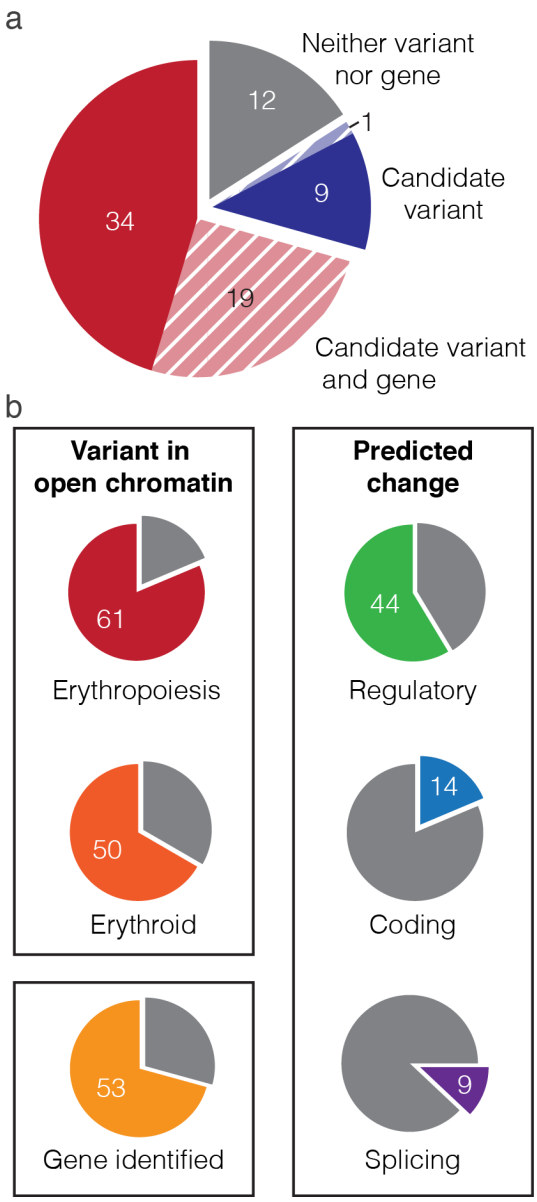

Fig. 4 | The integrated experimental and bioinformatics platform identifies candidate causal variants and effector genes at the majority of polygenic trait regions. a, Cumulative analysis of RBC trait associated variants at 75 GWAS chromosome regions identified candidate causal variants in 63 regions with three or fewer candidate causal variants at 43 regions (solid colouring) and more than three candidates at 20 regions (pale striped colouring). **b,** Pie charts with the number of regions, from a total of 75, with variants found in open chromatin, with variants predicted to alter a regulatory site (Sasquatch, deepHaem), or coding sequence (ANNOVAR), or splicing (SpiDEX, SpliceAI), and regions with identified candidate effector genes. Note, regions may have multiple candidate causal variants each with separate mechanisms of action.

