Article

# Causal modelling of gene effects from regulators to programs to traits

https://doi.org/10.1038/s41586-025-09866-3

Received: 28 January 2025

Accepted: 5 November 2025

Published online: 10 December 2025

Open access

Check for updates

Mineto Ota[1,2,3]✉, Jeffrey P. Spence[1,4,5], Tony Zeng[1], Emma Dann[1,2], Nikhil Milind[1], Alexander Marson[2,4,6,7,8,9,10]✉ & Jonathan K. Pritchard[1,11]✉

Genetic association studies provide a unique tool for identifying candidate causal links from genes to human traits and diseases. However, it is challenging to determine the biological mechanisms underlying most associations, and we lack genome-scale approaches for inferring causal mechanistic pathways from genes to cellular functions to traits. Here we propose approaches to bridge this gap by combining quantitative estimates of gene–trait relationships from loss-of-function burden tests[1] with gene-regulatory connections inferred from Perturb-seq experiments[2] in relevant cell types. By combining these two forms of data, we aim to build causal graphs in which the directional associations of genes with a trait can be explained by their regulatory effects on biological programs or direct effects on the trait[3]. As a proof of concept, we constructed a causal graph of the gene-regulatory hierarchy that jointly controls three partially co-regulated blood traits. We propose that perturbation studies in trait-relevant cell types, coupled with gene-level effect sizes for traits, can bridge the gap between genetic association and biological mechanism.

Genome-wide association studies (GWAS) and rare variant burden tests have identified tens of thousands of reproducible associations for a wide range of human traits and diseases. These signals have identified many genes that can serve as therapeutic targets[4–6]; driven discoveries of new molecular mechanisms[7,8], critical cell types[9] and physiological pathways of disease risks or traits[10–12]; and enabled genetic risk prediction for complex diseases[13].

But despite these successes, interpreting the vast majority of associations remains challenging. Aside from coarse-grained analyses such as identifying trait-relevant cell types and enriched gene sets, we lack genome-scale approaches for interpreting the molecular pathways and mechanisms through which hundreds, if not thousands, of genes affect a given phenotype.

One challenge for interpreting genetic associations is the observation that many hits act indirectly, via trans-regulation of other genes[14–19]. This observation is formalized in the omnigenic model[3,20], which proposes that, for any given trait, only a subset of genes, referred to as core genes, are located within key molecular pathways that act directly on the trait of interest. Meanwhile, many more genes affect the trait indirectly, by regulating core genes through links in gene-regulatory networks. In this model, we can interpret the effect size of a variant in terms of all paths through the network by which it affects core genes.

The central role of trans-regulation underlying many GWAS hits implies that fully understanding the genetic basis of complex traits requires tools to measure how genetic effects flow through networks. However, until recently, we have had very limited information about gene-regulatory networks in any human cell type, with the main information coming from observational data such as *trans*-expression quantitative trait locus (*trans*-eQTL) and co-expression mapping[14,16,21]. However, both approaches have important limitations including low power[20,22] and confounding effects of cell-type composition[14] in the case of *trans*-eQTLs, and ambiguous causality in co-expression analysis[23,24].

Advances in genome editing and single-cell RNA sequencing, including Perturb-seq, now provide new opportunities to measure causal gene-regulatory connections at genome scale[25–28]. In Perturb-seq experiments, a pool of cells is transduced with a library of guide RNAs, each of which causes knockdown (or other perturbation) of a single gene. After allowing the cells time to equilibrate, single-cell sequencing is used to determine which genes were knocked down in each cell and measure the transcriptome of the cell. Critically, Perturb-seq enables measurement of the trans-regulatory effects of each gene in a controlled experimental setting at the genome-wide scale. Recent work has shown that such approaches are a promising tool for interpreting GWAS data, finding that GWAS hits are often enriched in specific transcriptional programs identified by CRISPR perturbations of a subset of genes[29–33].

