bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MANUSCRIPT

Genome-wide association study on 13,167 individuals identifies regulators of hematopoietic stem and progenitor cell levels in human blood

Aitzkoa Lopez de Lapuente Portilla[1,2], Ludvig Ekdahl[1,2,*], Caterina Cafaro[1,2,*], Zain Ali[1,2,*], Natsumi Miharada[1,2], Gudmar Thorleifsson[3], Kristijonas Žemaitis[1,2], Antton Lamarca Arrizabalaga[1,2], Malte Thodberg[1,2], Maroulio Pertesi[1,2], Parashar Dhapola[1,2], Erik Bao[4,5,6], Abhishek Niroula[1,2,6], Divya Bali[1,2], Gudmundur Norddahl[3], Nerea Ugidos Damboriena[1,2], Vijay G. Sankaran[4,5,6], Göran Karlsson[1,2], Unnur Thorsteinsdottir[3], Jonas Larsson[1,2], Kari Stefansson[3], Björn Nilsson[1,2,6]

[1]Lund Stem Cell Center, Lund University, 221 84 Lund, Sweden. [2]Department of Laboratory Medicine, Lund University, 221 84 Lund, Sweden. [3]deCODE genetics/Amgen Inc., 101 Reykjavik, Iceland. [4]Division of Hematology and Oncology, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02215, USA. [5]Harvard Stem Cell Institute, Cambridge, MA 02215, USA. [6]Broad Institute, Cambridge, MA 02142, USA.

[*] Equal contribution.

Correspondence: Björn Nilsson, M.D. Ph.D., Hematology and Transfusion Medicine, Department of Laboratory Medicine, BMC B13, 221 84 Lund, Sweden; e-mail: bjorn.nilsson@med.lu.se, tel +46-46-2220738, fax +46-46-130064.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Keywords: hematopoietic stem and progenitor cells, genome-wide association study, stem cell transplantation.

No. of words: 3,226 in main text, 250 in abstract, 3,747 in methods.

No. of display items: 5 figures.

No. of supplements: 19 figures, 11 tables.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABSTRACT

Understanding how hematopoietic stem and progenitor cells (HSPCs) are regulated is of central importance for the development of new therapies for blood disorders and stem cell transplantation. To date, HSPC regulation has been extensively studied *in vitro* and in animal models, but less is known about the mechanisms *in vivo* in humans. Here, in a genome-wide association study on 13,167 individuals, we identify 9 significant and 2 suggestive DNA sequence variants that influence HSPC ($CD34^+$) levels in human blood. The identified loci associate with blood disorders, harbor known and novel HSPC genes, and affect gene expression in HSPCs. Interestingly, our strongest association maps to the *PPM1H* gene, encoding an evolutionarily conserved serine/threonine phosphatase never previously implicated in stem cell biology. *PPM1H* is expressed in HSPCs, and the allele that confers higher blood $CD34^+$ cell levels downregulates *PPM1H*. By functional fine-mapping, we find that this downregulation is caused by the variant rs772557-A, which abrogates a MYB transcription factor binding site in *PPM1H* intron 1 that is active in specific HSPC subpopulations, including hematopoietic stem cells, and interacts with the promoter by chromatin looping. Furthermore, rs772557-A selectively increases HSPC subpopulations in which the MYB site is active, and *PPM1H* shRNA-knockdown increases $CD34^+$ and $CD34^+90^+$ cell proportions in umbilical cord blood assays. Our findings represent the first large-scale association study on a stem cell trait, illuminating HSPC regulation *in vivo* in humans, and identifying *PPM1H* as a novel inhibition target that can potentially be utilized clinically to facilitate stem cell harvesting for transplantation.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Main text

Humans blood cells originate from HSPCs[1]. Although HSPCs primarily reside in the bone marrow, they continuously egress into the blood, where they constitute a tiny (~0.1%) subset of mononuclear white blood cells. Circulating HSPCs express the surface protein CD34, allowing quantification by flow cytometry (**Supplementary Fig. 1** and **2**)[2]. Epidemiological studies[3,4] show that the level of circulating $CD34^+$ cells varies between individuals, but remains relatively stable within individuals, pointing at a genetic component. However, the underlying genes and DNA sequence variants remain unknown[3].

From a clinical viewpoint, novel HSPC regulators are highly sought-after to improve the treatment of blood disorders. Firstly, stem cell transplantation is a cornerstone in the treatment of blood cell malignancies. Today, the preferred way of harvesting stem cells for transplantation is by leukapheresis of peripheral blood. This requires that the $CD34^+$ cell level in the donor's blood is sufficiently high, and existing ways to mobilize $CD34^+$ cells from the bone marrow into the blood are not always effective[5,6]. Finding genes that regulate $CD34^+$ levels can lead to new ways to mobilize HPSCs. Secondly, several hematologic malignancies, including acute myeloid leukemia (AML)[7,8], myelodysplastic syndrome (MDS)[9,10], and myeloproliferative neoplasms (MPN)[11] are caused by abnormal HSPC activity. Uncovering the regulatory circuits that control HSPCs can lead to novel anti-leukemic therapies.

Genome-wide association study

To search for regulators of blood $CD34^+$ cell levels *in vivo* in humans, we carried out a genome-wide association study (GWAS) in 13,167 individuals from southern Sweden (ages 18 to 71 years; **Supplementary Table 1** and **Supplementary Fig. 3**). To quantify blood $CD34^+$ cells at a large scale, we developed a high-throughput flow-cytometry workflow, and

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pattern recognition software for computer-assisted analysis of the large volumes of flow-cytometry data (**Online Methods**). In each sample, we analyzed up to 1 million white blood cells (**Supplementary Table 2**), and defined the $CD34^+$ cell level as the number of $CD34^+$ cells divided by the number of $CD45^+$ mononuclear cells (**Supplementary Fig. 1**). To assess reproducibility, we sampled 660 individuals twice, with 3 to 36 months between samplings, and found a strongly significant correlation between replicates (Spearman $P = 1.3\times10^{-84}$, $r^2 = 0.44$; **Supplementary Fig. 4**). We observed higher blood $CD34^+$ cell levels in males than in females, but no differences between age groups (**Supplementary Fig. 5**).

Participants were genotyped on single-nucleotide polymorphism (SNP) microarrays, and imputed with reference whole-genome sequencing data to a final resolution of 18 million SNPs and small insertions-deletions (INDELs). For association analysis, we split our data into a discovery set of 10,949 individuals of Swedish ancestry and a follow-up set of 2,218 non-Swedish Caucasians using principal component analysis of the genotype data. To correct for multiple testing, we partitioned variants into five classes based on genomic annotations and applied weighted Bonferroni adjustment, taking into account the predicted functional impact of variants within each class (**Online Methods**)[12].

In combined analysis of the two data sets, 6 loci reached significance and 2 were suggestive (within one order of magnitude from Bonferroni thresholds). Conditional analysis uncovered four independent signals at 2q22, and no underlying signals at the other loci (**Fig. 1)**. No significant heterogeneity was detected in the effect estimates between the discovery and follow-up set (**Supplementary Table 3**). The variance explained by the 9 significant variants was 4.6%. Using linkage disequilbrium (LD) score regression, we estimated the total heritability at 12.7%, which is comparable to other blood cell traits (*e.g.*, mature white blood cell counts)[13–16]. Moreover, we detected enrichments of heritability in genomic regions with accessible chromatin in HSPC subpopulations[17,18] (**Fig. 2a**), including hematopoietic stem

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cells (HSC), multi-potent progenitors (MPP), and common myeloid progenitors (CMP), indicating that our analysis preferentially identifies variants that act by altering the regulation of gene expression intrinsically in HSPCs.

Identification of candidate genes

To identify candidate genes based on HSPC-intrinsic gene-regulatory effects, we generated expression quantitative locus (eQTL) data for sorted $CD34^+$ cells from 155 blood donors by mRNA-sequencing (average 122,000 cells per sample). Additionally, to identify regulatory links between variants and genes that are not detectable at the mRNA level in peripheral blood, we retrieved promoter capture Hi-C (PCHi-C) data[16] for $CD34^+$ cells and ATAC-sequencing data for 19 sorted blood cell populations, including 7 HSPC populations[11]. We then defined the 99% credible sets of probable causal variants (**Supplementary Table 4**), and prioritized genes as candidate genes if they: (i) had a non-synonymous coding variant within the credible set, (ii) had a *cis*-eQTL in $CD34^+$ cells (**Fig. 2b** and **Supplementary Table 5**), or (iii) the credible set contained a regulatory variant that maps either to the promoter region, or to a region with a chromatin looping interaction with the promoter in $CD34^+$ cells, as determined by PCHi-C[19] (**Supplementary Fig. 6**). As regulatory variants, we considered variants in genomic regions with accessible chromatin in HSPCs, as determined by ATAC-sequencing[13] (**Supplementary Table 4**). If none of these criteria were fulfilled, we prioritized the closest gene. Using these criteria, we identified 7 candidate genes at the significant loci, including two known HSPC-relevant genes (*CXCR4, CEBPA*) and five genes that have never previously been implicated in hematopoietic stem cell biology (*PPM1H, ENO1, RERE, ARHGAP45)* or only studied minimally in this area (*ITGA9*)[20] (**Fig. 1**).

Genetic overlap with blood disorders and other blood cell traits

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To investigate further their impact, we asked if the variants that influence blood $CD34^+$ cell levels also associate with other traits and diseases. To address this, we searched for coincident associations among variants in high LD ($r^2 > 0.8$) with the lead variants using Phenoscanner[21,22]. Consistent with HSPCs producing blood and immune cells, we detected coincident associations with variants known to influence mature blood cell traits (6 significant signals, at 12q14/*PPM1H*, 1p36/*ENO1-RERE*, 19p13/*ARHGAP45*, 19q13/*CEBPA* and two signals at 2q22/*CXCR4*), as well as with hematologic malignancies and autoimmune disorders (2 signals, at 2p22/*CXCR4* and 19p13/*CEBPA*) (**Supplementary Table 6**).

Of note, some of the candidate genes underlie autosomal-dominant blood disorders characterized by aberrant HSPC function. Gain-of-function mutations in *CXCR4* cause the Warts, Hypogammaglobulinemia, Immunodeficiency and Myelokathexis (WHIM) syndrome, marked by impaired egression of HSPCs and other white blood cells out of the bone marrow[23,24]. *CEBPA* mutations cause familial acute myeloid leukemia[25]. *TERT* mutations cause dyskeratosis congenital, where impaired telomere maintenance leads to problems with HSPC regeneration and increased risk of MDS[26]. Furthermore, somatic acquired mutations in *CXCR4*[27], *CEBPA*[9,25], and *TERT*[11] have been reported in several hematologic malignancies.