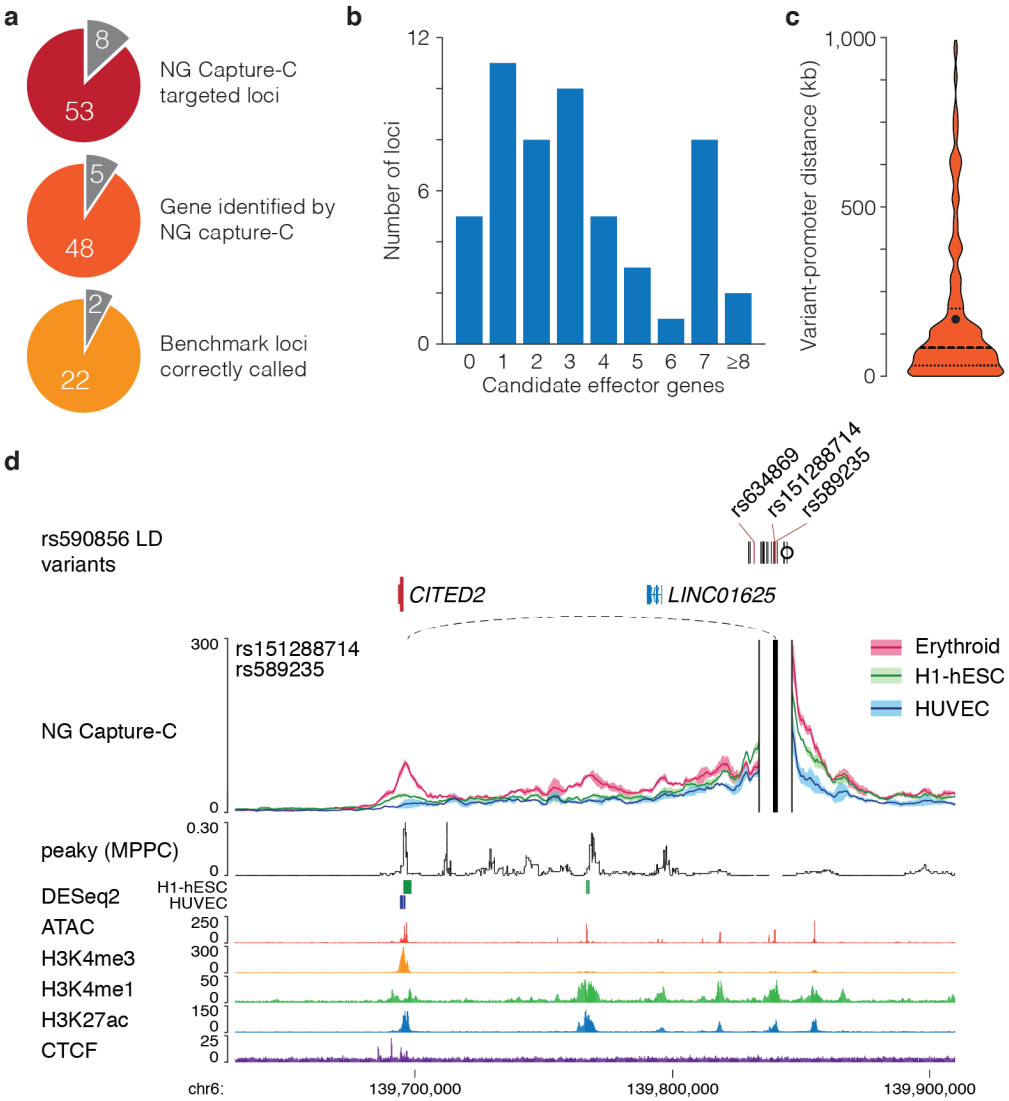

Fig. 5 I NG Capture-C can detect long range variant-promoter interactions. a, NG Capture-C oligonucleotides were designed for 61 chromosome regions with candidate regulatory causal variants, regions were excluded from targeting based on sheer number of targets (>100 target sites at a single locus, n=1), or where probes were impossible to design due to repetitive elements. Following analysis of generated 3C data genes were identified at 48 of the 53 targeted regions, 24 of which were used for accuracy benchmarking. **b,** Histogram of the number of candidate effector genes identified by NG Capture-C at each region. **c,** Violin plot of the distance between variants and the target transcription start sites. Median (83,944 bp) shown as a thick dashed line and mean (168,148 bp) shown as a black circle. **d,** 3C interaction profile for open chromatin containing rs151288714 and rs589235 in erythroid, human embryonic stem (H1-hESC) and human umbilical vein endothelial (HUVEC) cells (n=3). Capture viewpoints and proximity exclusion regions (solid vertical lines) were designed for open chromatin regions and profiles show mean interactions (solid line) with one standard deviation (shading). *CITED2*-variant interactions were identified as erythroid specific interactions (dashed loops; DESeq2 q-value < 0.05 shown as bars). Peaky values depict the Marginal Posterior Probability of Contact (MPPC) in erythroid cells. Variants within open chromatin are red, as are variant interacting genes, the index SNP is marked with a circle. FPKM normalised ATAC-seq and ChIP-seq tracks are from erythroid cells. Interaction was found witch *CITED2*, which encodes the Cbp/p300 Interacting Transactivator with Glu/Asp (E/D)-rich tail 2 protein and required for normal haematopoiesis.

