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.

BACKGROUND

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

84 genome; with the remaining variants assessed for effects on coding sequences and RNA
85 processing using established programmes[17–19]. The platform addresses the fact that both
86 causal and non-causal variants may lie in open chromatin. Using DNaseI footprinting and a
87 machine learning approach the platform prioritises variants predicted to directly affect the
88 binding of transcription factors or alter chromatin accessibility[20,21]. Having prioritised putative
89 regulatory causal variants, the platform then links the regulatory elements in which they
90 occur to genes using NG Capture-C, the highest resolution chromatin conformation capture
91 (3C) method currently available for targeting numerous loci[22,23]. To validate the predicted
92 molecular changes caused by such GWAS variants we use CRISPR/Cas9 facilitated
93 Homology Dependent Recombination (HDR) to directly model SNP alleles and determine
94 their effects.
95
96 Testing our platform against 75 chromosome regions from a previous GWAS of RBC traits[11]
97 we identified putative causal variants at ~80%, their candidate effector genes at ~70%, and
98 three or fewer candidate variants at ~60%. By benchmarking with the eight validated causal
99 variants from previous studies[12–16], and genes at well characterised erythroid loci, we
100 successfully predicted >87% of both causal variants and effector genes. Finally, we used
101 genome editing to directly determine the *in vivo* molecular effects of candidate SNPs in two
102 regions – showing both SNPs to be causal and verifying *JAK2* as a novel RBC trait effector
103 gene. As this platform was developed with methods appropriate for small numbers of cells,
104 and therefore rare cell types, the approach will enable researchers across a wide range of
105 traits or disorders to more readily identify causal variants, the cells in which they exert their
106 effects, their target genes, and the mechanisms by which they alter cell biology, and
107 ultimately, disease risk.

108 **RESULTS**

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

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

120

121 Previously, 1,114 index SNPs, each of which identifies a region of LD, have been associated
122 with specific RBC traits in two extensive GWASs of Asian and European populations[10,11].
123 These index SNPs are associated, via LD ($r^2 \geq 0.8$), with a total of 30,694 variants.
124 Approximately 8% of these variants, covering ~60% of RBC trait regions, intersected with
125 open chromatin in erythropoietic cells (n=2,590). Intersections were predominantly found in
126 fully committed, intermediate erythroid cells (days 10-13, cumulative binomial distribution,
127 $p=7x10^{-29}$) rather than in multipotent progenitor cells (Fig. 1a, Supplementary Figs. 3a,b).
128 Enrichment was trait specific as variants associated with immune diseases[25–27] showed
129 minimal enrichment for intersection with erythroid open chromatin but strong enrichment in
130 differentiated lymphocytes (Supplementary Fig. 3d-f) while non-haematological trait
131 variants[28–30] showed no enrichment in either red or white blood cells. For all traits, we saw
132 no enrichment when we analysed ATAC-seq profiles from two non-haematopoietic cell lines.

133

134 To further characterise the intersected regulatory elements, we generated ChIP-seq data to
135 distinguish promoters (Histone-3 Lysine-4 trimethylation, H3K4me3), enhancers (H3K4
136 monomethylation, H3K4me1), boundary elements (CTCF), and "active" sites (H3 Lysine-27
137 acetylation, H3K27ac) in committed erythroid cells. We applied GenoSTAN[31] to assign a
138 chromatin signature to each element and thereby generated a high-resolution map of open
139 chromatin in erythroid cells with seven functional classes (Fig. 1c, Supplementary Fig. 4).
140 Intersected elements were enriched for enhancers and promoters with high levels of
141 H3K27ac but not those with low levels of H3K27ac, nor ATAC-seq peaks with CTCF
142 enrichment (Fig. 1c). When putative enhancers were ranked for their levels of H3K27ac (Fig.
143 1d), elements containing RBC variants were significantly enriched amongst the highly
144 activity erythroid enhancers (χ^2: d.f.=3, $p=7x10^{-54}$).

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

182 adjacent to an AGATAA motif showed allelic imbalance in GATA1 binding (Fig. 2d-e). Often,
183 skew in enrichment was seen across more than one factor in elements affected by a single
184 variant, probably reflecting co-dependency in their binding (Supplementary Fig. 8). Such
185 imbalance is consistent with the alteration of binding predicted by Sasquatch, demonstrating
186 its ability to accurately detect causative variants.
187
188 Convolutional neural network based machine learning can predict open chromatin[36,37] and
189 was also used to identify causal variants. We used 936 chromatin-accessibility and
190 epigenetic datasets to train a deep convolutional neural network, deepHaem[21], to predict
191 chromatin accessibility based on DNA sequence across haematological cell-types (Fig. 3a,
192 Supplementary Fig. 9a-d). Using deepHaem it is possible to predict the effect of variants on
193 chromatin accessibility. DeepHaem identified 91 variants in open chromatin with changes
194 greater than 10% of the maximum accessibility score (1.0), with the strongest effects seen in
195 MEP and erythroid populations (Fig. 3b). 45 of the variants predicted to alter chromatin-
196 accessibility had scores greater than 0.1 in erythroid cells. Using ATAC-seq, we identified
197 heterozygous alleles for 15 of these 45 variants in three healthy individuals. Comparison of
198 sequencing from these alleles showed significant bias in ATAC-seq accessibility at seven of
199 the sites and skew at a further five sites (Fig. 3c, Supplementary Fig. 9e) indicating that
200 deepHaem can accurately predict variant-induced changes in chromatin accessibility.
201
202 To assess how well the platform performed at identifying causative regulatory variants we
203 used previously characterised RBC trait variants. Currently, no RBC trait variants have been
204 definitively shown to be causative using direct *in vivo* modelling; however, eight regulatory
205 variants have strong support from functional assays[12–16]. The approach established here
206 identified that seven of these eight variants lie in open chromatin in erythroid cells, and
207 therefore had the potential to be regulatory causal variants. Characterisation with Sasquatch
208 or deepHaem further prioritised six of these variants as likely to be causative
209 (Supplementary Table 1). Therefore, the platform accurately prioritises causal variants, and
210 thus identifies variants for functional analysis.
211
212 **A comprehensive search for all causal variants within an RBC GWAS.**
213 The first major RBC trait GWAS identified 75 index SNPs[11]; the associations identified are
214 likely to represent the most common variants with moderate effect sizes and some rare
215 variants with large effect sizes, therefore we focused specifically on this dataset for in-depth
216 follow-up. A comprehensive GWAS decoding platform must prioritise causal variants by
217 treating all mechanisms as plausible. By examination of the 75 index SNPs, as well as
218 variants in high LD with them ($r^2 \geq 0.8$, 1000 Genomes Project; n=6,420), we identified 486

219 candidate regulatory variants within 61 of the 75 chromosome regions (Supplementary Table
220 1). In addition to regulatory variants, we considered the possibility for coding and splicing
221 changes across these regions. Putative coding sequence changes were identified using
222 ANNOVAR[17] and then filtered for erythroid expressed genes. This identified 20 variants
223 predicted to alter protein sequence at 14 regions (Supplementary Table 1). Next, putative
224 alternative-splicing variants were identified using a combination of Splicing Index[18] and a
225 deep learning approach, SpliceAI[19]. Together, these programmes identified 13 putative
226 splice-altering SNPs in 11 erythroid expressed genes across nine regions; however, no
227 variants were highlighted by both algorithms (Supplementary Table 1). Using these
228 integrated analyses for coding, splicing and regulatory mechanisms we identified candidate
229 causal variants at 63 of the 75 chromosome regions, with 43 of these having three or fewer
230 strong candidates, and the majority of candidates being in tight linkage ($r^2 \geq 0.9$; n=394/515)
231 with their index SNP (Fig. 4, Supplementary Figs. 10,11a).
232
233 We next considered why no causal variants were predicted for 12 chromosome regions.
234 Immediate possibilities are that the variant affects gene function in a way that is currently
235 unrecognised, or exerts effects in an untested cell type. Consistently, rs855791, also
236 identified in multiple GWASs, including for iron status[38,39], is a missense variant of
237 *TMPRSS6*. *TMPRSS6* is expressed in the liver and encodes Matripase-2, a suppressor of
238 the iron homeostasis master regulator, Hepcidin[40,41]. It is also possible that the causal
239 variant affects mRNA stability. However, there are currently no good predictive software
240 programmes for this. Finally, it may be that causal variants were not identified because the
241 initial GWAS study was not sufficiently powered or used a sub-optimal catalogue of variants;
242 resulting in incompletely resolved genetics. Indeed, index SNPs in unresolved loci were less
243 likely to be replicated in subsequent RBC trait GWASs[10,42] than index SNPs at resolved loci
244 (Supplementary Fig. 11b). Additionally, in a region with multiple unlinked causal variants,
245 incompletely resolved genetics can lead to index SNPs being identified through weak
246 association ($r^2 < 0.8$) with two or more causal variants. Such index SNPs are referred to as
247 tag SNPs[43] (Supplementary Fig. 11a). At the *TMCC2* locus, where rs9660992 is an index
248 SNP[11], moderate linkage (r^2=0.51-0.82) is seen with two independent index SNPs from a
249 subsequent RBC trait GWAS[10]. While rs12137294 and rs1172129 are themselves unlinked
250 (r^2=0.46), each is in tight linkage with several variants in open chromatin (Supplementary
251 Fig. 12); suggesting rs9660992 may be a tag SNP. Therefore, both additional cell types and
252 incomplete genetics can explain unresolved regions.
253
254
255

256 **High resolution 3C mapping accurately identifies effector genes.**
257 The target or effector genes for splicing and coding variants can be directly inferred, but the
258 effector genes of regulatory variants must be identified experimentally. For enhancers to
259 regulate gene expression they often physically interact with target promoters, likely through
260 loop-extrusion and/or phase-separation[44,45]. The close proximity required for regulation can
261 be identified by chromosome conformation capture (3C) to map interactions[46] and this
262 provides a means by which to identify effector genes. NG Capture-C uses biotinylated
263 oligonucleotide probes to target specific loci at high resolution in multiplexed 3C samples[23];
264 allowing statistical comparison for identification of enhancer-promoter interactions. We
265 designed probes for 214 variant containing *cis*-regulatory elements covering 53/61
266 chromosome regions with putative regulatory variants; then simultaneously generated 3C
267 interaction data in intermediate erythroid cells, H1-hESCs and HUVECs to link *cis*-regulatory
268 elements with their effector genes. Using a combination of tissue-specificity (DESeq2)[23],
269 Bayesian modelling (Peaky)[47], and promoter proximity (≤5 kb) to call variant-promoter
270 interactions we identified 194 candidate effector genes at 48 of the 53 targeted regions (Fig.
271 5a-b, Supplementary Table 1). For each targeted region, NG Capture-C identified an
272 average of four genes, which is consistent with the predicted number of gene targets for
273 enhancers[48–50], though whether multiple genes contribute to a GWAS trait at a single locus
274 remains to be determined. Although some methods have indicated that GWAS variants are
275 most frequently found within 20 kb of their target genes[51] we detected interactions up to 992
276 kb, with a median distance of 83.9 kb (±9.3 kb SE, Fig. 5c). Such long-range interactions
277 were seen at several well characterised erythroid loci including *CITED2* (139 kb), *SLC4A1*
278 (47 kb), *RBM38* (24 kb), *ANK1* (25 kb), *MYB* (85 kb), and *HBA1/2* (63 kb), showing that
279 GWAS variants, as for other enhancer-promoter interactions, may act over large distances
280 (Fig. 5d, Supplementary Figs. 13-18).
281
282 The erythroid system and RBC traits have been intensively analysed and characterised;
283 therefore, we were able generate a set of the 24 “most likely” effector genes within the 53
284 targeted chromosome regions based on prior knowledge of their function (Supplementary
285 Fig. 19a). This set of genes allowed us to benchmark our approach; finding that with NG
286 Capture-C we correctly identified 22 of the 24 most-likely effector genes (Fig. 5a,
287 Supplementary Fig 19b). Of the remaining regions, no genes were identified at one (*miR-*
288 *181a*), and four incorrect candidates were identified in the region where *TAL1* is almost
289 certainly the effector. With these exceptions, NG Capture-C performs with a high rate of
290 success in identifying effector genes linked to potential causal variants. Three previous
291 attempts with diverse methods to identify effector genes associated with RBC traits have
292 been reported[5,6,11]. These were an annotation-based approach[11], Promoter Capture-HiC[5]

293 (PC-HiC), and a gene-centric shRNA screen[6]. We directly compared these different
294 approaches with NG Capture-C. There was little consistency between the candidate gene
295 lists from these approaches (Supplementary Fig. 19c), with NG Capture-C the only method
296 to identify *HBA-1/2*, the α-globin encoding genes, which are known to be associated with
297 anaemia and changes in RBC traits[52] (Supplementary Figs. 17,19c). Our approach was also
298 unique in identifying *RPL19*, of interest because ribosomal genes are known to cause
299 Diamond-Blackfan anaemia[53]. Across the 24 benchmark regions, NG Capture-C and the
300 annotation-based approach were the most sensitive methods, respectively identifying 91.7%
301 and 70.8% of the most-likely effector genes correctly (Supplementary Fig. 19 b,d). Overall,
302 the direct comparison of different gene identification methods shows that NG Capture-C is
303 the most successful tool.
304
305 **Direct modelling of rs9349205 shows reduced expression of its target gene *CCND3*.**
306 The most direct evidence that a particular variant alters gene expression comes from
307 introducing both alleles to an isogenic background and observing an appropriate change in
308 the relevant cell type. Previous studies characterising RBC trait variants have used reporter
309 assays and/or targeted deletions of the regulatory element[13–16]. However, these may not
310 faithfully recapitulate variants effects *in vivo.* Therefore, we used CRISPR/Cas9-facilitated
311 homology directed repair (HDR) to directly model prioritised variants at five GWAS regions in
312 the Human Umbilical Derived Erythroid Progenitor (HUDEP-2) cell line (Supplementary Fig.
313 20,21); a model of human erythroid differentiation and maturation[54,55]. As previous studies
314 have shown some clonal variation when using such cells[56] it is essential to analyse multiple
315 independently isolated clones. We were able to generate sufficient independent clones for
316 robust analysis of two regions (*CCND3* and *JAK2*).
317
318 Using our platform, rs9349205 was identified as tightly linked to the index SNP (rs9349204,
319 r^2=0.841); rs9349205 is the only one of ten linked variants which lies within open chromatin
320 in committed erythroid cells, and shows 3C interaction with *CCND3* (Fig. 6a), which is the
321 most-likely effector gene[16]. rs9349205 also had small effects on both the Sasquatch DNaseI
322 footprint and deepHaem chromatin openness scores (Supplementary Figs. 20,22a). Editing
323 was used to convert rs9349205$^{A/A}$ HUDEP-2 cells to rs9349205$^{G/G}$; a non-erythroid locus
324 was also edited to control for non-specific effects from editing (e.g. spontaneous
325 differentiation). Using ATAC-seq to assess chromatin accessibility, we found that the
326 identified regulatory element in the rs9349205$^{G/G}$ genotype was 54.5% less accessible than
327 in rs9349205$^{A/A}$ cells (Fig. 6b, Supplementary Fig. 22b). The rs9349205$^{G/G}$ clones also
328 showed significantly reduced *CCND3* expression during erythroid differentiation (Fig. 6c). As
329 previously discussed[16], *CCND3* regulates the G2 to S transition during erythropoiesis, and

330 thus knockout of *CCND3* in mice leads to an increased erythrocyte volume, consistent with
331 linkage to changes in mean cell volume (MCV) detected through GWAS.
332
333 **rs10758656 causes reductions in chromatin accessibility and *JAK2* expression.**
334 Using NG Capture-C we identified *JAK2*, which encodes Janus Kinase 2, as an effector
335 gene for variants in high LD with the index SNP rs2236496. We confirmed this interaction
336 using NG Capture-C from the *JAK2* promoter (Fig. 7a). Of the 18 linked variants, only two,
337 rs10758656 and rs10739069, intersect open chromatin. Of these two SNPs Sasquatch
338 characterised rs10758656 but not rs10739069 as having the potential to affect transcription
339 factor binding, with the motif strongly matching that of the GATA1 binding motif (Fig. 7b,
340 Supplementary Fig. 23a). DeepHaem also predicted that rs10758656 but not rs10739069
341 would affect chromatin accessibility. Therefore, HUDEP-2 cells, which are heterozygous A/G
342 for rs10758656, were edited to homozygosity. We generated 16 independent clones
343 homozygous for either A (n=10) or G (n=6). ATAC-seq of these cells during expansion and
344 differentiation showed up to 82% ablation of open chromatin in the rs10758656$^{G/G}$ clones,
345 associated with 86.2% and 58.4% reductions in GATA1 binding and H3K27ac, respectively
346 (Fig. 8a-b, Supplementary Figs. 24b-e). These findings match the prediction of both
347 Sasquatch and deepHaem. Despite being closer to the promoter of *RCL1* than *JAK2*,
348 rs10758656$^{G/G}$ specifically reduced expression of *JAK2* (Fig. 8c, Supplementary Fig. 24f,g),
349 consistent with the specificity of the rs10758656-*JAK2* interaction profile seen in NG
350 Capture-C. JAK2 functions as part of the erythropoietin signalling pathway[57]. Our results
351 demonstrate that *JAK2* is a GWAS effector gene and most likely results in changes to the
352 MCV noted in GWAS through altered signalling responses.

353 **DISCUSSION**

354 Here we have developed and validated a platform to identify causative GWAS variants and
355 link them to the genes whose function they affect. In our platform ATAC-seq analysis allows
356 researchers to identify relevant cell types using the fundamental regulatory elements of the
357 genome: enhancers, promoters and boundary elements. GWAS variants are then assessed
358 *in silico* to predict which variants are likely to alter gene expression or function. Candidate
359 regulatory variants are finally linked to their effector genes using NG Capture-C. Using this
360 method to analyse variants in high LD to RBC trait index SNPs resulted in identification of
361 candidate causal variants and effector genes at a majority of chromosome regions (>70%).
362 Benchmarking also shows that this approach is robust, with 88% of validated causal
363 variants, and 92% of most-likely effector genes identified. Application of this method to fine-
364 mapped GWAS variants is likely to further improve its success. Finally, the functional effects
365 of candidate polymorphisms can then be assessed using allele-specific assays of chromatin
366 accessibility and gene expression.
367
368 This platform has been developed and benchmarked using data from purified
369 haematopoietic cells at various stages of commitment, differentiation and maturation along
370 the erythroid pathway to producing RBC. Using haematopoiesis as a model, we show how
371 causal variants can be assigned to the cell types in which they exert their effects and the
372 genes whose expression is perturbed. In principle, this method could be used for any GWAS
373 datasets for which appropriate cell types are available. To ensure that this would apply to
374 rare cell types and a wide range of diseases, we have established a platform that can be
375 effectively applied using as few as 500 cells for ATAC-seq and 20,000 cells for NG Capture-
376 C[24,58]. These data can then be used to improve *in silico* processing and machine learning,
377 meaning that damaging changes can be predicted and prioritised using data from rare cells
378 and those grown under varying conditions of stimulation.
379
380 Linking variants to their effector genes using NG Capture-C can easily and reproducibly be
381 applied across a wide range of cell-types at hundreds of specifically targeted sites in either
382 gene- or enhancer- centric designs[22,23]. The ability to compare 3C data from multiple cell
383 types allows tissue-specific and tissue-invariant interactions to be called by a wide range of
384 statistical approaches[23,47,59–61], increasing the throughput of accurate effector gene
385 identification. These candidates can then be validated with functional follow-up, such as
386 screening approaches[6], or as shown here, *in vivo* modelling, to help to explain associated
387 cellular phenotypes.
388

389 In addition to elucidating the genes involved at GWAS regions, we have also addressed the
390 multiple molecular mechanisms that may underlie such signals. To date, strong evidence
391 supports a mixture of coding, splicing and regulatory mechanisms[62]. The approach
392 described here identifies enhancer, promoter, RNA processing and coding variants. Despite
393 this we were still unable to identify causative variants at 16% of chromosome regions. This
394 could partly have resulted from the fact that initial variant identification used linkage
395 disequilibrium, which could be improved with either fine mapping or whole genome
396 sequencing[62,63]. Nevertheless, other factors are also likely to contribute. The cell types
397 affected in any complex disease are not necessarily the most obvious candidates. For RBC
398 traits, the most likely affected lineage is erythropoiesis itself. However, other cell types
399 modify erythropoiesis, including those producing growth factors, cytokines, or mediating
400 cell/cell interactions such as macrophages that facilitate enucleation of RBC precursors.
401 Furthermore, causal variants may act in the identified cell type, but only in response to
402 specific environmental cues or signalling. Therefore, platforms such as this must be
403 implemented with a comprehensive appreciation of the systems involved. It is also important
404 to consider that additional untested molecular mechanisms may underlie GWAS signals.
405 Although we found no specific enrichment for variants in CTCF elements, numerous were
406 within CTCF peaks. Recent evidence has shown that disruption of CTCF binding by
407 common variants plays a role in determining the severity of influenza and breast cancer[64,65],
408 thus it likely represents a less common, yet important molecular mechanism. Additionally,
409 modelling of rs10758656 showed near complete loss of open chromatin. It is equally
410 possible that some variants generate, rather than abolish, open chromatin sites. Such a
411 variant has already been described as causing anaemia at the α-globin locus[66]. These sites
412 could only be detected by analysis of individuals with the correct genotype.
413
414 **CONCLUSIONS**
415 Although this integrative platform efficiently identifies variants and genes, it shifts the
416 bottleneck of GWAS follow-up to validation. Using direct *in vivo* modelling we have shown
417 how alleles can alter enhancers and gene expression to different extents. Although we have
418 shown in principle that prioritised variants can be proven to be functionally causative it
419 requires an HDR editable cell type, is labour intensive, and does not work at all loci; this step
420 will require rapid single base editing to enable significant progress. We expect that with the
421 implementation of integrative platforms such as this, and with ongoing advancement of
422 molecular techniques and editing technologies the benefits of GWAS for understanding
423 human physiology and improving health will accelerate.