Gene expression in human hematopoiesis

We next explored the expression of the candidate genes across hematopoietic cell types. First, in single-cell mRNA-sequencing data for 35,582 mononuclear blood cells from adult blood and bone marrow[28], we observed significant enrichment of expression in HSPC *vs* non-HSPC populations (Wilcoxon $P = 2.1\times10^{-10}$ for the candidate genes at the significant loci; **Fig. 2c** and **Supplementary Fig. 7**). The same observation was made in bulk mRNA-sequencing data for sorted blood cell populations (Wilcoxon $P = 7.1\times10^{-5}$; **Supplementary Fig. 8**).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To map gene expression within the CD34$^+$ compartment in better detail, we analyzed single-cell Cellular Indexing of Transcriptomes and Epitopes by sequencing (CITE-seq) data for 4,905 lineage-negative CD34$^+$ bone marrow cells (Sommarin *et al.*, in prep). This revealed distinct expression biases for *PPM1H*, *ITGA9*, *ENO1* and *RERE* (bias towards HSC, MPP, CMP and MEP), *CXCR4* and *ARHGAP45* (lymphoid bias), and *CEBPA* (GMP bias) (**Fig. 2d,e** and **Supplementary Fig. 9** and **10**). These data further support that the identified genes are relevant to HPSCs.

Associations with *CXCR4*

Strikingly, four of our most significant signals map to *CXCR4* (C-X-C chemokine receptor type 4) at 2q22 (**Fig. 1**). This receptor binds stromal-derived-factor-1 (SDF-1; also called CXCL12). Among its functions, the CXCR4/SDF-1 axis regulates HSPC and immune cell migration. Particularly, internalization of CXCR4 is required for HSPC egression[29]. In the WHIM syndrome, gain-of-function mutations in the C-terminal region result in an inability to internalize CXCR4 after stimulation, leading to retention of HSPCs and other white blood cells in the bone marrow and low white blood cell counts in blood[30]. CXCR4 inhibitors (Plerixafor/AMD3100) are one of the current methods to mobilize CD34$^+$ cells in stem cell donors for leukapheresis[31]. Thus, the fact that we find associations with *CXCR4* provides compelling proof-of-principle for the idea that bona fide regulators of blood CD34$^+$ cell levels can be found *in vivo* in humans by GWAS.

The four 2q22 signals represent three common (rs309137, rs11688530, rs10193623) and one rare variant (rs555647251), with a total of 69 variants within their 99% credible sets of probable causal variants (**Supplementary Table 4**). Within each of these sets, we identified a single regulatory variant that either maps to, or has a looping interaction with, the *CXCR4* promoter, making them plausible causal variants (**Fig. 3b**, **Supplementary Fig. 6a**,

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and **Supplementary Table 4**). In the credible sets of rs11688530 and rs555647251, we identified rs59222832 and rs770321415 as regulatory promoter variants, located 4.7 and 1.3 kb upstream of the transcription start site, respectively. In the other two credible sets, we identified the lead variants rs309137 and rs10193623 as regulatory variants with chromatin looping interactions with the *CXCR4* promoter in $CD34^+$ cells[19].

We asked if the 2p22 associations are caused by altered *CXCR4* expression in connection to egression, similar to the gain-of-function mutations in WHIM syndrome. In agreement with this hypothesis, multi-variate analysis of our $CD34^+$ cell mRNA-sequencing data for blood donors unveiled conditional *CXCR4 cis*-eQTLs for the three common variants (**Fig. 3c**), while the rare variant was not polymorphic in this data set. Notably, in all three cases, the effect on *CXCR4* expression was anti-directional to the effects on blood $CD34^+$ cell levels, consistent with CXCR4 being a negative regulator of egression. For further functional analysis, we used dual-sgRNA CRISPR/Cas9 to delete 587 to 1421-bp regions harboring each of the four putative causal variants in the acute myeloid leukemia cell line MOLM-13 (**Supplementary Table 7**; **Supplementary Fig. 11** and **12**). In all four cases, we observed downregulation of *CXCR4* (**Fig. 3d**). These data indicate that the 2q22 associations are caused by sequence variation in genomic regions that regulate *CXCR4* expression in HSPCs.

Association with *PPM1H*

Our most significant association signal maps to *PPM1H* (protein phosphatase, Mg^{2+}/Mn^{2+} dependent 1H) at 12q14. This gene encodes an evolutionarily conserved[32]serine/threonine phosphatase in the phosphatase 2C (PP2C) family[33]. Its biological role is unknown, and the few studies that have been done suggest that *PPM1H* could be involved in cellular signaling[34–36], trastuzumab resistance[37], lupus[38], and colon cancer[39].

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The 12q14 association is represented by 32 credible set variants in *PPM1H* intron 1 (**Fig. 4a**, **Supplementary Fig. 6b** and **Supplementary Table 4**). Consistent with the *PPM1H* *cis*-eQTL in our blood donor $CD34^+$ cell mRNA-sequencing data (**Fig. 2b**), the associated region shows chromatin looping interactions in $CD34^+$ cells, both with the standard *PPM1H* promoter and internal promoter (**Fig. 4a** and **Supplementary Fig. 6b**). Moreover, by integrating ATAC- and mRNA-sequencing data for 16 sorted blood cell populations[13], we discovered an approximately 500-bp-long chromosomal segment within the associated region where chromatin accessibility shows strong positive correlation with *PPM1H* expression, suggesting a regulatory role (**Fig. 4b**). This segment harbors four of the 32 credible set variants (rs772555, rs772556, rs772557, rs772559). Congruent with the expression pattern of *PPM1H* across hematopoietic cell types (**Fig. 2c-e** and **Supplementary Fig. 7** to **10**), the segment is selectively accessible in HSCs, MPPs, CMPs, and MEPs (**Fig. 4b**).

We hypothesized that the 12q14 association is caused by one of the four variants in the identified regulatory segment. For functional fine-mapping, we carried out luciferase experiments with constructs representing their reference and alternative alleles in the acute erythroleukemia cell line K562. We observed significantly higher luciferase signal with rs772557-G constructs compared with rs772557-A (**Fig. 4c**), consistent with the direction of the *PPM1H* *cis*-eQTL (**Fig. 2b**; rs772557[A>G] is in LD with the 12q14 lead variant rs669585[T>G] with $r^2 = 0.98$). Additionally, we observed higher chromatin accessibility at rs772557-G than at rs772557-A in ENCODE[40] DNAse-sequencing data for heterozygous primary adult $CD34^+$ cells (125 versus 47 reads; Binomial test $P = 2.0\times10^{-8}$).

Motif analysis predicted that rs772557 alters a binding site for the transcription factor MYB by altering a critical recognition base (**Fig. 4d** and **Supplementary Table 8**). Consistent with the *cis*-eQTL and luciferase data, the rs772557-G allele (which confers high *PPM1H* expression) creates the site, whereas rs772557-A abrogates it. This mechanism-of-

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

action was also supported by chromatin immunoprecipitation sequencing (ChIP-seq) data for MYB in Jurkat cells (heterozygous for rs772557), showing exclusive pull-down of reads harboring the rs772557-G allele (**Fig. 4e** and **Supplementary Fig. 13**). Moreover, luciferase experiments with co-transfected MYB siRNA showed selective attenuation of luciferase signal from rs772557-G but had no impact on rs772557-A, further supporting that MYB only drives transcription at rs772557 in the presence of the high-expressing allele (**Fig. 4f**).

To demonstrate causality, we perturbed rs772557 by CRISPR/Cas9 in K562 cells. These cells are triploid at *PPM1H*, having two copies of the rs772557-G allele and one copy of the rs772557-A allele. Serendipitously, we identified sgRNA sequences that overlap rs772557 and enable allele-specific disruption. These sgRNAs cut DNA only one bp upstream of rs772557, and within the MYB recognition site (**Supplementary Fig. 14** and **Supplementary Table 7**). Consistent with the other data, we observed downregulation of *PPM1H* with rs772557-G sgRNA, but no effect of rs772557-A sgRNA (**Fig. 4g**). Finally, we observed co-expression of *MYB* and *PPM1H* in hematopoietic cell types (**Fig. 4h**), and in our $CD34^+$ mRNA-sequencing data for blood donors, we found a highly significant correlation between *MYB* and *PPM1H* among rs772557-G carriers (**Fig. 4i**), but no correlation among rs772557-A homozygotes (**Fig. 4j**). These data identify rs772557 as a causal variant.

Because of the anti-correlation between *PPM1H* expression and blood $CD34^+$ cell levels (**Fig. 2b**), we hypothesized that *PPM1H* downregulation by rs772557-A increases the proportion of HPSC subpopulations in which rs772557 is accessible (**Fig. 4b**). To test this, we first calculated the correlation between rs772557 genotype and gene expression in our $CD34^+$ cell mRNA-sequencing data, and tested for enrichment of correlation among genes that are highly expressed in each $CD34^+$ subpopulation (**Supplementary Fig. 15**). We observed enrichments of positive correlations in the direction of rs772557-A for CMP, HSC, MEP, and MPP marker genes, and enrichments of negative correlations among marker genes

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

for lymphoid precursors (**Fig. 5a** and **Supplementary Fig. 15**). Second, we quantified 8 HSPC subpopulations in 642 umbilical cord blood samples (**Supplementary Fig. 16**). Consistent with the enrichment analysis, we observed associations between rs772552-A and increased proportion of CMPs and a lower proportion of B/NK progenitors (**Fig. 5b**). Third, also consistent with the enrichment analysis, shRNA-knockdown of *PPM1H* in primary $CD34^+$ cord blood cells (**Supplementary Fig. 17**), increased the proportion of $CD34^+$ and $CD34^+90^+$ HSPCs relative to control shRNA (**Supplementary Fig. 17** and **Fig. 5c**). In aggregate, our functional analysis of the 12q14 signal indicates that this association is driven by rs772557, with the rs772557-A allele abrogating a MYB binding site, leading to *PPM1H* downregulation and a selective increase of HSPC subpopulations in which rs772557 is active.

Additional associations

Among the remaining significant associations, 1p36/*ENO1-RERE* and 3p22/*ITGA9* displayed strong *cis*-eQTLs and preferential expression of candidate genes in $CD34^+$ cells (**Fig. 2b-c**). First, *ENO1* encodes alpha-enolase, a glycolytic enzyme. However, alternative translation also produces a shorter isoform that binds the *MYC* promoter and has been reported to act as a tumor suppressor[41]. *RERE* encodes a transcription factor that forms a complex with the retinoic acid receptor and increases transcription of its target genes[42]. Retinoic acid signalling has been linked to HSC self-renewal in mice[43]. In acute promyelocytic leukemia (APL), genetically perturbed retinoic acid signalling blocks HSPCs from differentiating into mature white blood cells; this block can be overcome by treatment with all-trans-retinoic acid[44]. Concordant with the *cis*-eQTLs, we identified putative causal regulatory variants in the region between *ENO1* and *RERE*, showing chromatin looping interaction with *ENO1* and *RERE* promoters (**Supplementary Fig. 6c** and **Supplementary Table 4**). Second, *ITGA9* encodes the cell surface protein integrin alpha-9. Although one previous study[20] has linked

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ITGA9 expression to HSPCs, its precise function in human hematopoiesis remains unclear. Our analysis now identifies ITGA9 as a functional cell surface marker that participates in the regulation of blood $CD34^+$ cell levels. The associated region maps to *ITGA9* introns 3 and 4, and several regulatory variants show looping interactions with the promoter (**Supplementary Fig. 6d** and **Supplementary Table 4**). In addition to the *cis*-eQTL data (**Fig. 2b**), flow cytometry analysis in 458 adults confirmed that the 3p22 variant impacts ITGA9 protein expression on blood $CD34^+$ cells (**Supplementary Fig. 18**).