Major challenges remain as we aim to move beyond identifying enriched programs to inferring genome-scale causal cascades of biological information. In this paper, we developed a new systematic approach to this problem. We demonstrate how, by combining loss-of-function (LoF) burden results with Perturb-seq, we can infer an internally coherent graph linking genes to functional programs to traits, and derive biological insight into the key genes and pathways that control these traits (Fig. 1a). The resulting graph helps us to understand not only the trait-relevant pathways but also the functions of genes and programs

[1]Department of Genetics, Stanford University, Stanford, CA, USA. [2]Gladstone-UCSF Institute of Genomic Immunology, San Francisco, CA, USA. [3]Department of Allergy and Rheumatology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [4]Institute for Human Genetics (IHG), University of California San Francisco, San Francisco, CA, USA. [5]Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA, USA. [6]Department of Medicine, University of California San Francisco, San Francisco, CA, USA. [7]UCSF Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, San Francisco, CA, USA. [8]Parker Institute for Cancer Immunotherapy, San Francisco, CA, USA. [9]Department of Microbiology and Immunology, University of California San Francisco, San Francisco, CA, USA. [10]Innovative Genomics Institute, University of California Berkeley, Berkeley, CA, USA. [11]Department of Biology, Stanford University, Stanford, CA, USA. ✉e-mail: mineto-ota@g.ecc.u-tokyo.ac.jp; alex.marson@gladstone.ucsf.edu; pritch@stanford.edu