424 **METHODS**

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

442

443 *Differentiation of $CD34^+$ cells:* Cells were differentiated under a three phase *ex vivo* protocol
444 adapted from that used for the BEL-A cell line[7,9,67]. Growth media are listed in
445 Supplementary Table 2. Briefly, for differentiation 0.5-2.5x10^5 cells were resuspended on
446 day 0 in Phase I media at 10^5 cells ml^{-1}. Cell counts were performed on days 3 and 5 with
447 additional Phase I media added to return the concentration to 10^5 cells ml^{-1}. On day 7, cells
448 were counted and pelleted (400 rcf, 5 min, RT) and resuspended in Phase II media at 3x10^5
449 cells ml^{-1}. Cells were counted on day 9 and diluted to 3x10^5 cells ml^{-1} Phase II media. On
450 day 11, cells were counted and pelleted (400 rcf, 5 min, RT) and resuspended in Phase III
451 media at 10^6 cells ml^{-1}. Cells were counted on days 13 and 15 and diluted to 10^6 cells ml^{-1} in
452 Phase III media. Reproducibility between differentiations was confirmed morphologically with
453 cytospins, immunologically with six FACS cell surface markers and epigenetically with
454 ATAC-seq enhancer staging[24]. For morphological analyses 10^5 cells were resuspended in
455 200 ml PBS and spun (5 min, 400 rpm) in a Cytospin 4 (ThermoFisher), before staining with
456 modified Wright's Stain on a Hematek (Bayer Health Care), and mounting with DPX (Sigma).
457 Images were taken on an Olympus BX 60 microscope at 10x and 20x magnification. For
458 differentiation and enucleation FACS analyses 10^5 cells were resuspended in FACS buffer
459 (90 % PBS, 10 % FBS) and stained with an erythroid differentiation panel of antibodies
460 (Supplementary Table 4) against CD34, CD36/Fatty acid translocase, CD235a/Glycophorin

461 A, CD71/Transferrin Receptor, CD233/Band3, CD49d/α-Integrin, and with Hoescht-33258
462 (ThermoFisher) for live/dead analysis, with Hoescht-33342 (ThermoFisher) for enucleation
463 assays. For immune cell purities, cells were stained with cell-type specific panels of
464 antibodies (Supplementary Fig. 25, Supplementary Table 4). FACS was carried out on an
465 Attune NxT (ThermoFisher), voltages and compensation were set using Ultra Comp eBeads
466 (ThermoFisher) for antibodies, and single stained cells for dyes. Gating was performed using
467 fluorescence minus one (FMO) controls. Analysis was performed using either Attune NxT
468 software (v3.0) or FlowJo (v10.4.2).
469
470 *Cell line culture and HUDEP-2 differentiation:* Human ESC line H1 (H1-hESC; WiCell) was
471 grown on Matrigel (Corning) coated plates in mTeSR1 medium (StemCell technologies).
472 Cells were harvested as a single cell suspension using Accutase (EDM Millipore); ATAC-seq
473 and fixation were carried out in mTeSR1 medium. Primary neonatal Human Umbilical Vein
474 Endothelial Cells (HUVEC) were sourced from three suppliers to provide genetic diversity
475 (Lonza, Gibco, PromoCell). HUVECs were expanded in endothelial cell growth medium
476 (Sigma) up to five passages following the manufacturer's protocol. Briefly, HUVECs were
477 grown to 60% confluence, washed with HBSS at room temperature and sub-cultured
478 following light trypsination using Trypsin-EDTA (Sigma) at room temperature and terminating
479 the reaction with trypsin inhibitor (Sigma) upon rounding of the cells and gentle release from
480 the flask. HUVECs were fixed in RPMI supplemented with 10 % FBS. Human Umbilical
481 Derived Erythroid Progenitor line 2 cells[54] (HUDEP-2; RIKEN) were maintained at 0.7-
482 1.5x10^6 cells ml^{-1} in HUDEP expansion media (SFEM, 50 ng/ml SCF, 3 IU/ml EPO, 10 μM
483 DEX, 1% L-Glu, 1% Penstrep) and changed into fresh media containing 2x doxycycline
484 (DOX) every two days. For differentiation we used a modified version of the CD34
485 differentiation protocol. 2-3x10^6 cells were resuspended at 0.3-05x10^6 cells ml^{-1} in HUDEP
486 Phase I media (IMDM, 200 μg/ml Holotransferrin, 10 g/ml Insulin, 3 IU/ml Heparin, 3%
487 Inactivated AB plasma, 2% FBS, 3 IU/ml EPO, 1 ng/ml IL-3, 10 ng/ml SCF, 1x Pen/Strep)
488 with 1x DOX on day 0. On days 1 and 3 cells were counted, pelleted (5 min, 250 rcf, RT)
489 and resuspended to 0.3-0.5x10^6 cells ml^{-1} in fresh HUDEP Phase I media supplemented with
490 2x DOX. On day 5, cells were counted, pelleted and resuspended to 0.5x10^6 cells ml^{-1} in
491 HUDEP Phase II media (IMDM, 500 μg/ml Holotransferrin, 10 g/ml Insulin, 3 IU/ml Heparin,
492 3% Inactivated AB plasma, 2% FBS, 3 IU/ml EPO, 1x Pen/Strep) without DOX. On days 7
493 and 9 cells were counted, pelleted and resuspended to 0.5x10^6 cells ml^{-1} in fresh HUDEP
494 Phase II media. Cytospins and FACS was carried out as for $CD34^+$ differentiation.
495
496 *HUDEP-2 genome editing:* Prior to guide RNA (gRNA) design HUDEP-2 SNPs were
497 genotyped by either Sanger or Next Generation sequencing at MRC WIMM Sequencing

498 Facility with locus specific primers (Supplementary Table 3). For introduction of SNPs by
499 CRISPR/Cas9 facilitated homology dependent repair (HDR), gRNAs were designed to cut in
500 close proximity to the SNP of interest with the PAM overlapping the SNP where possible,
501 additionally single stranded DNA donors (ssODN; IDT) were offset and thioated to promote
502 integration and reduce degradation. To control for global effects on HUDEP-2 cells caused
503 by CRISPR/Cas9 editing, gRNA and ssODN were designed for a homozygous SNP
504 (rs4508712) with no GWAS associations (www.ebi.ac.uk/gwas), which was not within an
505 erythroid regulatory element or expressed gene. HUDEP-2 genotype specific gRNAs
506 (Merck) were cloned into pX458 plasmid backbone[68] by the Genome Engineering Facility
507 (WIMM, University of Oxford) and purified using Plasmid Midi Kit (Qiagen). pX458
508 (pSpCas9(BB)-2A-GFP) was a gift from Feng Zhang (Broad Institute) and is available from
509 Addgene (plasmid #48138). HUDEP-2 cells were then transfected as previously described[55].
510 Briefly, ~$1x10^6$ cells were transfected with pairs of 5 µg gRNA plasmid and 4 µg ssODN
511 (Supplementary Table 3) using Amaxa™ Human CD34 Cell Nucleofector™ Kit (Lonza) in the
512 2B-Nucleofector™ on the U-08 setting. Cells were transferred to 2.5 ml HUDEP expansion
513 media supplemented with 2x DOX and 7.5 µM RAD51-stimulatory compound 1 (RS-1,
514 Sigma). After two days cells were pelleted (5 min, 250 rcf, room temp.) and resuspended in
515 2.5 ml HUDEP expansion media supplemented with 2x DOX with minimal light exposure. On
516 day 3 cells were single cell sorted on a BD FACSAria Fusion flow cytometer (BD Bioscience)
517 into terazaki plates containing 20 µl of expansion media (2x DOX). When colonies reached
518 more than 30 cells they were transferred to a 96-well plate and expanded over two weeks
519 with fresh media and DOX every 2 days until filling two to four wells of a 96-well plate. Half
520 of the cells for each expanded clone were frozen (90% FBS, 10% DMSO) as a stock for
521 recovery post genotyping. For genotyping we followed a 96-well barcoding approach with
522 Next Generation sequencing[69]. Clonally amplified cells were first lysed (50 mM Tris, 1 mM
523 EDTA, 0.5% Tween 20) and the targeted locus was amplified with primers containing a
524 modified m13 adaptor sequence (Supplementary Table 3), the adaptor was then used to for
525 priming with row and column specific primers in a second PCR to barcode each well. Finally,
526 all wells from a single plate were pooled and prepared for sequencing with the NEBNext
527 Ultra II DNA Library Prep kit for Illumina (New England Biolabs). Plates were multiplexed
528 and sequenced on the MiSeq platform (Illumina) using 250 bp paired-end reads (Nano kit,
529 v2 chemistry). Sequences were analysed using platescreen96[69] (v4.0.4,
530 github.com/Hughes-Genome-Group/plateScreen96/releases) to genotype clones. Screening
531 was carried out to exclude for rs9349205 and rs4508712 clones which appeared
532 homozygous due to microhomology driven large deletions (200-4,000 bp)[70,71], rather than
533 through HDR editing, by PCR with locus specific primers (Supplementary Table 3). For

534 rs10758656, two upstream heterozygous SNPs rs7870037 (+129 bp) & rs7855081 (+132
535 bp) allowed for the exclusion of loss of heterozygosity.
536
537 *Gene expression analyses:* 1-5x10^6 cells were fixed in 1 ml TRI-reagent (Sigma), snap
538 frozen and stored at -80ºC for less than one year. RNA was extracted by addition of 0.1 ml
539 1-bromo-3-chloropropane, pipette mixing and separation in a Phase Lock gel Heavy tube
540 (5Prime) and then precipitation with 1 µl of GlycoBlue and an equal volume (~500 µl)
541 isopropanol and centrifugation (10 min, 12,000 rcf, 4 ºC). The RNA pellet was washed with
542 75% ethanol, resuspended in DEPC-treated water, and stored at -80ºC for less than one
543 year. For RT-qPCR RNA was treated with 2U of rDNase I (Invitrogen) and then 1 µg of RNA
544 was used to generate cDNA using SuperScript III First Strand Synthesis SuperMix
545 (Invitrogen) following the manufacturers' instructions. Real-time RT-qPCR was performed on
546 a StepOne Thermocycler (ThermoFisher) using Taqman Universal PCR Master Mix II (Life
547 Tech) and commercially available expression assays (Supplementary Table 5; Life Tech).
548 For RNA-seq total RNA was treated with Turbo DNase (Invitrogen) at 25ºC for 60 min, then
549 RNA was separated using phenol-chloroform isoamylalcohol and a PhaseLock Light-gel
550 tube (5Prime). Treated RNA was precipitated at -80ºC overnight with sodium acetate,
551 glycoblue, and 75% ethanol, before centrifugation (12,000 rcf, 4ºC), 75% ethanol wash and
552 resuspension in DEPC-treated water. Globin and rRNA sequences were depleted from up to
553 5 µg of treated RNA using Globin-Zero Gold (Illumina), before PolyA selection with NEBNext
554 Poly(A) mRNA Magnetic Isolation module (New England Biolabs), and indexing with
555 NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs) following
556 the manufacturers' instructions. RNA-seq libraries were quantified by qPCR (KAPA) prior to
557 sequencing on the NextSeq platform (Illumina) with 39 bp paired-end reads. Reads were
558 mapped to hg19 (Supplementary Table 6) using STAR[72] (v2.4.2a; --outFilterMultimapNmax
559 1) and duplicates were filtered using samtools[73] (v1.3; rmdup). For visualisation directional
560 reads were normalised to RPKM using deepTools[74] with no windowing (v2.2.2;
561 bamCoverage --binSize 1 --normalizeUsingRPKM). Uniquely mapped reads were analysed
562 in DESeq2[61] using variance stabilising transformation and exclusion of genes lacking 5 total
563 reads. Expressed genes were classed as having more than $\log_2$(FPKM) greater than -5.
564
565 *Chromatin conformation capture and target gene identification:* For chromatin conformation,
566 1-2x10^7, H1-hESC, HUVEC or erythroid cells were crosslinked with 2% formaldehyde which
567 provides optimal *cis/trans* ratios and digestion efficiencies[58]. For each cell type triplicate 3C
568 libraries were prepared using *Dpn*II and standard methods[23] with the following modifications:
569 no douncing was performed, all spins were performed at 300 rcf, and after ligation intact
570 nuclei were pelleted (15 min, 300 rcf), supernatant was discarded, and nuclei were

571 resuspended in 300 µl Tris-EDTA (TE; Sigma) for phenol chloroform extraction. Digestion
572 efficiency was determined by RT-qPCR with TaqMan and custom oligonucleotides
573 (Supplementary Table 5), and ligation efficiency qualitatively determined by gel
574 electrophoresis. Only 3C libraries with >70% digestion efficiencies were used. 3C libraries
575 were sonicated to 200 bp in a Covaris S220 and indexed with NEB Next Illumina library Prep
576 reagents (NEB). Enrichment for specific viewpoints was performed with 70mer biotinylated
577 oligonucleotides designed using CapSequm[75] (http://apps.molbiol.ox.ac.uk/CaptureC/cgi-
578 bin/CapSequm.cgi). Double capture was performed in multiplexed reactions with pools of
579 oligonucleotides targeting either promoter proximal (within 5 kb of a transcription start site)
580 or promoter distal (enhancer centric) *DpnII* fragments (Supplementary Table 7) following the
581 described method[23] with each oligonucleotide at a working concentration of 2.9 nM.
582 Captured 3C libraries were sequenced on the NextSeq platform (Illumina) with 150 bp
583 paired-end reads. Reads were mapped and analysed using CCseqBasic5
584 (github.com/Hughes-Genome-Group/CCseqBasic5) as previously described[76] with the
585 following custom settings (--bowtie2 --globin 2). Briefly, CCseqBasic5 trims adaptor
586 sequences, flashes read pairs, *in silco* digests fragments and uses bowtie2 to map reads
587 before identifying capture and reporter reads. After primary analysis replicates were
588 compared using the comprehensive CaptureCompare software (github.com/Hughes-
589 Genome-Group/CaptureCompare). CaptureCompare normalises *cis* reporter counts per
590 100,000 *cis* reporters, generates per fragment mean counts for each cell type, calculates
591 difference in mean interactions between cell types, compares differences in raw interaction
592 counts per fragment using DESeq2[45] as previously described[23,77,78], and provides input for
593 peaky interaction calling[47]. Interaction calling using peaky was run with default settings
594 (omega -3.8) and interactions were filtered based upon the Marginal Posterior Probability of
595 Contact (MPPC) within local interaction domains (MPPC > 0.01) or within 1 Mb of the
596 viewpoint (MPPC > 0.1) and assigned to either Refseq transcription start sites (tss) or
597 variants within 500 bp of the significantly interacting fragment. Target genes were first
598 identified as those having a tss within 5kb of an intersecting variant (promoter proximity),
599 being within 500 bp of a significantly enriched erythroid fragment (FDR <0.05) or within 500
600 bp of a peaky identified interaction. Candidate genes were subsequently filtered for
601 detectable erythroid expression ($\log_2$(FPKM>-5)) using RNA-seq. 24 benchmarking genes
602 most likely to be effectors were identified based on published functional data (*IKZF1, KIT,*
603 *TAL1, RBM38, SMIM1, CD164, CCND3, MYB, HBA1, HBA2, BCL11A, JAK2*)[5,12–14,16,42,79],
604 presence in the Oxford Red Cell Panel for rare inherited anaemia (*KLF1, TFRC, ANK1, HK1,*
605 *SCL4A1*)[80], containing mutations causing hemochromatosis (*TFR2*)[81], having an erythroid
606 eQTL (*ATP2B4*)[15], and causing altered RBC phenotypes in mouse and zebrafish (*FBXO7,*
607 *CCNA2, miR-181a, PIEZO1, AKAP10, CITED2*)[82–88].

608

609 *Chromatin IP, ATAC-seq, and data processing:* For ChIP-seq, chromatin was crosslinked
610 with 1% formaldehyde (Sigma) by the addition of 1 ml 10x crosslinking buffer (50 mM
611 HEPES, 1 mM EDTA, 0.5 mM EGTA, 100 mM NaCl, 10% formaldehyde) to 10^7 cells in 9 ml
612 of media and incubation at room temperature for 10 minutes. Crosslinking was quenched
613 with 130 mM glycine, and cells were washed with cold PBS before snap freezing pelleted
614 cells. Fixed material was stored at -80ºC for less than 12 months. Chromatin
615 immunoprecipitation was performed using Agarose ChIP Assay Kit (Merck Millipore). Briefly,
616 10^7 cells were lysed by incubation on ice with 130 µl lysis buffer for 15 minutes. Lysed cells
617 were transferred to Covaris microtubes and sonicated on the Covaris S220 (Duty cycle: 2%,
618 Intensity: 3, Cycles per burst: 200, Power mode: Frequency sweeping, Duration: 480 sec,
619 Temp.: 6ºC) to generate 200-400 bp fragments. Insoluble material was removed by
620 centrifugation (15,000 rcf, 15 min, 4ºC) and soluble material was diluted to 4 ml with dilution
621 buffer. Immunoprecipitation was performed by incubation of 2 ml diluted chromatin
622 (equivalent to $5x10^6$ input cells) with antibodies for H3K4me1 (3 µg ab195391, lot:
623 GR304893-2; AbCam), H3K4me3 (1 µl 07-473, lot: 2664283; Millipore), H3K27ac (0.3 µg
624 ab4729, lot: GR3205523-1; AbCam), CTCF (10 µl 07-729, lot: 2836929; Millipore) or GATA1
625 (~7.2 µg ab11852, lots: GR208255-9, GR290167-7, GR239235-8; AbCam) overnight.
626 Chromatin binding to Protein A/agarose slurry, washes and elution were performed
627 according to the manufacturer's instructions. DNA was purified by phenol-chloroform
628 extraction with PhaseLock tubes (5Prime) and ethanol precipitation with NaOAc, and 2 µl
629 GlycoBlue (Invitrogen). ChIP enrichment was determined by RT-qPCR (Supplementary
630 Table 5) prior to addition of sequencing adaptors using NEBNext Ultra II DNA Library Prep
631 kit for Illumina (New England Biolabs). ATAC-seq was performed as previously described[76,89]
632 using two technical replicates with $7x10^5$ cells. ChIP-seq and ATAC-seq libraries were
633 quantified by RT-qPCR with the KAPA Library Quantification Complete Kit (KAPA) prior to
634 sequencing on the NextSeq platform (Illumina) with 39 bp paired-end reads. ATAC-seq,
635 DNaseI-seq and ChIP-seq reads were mapped to the hg19 genome using NGseqBasic[90]
636 (V20; --nextera --blacklistFilter --noWindow) which utilises bowtie. Sequence depth and
637 mapped reads for each sample are provided (Supplementary Table 6). Published GEO
638 repositories[24,66,91–98] were downloaded for ATAC-seq and DNaseI-seq from HSC, CMP,
639 MEP, MPP, Ery (GSE75384), and HUVEC (GSM736575, GSM736533), and ChIP-seq for
640 SCL/TAL (GSE95875, GSE93372, GSE42390, GSE70660, GSE59087, GSE52924), GATA1
641 (GSE32491, GSE36985, GSE107726, GSE29196), NF-E2 (GSE95875), BACH1 and MAFK
642 (GSE31477), and SPI1/PU1 (GSE70660), and were analysed by the same method. For
643 visualisation, PCR-duplicate filtered replicates were merged using samtools[73] (v1.3) and

644 converted to bigwigs with minimal smoothing using deepTools[74] (v2.2.2; bamCoverage --
645 binSize 10 --normalizeUsingRPKM --minMappingQuality 30).
646
647 *Imputation and in silico analysis of variants*: The original 75 anaemia index SNPs were
648 imputed with HapMap Phase 2 which is lower resolution than the 1000 Genomes Project
649 Phase 3 dataset[11,99]. Therefore variants in linkage disequilibrium (LD) with index SNPs were
650 identified using the rAggr proxy search online tool (raggr.usc.edu) with default settings
651 ($r^2 \geq 0.8$, distance limit: 500 kb, population panels: All European, All South Asian) for the 1000
652 Genomes Project Phase 3 database[100], which generated 6,420 variants. LD variants for
653 Astle *et al.* (2016) were provided by Lisa Schmunk, Tao Jiang, and Nicole Soranzo
654 (University of Cambridge). Summary statistics for Malaria[101], Multiple Sclerosis[25],
655 Inflammatory Bowel Disease[27], Type 1 Diabetes[26], Type 2 Diabetes[102], Intelligence[28], and
656 Central Corneal Thickness[29] were downloaded from the NHGRI-EBI GWAS Catalog[1]
657 (www.ebi.ac.uk/gwas). For comparison of linkage between index SNPs from van der Harst
658 *et al.* (2012)[11] and Astle *et al.* (2016)[10] we used LDmatrix on the LDlink web tool[103]
659 (http://ldlink.nci.nih.gov/) for European populations as this matched the cohort of the later
660 study. Variants were intersected with peak calls from ATAC-seq or DNaseI-seq for each cell
661 type of interest using bedtools[104]. Enrichment was calculated as the -log(p-value) of a
662 binomial cumulative distribution function b(x; n, p), describing the probability of x or more
663 successes from n Bernoulli trials, with the probability of success for each trial being p. P-
664 values were calculated using the R function pbinom (lower.tail=FALSE) where x was the
665 number of intersecting variants, n was the total number of variants and p was the total
666 number of base-pairs within cell specific peaks divided by the hg19 uniquely mappable
667 base-pairs (2,644,741,479 bp). Variants within the exons and introns of expressed coding
668 genes were tested for predicted damaging effects on coding ANNOVAR[17] or splicing
669 SPIDEX[18] (z-score ≥1.65) and SpliceAI[19] (AI score ≥0.2). Variants within open chromatin
670 were assessed for potential damage to transcription factor binding footprints using
671 Sasquatch[20] (7-mer, WIMM Fibach Erythroid, Exhaustive). Variants within open chromatin
672 were further classified based on their predicted effect on chromatin accessibility using a
673 deep convolutional neural net: deepHaem[21]. Model architecture and data encoding were
674 adapted from DeepSEA[36] with the following modifications: the number of convolutional
675 layers was increased from three to five; batch normalisation was excluded as it did not
676 improve convergence. The network was re-implemented in python using tensorflow (v1.8.0;
677 https://www.tensorflow.org/about/bib). The ENCODE data compendium previously used[36]
678 was supplemented with ATAC-seq and CTCF ChIP-seq data from erythroid differentiations
679 generated for this work, DNAseI-seq[20], and ATAC-seq from sorted progenitor populations[24].

680 Full model details and architecture are available on GitHub
681 (https://github.com/rschwess/deepHaem).
682
683 *Chromatin segmentation and enhancer based PCA analysis*: To ensure identification of all
684 ATAC-seq peaks a combination of the traditional MACS2 approach[105] (v2.0/10 callpeak -B -q
685 0.01) and digital signal processing with Ritornello[106] (v2.0 default settings) was used. Peak
686 summits from both calls were extended to 500 bp and intersected with bedtools[104] (v2.25.0),
687 and filtered for high ploidy regions in MIG viewer[107] to form peak calls for each cell type
688 (Supplementary Tables 8a-p). Chromatin segmentation was performed using the
689 GenoSTAN[31] hidden Markov model (HMM) which allows a more fine-tuned analysis than
690 ChromHMM[108] as it uses continuous rather than binary signal counts. Segmentation used a
691 peak centric approach, rather than signal across the whole genome, with triplicate
692 H3K4me1, H3K4me3, H3K27ac, and CTCF from day 10 of *ex vivo* CD34 differentiation.
693 Read coverage of each mark was calculated (deepTools v2.4.2) for 1 kb windows over open
694 chromatin peaks (bedtools merge -d 10) to capture histone modifications. The HMM model
695 was trained using Poisson log-normal distributions with 20 initial states. These were
696 manually curated to eight final states based on similarity of chromatin signature. For
697 Principle Component Analysis (PCA) trajectory plotting combined peak calls from sorted
698 hematopoietic populations covering 176,135 open chromatin regions not within 2 kb of
699 transcription start sites were first used to generate a PCA map of erythroid differentiation
700 from sorted populations of HSC, MPP, CMP, MEP and Erythroid populations[24]. Reads within
701 peaks were normalised (R scale) and the PCA was calculated using the R function prcomp.
702 The read counts from *ex vivo* differentiated cells within the same peak set were then used to
703 calculate sample mapping onto PC1 and PC2, and thus to map differentiation timepoints
704 onto the differentiation trajectory. Heatmaps of intersected peaks were generated with
705 pheatmap[109] (v1.0.8) using z-normalised counts of reads per basepair from all identified
706 peaks. For enhancer activity, peak calls were extended by 250 bp in both directions
707 (bedtools slop) to account for the spreading nature of H3K27ac ChIP-seq signal, enhancers
708 were then ranked based on reads per base pair. To determine the point of inflection between
709 low and high acting enhancers H3K27ac read counts were transformed so that the highest
710 value equalled the number of ranked peaks. The point of inflection where the gradient of the
711 curve became greater than one was used to define low and high enhancer activity. The
712 gradient was calculated for each peak as a window based on the local linear gradient of
713 ±200 peaks.
714
715
716

717 **DECLARATIONS**

718 *Ethics approval and consent to participate:* All donors provided consent to participate in this
719 study. Blood was collected with ethics approval by the NHS Health Research Authority
720 (National Research Ethics Service, REC 03/08/097) and stored according to Human Tissue
721 Authority guidelines (License 12433).