Finally, among the suggestive associations, we particularly noted the 8q24 locus. While this signal primary maps to *CCDC26*, PCHi-C analysis in blood $CD34^+$ cells revealed a long-distance looping interaction with the *MYC* promoter, located 1.86 Mb away (**Supplementary Fig. 6h**). The 8q24 signal is represented by 31 variants located between position 129,590,035 and 129,612,415 on chromosome 8 (**Supplementary Table 4**). This region was recently identified as an evolutionarily conserved "super-enhancer" required for *Myc* expression in mouse HSPCs[45]. The super-enhancer comprises eight enhancer modules (A to H; **Supplementary Table 9**), collectively called the Blood ENhancer Cluster (BENC). The 8p24 signal spans BENC module D. Deletion of this module in mice has been found to affect *Myc* expression mainly in HSCs and MPP cells[45]. Our data now identify BENC module D as critical for the regulation of blood $CD34^+$ cell levels in humans.

Discussion

In this work, we report the first large-scale genome-wide association study on a stem cell trait. To achieve this, we created a high-throughput flow-cytometry workflow, and custom pattern recognition software for data analysis. Hereby, we were able to quantify blood $CD34^+$ cell levels in unprecedented numbers of individuals.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Unlike studies in model systems, we exploit natural genetic variation to expose DNA HSPC regulators *in vivo* in humans. While there are many aspects of HSPC activity, searching for regulators of blood $CD34^+$ cell levels is motivated by the clinical importance of this parameter in stem cell harvesting. Moreover, mapping the genetic architecture of blood $CD34^+$ cell levels may expose genetic factors that influence several phenomena in stem cell biology, including HSPC pool size, migration, and early differentiation[1,46,47].

The validity of our approach is confirmed by several observations, including enrichment of heritability and gene expression in HSPCs, and discovery of variants at loci that are known to be HSPC-relevant, most notably multiple variants at *CXCR4*. Furthermore, we identify four novel regulators (*PPM1H*, *ENO1*, *RERE* and *ITGA9*), which we identify as target genes based on *cis*-eQTL, chromatin looping, and functional data. Dissecting our most significant association, we identify an anti-correlation between *PPM1H* expression and $CD34^+$ cell levels. This indicates that PPM1H could be exploitable clinically as an inhibition target to facilitate harvesting by leukapheresis. Accordingly, intriguing challenges ahead will be to map the signalling networks within the PPM1H pathway, and to search for PPM1H inhibitors. Towards this, a first step will be to map PPM1H dephosphorylation targets. While initial proteomic studies in non-hematopoietic cells have identified RAB8A, RAB10, RAB35, SMAD1/2 as tentative targets[34,36], the dephosphorylation network in HSPCs remains unexplored. Other notable findings include ENO1 as a regulatory enzyme with links to the MYC pathway, RERE as a transcription factor with links to retinoic acid signalling, and *ITGA9* as a cell surface marker with a functional role in the regulation of blood $CD34^+$ levels.

In conclusion, we report the first large-scale association study on a stem cell trait. We demonstrate that HSPC regulators can be exposed through natural genetic variation, and identify novel regulatory pathways that can be investigated further with the ultimate aim of improving the treatment of blood disorders.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary information

Supplementary information is available in the online version of the paper.

Acknowledgements

This work was supported by grants from the European Research Council (ERC 770992-BloodVariome), Knut and Alice Wallenberg Foundation (2012.0193, 2014.0071 and 2017.0436), the Swedish Research Council (2017-02023 and 2018-00424), the Swedish Cancer Society (2017/265 and 20.0694), the Swedish Children's Cancer Fund (PR2018-0118 and TJ2017-0042), and Inga-Britt and Arne Lundberg's Foundation (2017-0055). We thank Ellinor Johnsson, Christina Jönsson, and Camilla Streimer for their assistance. We are indebted to the personnel at the Clinical Chemistry and Clinical Immunology and Transfusion Medicine units in Region Skåne for their help with sample collection, and to the patients and blood donors who participated in the study.

Author contributions

A.L.d.L.P, U.T., J.L., K.S., and B.N. conceived the study. A.L.d.L.P, L.E., C.C., Z.A., M.P., K.Ž., U.T., J.L., K.S., and B.N. designed experiments. A.L.d.L.P., C.C., and N.M. developed the flow cytometry platform. L.E. and A.L.A. developed software for flow cytometry data analysis. A.L.d.L.P., L.E., C.C., Z.A., N.M., N.U.D., and D.B. carried out large-scale phenotyping. A.L.d.L.P., C.C., G.T., M.P., U.T. and K.S. carried out genotyping and genomic analyses. C.C., Z.A. N.M., and K.Ž. carried out functional experiments. A.L.d.L.P., L.E., C.C., Z.A., N.M., G.T., K.Ž., A.L.A., M.T., M.P., P.D., E.B., A.N., V.S., G.K., and B.N. carried out bioinformatic and statistical analyses. A.L.d.L.P., L.E., C.C., Z.A. and B.N. drafted the manuscript. All authors contributed to the final manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Conflicts-of-interest

The authors have no relevant conflicts-of-interest. G.T., G.N., U.T., and K.S. are employed by deCODE Genetics/Amgen Inc.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure legends

Figure 1: Genome-wide association study. Sequence variants influencing blood $CD34^+$ cell levels identified in combined analysis of association data for 10,949 Swedes and 2,218 Caucasian non-Swedes. We identified 9 significant and 2 suggestive (*) associations (**Supplementary Table 3**). The listed variants are the most significant (lead) variants for each association. We prioritized genes as candidate genes if they: (i) had a coding variant within the 99% credible set of probable causal variants (**Supplementary Table 4**); (ii) had a *cis*-eQTL in $CD34^+$ cells from 155 blood donors (**Supplementary Table 5**); or (iii) the credible set contained a regulatory variant that maps either to the promoter, or to a region with a chromatin looping interaction with the promoter in $CD34^+$ cells, as determined by PCHi-C. As regulatory variants, we considered variants in genomic regions whose chromatin is accessible in HSPCs, as determined by ATAC-sequencing. If none of these criteria were fulfilled, we prioritized the closest gene. The criterion used to call each gene a candidate gene is indicated in the matrix.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2: Effects on gene expression in HSPCs. (a) LD score regression shows enrichments of heritability in regions with accessible chromatin in HSPC subpopulations. **(b)** To identify candidate genes with HSPC-intrinsic gene-regulatory effects, we generated eQTL data for sorted $CD34^+$ cells from 155 blood donors. These figures illustrate strong *cis*-eQTLs identified at *PPM1H*, *ENO1*, *RERE* and *ITGA9* (**Supplementary Table 5**). Data are residual FPKM values after correction for 10 expression principal components. Wedges indicate directions of effects on blood $CD34^+$ cell levels for the same variant. Notably, we detected an anti-correlation between *PPM1H* expression and $CD34^+$ levels for the 12q14 variant. **(c)** Candidate gene expression in scRNA-seq data from blood and bone marrow mononuclear cells[28], showing enriched expression in HSPC populations (**Supplementary Fig. 7** and **8**). **(c,d)** To map expression within the $CD34^+$ compartment in better detail, we analyzed CITE-seq (*i.e.*, mRNA-sequencing with antibody-derived tags) data for 4,905 lineage-negative $CD34^+$ cells from adult bone marrow: **(d)** bulked expression in cell clusters inferred from mRNA levels; **(e)** bulked expression in clusters inferred from antibody-derived tags representing classical HSPC surface markers (**Supplementary Fig. 9** and **10**). Abbreviations: Hematopoietic stem cells (HSC), multi-potent progenitors (MPP), common myeloid progenitors (CMP), granulocyte-monocyte progenitors (GMP), common lymphoid progenitors CLP), lymphoid-primed multipotent progenitors (LMPP), erythroid progenitors (ERY), megakaryocyte-erythrocyte progenitors (MEP), mast cell/basophil progenitors, (MB), dendritic cells (DC), plasma cells (PC), $CD4^+$ T-cells (CD4), $CD8^+$ T-cells (CD8), B-cells (B), pre B-cells (PreB), lymphoid progenitors (Ly), natural killer cells (NK), basophil (Baso), neutrophil (Neut), monocyte (Mono), cycling cells (Cyc).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3: Associations with *CXCR4*. (a) We detected four conditionally independent associations at 2p22, represented by a total of 69 credible set variants clustered around *CXCR4* (**Supplementary Table 4**). The credible sets are indicated in red (lead variant rs309137), green (rs11688530), cyan (rs555647251) and blue (rs10193623). The rs309137, rs11688530 and rs10193623 credible sets represent common variants, whereas rs555647251 represents a rare variant. By integrating ATAC-sequencing and PCHi-C data for $CD34^+$ cells, we identified a single plausible causal variant within each credible set. rs309137 ("V1") and rs10193623 ("V4") have chromatin looping interactions with the *CXCR4* promoter (red and blue arches; *y*-axis indicates PCHi-C *P*-score). rs59222832 ("V2"; credible set of rs11688530), and rs770321415 ("V3"; credible set of rs555647251) map to the *CXCR4* promoter. **(b)** Chromatin accessibility at the four plausible causal variants; *y*-axis indicates ATAC-sequencing signal. **(c)** Using multi-variate regression, we detected conditional *CXCR4* *cis*-eQTLs for the three common variants in our $CD34^+$ cell mRNA-sequencing data from blood donors. Wedges indicate directions of effects on blood $CD34^+$ cell levels. Of note, the effects of these three variants on *CXCR4* expression are anti-directional to their effects on blood $CD34^+$ cell levels. Data are residual FPKM values after correction for covariate SNPs and 10 principal components. **(e)** Using dual-sgRNA CRISPR/Cas9, we deleted 587 to 1421-bp regions harboring the four putative causal variants in MOLM-13 cells, resulting in downregulation of *CXCR4*.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 4: Association with *PPM1H*. (a) Top: the 12q14 signal is represented by a credible set of 32 variants in *PPM1H* intron 1. Middle: chromatin looping interactions in $CD34^+$ cells with standard and internal promoter (red arches; *y*-axis indicates PCHi-C *P*-score). Bottom: we identified an approximately 500-bp-long chromosomal segment (red peak) where ATAC-sequencing signal (100-bp sliding window) shows strong positive correlation with *PPM1H* expression across 16 sorted blood cell populations[13] (*y*-axis indicates false discovery rate for Pearson correlation). **(b)** Four credible set variants map to the identified regulatory segment that is accessible in HSC, MPP, CMP and MEPs (*y*-axis indicates ATAC-seq signal). **(c)** Luciferase analysis revealed higher activity with rs772557-G compared to rs772557-A constructs in the *cis*-eQTL direction (*P*-value for one-sided Student's t-test; **Fig. 2b**). Signals normalized to hg38 reference (left) alleles. **(d)** MYB binding site that is altered by rs772557; rs772557-G creates binding site, while rs772557-A abrogates it, by changing a critical recognition base (arrow). **(e)** ChIP-seq data for MYB in Jurkat cells (rs772557-heterozygous) show exclusive pull-down of reads harboring rs772557-G. **(f)** siRNA knockdown of MYB in K562 cells selectively attenuates rs772557-G luciferase activity. **(g)** Allele-specific CRISPR-Cas9 disruption at rs772557 (**Supplementary Fig. 14**) in K562 cells, heterozygous for rs772557. We observed *PPM1H* downregulation with rs772557-G, but not with rs772557-A, sgRNA. **(h)** *PPM1H* and *MYB* are co-expressed in hematopoiesis[28]. **(i,j)** Analysis of our $CD34^+$ mRNA-sequencing data for blood donors revealed a correlation between *MYB* and *PPM1H* expression in rs772557-G carriers, and no correlation in non-carriers. Data are $\log_2$-transformed FPKM values, median-centered per genotype group.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 5: Effects of *PPM1H* downregulation. Since our *cis*-eQTL analysis identified an anti-correlation between blood $CD34^+$ cell levels and *PPM1H* expression (**Fig. 2**), we further explored the effects of *PPM1H* downregulation on human hematopoiesis. **(a)** We first searched for effects of rs772557 on cell type composition within the $CD34^+$ compartment in adults. For this, we calculated correlations between rs772557 genotype and gene expression in our $CD34^+$ mRNA-sequencing data for blood donors, and tested for enrichment of correlation within sets of marker genes for HPSC subpopulations (**Supplementary Fig. 15**). This composite plot shows the distributions of Pearson correlation coefficients for the top 250 marker genes inferred using mRNA-sequencing data for sorted cells[13]. We detected enrichments of positive correlations in the direction of rs772557-A allele for CMP, HSC, MEP, and MPP (red), and of negative correlations for CLP and LMPP (blue), compared to other genes in the genome (black) (details in **Supplementary Fig. 15**). This finding is consistent with rs772557-A increasing the relative abundance of $CD34^+$ subpopulations in which the rs772557 region is accessible (**Fig. 4b**). **(b)** Quantifying HSPC subpopulations in 642 umbilical cord blood samples (**Supplementary Fig. 16**), we observed association between rs772552-A and increased proportion of CMP and lower proportion of B/NK progenitors. **(c)** shRNA-knockdown of *PPM1H* induced an increase in the proportions of $CD34^+$ and $CD34^+90^+$ primary cord blood cells out of green fluorescent protein (GFP)-positive. Data are proportion at day 7, 14 and 21 after transduction, normalized to shRNA-control. *P*-value is for permutation testing, taking into account the structure of the experimental design (**Online Methods**).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Online methods