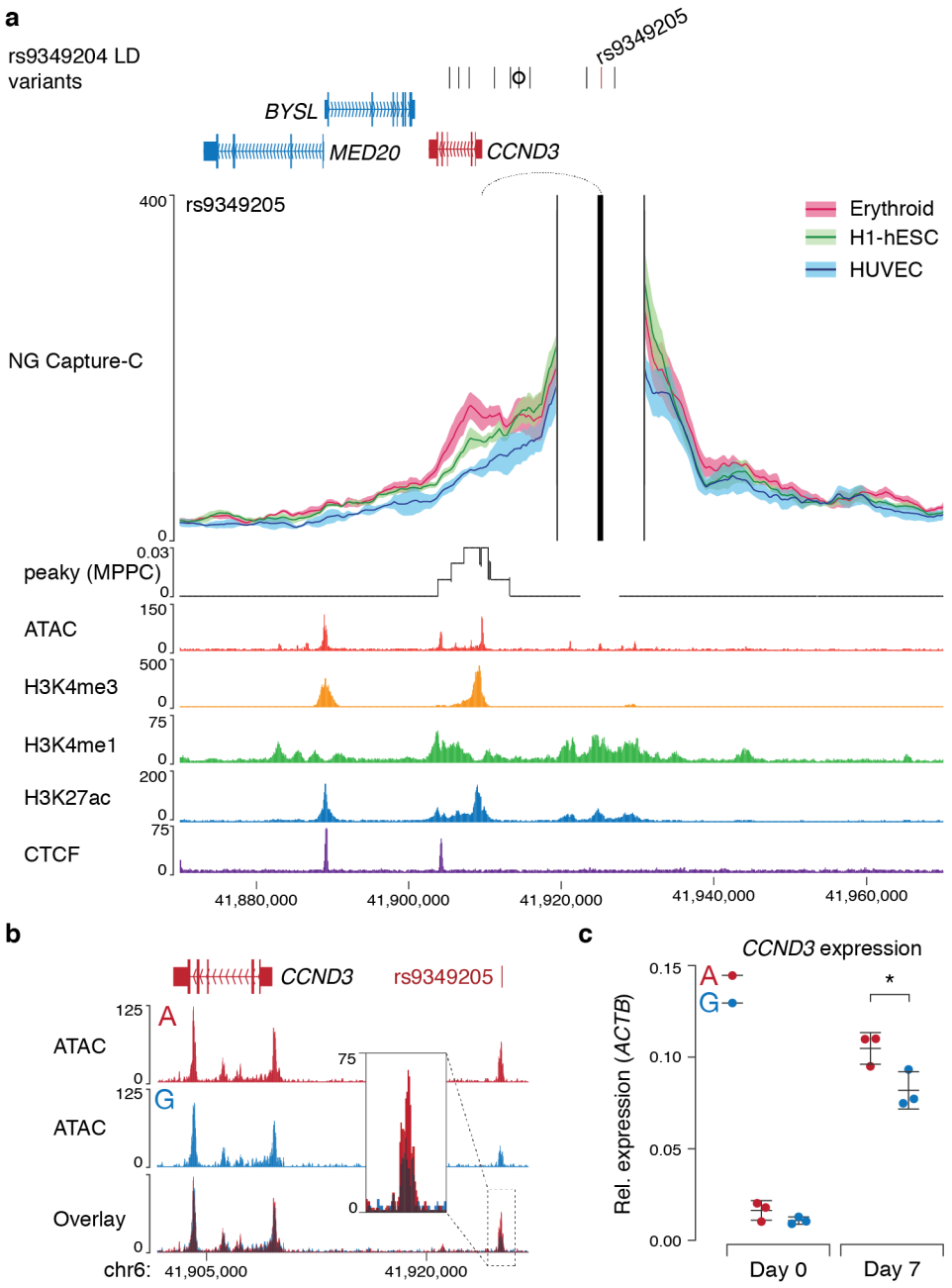

Fig. 6 | rs9349205 interacts with, and regulates CCND3. a, 3C interaction profile for rs9349205 in erythroid, embryonic stem (H1-hESC) and umbilical vein endothelial (HUVEC) cells (n=3). Profiles show windowed mean interactions (solid lines) with one standard deviation (shading). Peaky values depict the MPPC in erythroid cells. Interaction with *CCND3* was detected by peaky (dotted loop; MPPC > 0.01). Variants within open chromatin are red, as are variant interacting genes, the index SNP is marked with a circle. FPKM normalised ATAC-seq and ChIP-seq tracks are from erythroid cells. **b,** Merged FPKM normalised ATAC-seq (n=3) from HUDEP-2 cells homozygous for either rs9349205 allele with overlaid track showing high similarity, and a slight reduction at the intersected peak for homozygous G clones (inset). **c,** Real time reverse-transcriptase PCR of *CCND3* in differentiating HUDEP-2 clones (n=3) showed lower expression in G clones at day 7 (Student's two-tailed t-test, *p=0.0387). Bars show mean and one standard deviation of independent clonal populations (circles).