722

723 *Consent for publication:* Not applicable

724

725 *Availability of data and materials:* Sequence reads and processed data generated for this
726 work have been archived with GEO; H1-hESC ATAC-seq (GSE124853), Immune cell ATAC-
727 seq (GSE125164), HUDEP-2 data (GSE125715), primary human erythroid data
728 (GSE125926) and processed data are available via a UCSC hub
729 (http://sara.molbiol.ox.ac.uk/public/hugheslab/GWAS_Pipeline/hub.txt).

730

731 *Competing Interests:* J.R.H and J.O.J.D. are founders and shareholders of Nucleome
732 Therapeutics.

733

734 *Funding:* This work was carried out as part of the WIGWAM Consortium (Wellcome
735 Investigation of Genome Wide Association Mechanisms) funded by a Wellcome Trust
736 Strategic Award (106130/Z/14/Z). This work was also supported by Medical Research
737 Council (MRC) Core Funding (MC_UU_12009). D.J.D. received funding from the Oxford
738 University Medical Science Internal Fund: Pump Priming (0006152). Wellcome Trust
739 Doctoral Programmes supported R.S. (203728/Z/16/Z), C.Q.E. (203141/Z/16/Z), M.C.S
740 (097309/Z/11/Z), A.M.O. (105281/Z/14/Z), and R.C.B. (203141/Z/16/Z). C.S. was supported
741 by the Congenital Anaemia Network (CAN). A.M.O was supported by the Stevenson Junior
742 Research Fellowship (University College, Oxford). J.O.J.D. is funded by an MRC Clinician
743 Scientist Award (MR/R008108) and received Wellcome Trust Support (098931/Z/12/Z). G.L.
744 is supported by the Wellcome Trust (090532/Z/09/Z) and MRC Strategic Alliance Funding
745 (MC_UU_12025).

746

747 *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
748 J.R.H. designed and planned experiments. D.J.D., S.J.H., L.N., C.S., M.E.G., P.P.H.,
749 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
750 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
751 analysed data. R.K., and Y.N. provided essential reagents. D.J.D., J.A.T., D.R.H, J.R.H.
752 wrote the manuscript.

753

754 *Acknowledgements:* We appreciate the time and support of Kevin Clarke, Sally-Anne Clarke
755 and Paul Sopp (WIMM Flow Cytometry Facility, Oxford University) during cell sorting.

756 **REFERENCES**

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

793 variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
794 18. Xiong, H. Y. *et al.* The human splicing code reveals new insights into the genetic
795 determinants of disease. *Science (80-.).* **347**, 1254806 (2015).
796 19. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning.
797 *Cell* **176**, 535-548.e24 (2019).
798 20. Schwessinger, R. *et al.* Sasquatch : predicting the impact of regulatory SNPs on
799 transcription factor binding from cell- and tissue-specific DNase footprints. *Genome*
800 *Res.* **27**, 1730–1742 (2017).
801 21. Schwessinger, R. *et al.* DeepC : Predicting chromatin interactions using megabase
802 scaled deep neural networks and transfer learning. *bioRxiv* **724005**, (2019).
803 22. Davies, J. O. J., Oudelaar, A. M., Higgs, D. R. & Hughes, J. R. How best to identify
804 chromosomal interactions: A comparison of approaches. *Nat. Methods* **14**, 125–134
805 (2017).
806 23. Davies, J. O. J. *et al.* Multiplexed analysis of chromosome conformation at vastly
807 improved sensitivity. *Nat. Methods* **13**, 74–80 (2016).
808 24. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts
809 human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
810 25. Beecham, A. H. *et al.* Analysis of immune-related loci identifies 48 new susceptibility
811 variants for multiple sclerosis. *Nat. Genet.* **45**, 1353–1360 (2013).
812 26. Onengut-gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and
813 evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat.*
814 *Genet.* **47**, 381–386 (2015).
815 27. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory
816 bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**,
817 979–986 (2015).
818 28. Sniekers, S. *et al.* Genome-wide association meta-analysis of 78,308 individuals
819 identifies new loci and genes influencing human intelligence. *Nat. Genet.* **49**, 1107–
820 1112 (2017).
821 29. Iglesias, A. I. *et al.* Cross-ancestry genome-wide association analysis of corneal
822 thickness strengthens link between complex and Mendelian eye diseases. *Nat.*
823 *Commun.* **9**, 1–11 (2018).
824 30. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic
825 architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
826 31. Zacher, B. *et al.* Accurate promoter and enhancer identification in 127 ENCODE and
827 roadmap epigenomics cell types and tissues by GenoSTAN. *PLoS One* **12**, 1–25
828 (2017).
829 32. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated

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

867 50. Hait, T. A., Amar, D., Shamir, R. & Elkon, R. FOCS : a novel method for analyzing
868 enhancer and gene activity patterns infers an extensive enhancer – promoter map.
869 *Genome Biol.* **19**, 1–14 (2018).
870 51. Kumasaka, N., Knights, A. J. & Gaffney, D. J. High-resolution genetic mapping of
871 putative causal interactions between regions of open chromatin. *Nat. Genet.* **51**,
872 (2019).
873 52. Kan, Y. W. *et al.* Molecular basis of hemoglobin-H disease in the Mediterranean
874 population. *Blood* **54**, 1434 LP – 1438 (1979).
875 53. Tolosano, E. & Chiabrando, D. Diamond Blackfan anemia at the crossroad between
876 ribosome biogenesis and heme metabolism. *Adv. Hematol.* **2010**, (2010).
877 54. Kurita, R. *et al.* Establishment of Immortalized Human Erythroid Progenitor Cell Lines
878 Able to Produce Enucleated Red Blood Cells. *PLoS One* **8**, (2013).
879 55. Moir-Meyer, G. *et al.* Robust CRISPR/Cas9 Genome Editing of the HUDEP-2
880 Erythroid Precursor Line Using Plasmids and Single-Stranded Oligonucleotide
881 Donors. *Methods Protoc.* **1**, 28 (2018).
882 56. Vinjamur, D. S. & Bauer, D. E. Growing and Genetically Manipulating Human
883 Umbilical Cord Blood-Derived Erythroid Progenitor (HUDEP) Cell Lines. in
884 *Erythropoiesis: Methods and Protocols* (ed. Lloyd, J. A.) 275–284 (Springer New
885 York, 2018). doi:10.1007/978-1-4939-7428-3_17
886 57. Bittorf, T., Jaster, R., Lüdtke, B., Kamper, B. & Brock, J. Requirement for JAK2 in
887 erythropoietin-induced signalling pathways. *Cell. Signal.* **9**, 85–89 (1997).
888 58. Oudelaar, A. M., Davies, J. O. J., Downes, D. J., Higgs, D. R. & Hughes, J. R. Robust
889 detection of chromosomal interactions from small numbers of cells using low-input
890 Capture-C. *Nucleic Acids Res.* **45**, (2017).
891 59. Geeven, G., Teunissen, H., Laat, W. De & Wit, E. De. peakC : a flexible , non-
892 parametric peak calling package for 4C and Capture-C data. **46**, (2018).
893 60. Klein, F. A. *et al.* FourCSeq: Analysis of 4C sequencing data. *Bioinformatics* **31**,
894 3085–3091 (2015).
895 61. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and
896 dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).
897 62. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and
898 Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
899 63. Spain, S. L. & Barrett, J. C. Strategies for fi ne-mapping complex traits. *Human* **24**,
900 111–119 (2015).
901 64. Liu, Y. *et al.* Identification of breast cancer associated variants that modulate
902 transcription factor binding. *PLoS Genet.* 1–21 (2017).
903 65. Allen, E. K. *et al.* SNP-mediated disruption of CTCF binding at the IFITM3 promoter is

904 associated with risk of severe influenza in humans. *Nat. Med.* **23**, 975–983 (2017).
905 66. Gobbi, M. De *et al.* A Regulatory SNP Causes a Human Transcriptional Promoter.
906 1215–1218 (2006).
907 67. Trakarnsanga, K. *et al.* An immortalized adult human erythroid line facilitates
908 sustainable and scalable generation of functional red cells. *Nat. Commun.* **8**, 1–7
909 (2017).
910 68. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**,
911 2281–2308 (2013).
912 69. Nussbaum, L. *et al.* High-Throughput Genotyping of CRISPR/Cas Edited Cells in 96-
913 Well Plates. *Methods Protoc.* **1**, 29 (2018).
914 70. Kosicki, M., Tomberg, K. & Bradley, A. Repair of CRISPR–Cas9-induced double-
915 stranded breaks leads to large deletions and complex rearrangements. *Nat.*
916 *Biotechnol.* **36**, 765–771 (2018).
917 71. Owens, D. D. G. *et al.* Microhomologies are prevalent at Cas9-induced larger
918 deletions. *Nucleic Acids Res.* **47**, 7402–7417 (2019).
919 72. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21
920 (2013).
921 73. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
922 2078–2079 (2009).
923 74. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing
924 data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
925 75. Hughes, J. R. *et al.* Analysis of hundreds of cis-regulatory landscapes at high
926 resolution in a single, high-throughput experiment. *Nat. Genet.* **46**, 205–12 (2014).
927 76. Hay, D. *et al.* Genetic dissection of the α-globin super-enhancer in vivo. *Nat. Genet.*
928 1–12 (2016). doi:10.1038/ng.3605
929 77. Hanssen, L. L. P. *et al.* Tissue-specific CTCF – cohesin-mediated chromatin
930 architecture delimits enhancer interactions and function in vivo. *Nat. Cell Biol.* **19**,
931 952–961 (2017).
932 78. Simon, C. S. *et al.* Functional characterisation of cis -regulatory elements governing
933 dynamic Eomes expression in the early mouse embryo. *Development* **144**, 1249–
934 1260 (2017).
935 79. Sankaran, V. G. *et al.* Human Fetal Hemoglobin Expression Is Regulated by the
936 Developmental Stage-Specific Repressor BCL11A. *Science (80-.).* **322**, 1839–1842
937 (2008).
938 80. Roy, N. B. A. *et al.* A novel 33-Gene targeted resequencing panel provides accurate,
939 clinical-grade diagnosis and improves patient management for rare inherited
940 anaemias. *Br. J. Haematol.* **175**, 318–330 (2016).

941 81. Camaschella, C. *et al.* The gene TFR2 is mutated in a new type of haemochromatosis
942 mapping to 7q22. *Nat. Genet.* **25**, 14–15 (2000).
943 82. Ludwig, L. S. *et al.* Genome-wide association study follow-up identifies cyclin A2 as a
944 regulator of the transition through cytokinesis during terminal erythropoiesis. *Am. J.*
945 *Hematol.* **90**, 386–391 (2015).
946 83. Jayapal, S. R. *et al.* Cyclin A2 regulates erythrocyte morphology and numbers. *Cell*
947 *Cycle* **15**, 3070–3081 (2016).
948 84. Bielczyk-Maczyńska, E. *et al.* A Loss of Function Screen of Identified Genome-Wide
949 Association Study Loci Reveals New Genes Controlling Hematopoiesis. *PLoS Genet.*
950 **10**, (2014).
951 85. Chen, F. *et al.* High-frequency genome editing using ssDNA oligonucleotides with
952 zinc-finger nucleases. *Nat. Methods* **8**, 753–757 (2011).
953 86. Randle, S. J., Nelson, D. E., Patel, S. P. & Laman, H. Defective erythropoiesis in a
954 mouse model of reduced Fbxo7 expression due to decreased p27 expression. *J.*
955 *Pathol.* **237**, 263–272 (2015).
956 87. Cahalan, S. M. *et al.* Piezo1 links mechanical forces to red blood cell volume. *Elife* **4**,
957 1–12 (2015).
958 88. Figueroa, A. A. *et al.* miR-181a regulates erythroid enucleation via the regulation of
959 Xpo7 expression. *Haematologica* **103**, e341–e344 (2018).
960 89. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J.
961 Transposition of native chromatin for fast and sensitive epigenomic profiling of open
962 chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–8
963 (2013).
964 90. Telenius, J. M. & Hughes, J. R. NGseqBasic - a single-command UNIX tool for ATAC-
965 seq, DNaseI-seq, Cut-and-Run, and ChIP-seq data mapping, high-resolution
966 visualisation, and quality control. *bioRxiv* 393413 (2018). doi:10.1101/393413
967 91. Canver, M. C. *et al.* Variant-aware saturating mutagenesis using multiple Cas9
968 nucleases identifies regulatory elements at trait-associated loci. *Nat. Genet.* **49**, 625–
969 634 (2017).
970 92. ENCODE. An integrated encyclopedia of DNA elements in the human genome.
971 *Nature* **489**, 57–74 (2012).
972 93. Huang, J. *et al.* Dynamic Control of Enhancer Repertoires Drives Lineage and Stage-
973 Specific Transcription during Hematopoiesis. *Dev. Cell* **36**, 9–23 (2016).
974 94. Xu, J. *et al.* Developmental control of Polycomb subunit composition by GATA factors
975 mediates a switch to non-canonical functions Jian. *Mol. Cell* **57**, 304–316 (2015).
976 95. Pinello, L., Xu, J., Orkin, S. H. & Yuan, G. Analysis of chromatin-state plasticity
977 identifies cell-type – specific regulators of H3K27me3 patterns. *Proc. Natl. Acad. Sci.*

978 344-E353 (2014). doi:10.1073/pnas.1322570111
979 96. Kang, Y. *et al.* Autophagy Driven by a Master Regulator of Hematopoiesis. *Mol. Cell.*
980 *Biol.* 226–239 (2012). doi:10.1128/MCB.06166-11
981 97. Trompouki, E. *et al.* Lineage Regulators Direct BMP and Wnt Pathways to Cell-
982 Specific Programs during Differentiation and Regeneration. *Cell* **147**, 577–589 (2011).
983 98. Mciver, S. C. *et al.* The exosome complex establishes a barricade to erythroid
984 maturation. *Blood* **124**, 2285–2298 (2019).
985 99. Vries, P. S. De *et al.* Comparison of HapMap and 1000 Genomes Reference Panels
986 in a Large-Scale Genome- Wide Association Study. *PLoS One* 1–22 (2017).
987 doi:10.1371/journal.pone.0167742
988 100. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74
989 (2015).
990 101. Ravenhall, M. *et al.* Novel genetic polymorphisms associated with severe malaria and
991 under selective pressure in North-eastern Tanzania. *PLoS Genet.* **14**, e1007172
992 (2018).
993 102. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47
994 (2016).
995 103. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring
996 population-specific haplotype structure and linking correlated alleles of possible
997 functional variants. *Bioinformatics* **31**, 3555–3557 (2015).
998 104. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing
999 genomic features. *Bioinformatics* **26**, 841–842 (2010).
1000 105. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using
1001 MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).
1002 106. Stanton, K. P., Jin, J., Lederman, R. R., Weissman, S. M. & Kluger, Y. Ritornello: high
1003 fidelity control-free chromatin immunoprecipitation peak calling. *Nucleic Acids Res.*
1004 **45**, (2017).
1005 107. McGowan, S. J., Hughes, J. R., Han, Z. P. & Taylor, S. MIG: Multi-Image Genome
1006 viewer. *Bioinformatics* **29**, 2477–2478 (2013).
1007 108. Ernst, J. & Kellis, M. ChromHMM: Automating chromatin-state discovery and
1008 characterization. *Nat. Methods* **9**, 215–216 (2012).
1009 109. Kolde, R. Pheatmap: pretty heatmaps. (2012).
1010

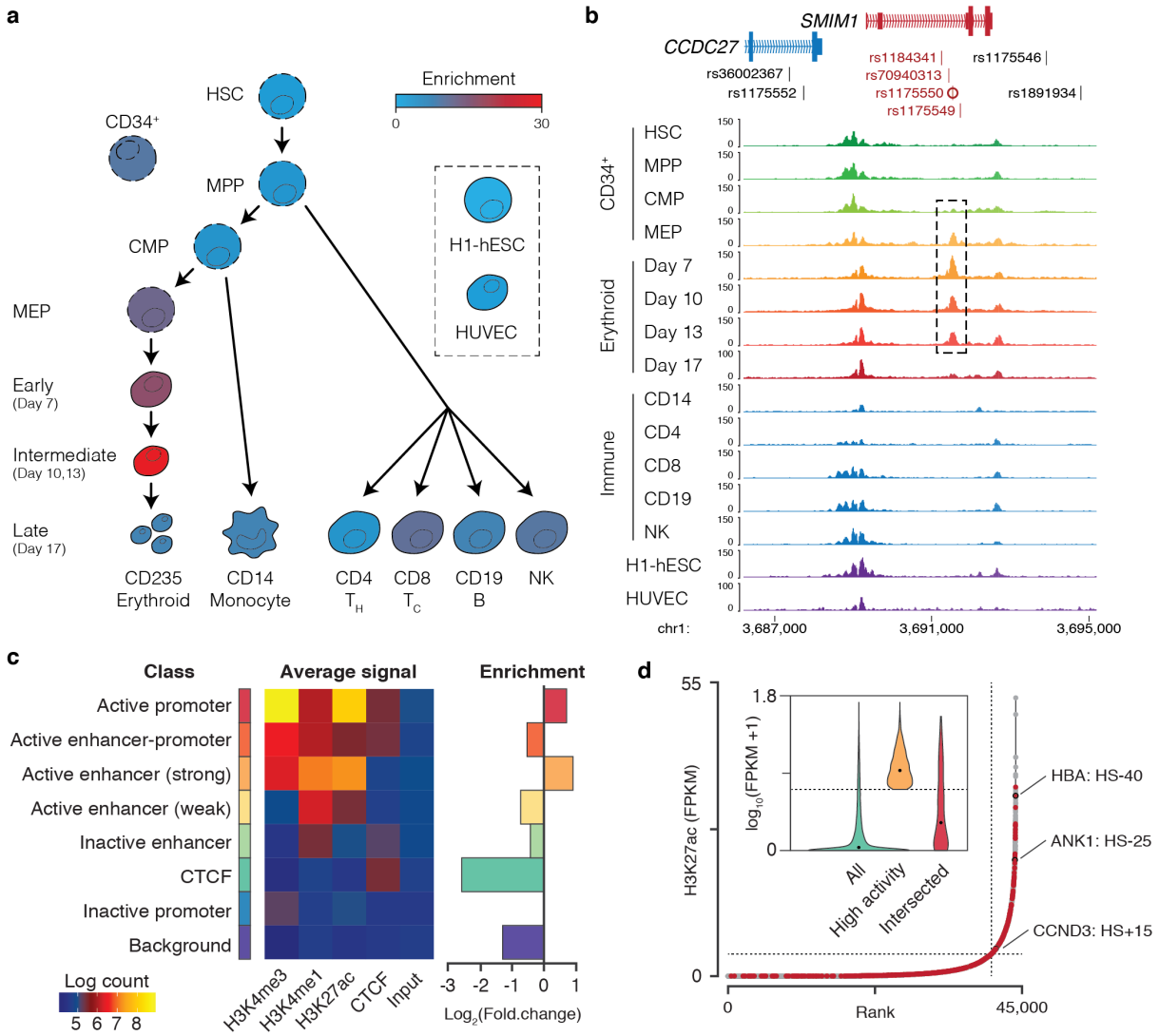


Fig. 1) Variants associated with RBC traits lie within highly active enhancers. a, Schematic of selected cells from human haematopoiesis showing enrichment (-log(p) of a cumulative Binomial Distribution) for RBC associated variants within open chromatin regions of haematopoietic stem cells (HSC), multi-potent 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).

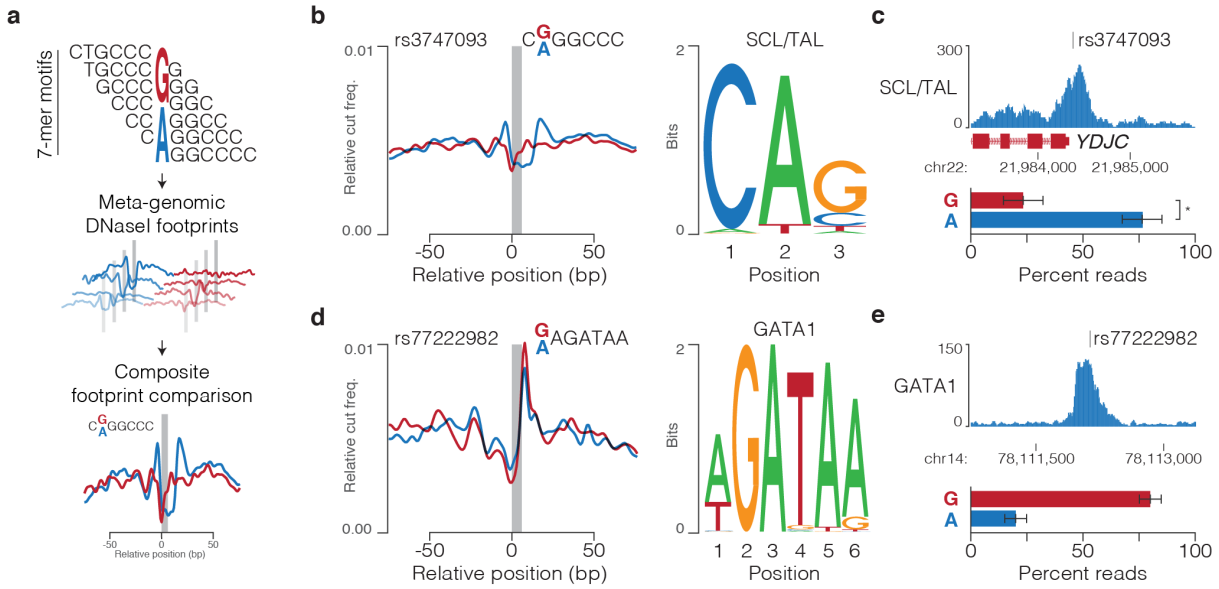


Fig. 2) Sasquatch provides an unbiased prediction of variant effect. a, Sasquatch analyses *in vivo* generated DNaseI footprints over 7-mer motifs within open chromatin regions to generate meta-genomic footprints. Comparison of Relative cut frequency for each profile is used to generate predictive footprint-change 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.