Subjects

For the genome-wide association study, we collected 16,931 peripheral blood samples from random blood donors (n=7,773) and primary care patients (n=9,158) during three time periods (November 2015 to April 2016, "Phase I", January 2017 to November 2017, "Phase II", and August 2018 to April 2019 "Phase III"; **Supplementary Table 1**). Samples from blood donors were obtained from Clinical Immunology and Transfusion Medicine, Skåne University Hospital, Lund, Sweden. By Swedish law, men are allowed to donate blood every 3 months, women every 4 months. Hence, by collecting samples for slightly longer periods of time, we ensured collection of repeat donors samples, allowing us to assess reproducibility (**Supplementary Fig. 4**). Samples from primary care patients were surplus material from the clinical routine blood chemistry pipeline at Clinical Chemistry, Skåne University Hospital, Lund, Sweden. By programming the sample processing robot at Clinical Chemistry, samples from patients between 18 and 71 years-of-age from primary care clinics in the Lund area were selected automatically. Making use of the fact that the processing robot had a 100,000-sample history buffer, we minimized the number repeat samples from primary care patients. Samples were collected subject to ethical approval (Lund University Ethical Review Board; dnr 2018/2). Except for information about age and sex (**Supplementary Fig. 3**), the samples were irreversibly anonymized by clinical personnel before being forwarded to us for analysis. Following genotype and phenotype data quality control, and removal of duplicate individuals, a total of 13,167 unique individuals remained (**Supplementary Table 1**). Out of the 9 significant variants, only the rare variant rs555647251 showed Bonferroni-significant heterogeneity between blood donors and primary care patients (**Supplementary Table 10**).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Flow-cytometry phenotyping for the association study

Collected samples were first diluted 1:10 in 0.84% ammonium chloride and incubated at room temperature for 10 min for erythrocyte lysis. After centrifugation at 1,200 g x 5 min and supernatant removal, leukocytes were resuspended in 20 mL of wash buffer (PBS with 6 mM EDTA) and centrifuged at 1,200 g x 5 min. Washing was done twice. After supernatant removal, cells were resuspended in residual wash buffer, and 60 uL of cell suspension were plated in 96-well plates and centrifuged at 1,200 g x 1 min. Supernatant was removed and pelleted cells were resuspended in 30 uL of staining buffer (PBS containing 6 mM EDTA and 0.1% BSA) including 0.2 µL of APC-H7 mouse anti-human CD45 (clone 2D1; BD #560178) and 0.09 µL of PE-CF594 mouse anti-human CD34, clone 563 (BD #562449), for 15 min at room temperature, in the dark. After a wash step with 100 µL of wash buffer, plates were centrifuged at 1200 g x 1 min, supernatant was removed and pelleted stained cells were re-suspended in 250 µL of staining buffer. Flow-cytometry analysis was performed using a BD FACS Canto II™ (Phase I), BD LSR Fortessa™ (Phase II), and BioRad ZE5™ (Phase III). The numbers of events recorded per sample are summarized in **Supplementary Table 2**.

Gating of flow-cytometry data for association study

To quantify blood $CD34^+$ cell levels, we first gated singlet cells based on forward scatter area (FSC-A) and forward scatter height (FSC-H). From singlets, we then gated peripheral blood mononuclear cells (PBMC) based on FSC-A and side scatter area (SSC-A). Finally, from PBMCs, we gated $CD34^+45^{low}$ cells and $CD45^+$ cells. In CD34 and CD45 intensity space, $CD34^+$ cells form a discrete cluster. We defined the $CD34^+$ level (frequency) as the number of $CD34^+45^{low}$ cells divided by the number of $CD45^+$ PBMCs (**Supplementary Fig. 1**).

To facilitate the analysis of our flow cytometry data, we developed pattern recognition software (https://github.com/LudvigEk/HSPC-regulators-in-human-blood) that

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mimics the manual gating strategy. First, to gate singlets, we used an ellipsoid boundary calculated using principal component analysis in FSC-A and FSC-H space (**Supplementary Fig. 1a**). Second, to gate PBMCs, we employed Dijkstraa's shortest path algorithm to infer an optimal FSC-A-dependent SSC-A boundary between PBMCs and granulocytes by searching for the lowest total event density path from the left edge (minimal FSC-A) to right edge (maximal FSC-A). Having set the Dijkstraa boundary, the definition of PBMCs was refined by overlaying an ellipsoid gate defined by principal component analysis of FSC-A and SSC-A values (**Supplementary Fig. 1b**). Third, in CD34 and CD45 space, debris and $CD45^+$ cells were separated by finding the CD45 intensity value corresponding to the lowest event density among $CD34^-$ events. Having found a CD45 threshold, the CD34 intensity standard deviation of the $CD45^+$ cluster was estimated. Using the estimated parameters, CD34 and CD45 intensity boundaries were inferred for the $CD34^+45^{low}$ cluster.

To validate our gating software, we compared blood $CD34^+$ levels obtained with computer gating to blood $CD34^+$ levels obtained with manual gating for 2,838 of our samples, and observed a highly significant correlation (Spearman correlation $P < 4.94\times10^{-324}$ and $r^2 = 0.836$; **Supplementary Fig. 19**). Phase I and II samples were gated using computer gating. For extra quality control, we plotted the computer gating results (all three gating steps) and visually inspected each of these plots to verify high-quality gating. Phase III samples were gated manually using Flojwo V10.6.1 (Becton, Dickinson & Company).

Genotyping and association analysis

The Swedish samples were genotyped with Illumina single-nucleotide polymorphism microarrays and phased together with 570,100 samples from North-Western Europe using Eagle2[48]. Samples and variants with <98% yield were excluded. We created a haplotype reference panel by phasing the whole-genome sequence (WGS) genotypes for 15,575

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

individuals from North-Western Europe, including 3,012 Swedish samples, together with the phased microarray data, and to impute the genotypes from the haplotype reference panel into the phased microarray data using methods described previously[49,50].

Genetic ancestry analysis was done in two stages for the Swedish sample sets. Firstly, ADMIXTURE v1.23[51] was run in supervised mode with 1,000 Genomes populations CEU, CHB, and YRI[52] as training samples and Swedish individuals as test samples. Input data for ADMIXTURE had long-range LD regions removed[53] and was then LD-pruned with PLINK v.190b3a[54] using the --indep-pairwise 200 25 0.3 option. Samples with < 0.9 CEU ancestry were excluded. Secondly, remaining samples were projected onto a principal component analysis (PCA), calculated with an in-house European reference panel to calculate the 20 first principal components for each population. UMAP[55] was used to reduce the coordinates of test samples on 20 principal components to two dimensions. Additional European samples not in the original reference set were also projected onto the PCA and UMAP components to identify ancestries, and samples with Swedish ancestry were identified. This included 10,949 unique individuals with blood $CD34^+$ cell level measurements. After excluding samples of Swedish ancestry, the process was repeated for the remaining samples with blood $CD34^+$ cell level measurements and 2,218 unique individuals with > 0.7 CEU ancestry were identified and principle components were created separately for those individuals.

We used linear regression to test for association between blood $CD34^+$ cell levels and genotypes for the discovery and follow-up sets separately using deCODE software[49]. Prior to association the blood $CD34^+$ level measurements were adjusted for gender and 20 principal components and standardized using an inverse normal transform. The association was restricted to 18,061,173 variants with MAF > 0.05%, imputation info > 0.9 and that passed various quality test. We used LD score regression to account for distribution inflation due to cryptic relatedness and population stratification[56].