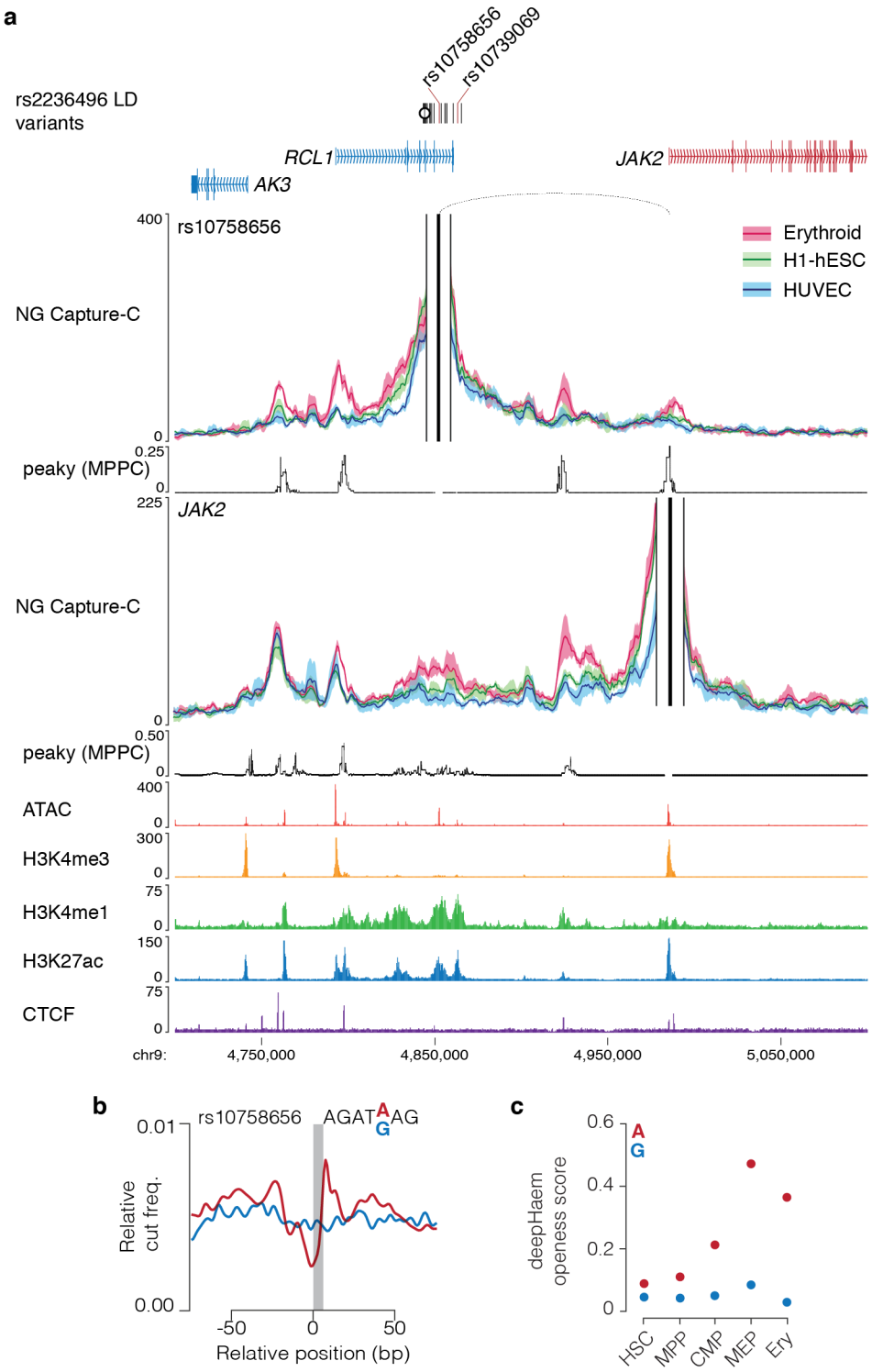

Fig. 7 | rs10758656 interacts with *JAK2* and is predicted to alter chromatin accessibility. a, 3C interaction profiles for rs10758656 and *JAK2* in erythroid, embryonic stem (H1-hESC) and umbilical vein endothelial (HUVEC) cells (n=3). Profiles show windowed mean interactions (solid lines) with one standard deviation (shading). Peaky values depict the MPPC in erythroid cells. *JAK2*-rs10758656 interaction was detected by peaky (dotted loop; MPPC > 0.01). Variants within open chromatin are red, as are variant interacting genes, the index SNP is marked with a circle. FPKM normalised ATAC-seq and ChIP-seq tracks are from erythroid cells. **b,** Sasquatch profiles for rs10758656 show loss of a GATA footprint. **c,** deepHaem openness scores rs10758656 predict a loss of chromatin accessibility in erythroid cells.