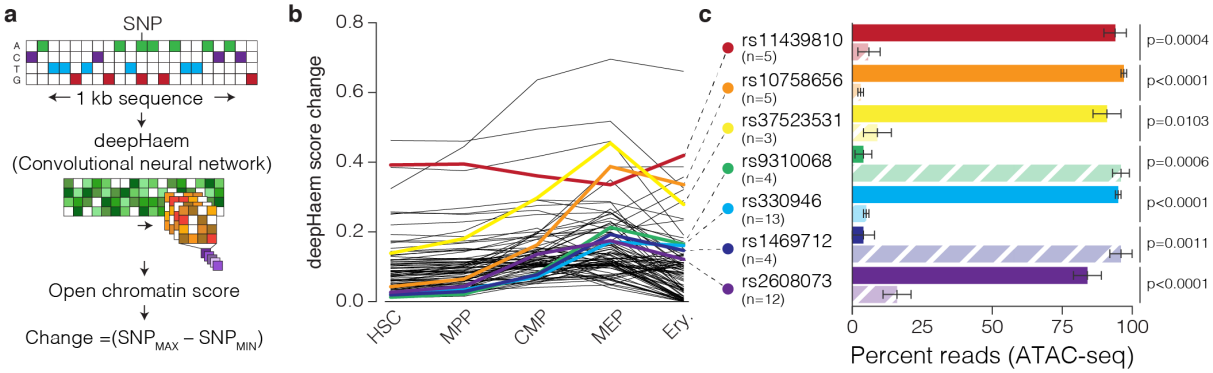


Fig. 3) Deep Learning predicts variant driven changes in chromatin accessibility. a, deepHaem, a deep convolutional neural network, calculates a chromatin openness score using 1 kb of DNA sequence which can be used to compare variant alleles. **b**, Comparison of alleles for all RBC trait variants in 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.

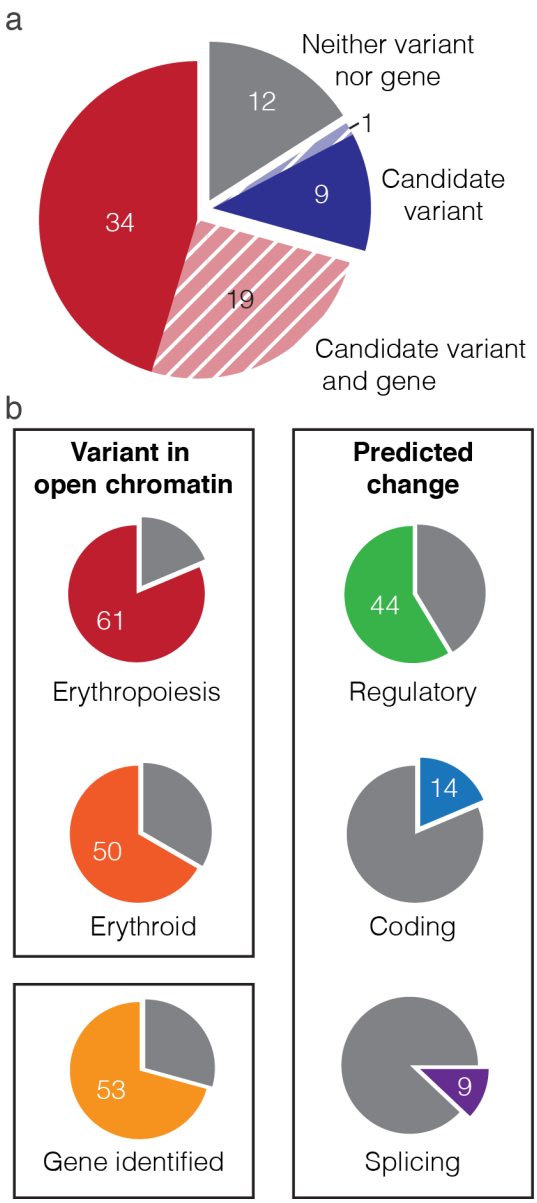


Fig. 4) The integrated experimental and bioinformatics platform identifies candidate causal variants and effector genes at the majority of polygenic trait regions. a, Cumulative analysis of RBC trait associated variants at 75 GWAS chromosome regions identified candidate causal variants in 63 regions 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. 16 regions have candidate causal variants with separate mechanisms of action, five of which have regulatory, splicing and coding candidates.

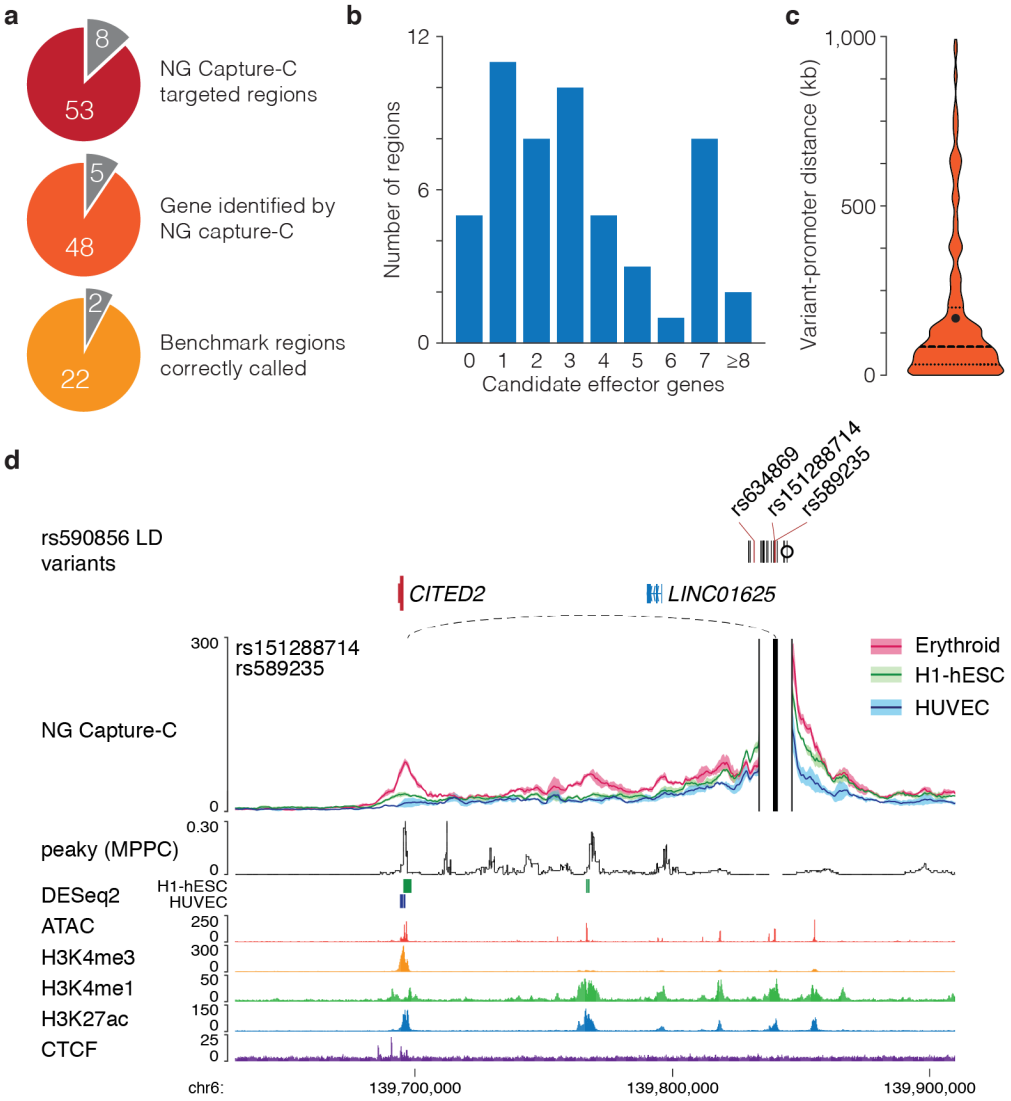


Fig. 5) NG Capture-C can detect long range variant-promoter interactions. a, NG Capture-C oligonucleotides were designed for 61 chromosome regions with candidate regulatory causal variants, regions were excluded from targeting based on sheer number of targets (>100 target sites at a single locus, n=1), 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 with *CITED2*, which encodes the Cbp/p300 Interacting Transactivator with Glu/Asp (E/D)-rich tail 2 protein and required for normal haematopoiesis.

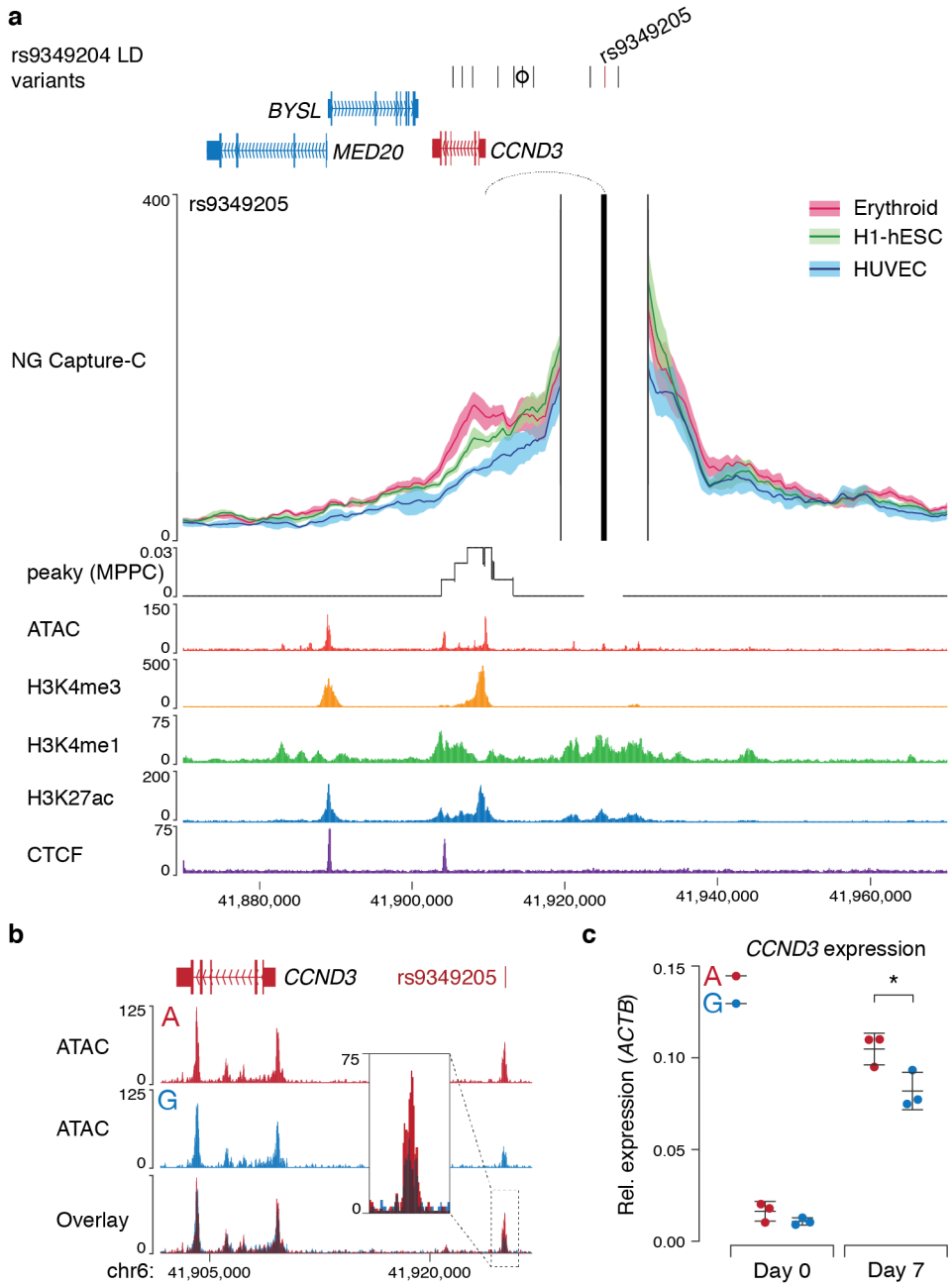


Fig. 6) rs9349205 interacts with, and regulates CCND3. a, 3C interaction profile for rs9349205 in erythroid, embryonic stem (H1-hESC) and umbilical vein endothelial (HUVEC) cells (n=3). Profiles show windowed mean interactions (solid lines) with one standard deviation (shading). Peaky values depict the 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).

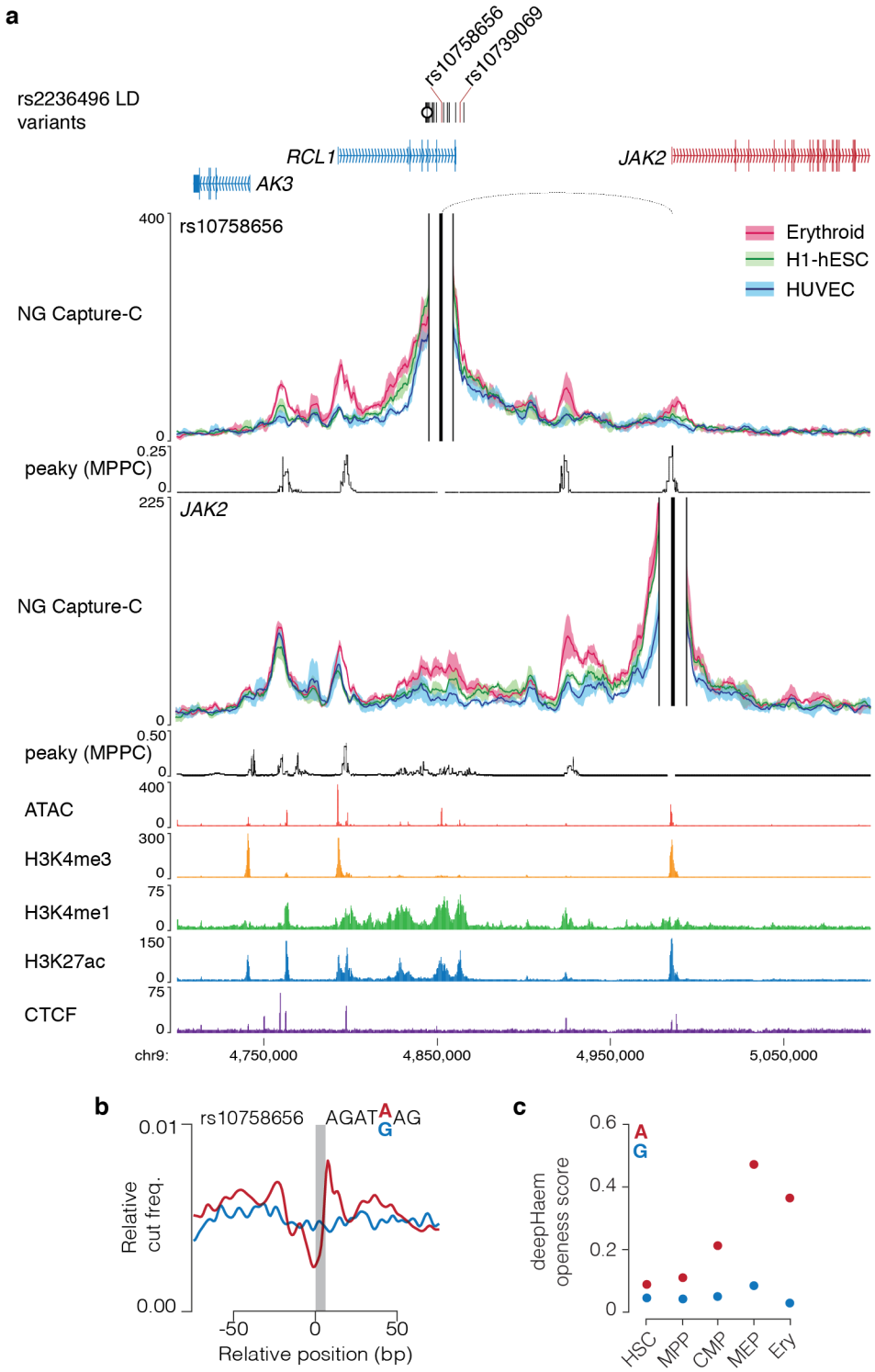


Fig. 7) rs10758656 interacts with *JAK2* and is predicted to alter chromatin accessibility. a, 3C interaction profiles for rs10758656 and *JAK2* in erythroid, embryonic stem (H1-hESC) and umbilical vein endothelial (HUVEC) cells (n=3). Profiles show windowed mean interactions (solid lines) with one standard 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.

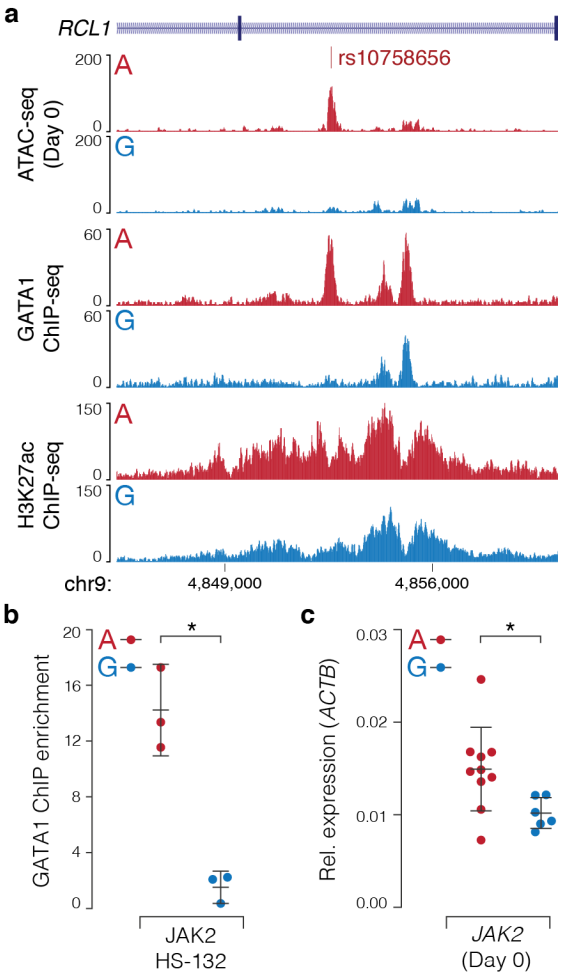
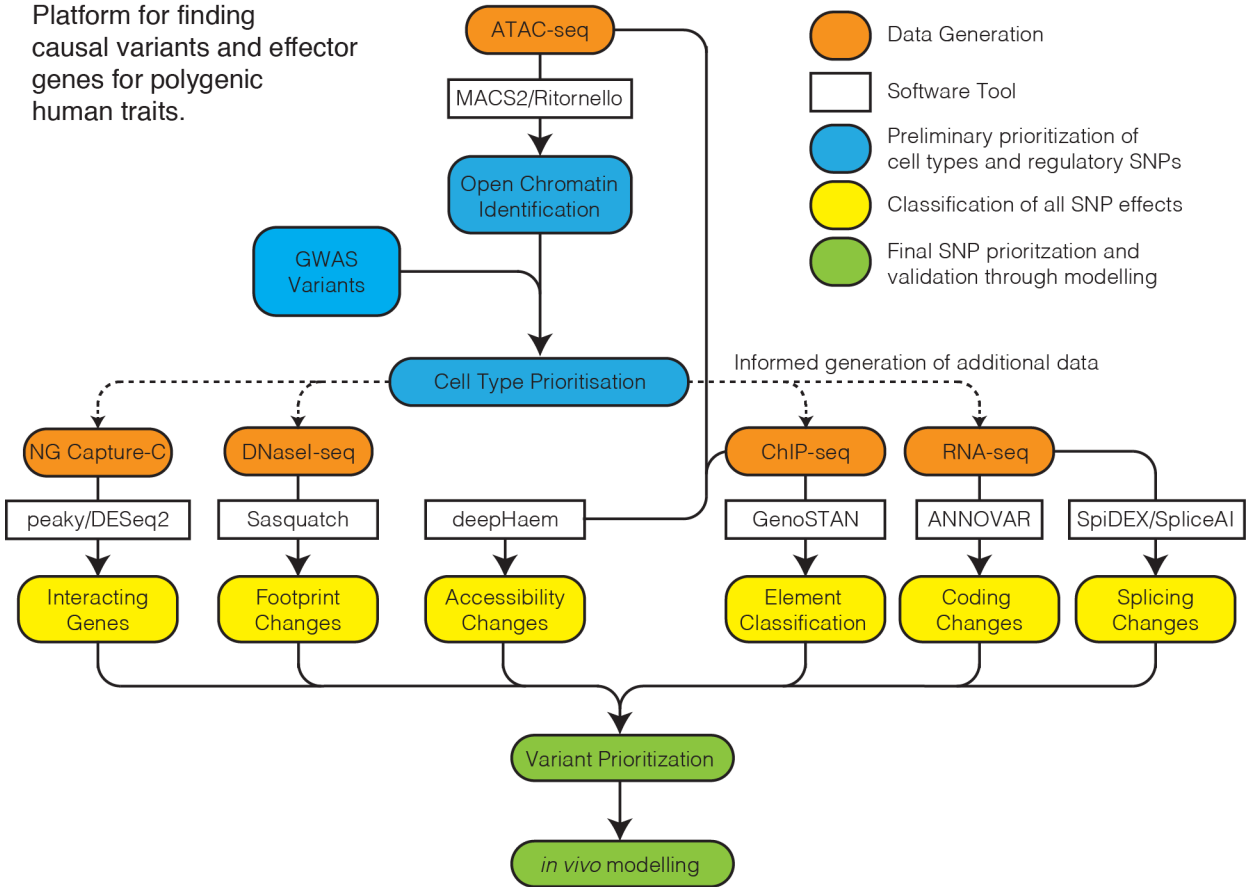


Fig. 8) rs10758656 causes loss of open chromatin and reduced *JAK2* expression. a, FPKM normalised ATAC-seq (n=3), GATA1 ChIP-seq (n=3) and H3K27ac ChIP-seq (n≥1) from differentiating HUDEP-2 clones homozygous for either rs10758656 allele. **b,** Real time quantitative PCR for GATA1 ChIP at rs10758656 (*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). Each circle reflects an independently isolated clonal population, bars show mean and one standard deviation.