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For the meta-analysis we used a fixed-effects inverse variance method to combine the results from the discovery and follow-up dataset[57]. We tested for heterogeneity by comparing the null hypothesis of the effect being the same in both sets to the alternative hypothesis of each set having a different effect using a likelihood ratio test (Cochran´s Q) reported as P_{het}. Genome-wide significance was determined using class-based Bonferroni significance thresholds adjusting for the 18 million variants tested. Sequence variants were split into classes based on their genome annotation, and the significance threshold for each class was based on the number of variants in that class[58]. The adjusted significance thresholds used are 4.5×10^{-7} for variants with high impact (including stop-gained and stop-loss, frameshift, splice acceptor or donor and initiator codon variants), 9.0×10^{-8} for variants with moderate impacts (including missense, splice-region variants, in-frame deletions and insertions), 8.2×10^{-9} for low-impact variants (including synonymous, 3′ and 5′ UTR, and upstream and downstream variants) and 1.4×10^{-9} for all other variants, including those in intergenic regions.

We used step-wise conditional analysis using genotype data to look for additional signals at each locus. The conditional analysis was done for the discovery and follow-up dataset separately, and the results combined. At each locus, we tested all variants in a 2 Mb region centred at the lead variant and significance threshold was set at 1×10^{-6} for a secondary signal. We calculated 99% credible sets of plausible causal variants for each independent association signal[59], weighting the variants using the same weights as were used to define the class specific genome-wide significance thresholds.

Chromatin accessibility data

Previously published ATAC-seq data for sorted hematopoietic cell types were downloaded from the Sequence Read Archive[17,18]. Reads were processed as the MM ATAC-seq libraries using the hg38 reference genome. Next, we created a master peak file by aggregating the

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

summits of each population and enumerating the fragments overlapping each peak for each population as previously described. In addition to the ATAC-seq data for sorted blood cells, we used DNA DNAse-sequencing data for heterozygous primary adult $CD34^+$ cellsfrom ENCODE Phase 3 (accession no. ENCSR098PTC, ENCFF971ZOL, ENCFF814EOK).

For MOLM-13 cells, we generated ATAC-seq libraries from 50,000 cells as described in ref[60]. Libraries were prepared using the Nextera DNA Library Prep kit and sequenced on an Illumina Novaseq 6000 Sequencing System with a read length of 50 bases in paired-end mode. Adapter sequences in sequence reads were removed using Trimmomatic (v0.36[61]) and aligned using Bowtie2 to hg38. Duplicate and mitochondrial reads were filtered out using SAMtools[62] and Picard (http://broadinstitute.github.io/picard).

ChIP-seq data

To test of allele-specific binding of MYB to rs772557, we used MYB ChIP-sequencing data for Jurkat cells[63]. Raw SRA files were obtained from the Sequence Read Archive (accession no. SRR1603653) and converted to FASTQ format using fastq-dump. The raw FASTQ file was aligned using bowtie2, and the resulting SAM file was sorted and indexed using samtools. For Jurkat WGS, a SAM file containing the region of interest was downloaded from the Sequence Read Archive (accession no. SRR5349449[64]).

Heritability estimation

Total SNP heritability was estimated with the LDSC software[56,65] with default settings, intercept 1.0, and Baseline model v1.1 as control. To estimate cell type-specific partitioned heritability based on chromatin accessibility, we used LD-scores based on ATAC-sequencing data for sorted blood cell types available for LDSC[56], extended with ATAC-sequencing data for mDC and pDC[18] (NCBI Gene Expression Omnibus accession no. GSE119453).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Promoter capture Hi-C data analysis

A table of significant PCHi-C contacts for CD34+ cells was obtained from the ArrayExpress repository (accession number: E-MTAB-2323[66]). These data had been previously processed using the GOTHiC Bioconductor package and filtered to interactions with FDR < 0.05 in two biological replicates. The log of the observed/expected ratio shown in locus plots is an estimate of effect size or interaction frequency.

Expression quantitative trait locus (eQTL) analysis in blood CD34$^+$ cells

To generate eQTL data for blood CD34$^+$ cells, we purified CD34$^+$ cells from 155 leukocytes depletion filters (Reveos^{TM}) from random blood donors within 8 h of donation. First, we pre-enriched CD34$^+$ cells using the MACSprep™ Chimerism CD34$^+$ MicroBead kit (Miltenyi Biotech; #130-120-673). The enriched cells were stained using 5 uL of APC-H7 mouse anti-human CD45, clone 2D1 (BD #560178) and 22 uL of PE-CF594 mouse anti-human CD34, clone 563 (BD #562449) for 15 min at room temperature. Cells were washed with 500 µL of wash buffer and centrifuged at 900 rpm g x 5 min, supernatant was removed and pelleted stained cells were re-suspended in 350 µL of staining buffer. CD34$^+$CD45^{dim} mononuclear cells were sorted using a BD FACSAria^{TM} III (BD Biosciences) into lysis buffer (Qiagen; RLT buffer #74004; containing 1% β-mercaptoethanol), vortexed, and snap frozen. From each sample, 14,000 to 358,000 cells (average 122,000) were sorted. The samples were thawed on ice and RNA was extracted using the RNeasy plus micro kit (Qiagen; 74034) and treated with RNAse-Free DNase Qiagen; #79256).

cDNA libraries were prepared using the SMARTer Stranded Total RNA-Seq Kit v2 Pico Input Kit (Clontech, Takara) and sequenced on an Illumina NovaSeq 6000 Sequencing System with a read length of 100 bases in paired-end mode. After trimming the first three

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bases using Trimmomatic v0.36[61], the sequenced reads were aligned to the human reference genome (human GRCh38 primary assembly) using STAR (v2.5.2b[67]). Fragments aligned to human genes were quantified using featureCounts[68], with the following settings: paired-end mode (-p), strand specific (-s 2), no multimapping reads counted, counting exonic reads. Raw gene counts were transformed to FPKM using in-house Python scripts.

To test for associations between variants and gene expression, we used linear modeling with the variant genotype is independent variable and 10 principal-component covariates, calculated using genes with average FPKM > 1.0 in our mRNA-sequencing data set. For *CXCR4*, we included the three common variants as independent variables in the same model. *P*-values were calculated using Student's *t*-test for the independent variables.

Gene expression data and analysis

To map candidate gene expression in hematopoietic cell types, we used published single-cell mRNA sequencing data for 35,582 mononuclear cells from blood and bone marrow[28], and bulk mRNA-sequencing data for sorted blood cell populations[18]. The definitions of the cell populations in these data sets are detailed in the original publications. Additionally, we used CITE-seq data on 4,905 $CD34^+$ cells from adult bone marrow. Definitions of cell populations are detailed in Sommarin *et al.* (in prep). For dimension-reduction of single-cell data, we used uniform manifold approximation and projection (UMAP)[69], and imputed gene expression using the MAGIC tool[70]. Plots were generated using SCANPY[71].

CRISPR/Cas9 perturbation of variant regions

We used a dual-sgRNA CRISPR/Cas9 approach to delete the regions harboring the four candidate causal variants at *CXCR4* (**Supplementary Table 7**), and an allele-specific, single-sgRNA CRISPR/Cas9 approach to abrogate the MYB binding site at rs772557 at *PPM1H*

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(**Supplementary Fig. 14**). sgRNAs were cloned into the pSpCas9(BB)-2A-GFP PX458 vector (Addgene; #48138), and transfected into MOLM13 cells (*CXCR4* variants) or K562 cells (*PPM1H* rs772557 variant) using the Neon system (Thermo Fisher Scientific). At 24 hours post transfection, GFP-positive cells were isolated using fluorescence-activated cell sorting. To verify CRISPR/Cas9 deletions for the *CXCR4* variants in MOLM13 cells, DNA was extracted, the targeted regions amplified by PCR (**Supplementary Table 7**), purified using the NucleoSpin® Gel and PCR Clean-up kit, loaded onto a 1% agarose gel. The deletion efficiency was estimated as the intensity of the deletion band divided by the sum of the intensity of the deletion band and the intensity of the wild type band. For K562 cells treated with allele-specific CRISPR/Cas9 towards rs772557, perturbation was verified using Sanger sequencing and the Tracking of Indels by Decomposition (TIDE, https://tide.nki.nl/). As control, we used sgRNAs targeting a randomly selected non-coding region on a different chromosome (here an intronic region in the *WAC* gene on chromosome 10). RNA was extracted from the same cells using the RNeasy plus micro kit (Qiagen; #74034) reverse-transcribed using the Sensiscript RT Kit (Qiagen; #205213) and Oligo-dT Primers (Qiagen; #79237), and quantified using Taqman assays (Thermo Fisher; #Hs00324748_m1 for *PPM1H* and Hs01060665_g1 for *ACTB*, as control).

Luciferase experiments

For the *PPM1H* variants, 200-bp sequences representing the reference and alternative allele of rs772555, rs772556, rs772557 and rs772559 in their genomic contexts were synthesized as gBlocks and cloned into the pGL3-Basic plasmid using KpnI and BglII restriction sites. Each sequence was centered on the variant and the two constructs differed only for the variant. 240 ng of Renilla luciferase construct were co-transfected with 10 ug of firefly construct in 5 million K562 ells, using the Neon electroporation system (Thermo Fisher Scientific). At 24 h

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

post-electroporation, luciferase and renilla activity were measured using DualGlo Luciferase (Promega; #E1960) on a GLOMAX 20/20 Luminometer. Based on luciferase/renilla readings, $\log_2$ scores were calculated for each variant reflecting the luciferase activity of the alternative allele relative to the corresponding reference allele. For co-electroporation experiments with *MYB* siRNA, luciferase plasmids were co-transfected with 400 nM Qiagen FlexiTube siRNA solution targeting *MYB* (Qiagen; #1027415) or 400nM Qiagen Negative Control siRNA (Qiagen; #1022076). Lysates for luciferase activity measurement and Western blot were collected simultaneously, 24 hours after electroporation. Electroporation conditions were not modified for co-transfection with siRNA and knockdown was confirmed by Western Blot using a recombinant c-Myb antibody (Abcam; #ab109127).

Motif analysis

To identify differentially binding transcription factors, we used the PERFECTOS-APE tool (http://opera.autosome.ru/perfectosape) with the HOCOMOCO-10, JASPAR, HT-SELEX, SwissRegulon and HOMER motif databases.

Enrichment analysis

To identify effects of rs772557 on specific $CD34^+$ populations, we calculated the correlation between rs772557 genotype and the expression of all genes with median FPKM > 1.0 in our $CD34^+$ mRNA-sequencing data set for blood donors. We then tested for enrichment for correlations within sets of marker genes for different $CD34^+$ HSPC subpopulations, inferred by comparing the expression profile of each cell type to other $CD34^+$ cell types in three data sets: **(a)** bulk RNA-sequencing data for sorted blood cells[18]; **(b)** single-cell RNA-sequencing data for $CD34^+$ cells (Sommarin *et al.*, in prep); and **(c)** single-cell RNA sequencing data for mononuclear white blood cells from adult blood and bone marrow[28]. For analysis we used

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RenderCat[72] with default settings. For completeness, we carried out the analysis with varying numbers of marker genes (25 to 500), with agreeing results.