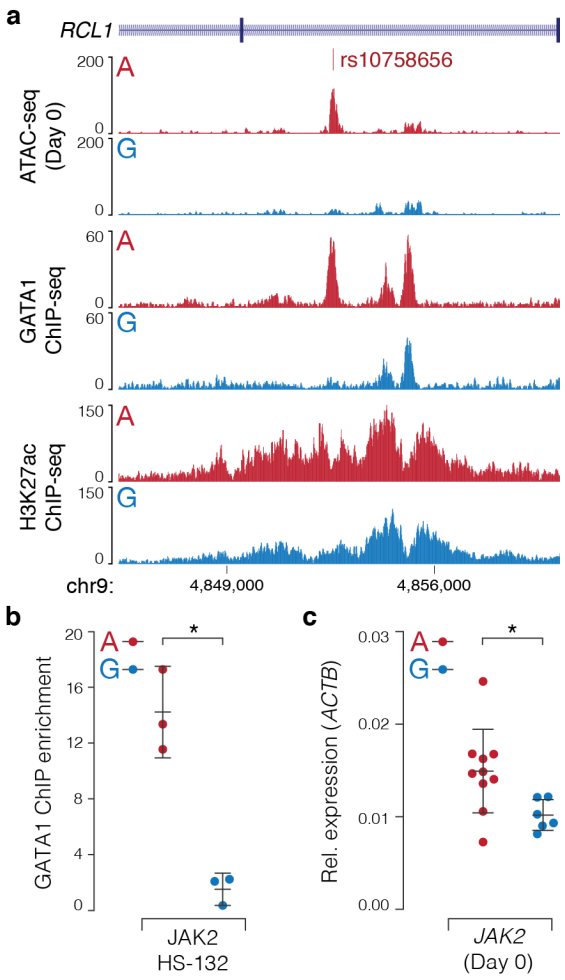

Fig. 8 | rs10758656 causes loss of open chromatin and reduced *JAK2* expression. a, Overlaid FPKM normalised ATAC-seq (n=3) and H3K27ac ChIP-seq (n≥1) from differentiating HUDEP-2 clones homozygous for either rs10758656 allele. Dark shading indicated overlapping signal **b,** Real time quantitative PCR for GATA1 ChIP at rs10758656 (*Student's two-tailed t-test, p=0.0136). **c,** Real time reverse-transcriptase PCR of *JAK2* in differentiating HUDEP-2 clones (n≥6) showed lower expression in G clones (Mann-Witney test, *p=0.0160). Bars show mean and one standard deviation of independent clonal populations (circles).