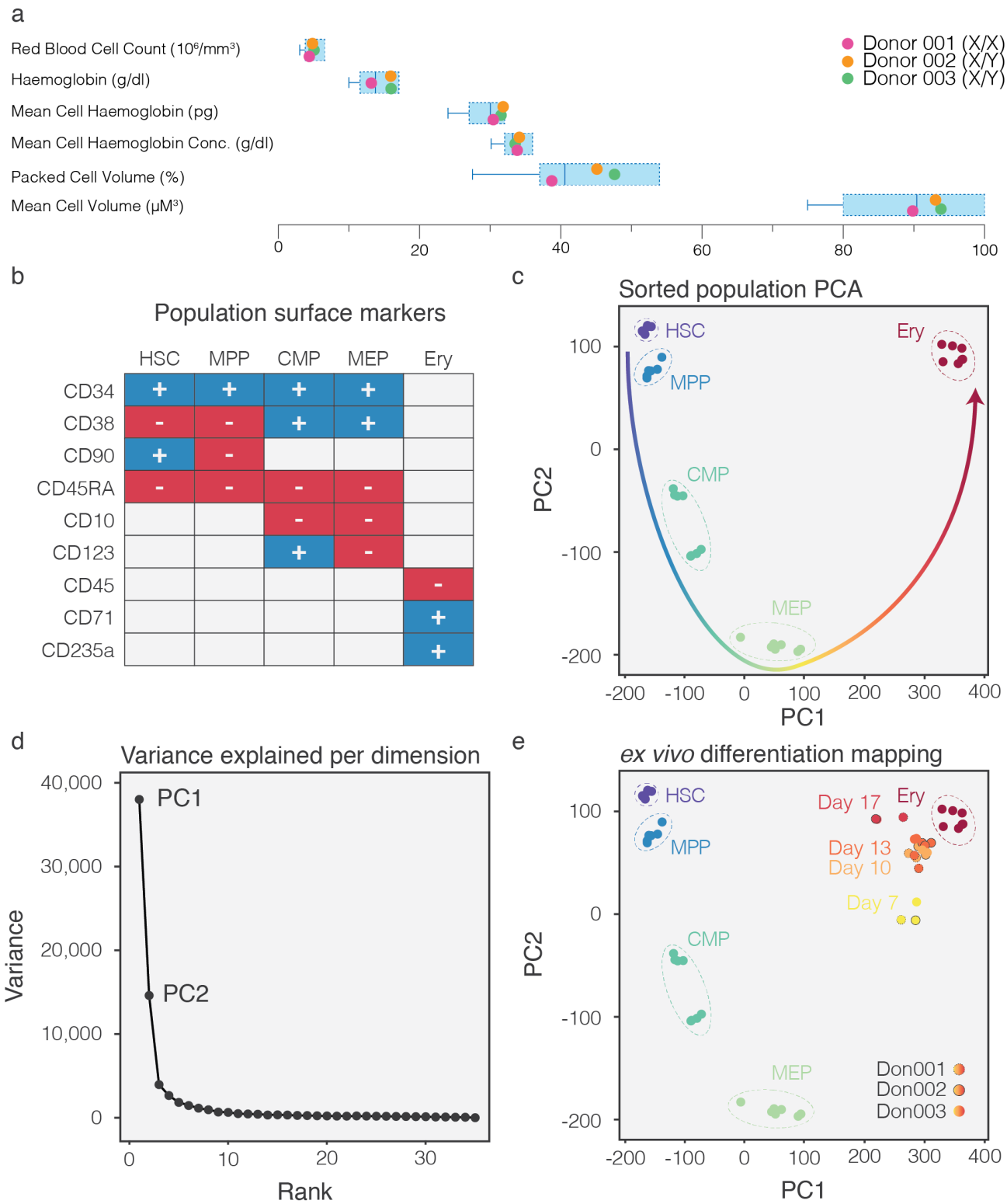
Supplementary Information for: An integrated platform to systematically find causal variants and genes for polygenic human traits.

Contents

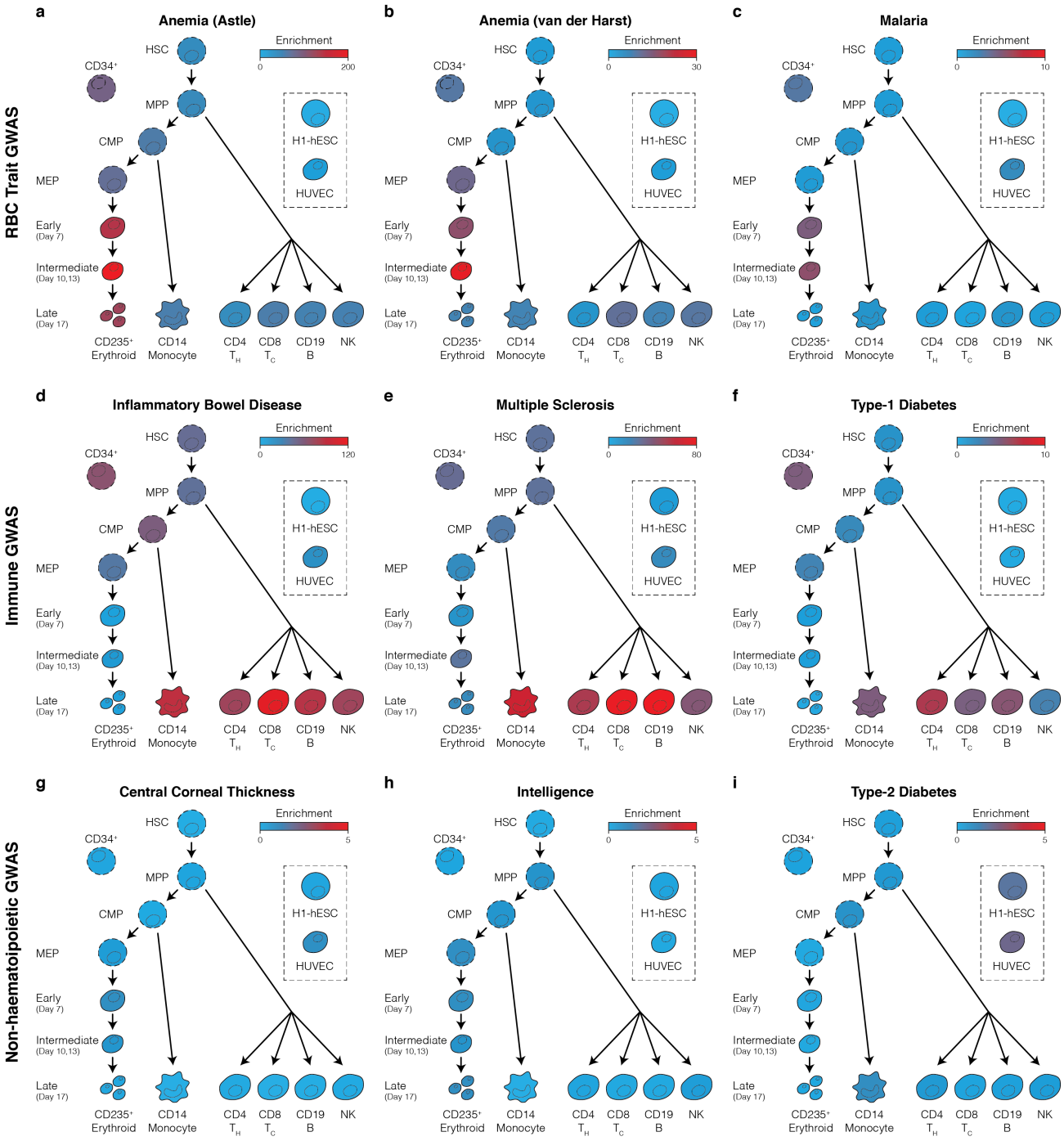
Supplementary Figures
Supplementary Table Legends



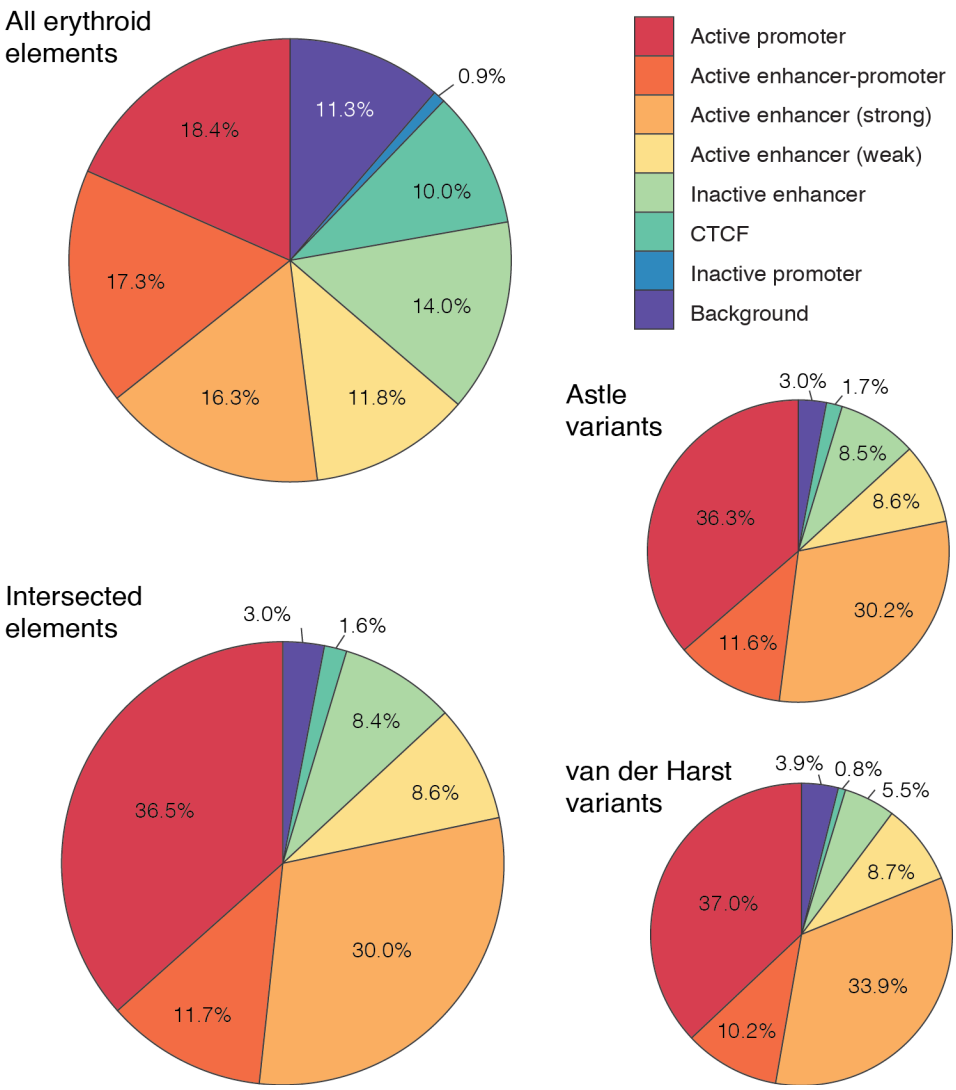
Supplementary Fig. 1) Flowchart of an integrated platform for dissecting multigenic traits using a multi-omics and machine learning approach. ATAC-seq is generated for multiple candidate cell-types and open chromatin regions are identified by MACS2 and Ritornello peak calling. Open chromatin regions for all cell types are intersected with GWAS variants to identify the key cell type(s) for further study. Additional -omics data is then generated in the key cell type(s) to characterise open chromatin regions (ChIP-seq), transcription factor footprints (DNaseI-seq), chromatin interactions (NG Capture-C) and expressed genes (RNA-seq). Note, Omics data is integrated through a suite of traditional and machine learning software (GenoSTAN, Sasquatch, deepHaem, peaky, DESeq2, ANNOVAR, SpiDEX, SpliceAI) to determine candidate causative variants and effector genes. Sasquatch currently has 314 available datasets for analysis, so DNaseI-seq would only need to be generated if the relevant cell type was not available. This wholistic approach allows the prioritization of variants for which can then be tested functionally by *in vivo* modelling.



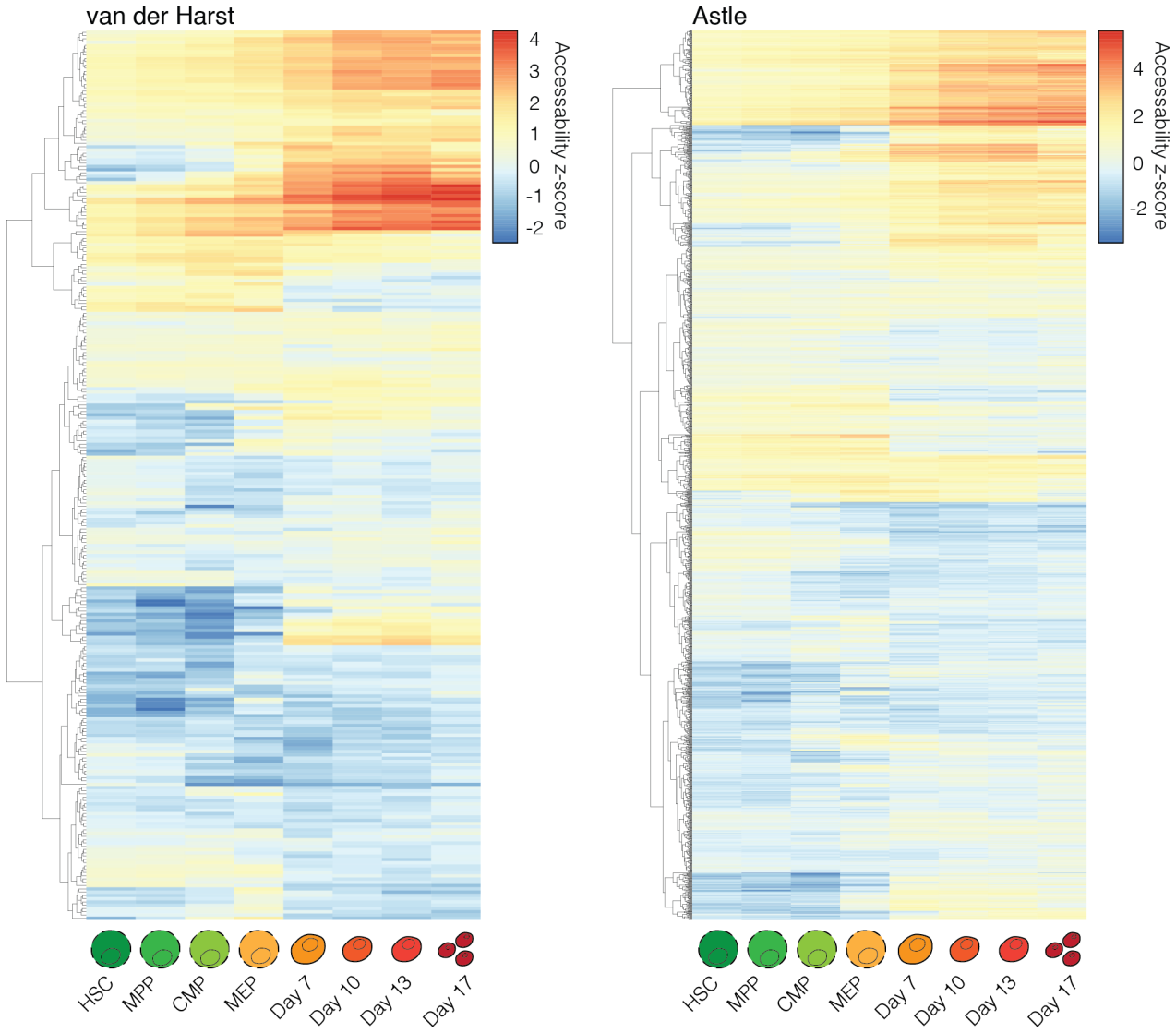
Supplementary Fig. 2) Differentiation of CD34⁺ cells from healthy donors. a, Whole cell counts for three healthy donors were compared against Oxford University Hospital, NHS Foundation Trust Normal ranges for adults (boxes) and GWAS exclusion values (whiskers) prior to differentiation. **b,** Cell surface markers used to define populations from published sorted cell population ATAC-seq (Corces *et al.* 2016). **c,** Principal Component Analysis (PCA) of ATAC-seq peaks outside 2 kb of transcription start sites maps the erythropoietic differentiation trajectory using the first two principal components (PC) which account for the majority of the variance (**d**). **e,** ATAC-seq from *ex vivo* CD34+ differentiation can be used to map progress through erythropoiesis by calculating PC1 and PC2 and shows high reproducibility between donors and replicate differentiations.



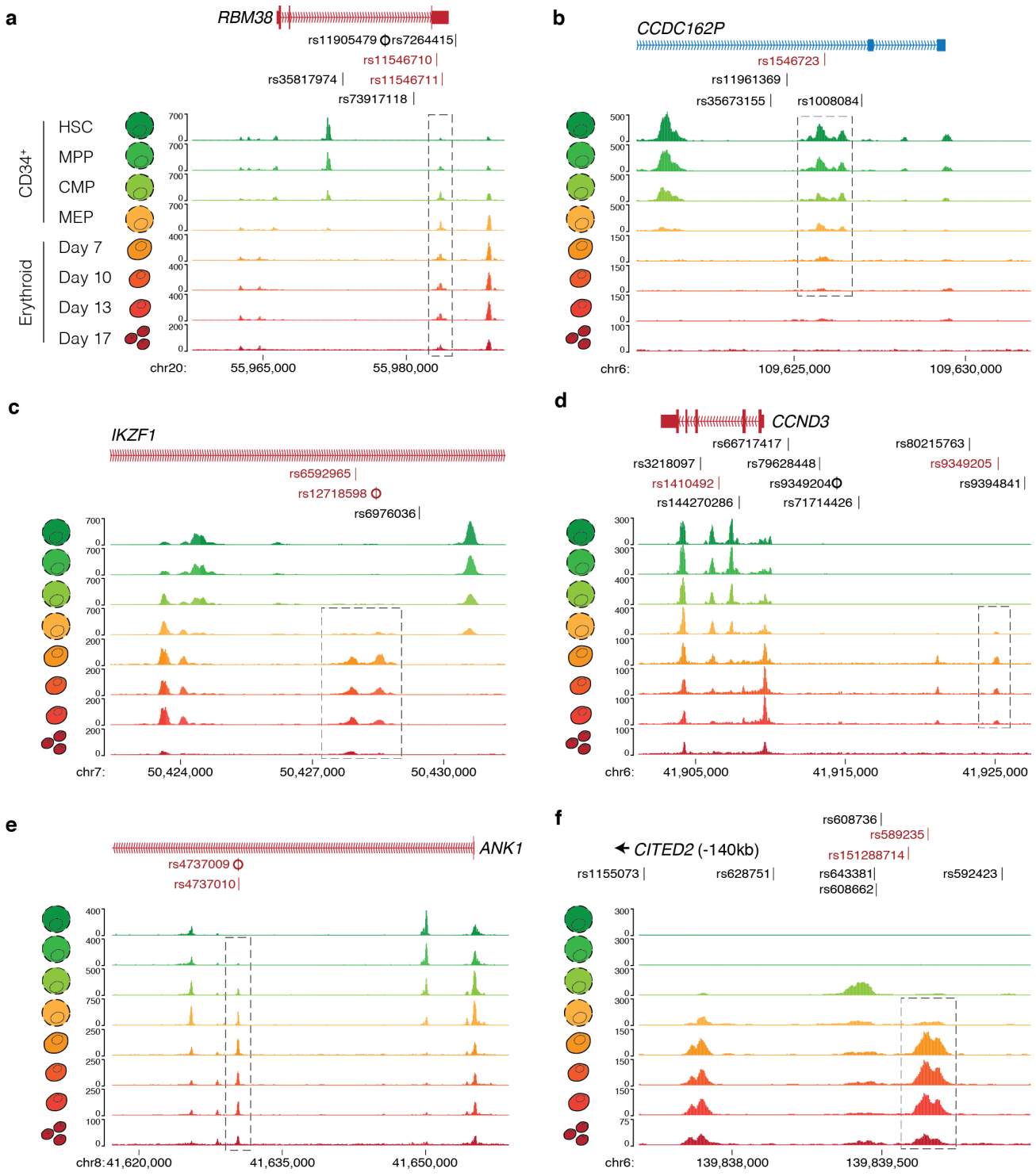
Supplementary Fig. 3) ATAC-seq peak intersections show disease specific profiles. Schematic of human haematopoiesis with enrichment (-log(p) of a cumulative Binomial Distribution) for red blood cell (**a-c**), monocytes and lymphocytes (**d-f**) and non-haematological (**g-i**) GWAS associated variants within open chromatin of haematopoietic stem cells (HSC), multi-potent progenitors (MPP), common myeloid progenitors (CMP), megakaryocyte-erythroid progenitors (MEP), 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).



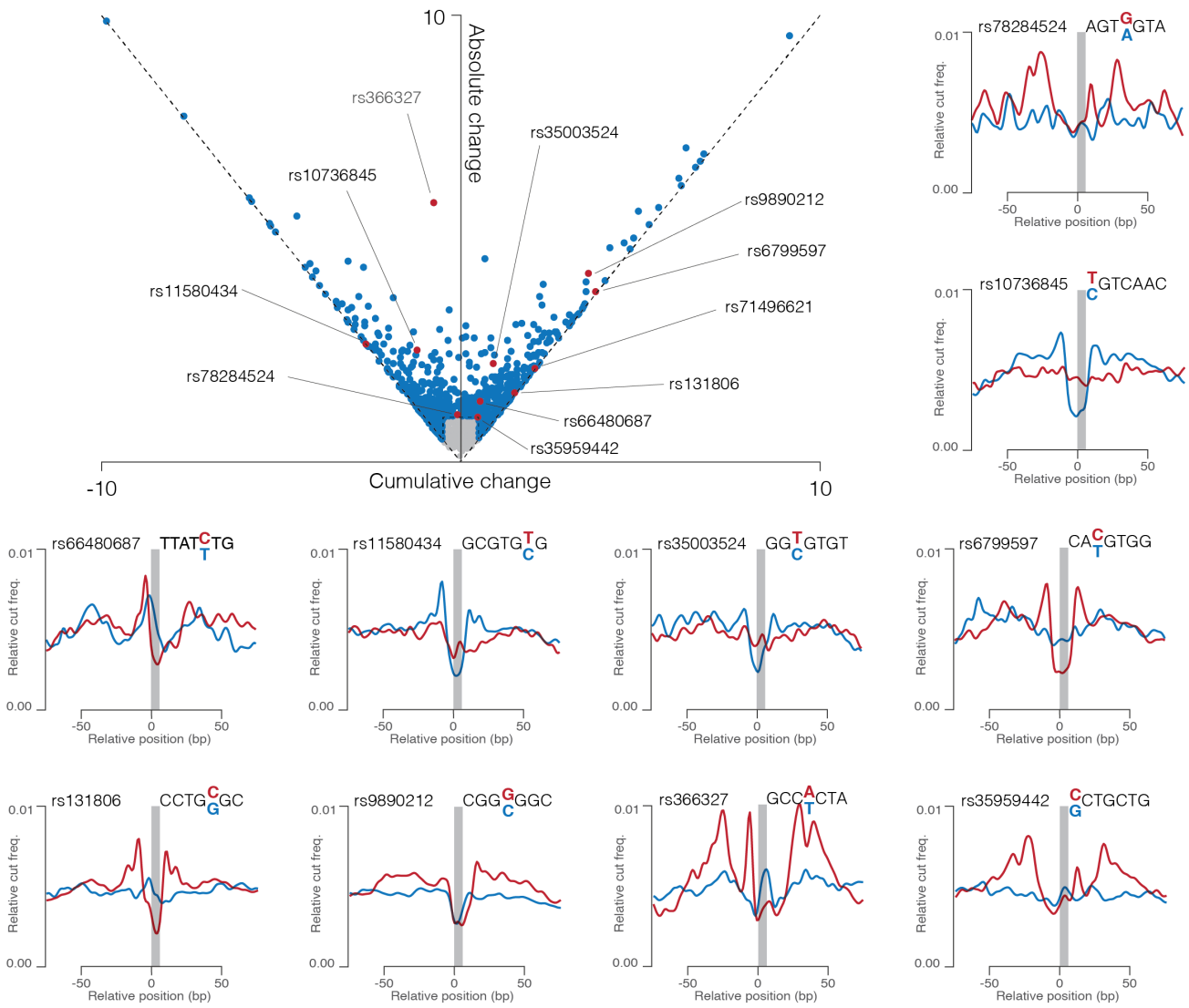
Supplementary Fig. 4) RBC Trait containing ATAC-seq peaks are enriched for strong enhancers and promoters. GenoSTAN uses a hidden Markov model and was implemented to classify defined chromatin regions into "states" using ChIP-seq read counts. Erythroid open chromatin regions were subclassified into 7 functional classes and a background class using H3K4me3 (promoter mark), H3K4me1 (enhancer mark), H3K27ac (active transcription mark) and CTCF (boundary element) ChIP-seq. Following intersection of open chromatin regions from two RBC trait GWAS studies, the re-distribution of GenoSTAN states of open chromatin regions containing variants was determined. Intersected elements are enriched for the Promoter and Strong Enhancer (high levels of H3K27ac) classes.