Quantification of CD34$^+$ cell populations in cord blood

Umbilical cord blood (CB) samples were obtained from newborns at Skåne University Hospital (Lund and Malmö, Sweden) and Helsingborg Hospital (Helsingborg, Sweden), in compliance with regulations set by the regional ethical committee and informed consent. Samples were collected in Dulbecco's modified Eagle medium (DMEM) (GE Healthcare Life Sciences; #SH30022.01) with 1% fetal bovine serum (Fisher Scientific; #SV30160), 1% penicillin-streptomycin (Cytiva; #SV30010) and 4% heparin (Leo Pharmaceutics; 20U/ml). Samples were processed within 48 hours by isolating mononuclear cells using the density-gradient method (Abbott; #1019817). Mononuclear cells were isolated and frozen in FBS (Fisher Scientific; #SV30160) with 10% dimethyl sulfoxide (Sigma-Aldrich; #D5879), and kept at -80^{o}C until analysis. For flow cytometry analysis, cells were thawed, incubated with 0.8% ammonium chloride (Stem Cell Technologies; #07850) for 5 minutes at room temperature to remove remaining red blood cells, washed, filtered (BD #340632), and transferred to a polypropylene 96-well plate. After pelleting by centrifuging at 300 g for 5 min, cells were resuspended in 30ul antibody mix (**Supplementary Table 11**), diluted in staining buffer (PBS with 4mM EDTA and 2% BSA), and incubated for 60 min at 4^{o}C in the dark with shaking. Cells were then washed, resuspended in 200 µL PBS with 4mM EDTA and 2% BSA and analyzed using a four-laser BioRad ZE5^{TM} .

Gating of flow-cytometry data for cord blood

Cord blood samples were gated using pattern recognition software developed in-house (https://github.com/LudvigEk/HSPC-regulators-in-human-blood) that implements the

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

strategy described in **Supplementary Fig. 16**. Algorithmically, similar to the strategy used in adult blood, singlets in cord blood were separated using principal component analysis, and the PBMC gate also used the Dijkstraa's shortest path algorithm to separate granulocytes from PBMCs. Debris with low scatter values were also removed at this gate. The $CD34^+$ cluster was identified by finding the highest intensity value among the $CD45^+34^+$ cells, which often corresponds to the center of the cluster. The multiple gates involving lineage markers used the lowest density point in the corresponding marker-axis to separate the cells into two groups each time. The CD38 gate was fixed, considering the 30% of cells with the lowest CD38 values to be CD38- cells, while the rest were determined to be $CD38^+$. B/NK progenitors, CMPs, GMPs, MEPs, HSCs, MLPs and MPPs were determined in each case by different thresholds on CD10, CD135 and CD45RA values. These thresholds were determined by first identifying the highest density point in the distribution of the relevant marker values, and then locating a point in a predefined interval where the density dropped below a specific fraction of that value.

shRNA-knockdown in umbilical cord blood cells

Umbilical cord blood samples were collected as in the previous section. Thawed CB-derived $CD34^+$ cells were sorted and cultured in tissue-treated plates. Cells were maintained in serum-free expansion medium (SFEM) (Stem Cell Technologies; #09650) with penicillin-streptomycin (Cytiva; #SV30010), supplemented with stem cell factor, thrombopoietin, and FLT3-ligand at 100 ng/ml (Peprotech; #130-096-696, #130-095-754 and #130-096-480). pLKOpuro clones containing these shRNA sequences targeting *PPM1H* were obtained: CCCTCATTGTGATTTGCCTTT (TRCN0000331708), GCTTGGAAAGAAGATGCTCTA (TRCN0000052768), CCCAAACAGGAACTCATCCAA (TRCN0000052771) (Sigma), and recloned into pLKO.1-EGFP vector. Lentiviruses were produced in 293T cells as previously

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

described[73]. Knockdown efficiency was assessed in K562 cells cultured in RPMI media with 10% FBS. GFP^+ cells were sorted 2 days after transduction using a FACSAria™ III (BD Biosciences). RNA was extracted with the RNeasy Plus Micro Kit (Qiagen; #74034). cDNA was generated using the Sensiscript RT Kit (Qiagen; #205211) with poly-dT primers. *PPM1H* expression was quantified with TaqMan probes from ThermoFisher Scientific with ACTB as an internal control (*PPM1H* assay ID: Hs00324748_m1; *ACTB* assay ID: Hs01060665_g1). In the subsequent experiments on primary umbilical cord blood cells, transduced cells were analysed using a BD LSR Fortessa flow cytometer after staining with 1:200 CD34-PeCy7 (BioLegend, #343516), 1:100 CD90-APC (BioLegend, #328114), and resuspended with CountBright™ Absolute Counting Beads (Fisher Scientific; #C36950) and 7-AAD for dead cell exclusion. As readout, we measured the percentage and absolute count (using the CountBright beads) of $CD34^+$ and $CD34^+90^+$ cells. Cell enrichment was calculated by dividing $CD34^+$ and $CD34^+90^+$ cells counts at each time point by initial cell counts, and then normalizing the enrichment values to the control shRNA transduced cells at the same time point. At each time point, three transduction replicates were recorded. The experiment was repeat three times. For statistical analysis, we compared normalized enrichment values at day 7, 14 and 21 for shRNA-transduced cells vs non-targeting shRNA control. To ensure conservative statistical analysis, we first averaged the transduction replicates for each experiment, then calculated *P*-values using permutation testing with the data the day 7, 14, and 21 data for the same experiment being permuted together (100,000 permutations).

Protein quantitative trait locus (pQTL) analysis

Using flow cytometry, we quantified ITGA9 protein expression on the surface of $CD34^+$ cells in blood samples from 458 primary care patients, who were not included in the association study. Samples were obtained and processed as described above. Cells were stained with

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

CD34-PE-CF594 and CD45-APC-H7 (**Supplementary Table 11**), in addition to ITGA9-PE (Biolegend # 351606) at a concentration of 0.02µl of antibody/30µl of total staining volume. Gating and calculcation of median fluorescence intensity was done with FlowJo V10.6.1 (Becton, Dickinson & Company).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Notta F, Zandi S, Takayama N, et al. Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science (80-)*. 2016;351(6269). doi:10.1126/science.aab2116

2. Barnett D, Janossy G, Lubenko A, et al. Guideline for the flow cytometric enumeration of CD34+ haematopoietic stem cells. *Clin Lab Haematol*. 1999;21(5):301-308. doi:10.1046/j.1365-2257.1999.00253.x

3. Cohen KS, Cheng S, Larson MG, et al. Circulating CD34+ progenitor cell frequency is associated with clinical and genetic factors. *Blood*. 2013;121(8):3-5. doi:10.1182/blood-2012-05-424846

4. Eidenschink L, Dizerega G, Rodgers K, Bartlett M, Wells DA, Loken MR. Basal levels of CD34 positive cells in peripheral blood differ between individuals and are stable for 18 months. *Cytom Part B - Clin Cytom*. 2012;82 B(1):18-25. doi:10.1002/cyto.b.20611

5. Hübel K. Mobilization and Collection of HSC. In: Carreras E, Dufour C, Mohty M, Kröger N, eds. Cham (CH); 2019:117-122. doi:10.1007/978-3-030-02278-5_15

6. Ataca Atilla P, Bakanay Ozturk SM, Demirer T. How to manage poor mobilizers for high dose chemotherapy and autologous stem cell transplantation? *Transfus Apher Sci Off J World Apher Assoc Off J Eur Soc Haemapheresis*. 2017;56(2):190-198. doi:10.1016/j.transci.2016.11.005

7. Hope KJ, Jin L, Dick JE. Acute myeloid leukemia originates from a hierarchy of leukemic stem cell classes that differ in self-renewal capacity. *Nat Immunol*. 2004;5(7):738-743. doi:10.1038/ni1080

8. Lapidot T, Sirard C, Vormoor J, et al. A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature*. 1994;367(6464):645-648.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

doi:10.1038/367645a0

9. Nilsson L, Edén P, Olsson E, et al. The molecular signature of MDS stem cells supports a stem-cell origin of 5q-myelodysplastic syndromes. *Blood*. 2007;110(8):3005-3014. doi:10.1182/blood-2007-03-079368

10. Nilsson L, Astrand-Grundstrom I, Arvidsson I, et al. Isolation and characterization of hematopoietic progenitor/stem cells in 5q-deleted myelodysplastic syndromes: Evidence for involvement at the hematopoietic stem cell level. *Blood*. 2000;96(6):2012-2021. doi:10.1182/blood.v96.6.2012

11. Bao EL, Nandakumar SK, Liao X, et al. *Inherited Myeloproliferative Neoplasm Risk Affects Haematopoietic Stem Cells*. Vol 586.; 2020. doi:10.1038/s41586-020-2786-7

12. Sveinbjornsson G, Albrechtsen A, Zink F, et al. Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat Genet*. 2016;48(3):314-317. doi:10.1038/ng.3507

13. Ulirsch JC, Nandakumar SK, Wang L, et al. Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell*. 2016;165(6):1530-1545. doi:10.1016/j.cell.2016.04.048

14. Astle WJ, Elding H, Jiang T, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease Resource The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*. 2016:1415-1429. doi:10.1016/j.cell.2016.10.042

15. Vuckovic D, Bao EL, Akbari P, et al. The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell*. 2020;182(5):1214-1231.e11. doi:10.1016/j.cell.2020.08.008

16. Van Der Harst P, Zhang W, Mateo Leach I, et al. Seventy-five genetic loci influencing the human red blood cell. *Nature*. 2012;492(7429):369-375. doi:10.1038/nature11677

17. Corces MR, Buenrostro JD, Wu B, et al. Lineage-specific and single-cell chromatin

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

accessibility charts human hematopoiesis and leukemia evolution. 2016;48(10). doi:10.1038/ng.3646

18. Ulirsch JC, Lareau CA, Bao EL, et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat Genet*. 2019;51(4):683-693. doi:10.1038/s41588-019-0362-6

19. Mifsud B, Tavares-Cadete F, Young AN, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet*. 2015;47(6):598-606. doi:10.1038/ng.3286

20. Peng Y, Wu D, Li F, Zhang P, Feng Y, He A. Identification of key biomarkers associated with cell adhesion in multiple myeloma by integrated bioinformatics analysis. *Cancer Cell Int*. 2020;20(1):1-16. doi:10.1186/s12935-020-01355-z

21. Staley JR, Blackshaw J, Kamat MA, et al. PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics*. 2016;32(20):3207-3209. doi:10.1093/bioinformatics/btw373

22. Kamat MA, Blackshaw JA, Young R, et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics*. 2019;35(22):4851-4853. doi:10.1093/bioinformatics/btz469

23. Milanesi S, Locati M, Borroni EM. Aberrant CXCR4 signaling at crossroad of WHIM syndrome and Waldenstrom's macroglobulinemia. *Int J Mol Sci*. 2020;21(16):1-15. doi:10.3390/ijms21165696

24. Heusinkveld LE, Majumdar S, Gao JL, McDermott DH, Murphy PM. WHIM Syndrome: from Pathogenesis Towards Personalized Medicine and Cure. *J Clin Immunol*. 2019;39(6):532-556. doi:10.1007/s10875-019-00665-w

25. Smith ML, Cavenagh JD, Lister TA, Fitzgibbon J. Mutation of CEBPA in familial acute myeloid leukemia. *N Engl J Med*. 2004;351(23):2403-2407.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

doi:10.1056/NEJMoa041331

26. Vulliamy T, Marrone A, Goldman F, et al. The RNA component of telomerase is mutated in autosomal dominant dyskeratosis congenita. *Nature*. 2001;413(6854):432-435. doi:10.1038/35096585

27. Ngo HT, Leleu X, Lee J, et al. SDF-1/CXCR4 and VLA-4 interaction regulates homing in Waldenstrom macroglobulinemia. *Blood*. 2008;112(1):150-158. doi:10.1182/blood-2007-12-129395

28. Granja JM, Klemm S, McGinnis LM, et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat Biotechnol*. 2019;37(12):1458-1465. doi:10.1038/s41587-019-0332-7

29. Karpova D, Bonig H. Concise review : CXCR4 / CXCL12 signaling in immature hematopoiesis — lessons from pharmacological and genetic models. *Stem Cells*. 2015;33(8):2391-2399.