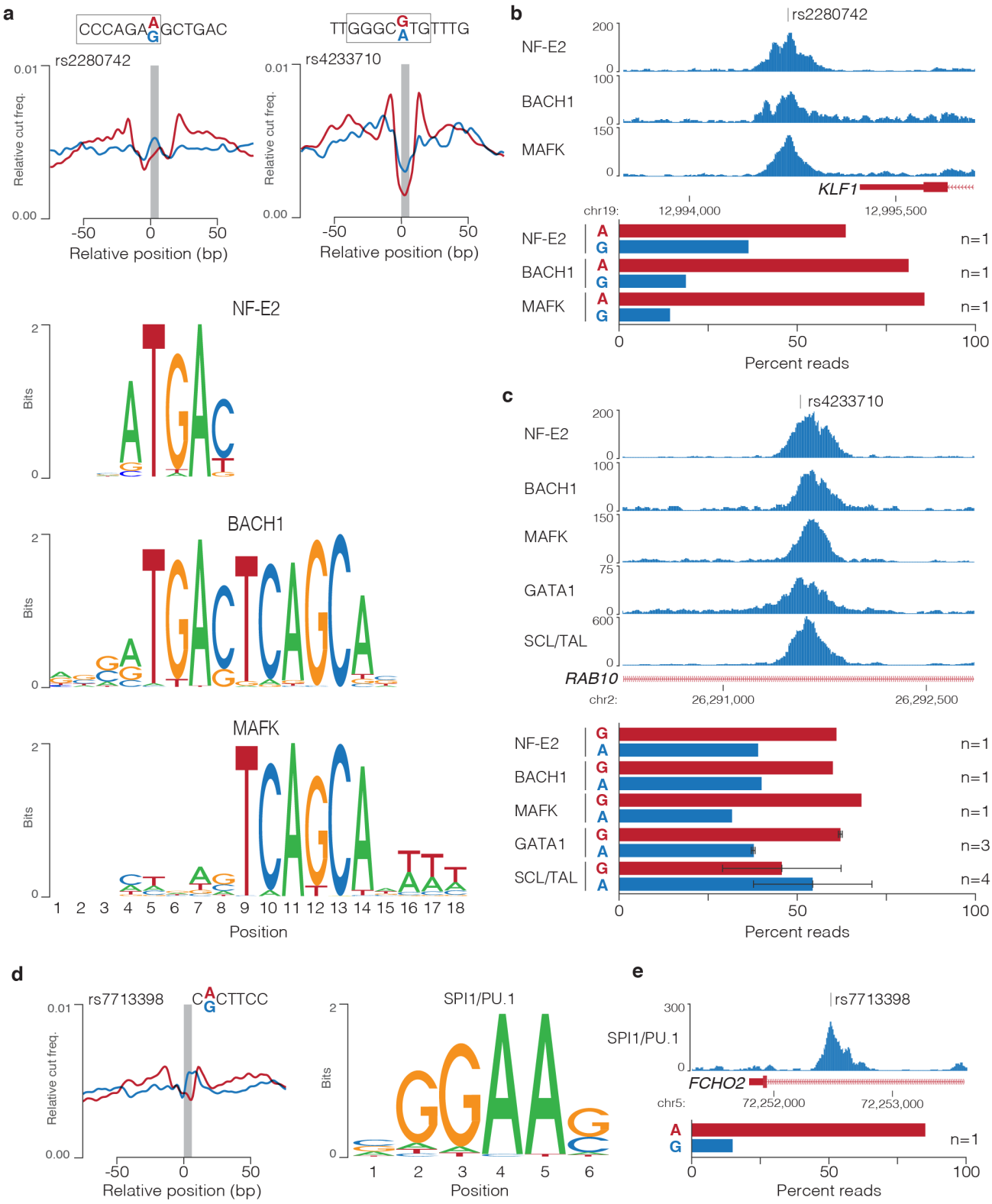
Supplementary Fig. 5) Intersected peaks show multiple trajectories of openness. pHeatmap clustering of z-score normalised ATAC-seq read counts for RBC variant containing peaks from haematopoietic stem cells (HSC), multi-potent progenitors (MPP), common myeloid progenitors (CMP), megakaryocyte-erythroid progenitors (MEP), *in vitro* differentiated erythroid cells (Day7-17).



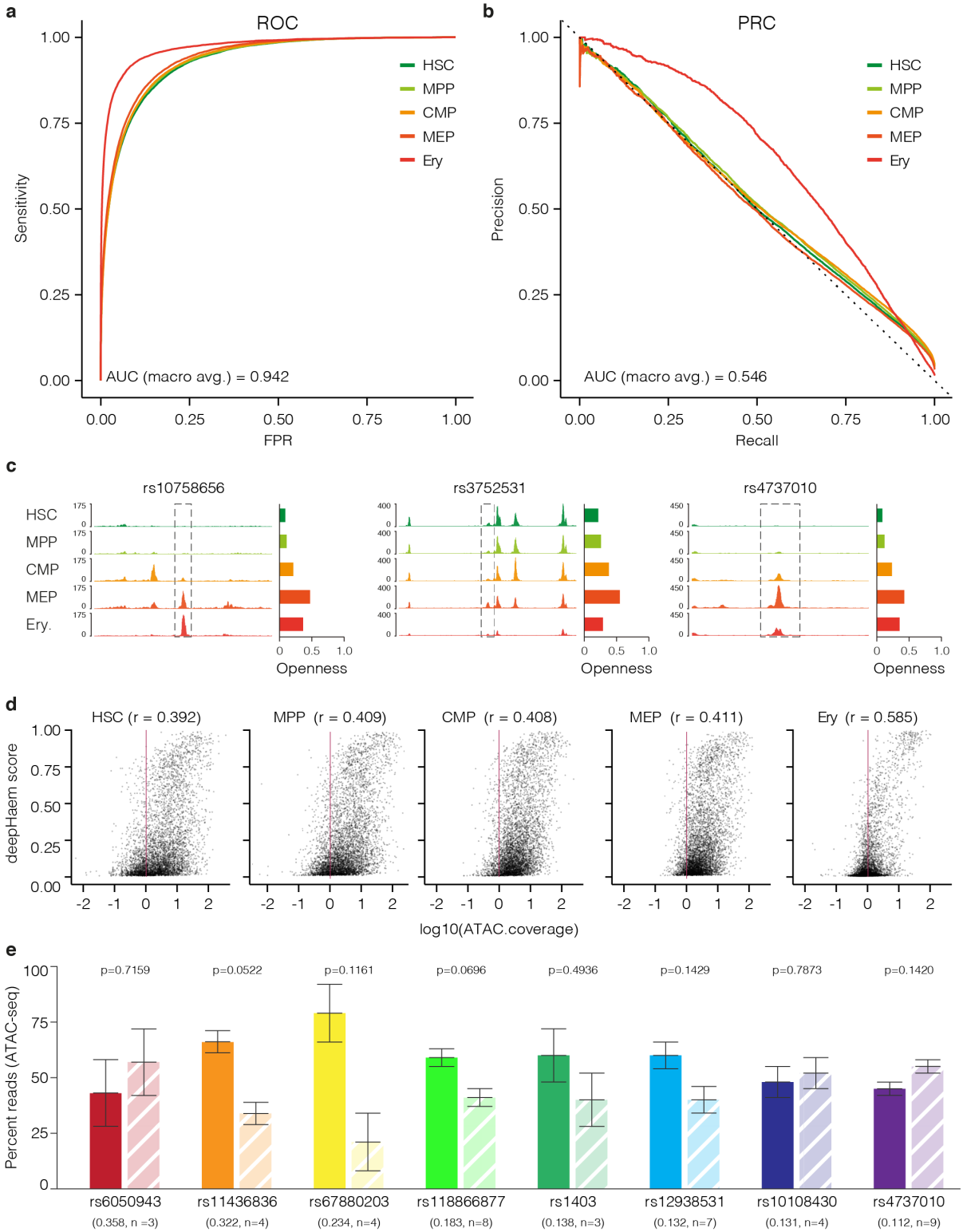
Supplementary Fig. 6) Examples of ATAC-seq peak intersection trajectories. ATAC-seq from haematopoietic stem cells (HSC), multi-potent progenitors (MPP), common myeloid progenitors (CMP), megakaryocyte-erythroid progenitors (MEP), *in vitro* differentiated erythroid cells was intersected with trait associated variants. Trajectories of intersection included continual (**a**), progenitor (**b**), transient (**c,d**) and terminal (**e,f**) trajectories (boxed). Variants within open chromatin are red, the index SNP is marked with a circle.



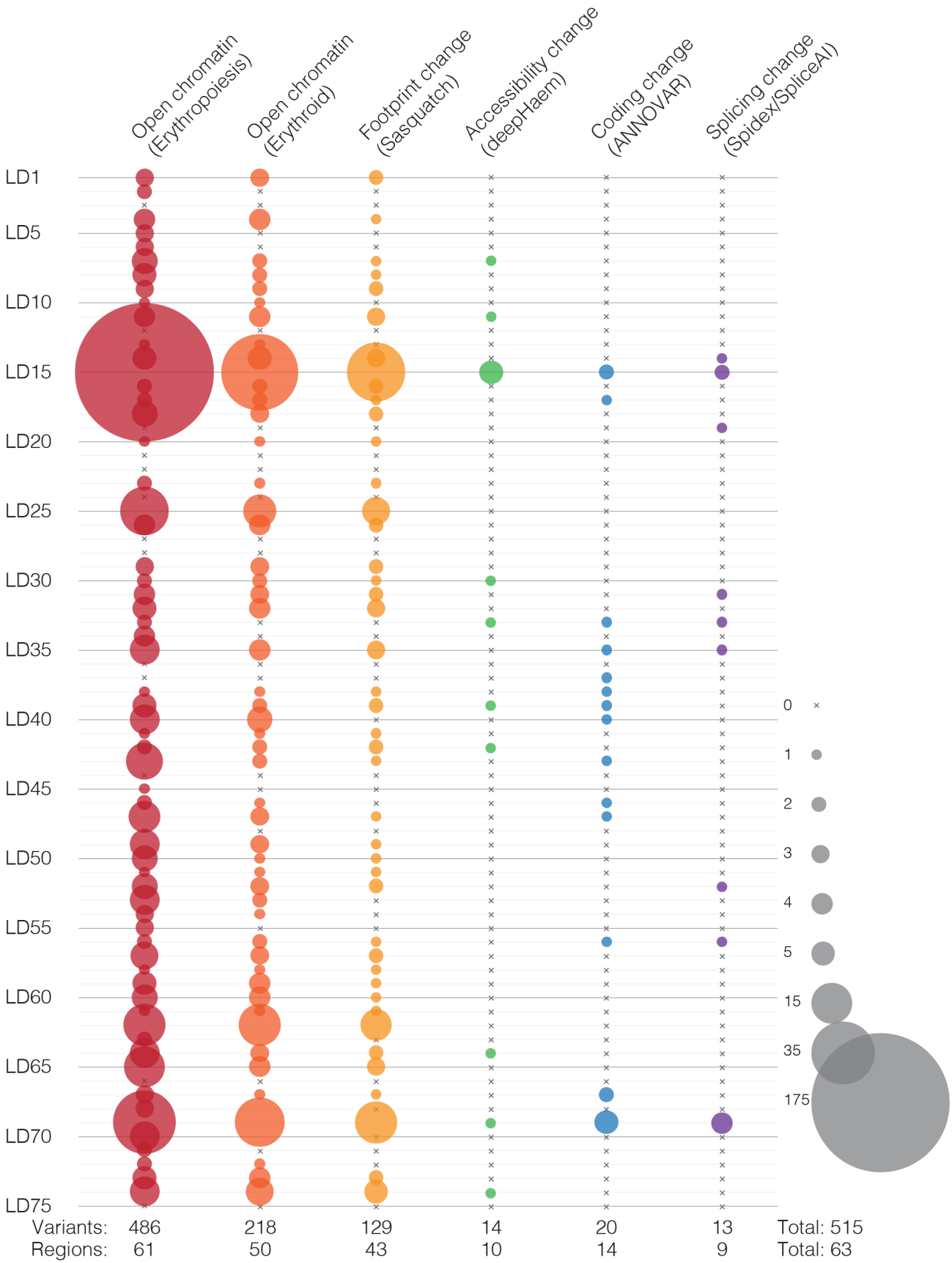
Supplementary Fig. 7) Sasquatch predicts changes in transcription factor footprints. Sasquatch generates meta-genomic footprints of 7-mer motifs using DNaseI signal from open chromatin regions to measure relative cut frequency. For variants in open chromatin, footprints and relative cut frequencies were generated with all seven possible 7-mer sequences for both alleles. Cumulative change and Absolute change for a variant are calculated by comparing the difference in relative cut frequency for each 7-mer and calculating the cumulative (directional) and absolute (total) change. Positive cumulative change scores can be interpreted as loss of a footprint, while negative values represent a gain of a footprint. Absolute change scores much larger than the cumulative change score (e.g. rs366327) represent a change in footprint type rather than a simple loss/gain. Variants with a cumulative change ≥0.5 or an absolute change ≥1.0 were considered damaging (blue dots). Representative footprints are shown for a range of SNPs (red dots).



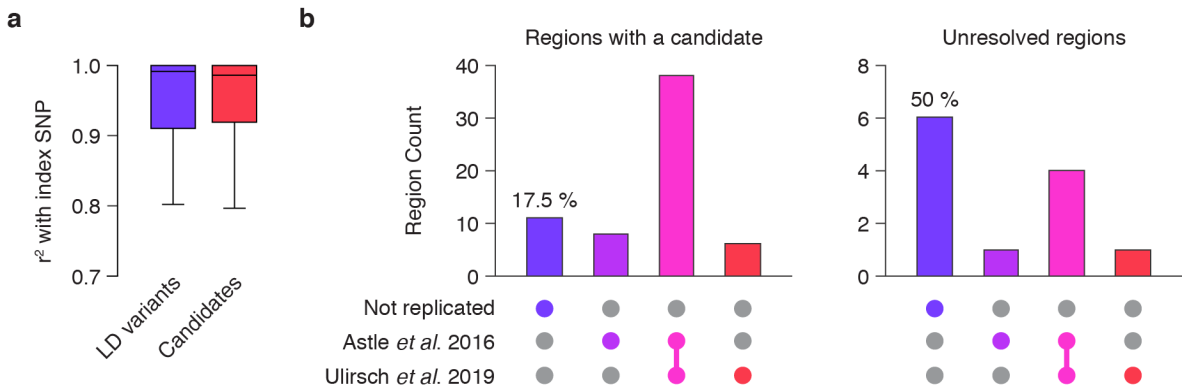
Supplementary Fig. 8) Sasquatch predicted changes correlate with skew in binding of multiple transcription factors. a, rs2280742 and rs4233701 lie within 7-mer motifs (boxed) that are adjacent to binding motifs for NF-E2, BACH1 and MAFK and are predicted to alter DNaseI footprints at these sites. ChIP-seq for NF-E2, BACH1, MAFK, GATA1, and SCL/TAL over rs2280742 (**b**) and rs4233701 (**c**) show allelic skew in multiple factors as shown by percent of reads containing either allele. Error bars depict standard error of the mean. **d,** rs7713398 alters the predicted DNaseI footprint for a 7-mer motif which contains an SPI1/PU.1 consensus motif. **e,** ChIP-seq of SPI1/PU.1 shows allelic skew over rs7713398, as shown by percent of reads containing either allele.



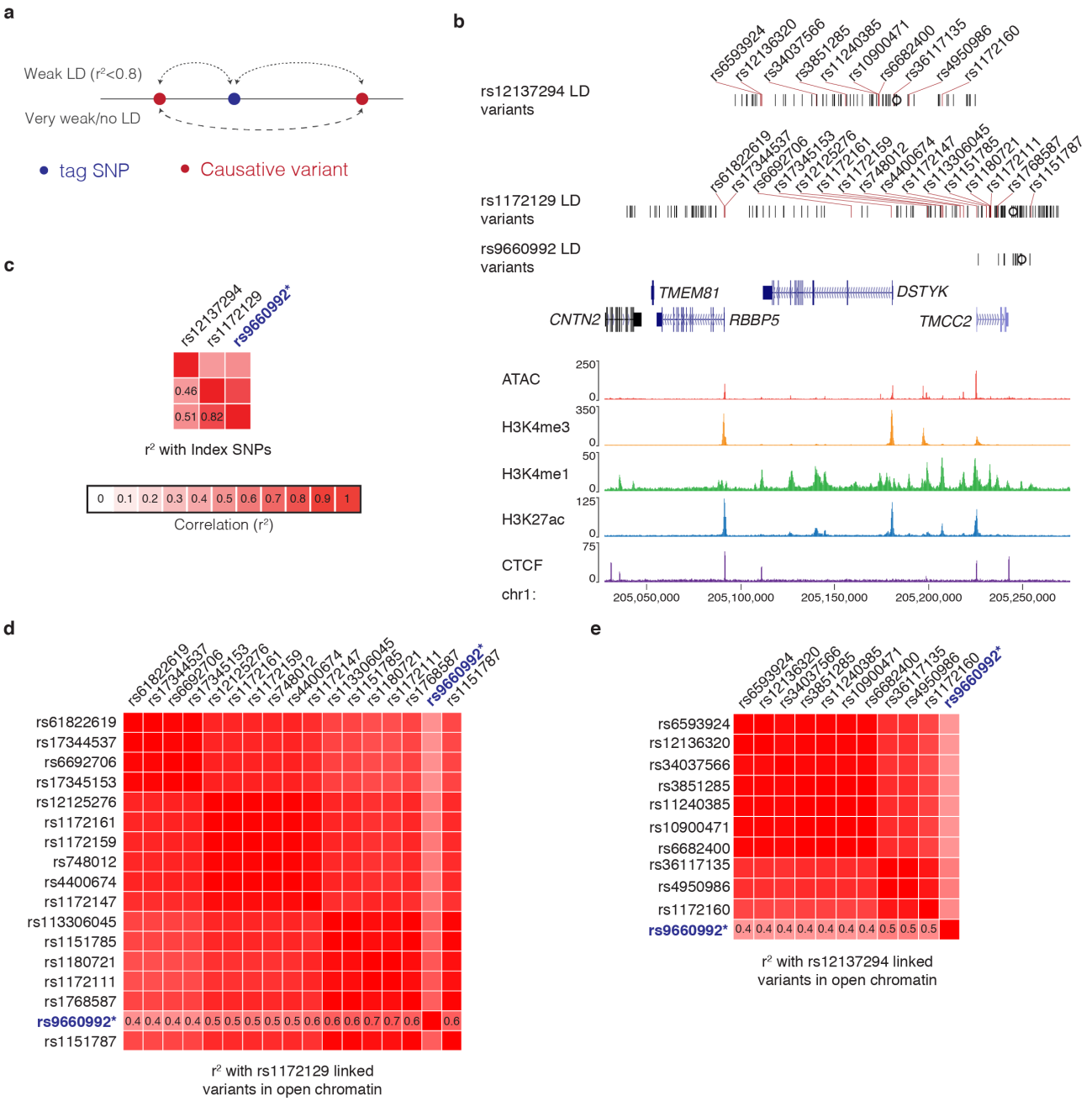
Supplementary Fig. 9) deepHaem prediction chromatin openness throughout erythropoiesis. Receiver operating characteristic (ROC) curve (**a**) and Precision-Recall Curve (PRC) (**b**) for deepHaem in haematopoietic stem cells (HSC), multi-potent progenitors (MPP), common myeloid progenitors (CMP), megakaryocyte-erythroid progenitors (MEP) and erythroid (Ery). **c**, FPKM normalised ATAC-seq around three SNPs predicted to be damaging by deepHaem, with openness scores shown for the hg19 reference sequence calculated for the 1kb box (dashed lines) centred on the SNP. **d**, Pearson correlation between deepHaem open chromatin scores and ATAC-seq coverage in corresponding cell types in a random sampling (1/50th) of MACS2 identified peaks. **e,** Mean percentage of day 10 and day 13 erythroid ATAC-seq reads on either the reference (dark bars) or variant (light dashed bars) allele from heterozygous individuals with a minimum of 5 reads. Error bars depict the standard error of the mean. deepHaem change score and number of independent replicates (including multiple donors and differentiations) are shown in parenthesis. p-values shown are for a ratio paired t-test.



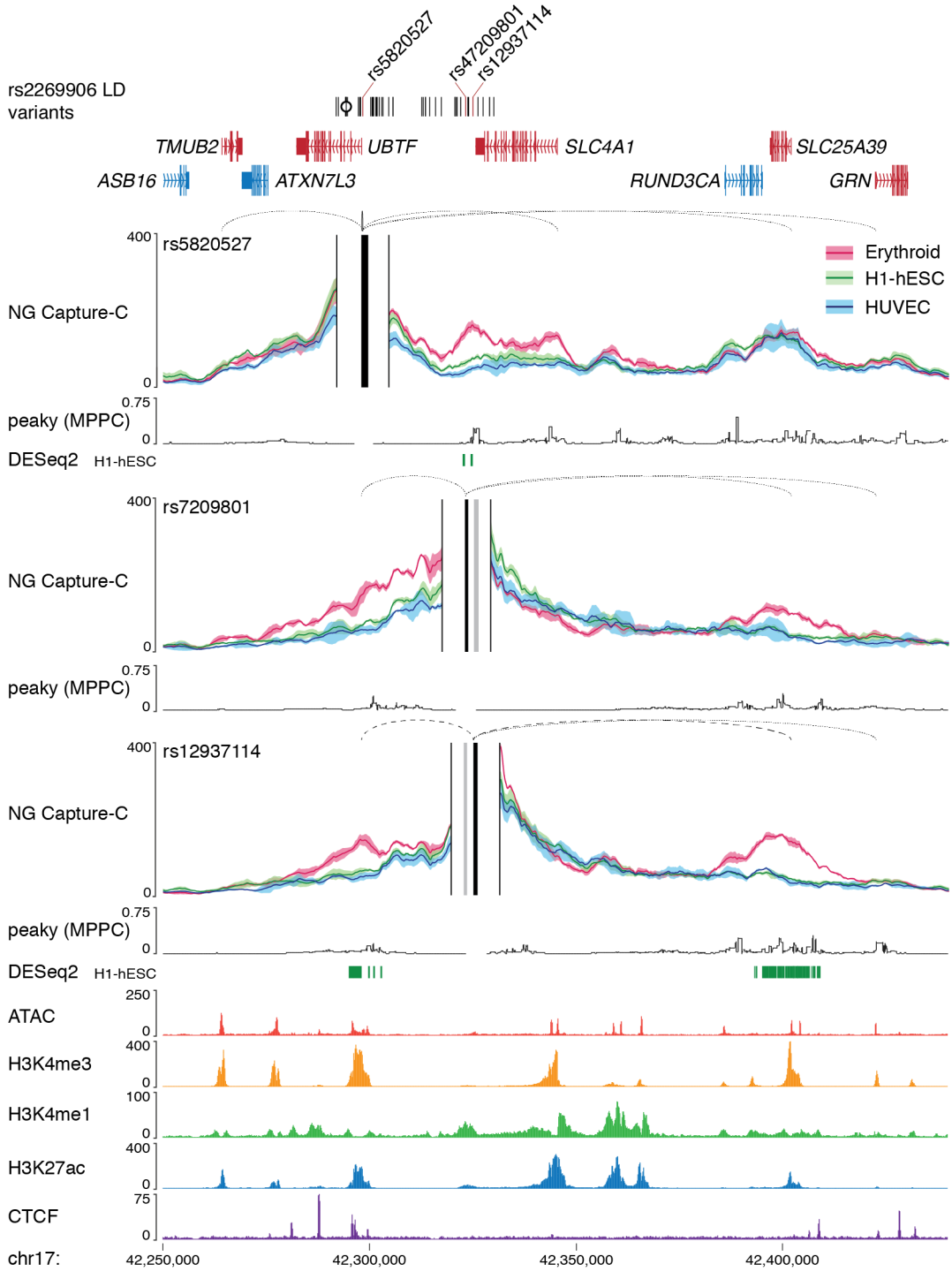
Supplementary Fig. 10) Pipeline summary for prioritization GWAS variants. RBC GWAS variants in linkage disequilibrium (LD) at 75 chromosome regions (n=6,420) were examined for presence in open chromatin, and putative changes in DNaseI footprint, chromatin accessibility, coding sequence, splicing potential. The total number of variants in each class is shown at the bottom. Circle size indicated the number of variants included for each class in a chromosome region.