30. Hernandez PA, Gorlin RJ, Lukens JN, et al. Mutations in the chemokine receptor gene CXCR4 are associated with WHIM syndrome, a combined immunodeficiency disease. *Nat Genet*. 2003;34(1):70-74. doi:10.1038/ng1149

31. Liles WC, Broxmeyer HE, Rodger E, et al. Mobilization of hematopoietic progenitor cells in healthy volunteers by AMD3100, a CXCR4 antagonist. *Blood*. 2003;102(8):2728-2730. doi:10.1182/blood-2003-02-0663

32. Madeira F, Park YM, Lee J, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res*. 2019;47(W1):W636-W641. doi:10.1093/nar/gkz268

33. Toft N, Birgens H, Abrahamsson J, et al. Results of NOPHO ALL2008 treatment for patients aged 1-45 years with acute lymphoblastic leukemia. *Leukemia*. 2018;32(3):606-615. doi:10.1038/leu.2017.265

34. Sugiura T, Noguchi Y. Substrate-dependent metal preference of PPM1H, a cancer-

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

associated protein phosphatase 2C: Comparison with other family members. *BioMetals*. 2009;22(3):469-477. doi:10.1007/s10534-009-9204-9

35. Chen MJ, Dixon JE, Manning G. Genomics and evolution of protein phosphatases. *Sci Signal*. 2017;10(474). doi:10.1126/scisignal.aag1796

36. Berndsen K, Lis P, Yeshaw WM, et al. PPM1H phosphatase counteracts LRRK2 signaling by selectively dephosphorylating rab proteins. *Elife*. 2019;8:1-37. doi:10.7554/eLife.50416

37. Lee-Hoeflich ST, Pham TQ, Dowbenko D, et al. PPM1H Is a p27 phosphatase implicated in trastuzumab resistance. *Cancer Discov*. 2011;1(4):326-337. doi:10.1158/2159-8290.CD-11-0062

38. Ghodke-Puranik Y, Imgruet M, Dorschner JM, et al. Novel genetic associations with interferon in systemic lupus erythematosus identified by replication and fine-mapping of trait-stratified genome-wide screen. *Cytokine*. 2020;132:154631. doi:10.1016/j.cyto.2018.12.014

39. Sugiura T, Noguchi Y, Sakurai K, Hattori C. Protein phosphatase 1H, overexpressed in colon adenocarcinoma, is associated with CSE1L. *Cancer Biol Ther*. 2008;7(2):285-292. doi:10.4161/cbt.7.2.5302

40. Davis CA, Hitz BC, Sloan CA, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res*. 2018;46(D1):D794-D801. doi:10.1093/nar/gkx1081

41. Ghosh AK, Steele R, Ray RB. Functional Domains of c-myc Promoter Binding Protein 1 Involved in Transcriptional Repression and Cell Growth Regulation. *Mol Cell Biol*. 1999;19(4):2880 LP - 2886. doi:10.1128/MCB.19.4.2880

42. Vilhais-Neto GC, Maruhashi M, Smith KT, et al. Rere controls retinoic acid signalling and somite bilateral symmetry. *Nature*. 2010;463(7283):953-957.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

doi:10.1038/nature08763

43. Cabezas-Wallscheid N, Buettner F, Sommerkamp P, et al. Vitamin A-Retinoic Acid Signaling Regulates Hematopoietic Stem Cell Dormancy. *Cell*. 2017;169(5):807-823.e19. doi:10.1016/j.cell.2017.04.018

44. Geoffroy M-C, Esnault C, de Thé H. Retinoids in haematology : a timely revival? *Blood*. March 2021. doi:10.1182/blood.2020010100

45. Bahr C, von Paleske L, Uslu V V, et al. A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. *Nature*. 2018;553(7689):515-520. doi:10.1038/nature25193

46. Doulatov S, Notta F, Eppert K, Nguyen LT, Ohashi PS, Dick JE. Revised map of the human progenitor hierarchy shows the origin of macrophages and dendritic cells in early lymphoid development. *Nat Immunol*. 2010;11(7):585-593. doi:10.1038/ni.1889

47. Notta F, Doulatov S, Laurenti E, Poeppl A, Jurisica I, Dick JE. Isolation of Single Human Hematopoietic. *Science (80-).* 2011;333(July):218-222. doi:10.1126/science.1201219

48. Loh PR, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet*. 2016;48(11):1443-1448. doi:10.1038/ng.3679

49. Gudbjartsson DF, Helgason H, Gudjonsson SA, et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet*. 2015;47(5):435-444. doi:10.1038/ng.3247

50. Kong A, Masson G, Frigge ML, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet*. 2008;40(9):1068-1075. doi:10.1038/ng.216

51. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;19(9):1655-1664. doi:10.1101/gr.094052.109

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

52. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393

53. Price AL, Weale ME, Patterson N, et al. Long-Range LD Can Confound Genome Scans in Admixed Populations. *Am J Hum Genet*. 2008;83(1):132-135. doi:10.1016/j.ajhg.2008.06.005

54. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575. doi:10.1086/519795

55. Diaz-Papkovich A, Anderson-Trocmé L, Gravel S. A review of UMAP in population genetics. *J Hum Genet*. 2020. doi:10.1038/s10038-020-00851-4

56. Bulik-Sullivan B, Loh PR, Finucane HK, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet*. 2015;47(3):291-295. doi:10.1038/ng.3211

57. MANTEL N, HAENSZEL W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst*. 1959;22(4):719-748.

58. Sveinbjornsson G, Albrechtsen A, Zink F, et al. Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat Genet*. 2016;48(3):314-317. doi:10.1038/ng.3507

59. Maller JB, McVean G, Byrnes J, et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet*. 2012;44(12):1294-1301. doi:10.1038/ng.2435

60. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol*. 2015;109:21.29.1-21.29.9. doi:10.1002/0471142727.mb2129s109

61. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sequence data. *Bioinformatics*. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170

62. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987-2993. doi:10.1093/bioinformatics/btr509

63. Mansour MR, Abraham BJ, Anders L, et al. Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science*. 2014;346(6215):1373-1377. doi:10.1126/science.1259037

64. Gioia L, Siddique A, Head SR, Salomon DR, Su AI. A Genome-wide Survey of Mutations in the Jurkat Cell Line. *bioRxiv*. 2017:1-13. doi:10.1101/118117

65. Finucane HK, Bulik-Sullivan B, Gusev A, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet*. 2015;47(11):1228-1235. doi:10.1038/ng.3404

66. Mifsud B, Tavares-Cadete F, Young AN, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet*. 2015;47(6):598-606. doi:10.1038/ng.3286

67. Dobin A, Davis CA, Schlesinger F, et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635

68. Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923-930. doi:10.1093/bioinformatics/btt656

69. Becht E, McInnes L, Healy J, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*. 2019;37(1):38-47. doi:10.1038/nbt.4314

70. van Dijk D, Sharma R, Nainys J, et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*. 2018;174(3):716-729.e27.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

doi:10.1016/j.cell.2018.05.061

71. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. 2018;19(1):15. doi:10.1186/s13059-017-1382-0

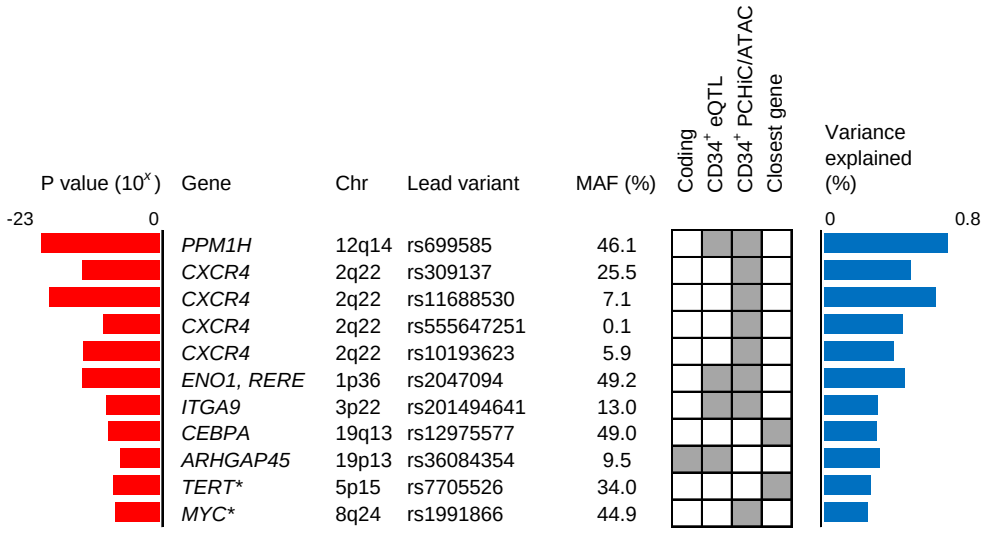
72. Nilsson B, Håkansson P, Johansson M, Nelander S, Fioretos T. Threshold-free high-power methods for the ontological analysis of genome-wide gene-expression studies. *Genome Biol*. 2007;8(5). doi:10.1186/gb-2007-8-5-r74

73. Ali N, Karlsson C, Aspling M, et al. Forward RNAi screens in primary human hematopoietic stem/progenitor cells. *Blood*. 2009;113(16):3690-3695. doi:10.1182/blood-2008-10-176396