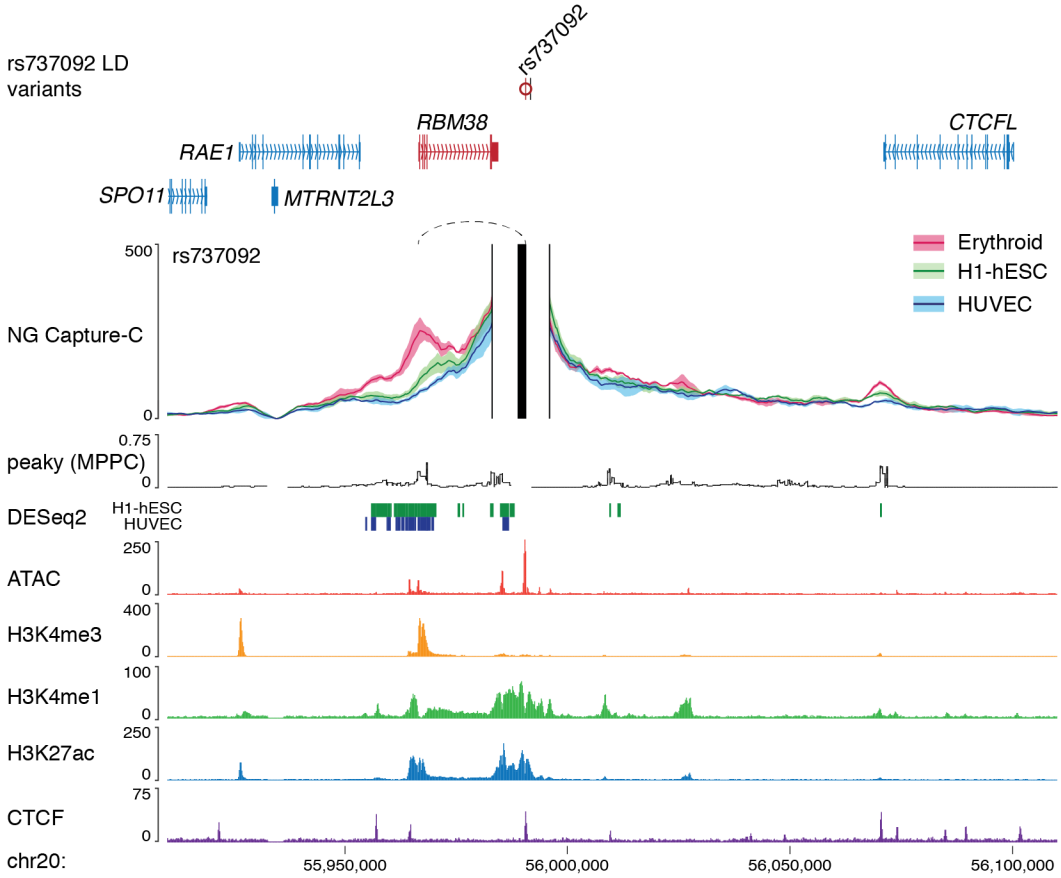
Supplementary Fig. 11) Analysis of resolved and unresolved regions. a, r^2 for linkage disequilibrium analysis of RBC trait index SNPs with all analysed variants (n=6,420) and candidate variants identified by the platform (n=515). **b,** Counts of regions where index SNPs identified in van der Harst *et al.* (2012) was replicated in either of two subsequent RBC trait GWAS. Regions are separated into those with candidate variants following analysis in the described platform (n=63), and those without (n=12).



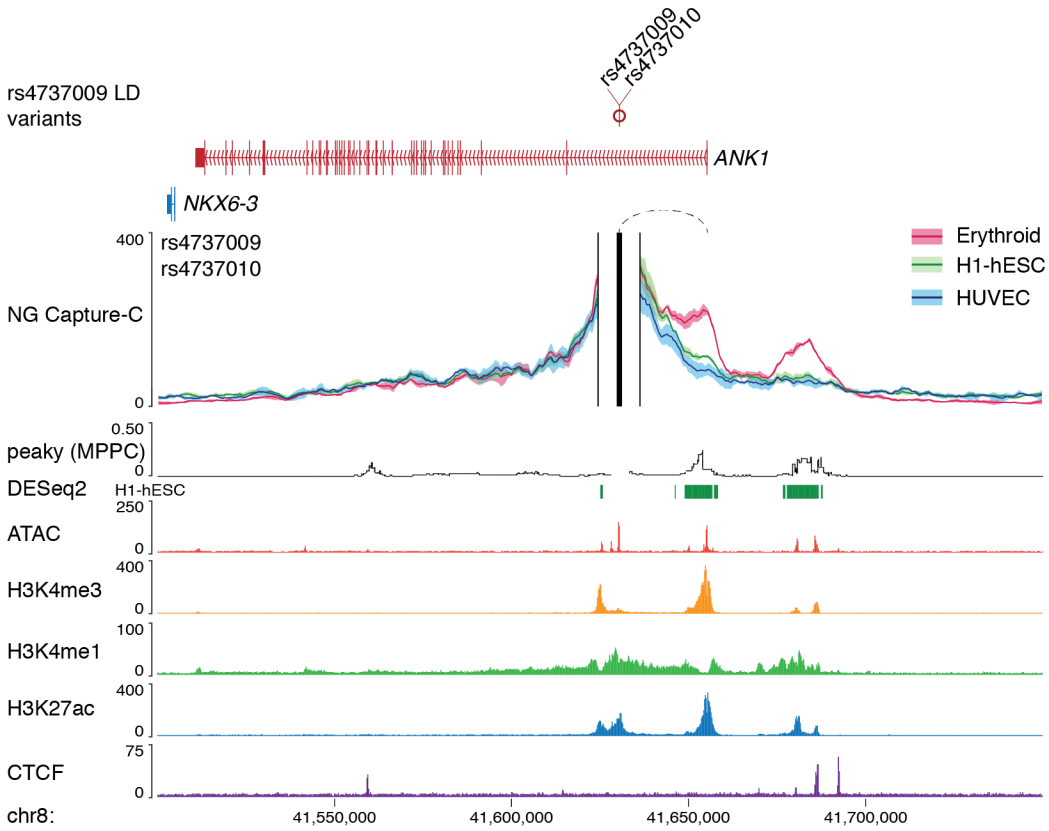
Supplementary Fig. 12) rs9660992 may be a tag SNP at the TMCC2 locus. a, Model for tag SNPs, whereby an index SNP may be identified through weak linkage disequilibrium (LD) to multiple unlinked causative variants. **b**, Map of RBC trait associated variants at the TMCC2 locus, with ATAC-seq and ChIP-seq chromatin marks from erythroid cells. Variants in open chromatin are red and labelled, index SNPs are marked with a circle. TMCC2 is a transmembrane and coiled-coil domain protein required for terminal erythropoiesis. **c**, Linkage analysis between index SNPs at the TMCC2 locus, the putative tag SNP rs9660992 is in blue. Linkage analysis between rs9660992 and variants in open chromatin that are in strong LD ($r^2 \geq 0.8$) with rs1172129 (**d**) and rs12137294 (**e**).



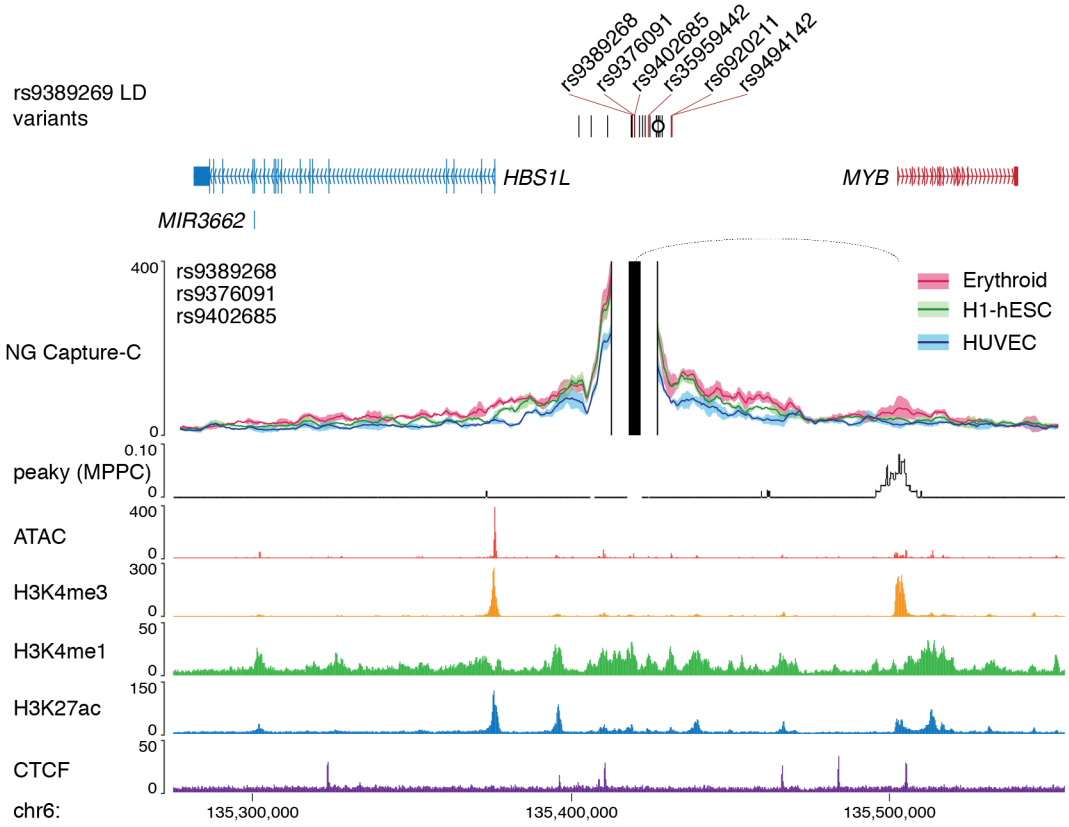
Supplementary Fig. 13) 3C interactions at the Band 3 locus. NG Capture-C interaction profiles for three open chromatin intersecting variants at the SLC4A1 (which encodes Band 3) locus performed in erythroid, embryonic stem (H1-hESC) and umbilical vein endothelial (HUVEC) cells (n=3). Capture viewpoints and proximity exclusion regions (solid vertical lines) were designed for each open chromatin region and profiles show mean interactions (solid line) with one standard deviation (shading). Peaky values depict the Marginal Posterior Probability of Contact (MPPC) in erythroid cells. Promoter-variantsP interactions were identified by promoter proximity (solid loops), erythroid specific interactions (dashed loops; DESeq2 q-value < 0.05 shown as bars), and Bayesian modelling (dotted loops; peaky MPPC > 0.01). SNPs 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. Key interactions were seen with both *SLC4A1*, an erythrocyte membrane anion transport protein and, and *SLC25A39* is a mitochondrial iron transport protein, both of which are required for erythropoiesis.



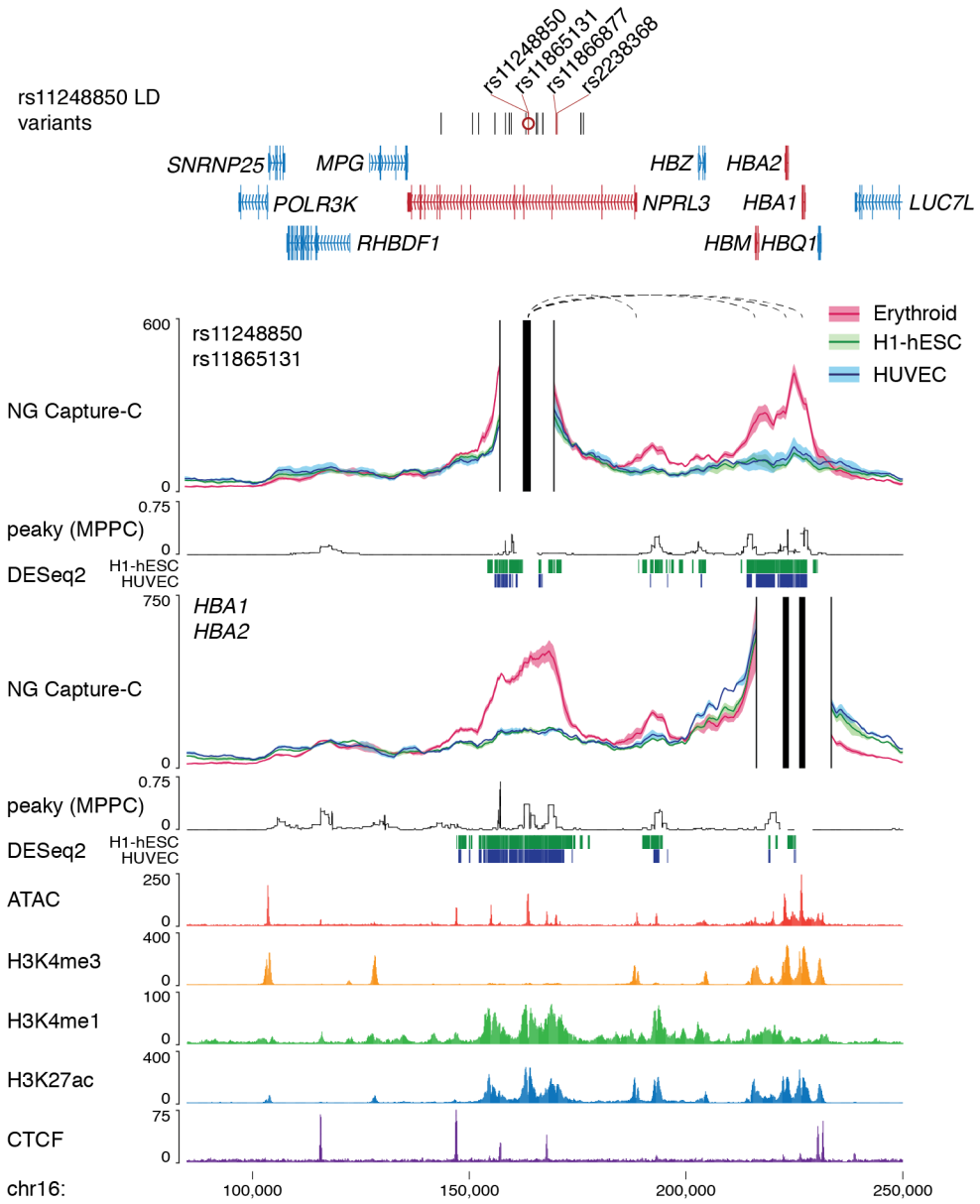
Supplementary Fig. 14) 3C interaction profile for RNPC1 locus. 3C interaction profile for rs737092 from erythroid, embryonic stem (H1-hESC) and umbilical vein endothelial (HUVEC) cells (n=3). Capture viewpoints and proximity exclusion regions (solid vertical lines) were designed for each open chromatin regions and profiles show mean interactions (solid line) with one standard deviation (shading). Peaky values depict the MPPC in erythroid cells. Promoter-variant interactions were identified as erythroid-specific interactions (dashed loops; DESeq2 q-value < 0.05 shown as bars). 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 seen with *RBM38*, which encodes the erythroid splicing regulator RNPC1.



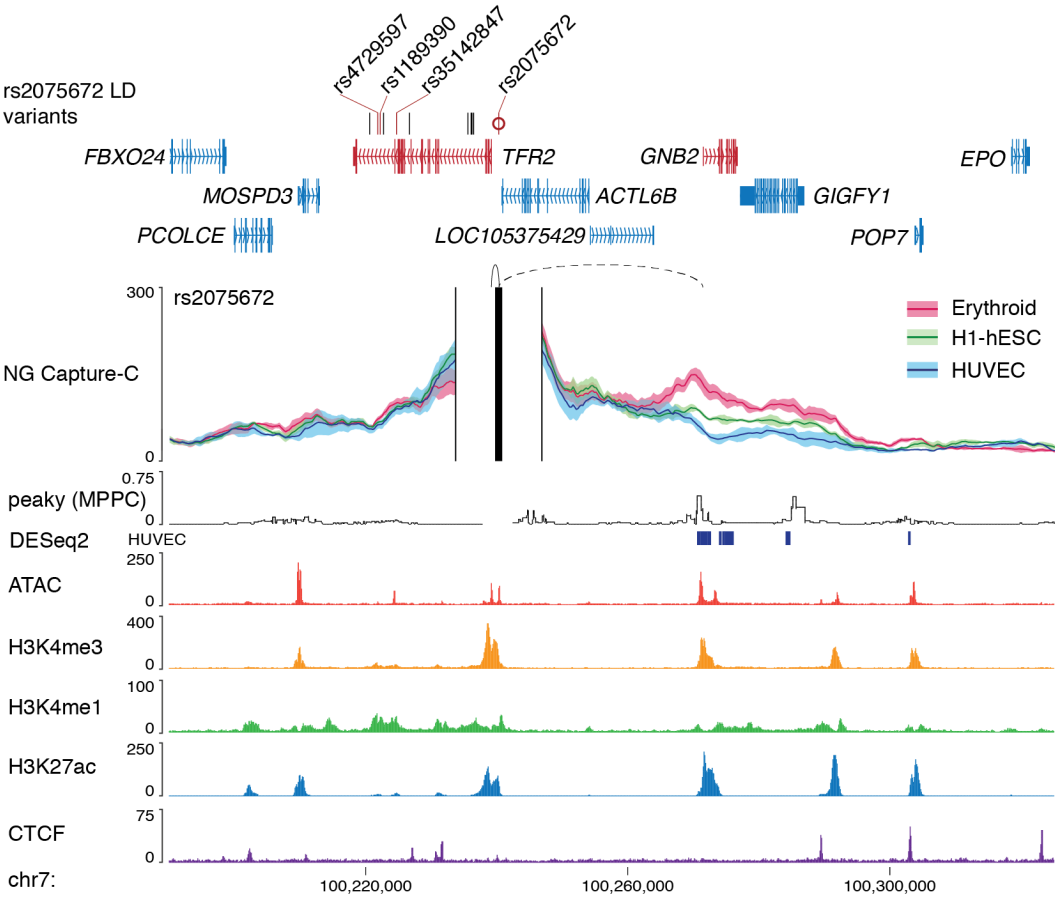
Supplementary Fig. 15) 3C interaction profile for the Ankyrin locus. 3C interaction profile for rs4737009 and rs4737010 in erythroid, embryonic stem (H1-hESC) and umbilical vein endothelial (HUVEC) cells (n=3). Capture viewpoints and proximity exclusion regions (solid vertical lines) were designed for each open chromatin regions and profiles show mean interactions (solid line) with one standard deviation (shading). Peaky values depict the MPPC in erythroid cells. Promoter-variant interactions were identified as erythroid specific (dashed loops; DESeq2 q-value < 0.05 shows as bars). 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 seen with *ANK1*, which encodes Ankyrin-1 – an integral erythroid membrane protein often missing in hereditary spherocytosis patients.



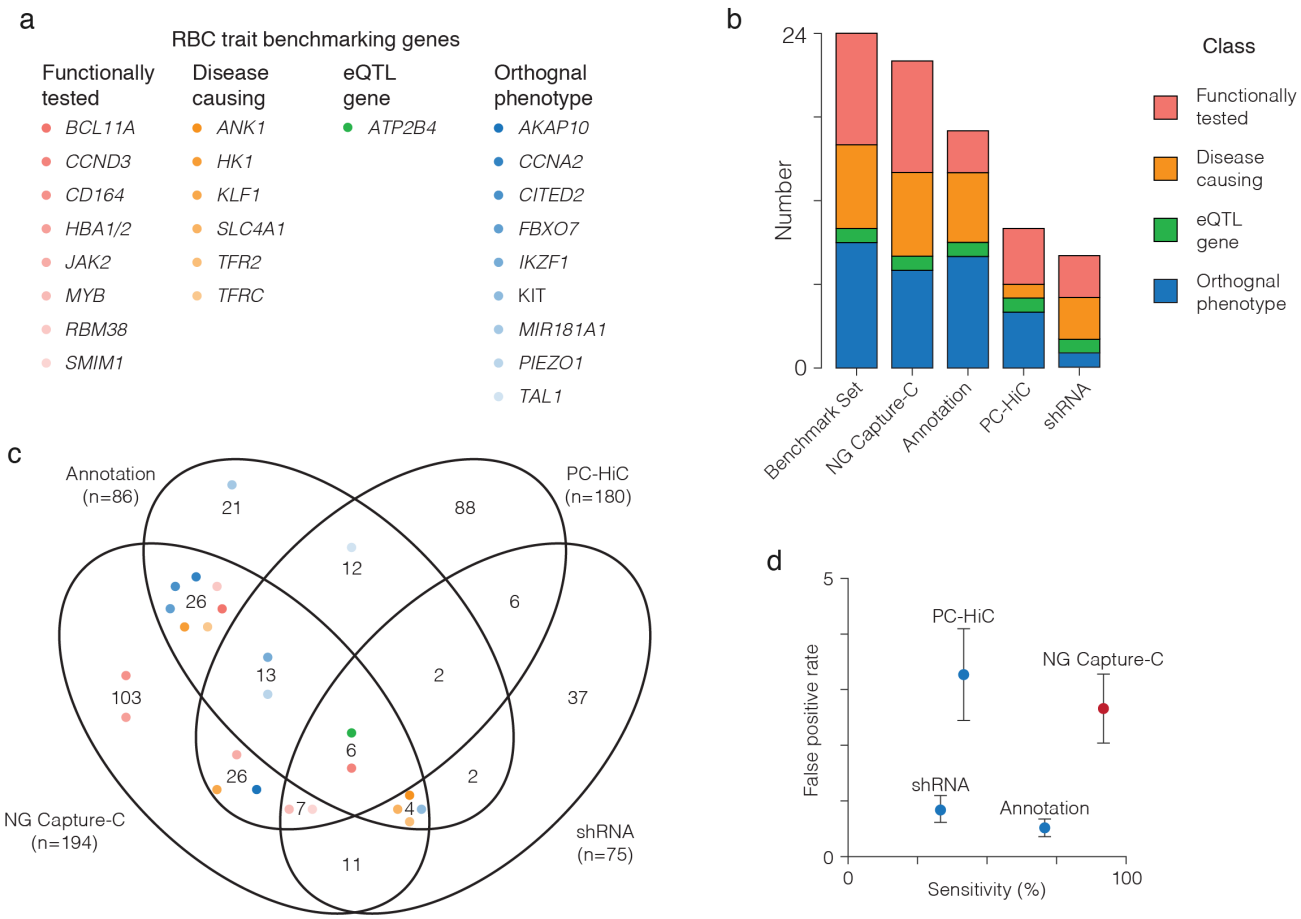
Supplementary Fig. 16) 3C interaction profile for the c-Myb locus. 3C interaction profile for rs9389268, rs9376091, and rs9402685 in erythroid, embryonic stem (H1-hESC) and umbilical vein endothelial (HUVEC) cells (n=3). Capture viewpoints and proximity exclusion regions (solid vertical lines) were designed for each open chromatin regions and profiles show mean interactions (solid line) with one standard deviation (shading). Peaky values depict the MPPC in erythroid cells. Interaction with *MYB* 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. Interaction was seen with the *MYB*, which encodes c-Myb a key regulatory transcription factor involved in erythroid progenitor proliferation and differentiation.



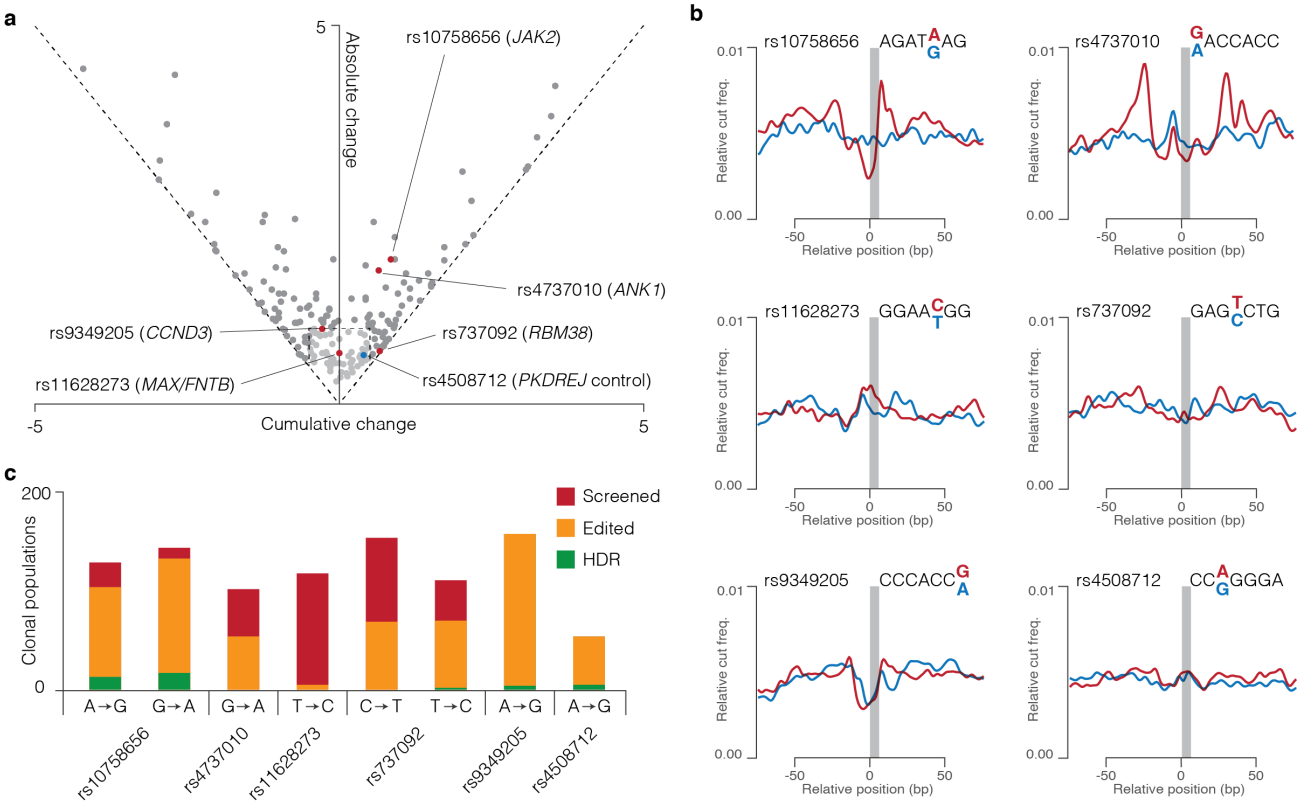
Supplementary Fig. 17) 3C interaction profile for the α-globin locus. 3C interaction profile for rs11248850 and rs11865131, which lie in a conserved HBA1 and HBA2 enhancer, as well as reciprocal promoter capture in erythroid, embryonic stem (H1-hESC) and umbilical vein endothelial (HUVEC) cells (n=3). Capture viewpoints and proximity exclusion regions (solid vertical lines) were designed for each open chromatin regions and profiles show mean interactions (solid line) with one standard deviation (shading). Peaky values depict the MPPC in erythroid cells. Promoter-variant interactions were identified as erythroid specific (dashed loops; DESeq2 q-value < 0.05 show as bars). 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.



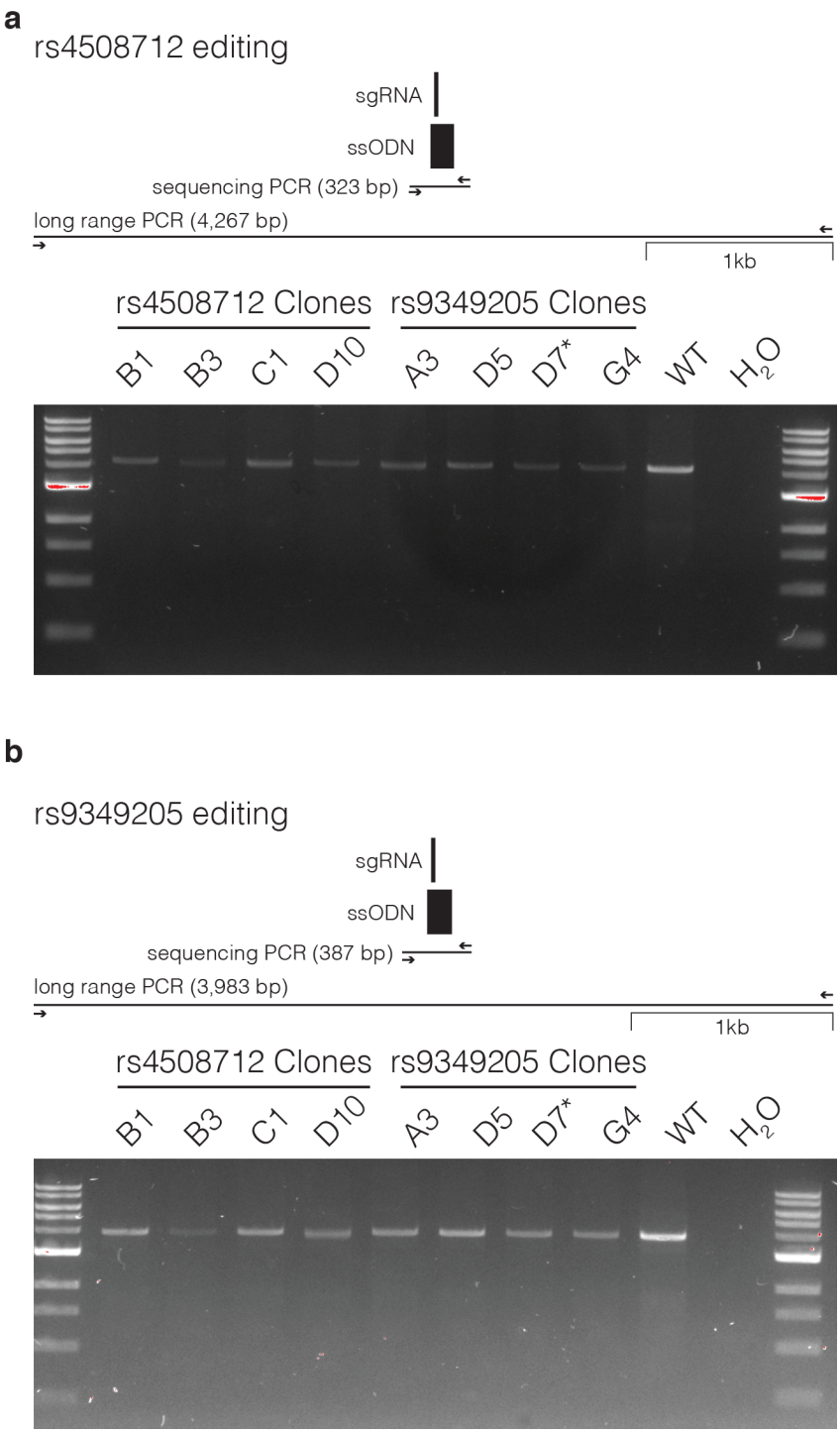
Supplementary Fig. 18) 3C interaction profile for the TfR2 locus. 3C interaction profile for rs2075672 in erythroid, embryonic stem (H1-hESC) and umbilical vein endothelial (HUVEC) cells (n=3). Capture viewpoints and proximity exclusion regions (solid vertical lines) were designed for each open chromatin regions and profiles show mean interactions (solid line) with one standard deviation (shading). Peaky values depict the MPPC in erythroid cells. Promoter-variant interactions were identified by promoter proximity (solid loops) and erythroid specific interactions (dashed loops; DESeq2 q-value < 0.05 shows as bars). Variants within open chromatin are red, as are variant interacting genes, the index SNP is are marked with a circle. FPKM normalised ATAC-seq and ChIP-seq tracks are from erythroid cells. Interaction was seen with *TFR2*, which encodes the erythroid and liver hepcidin receptor Transferrin receptor 2 (TfR2), and *GNB2*, which encodes the signalling G-protein subunit B.



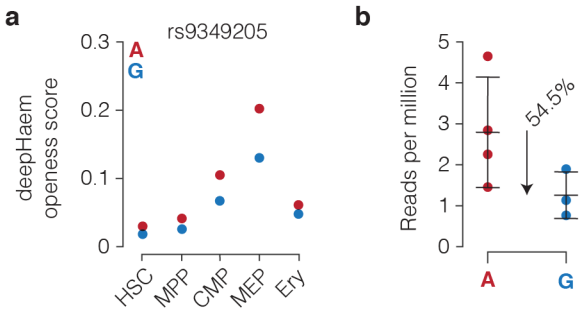
Supplementary Fig. 19) Comparison of methods for identification of GWAS effector genes. a, Classification of 24 most-likely effector genes used for benchmarking as identified through known function in erythropoiesis. **b,** Number of benchmarking genes identified by four different methods including NG Capture-C, Promoter Capture-HiC (PC-HiC), a gene centric shRNA screen, and an annotation-based method utilising coding variants, eQTLS, nearby genes and gene relationships among implicated loci (GRAIL). **c,** Venn comparison of all genes identified by the four different methods across the 53 NG Capture-C targeted chromosome regions with dots for each of the benchmarking genes. **d**, Percent of 24 benchmarking genes identified (sensitivity) and average number of false positives per locus (false positive rate) for each of the four methods. Error bars show standard error of the mean (n = 24).



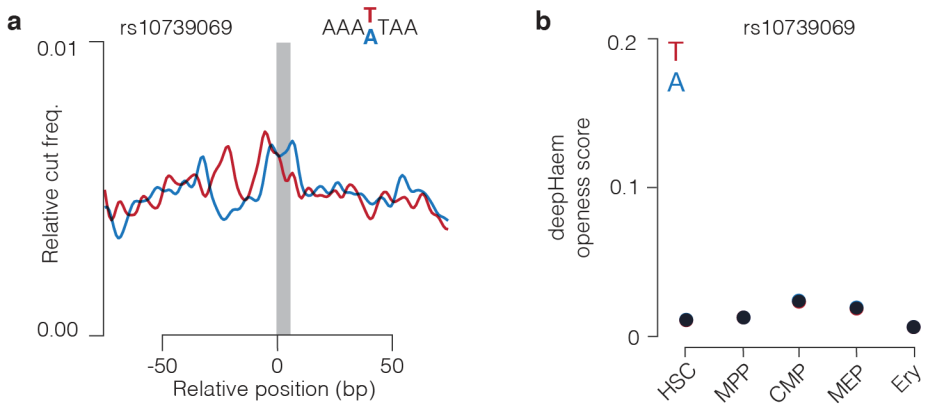
Supplementary Fig. 20) Pipeline prioritized SNPs were edited by CRISPR/Cas9. a, Five SNPs with candidate effector genes (red dots) and one control SNP (blue dot) were selected for *in vivo* modelling. **b,** Selected SNPs showed a range of predicted Sasquatch effects and alleles were directly modelled by CRISPR/Cas9 facilitated homology directed repair (HDR). **c,** Clonal HUDEP-2 populations were derived by editing and single-cell sorting and screened by next generation sequencing to detect changes. Bar height shows total number of clonal populations screened, with variable rates of editing (indels and HDR, yellow) and specific changes (HDR, green) observed across the loci.



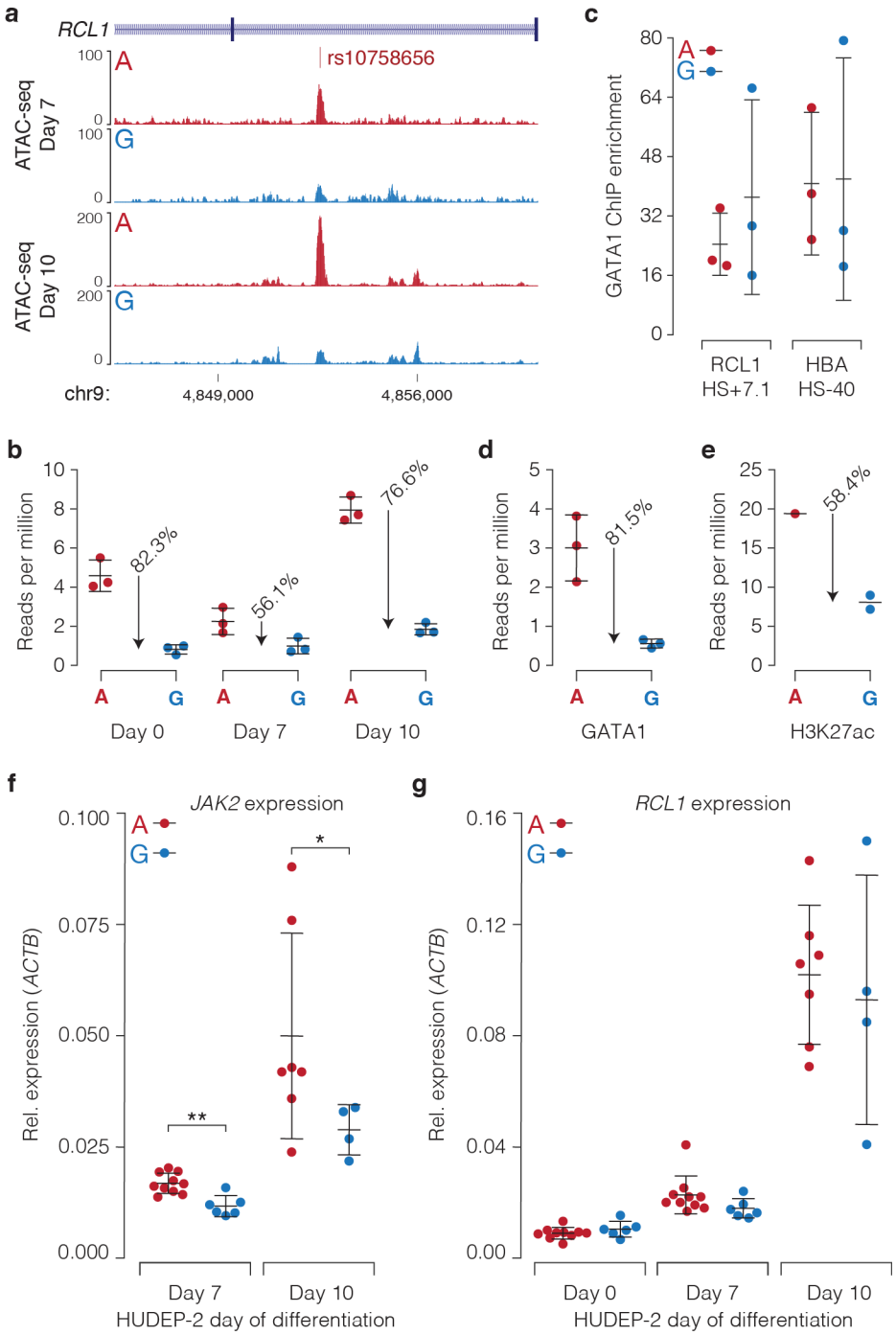
Supplementary Fig. 21) Long range PCR to detect loss of heterozygosity. Long range PCR was carried out for the both rs4508712 (**a**) and rs9349205 (**b**) loci on clones identified by sequencing as homozygous for the desired editing events. *Re-sequencing identified a 2bp deletion leading to the exclusion of this clone.



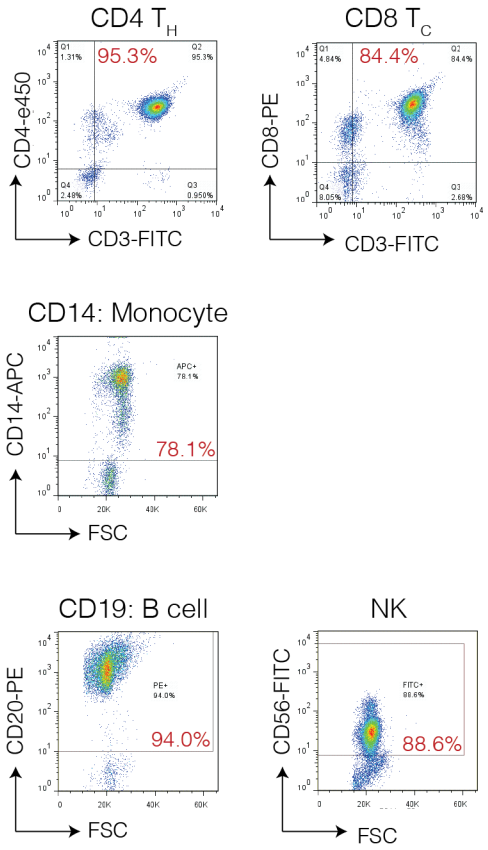
Supplementary Fig. 22) deepHaem openness score for rs9349205. a, DeepHaem uses a deep convolutional neural network machine learning approach to predict the likelihood of chromatin openness (or "openness score") ranging from 0-1. These prediction scores were generated through training on a binary yes or no for openness, but do have a positive correlation with observed levels ATAC-seq signal (see Supplementary Figure 7). For this study, differences between variant alleles greater than 0.1 (or 10% of the maximum openness score) were classed as damaging. For rs9349205 a slight difference between the two alleles is seen with deepHaem, though this does no cross the strict threshold for classification as damaging. **b,** Count of ATAC-seq reads per million mapped reads in the rs9349205 containing peak (chr6:41,924,682-41,925,483, hg19) in undifferentiated human umbilical derived erythroid progenitors (HUDEP-2) cells. Each circle is an independently isolated clone homozygous for either A or G at rs9349205, bars show mean and one standard deviation.



Supplementary Fig. 23) rs10739069 is not predicted to be damaging. rs10739069 was found within an open chromatin region, however the Sasquatch footprint (**a**) and deepHaem openness scores (**b**) show it is unlikely to be a damaging RBC trait SNP.



Supplementary Fig. 24) rs10758656 is specific for regulation of *JAK2*. a, FPKM normalised ATAC-seq (n=3) from differentiating HUDEP-2 clones homozygous for either rs10758656 allele. **b,** Count of ATAC-seq reads per million mapped reads in the rs10758656 containing peak (chr9:4,852,237-4,853,036, hg19) in undifferentiated human umbilical derived erythroid progenitors (HUDEP-2) cells. **c,** Real-time quantitative PCR for GATA1 ChIP enrichment at two control open chromatin regions, one upstream of rs10758656 near the RCL1 promoter (chr9:4799898-4799984, hg19) and a well characterized *HBA1/2* enhancer (HS-40; chr16:163615-163730, hg19) in rs10758656$^{A/A}$ and rs10758656$^{G/G}$ clones. Enrichment was calculated using the ΔΔCt method against a background region. Count of reads per million mapped reads in the rs10758656 containing peak from ChIP-seq of GATA1 (**d**) and H3K27ac (**e**) at day 7 of differentiation**.** Real time reverse-transcriptase PCR of *JAK2* (**f**) and the gene nearest to rs10758656, *RCL1* (**g**) – which encodes RNA terminal phosphate cyclase-like 1 – in differentiating HUDEP-2 clones edited to be homozygous for both alleles of rs10758656 (n≥4). Mann-Witney test, **p=0.0030, *p=0.0364. For all graphs each circle is an independently isolated clone homozygous for either A or G at rs10758656, bars show mean and one standard deviation.



Supplementary Fig. 25) Population purities for MACS isolated immune cell populations. MACS sorted populations were stained with a specific cocktail of cell markers to determine purity. Gates were set using a negative staining cell population.

Supplementary Table 1. RBC Prioritized variants.

a. Overview of number of variants analysed by the integrated platform
b. List of prioritized variants with supporting evidence and gene interactions.
c. List of previously validated variants.
d. List of benchmarking RBC trait genes.

Supplementary Table 2. Growth media recipes.

a. CD34 Differentiation.
b. HUDEP-2 Differentiation.

Supplementary Table 3. Oligonucleotides used for genome editing.

a. Genotyping primers.
b. Guide RNA.
c. ssODN.
d. Long range PCR primers.

Supplementary Table 4. List of antibodies.

a. Flow Cytometry.
b. Chromatin Immunoprecipitation.

Supplementary Table 5. Quantitative PCR primers

a. TaqMan Expression.
b. DpnII 3C library digestion.
c. ChIP Enrichment.

Supplementary Table 6. Sample Sequencing Depth.

a. RNA-seq.
b. NG Capture-C.
c. ATAC-seq.
d. ChIP-seq.

Supplementary Table 7. List of biotinylated NG Capture-C oligonucleotides.

a. Enhancer Oligonucleotides.
b. Promoter Oligonucleotides.

Supplementary Table 8. Bed files of open chromatin peak calls.

a. HSC.
b. MPP.
c. CMP.
d. MEP.
e. Erythroid Day 7.
f. Erythroid Day 10.
g. Erythroid Day 13.
h. Erythroid Day 17.
i. Erythropoiesis Staging.
j. H1-hESC.
k. HUVEC.
l. CD4.
m. CD8.
n. CD14.
o. CD19.
p. Natural Killer.