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

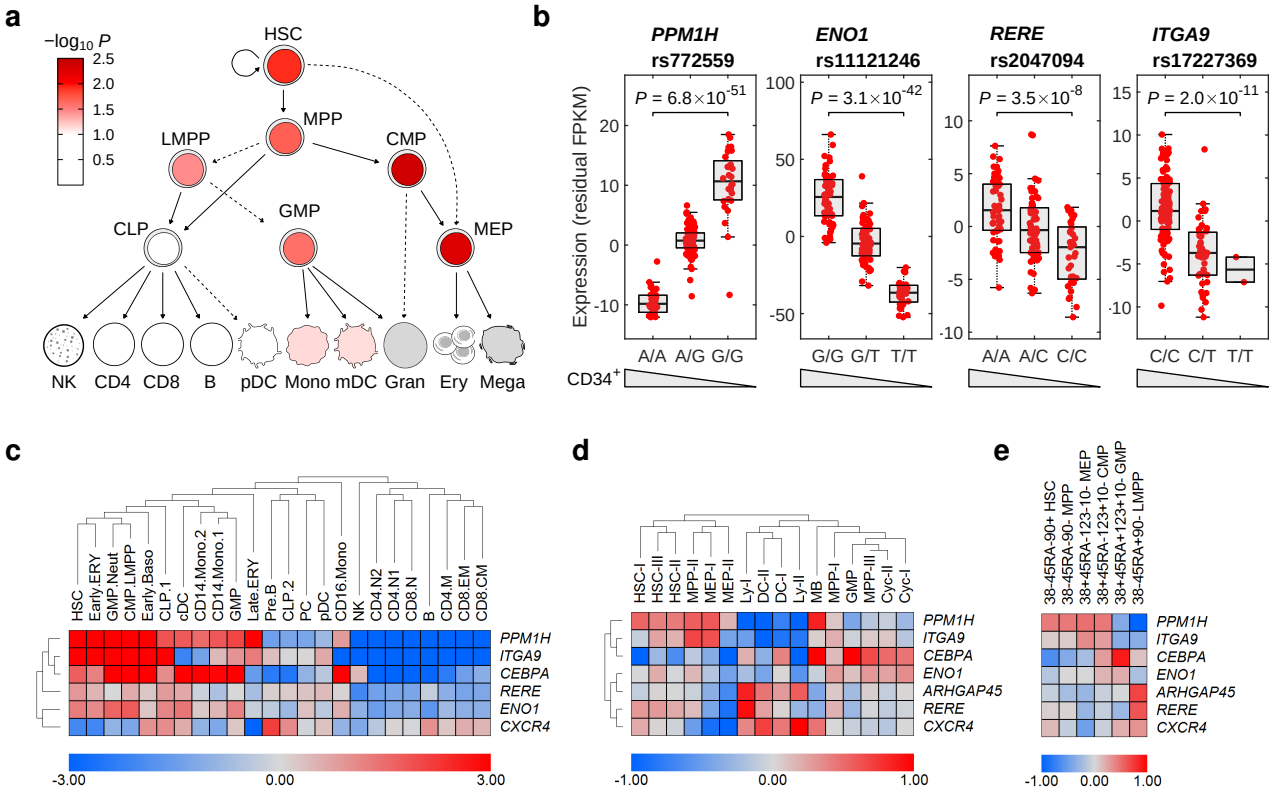
bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 1



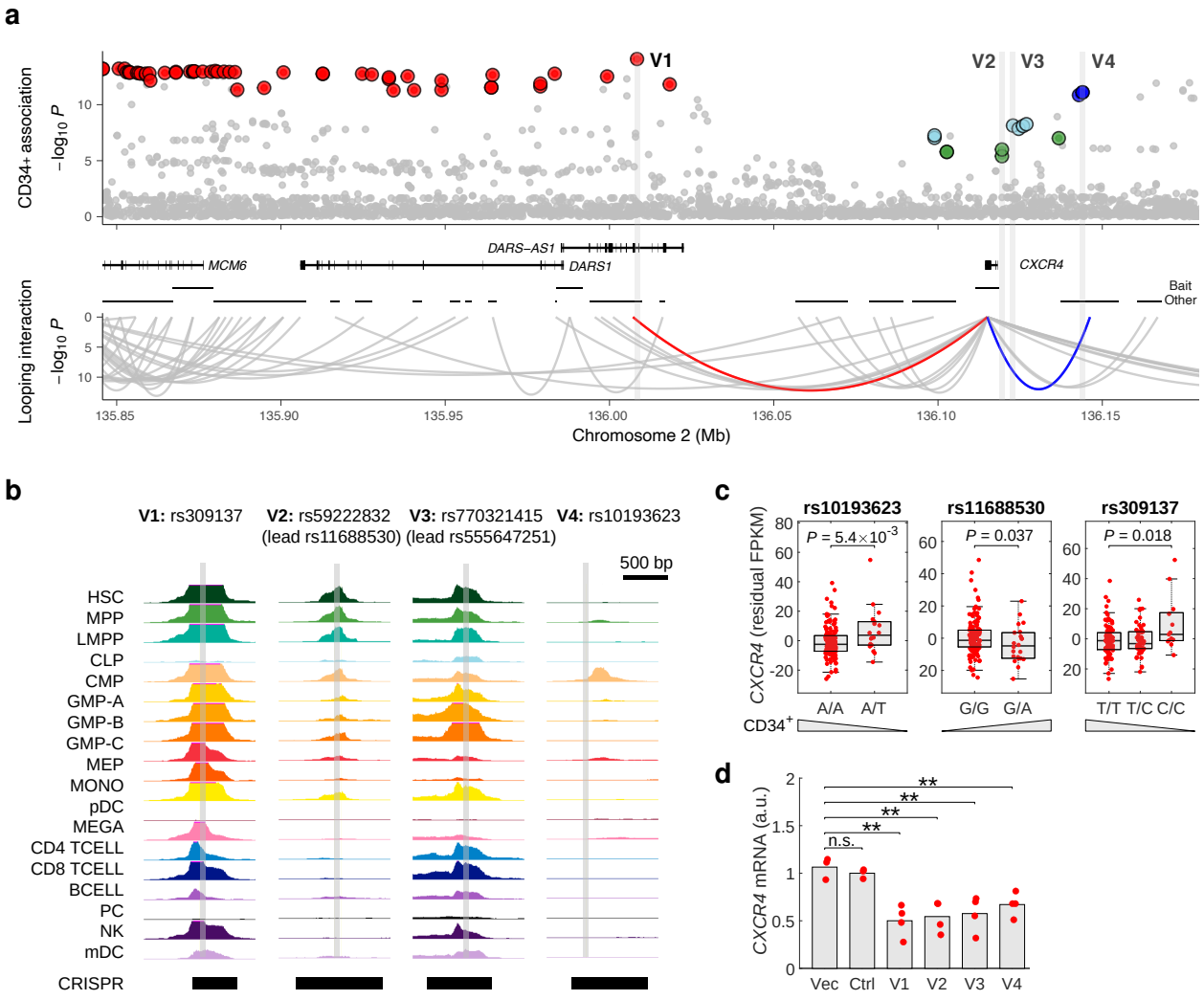
bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2



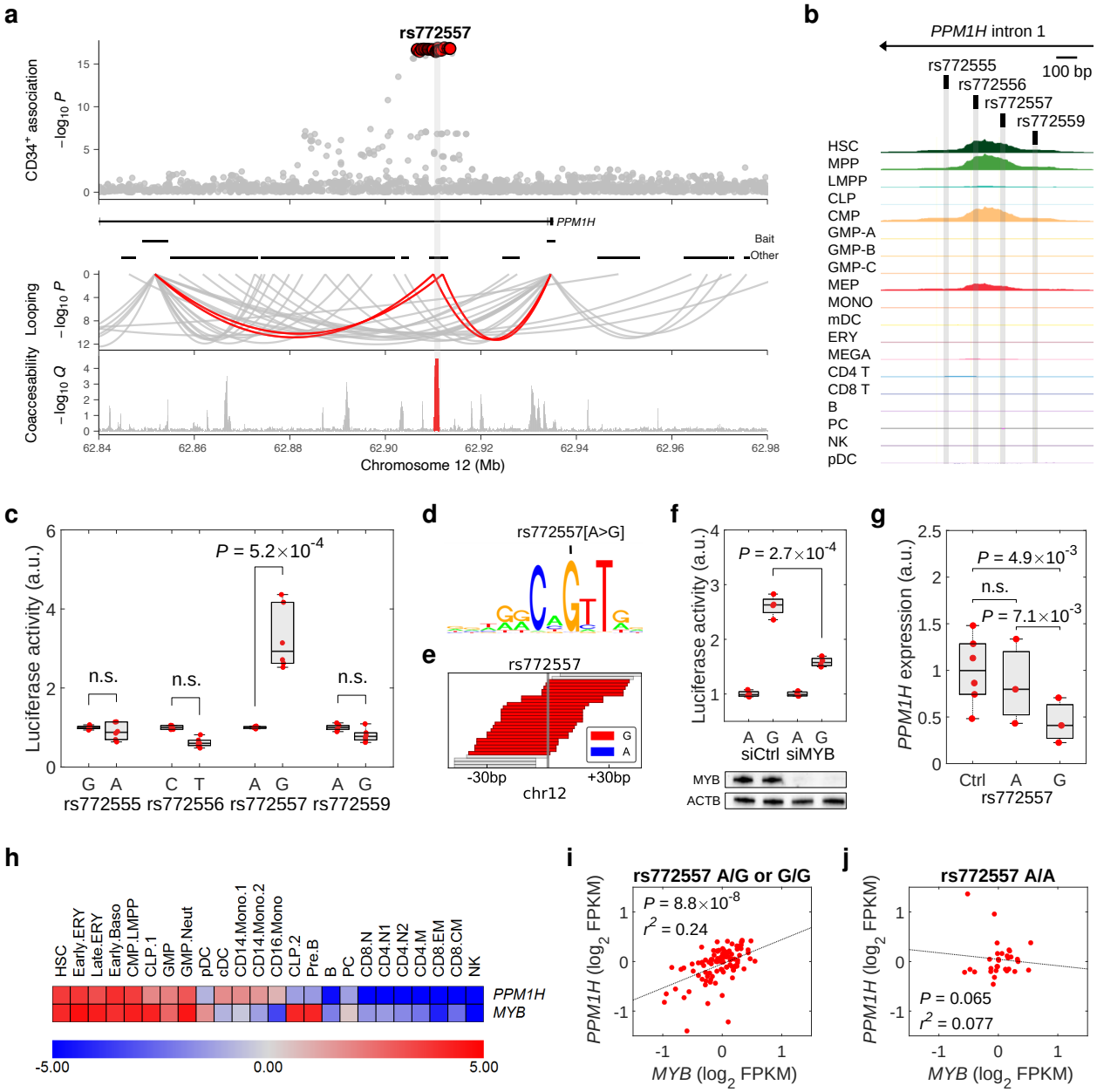
bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3



bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 4



bioRxiv preprint doi: https://doi.org/10.1101/2021.03.31.437808; this version posted April 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 5