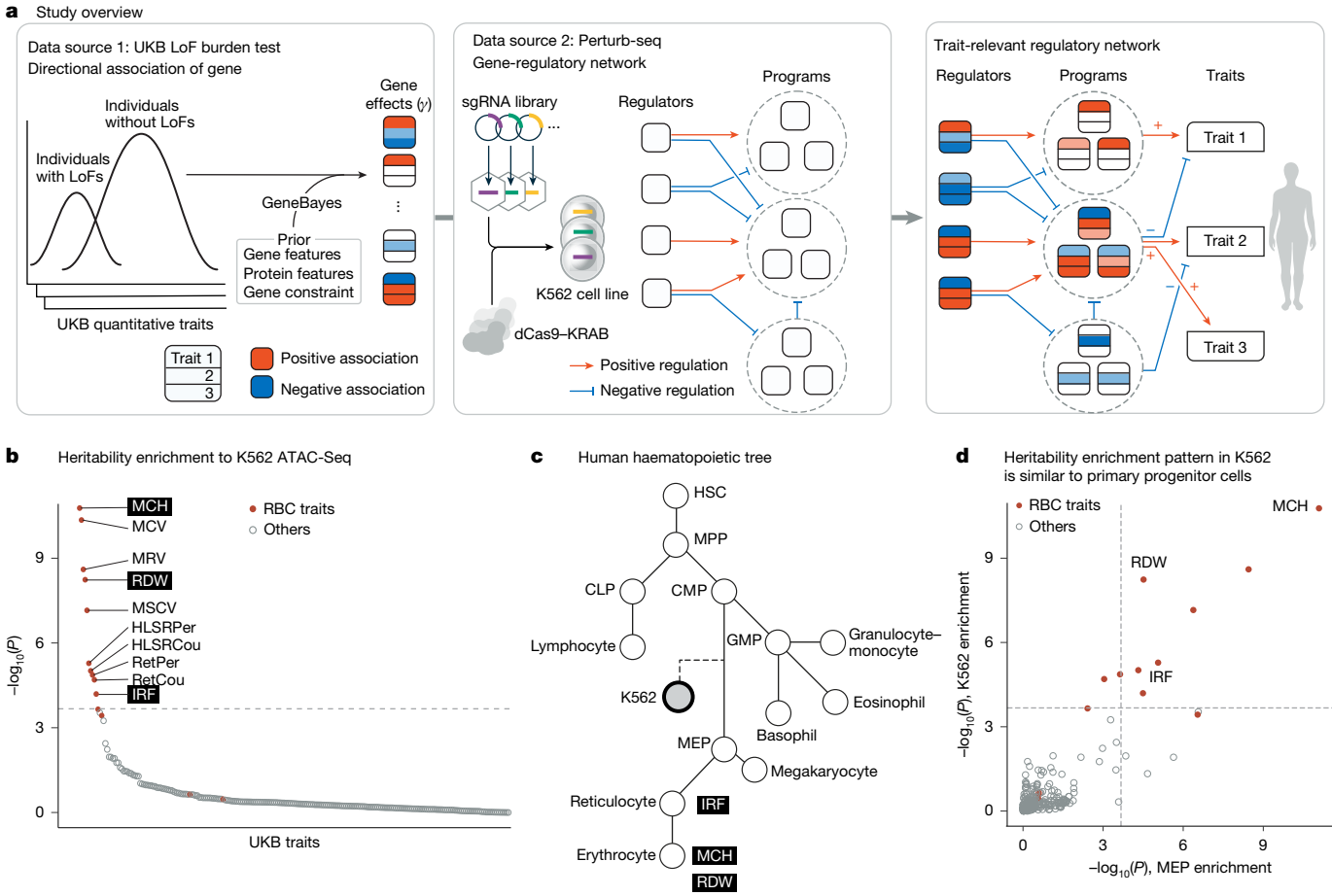



**Fig. 1 | Study overview and selection of model traits. a**, Overview of the study. The square nodes represent genes, the coloured arrows between genes represent regulatory effects and the arrows from genes to traits represent associations. sgRNA, single guide RNA. **b**, Heritability enrichment of UKB traits to open chromatin regions in K562. Traits are ordered based on the *P* value of enrichment, which was estimated using the Jackknife test in S-LDSC. The dashed line indicates the threshold for Bonferroni significance. ATAC-seq, assay for transposase-accessible chromatin using sequencing; Cou, count; HLSR, high light scatter reticulocyte count; MCV, mean corpuscular volume; MRV, mean reticulocyte volume; MSCV, mean sphered corpuscular volume; Per, percentage; RBC, red blood cell; Ret, reticulocyte. **c**, Schematic of the human haematopoietic tree. Traits of interest are annotated near their relevant cell types. CLP, common lymphoid progenitor; CMP, common myeloid progenitor; GMP, granulocyte–monocyte progenitor; HSC, haematopoietic stem cell; MPP, multipotent progenitor. **d**, Comparison of heritability enrichment to UKB traits, between MEP and K562 open chromatin regions. *P* values were estimated using the Jackknife test in S-LDSC. The dashed line indicates the threshold for Bonferroni significance.

within the graph, to explain why those genes are associated with the traits. On the basis of our results, we expect that forthcoming efforts to generate perturbation data in a wide variety of cell types will provide a critical interpretative framework for human genetics.

## Selection of model traits

To integrate genetic association data with Perturb-seq, our first step was to evaluate whether there are any traits with high-quality genetic data where the most relevant cell type (or types) can be well modelled by existing Perturb-seq data. At the time of writing, the only published genome-wide Perturb-seq dataset was collected in a leukemia cell line: K562 (ref. 2). In that experiment, every expressed gene was knocked down using CRISPR interference, one gene per cell, before single-cell RNA sequencing.

To determine which traits could reasonably be modelled in terms of the gene-regulatory networks of K562 cells, we compiled published GWAS and LoF burden test data for a wide range of traits measured in the UK Biobank (UKB)[1,34]. Of these, we selected 234 quantitative traits with single-nucleotide polymorphism (SNP) heritability > 0.04 for further consideration (Supplementary Table 1) and performed stratified linkage disequilibrium score regression (S-LDSC)[9] across all 234 traits. We observed that open chromatin regions in K562 exhibited significant heritability enrichment exclusively for traits related to morphology or quantity of erythroid lineage cells (Fig. 1b).

This result is intuitive, as the K562 cell line was derived from erythroleukaemia cells, which are a neoplastic form of erythroid progenitors (Fig. 1c), and K562 cells retain multipotency and can differentiate into erythroid cells[35].

We also performed S-LDSC across the same set of traits for various primary cell types, and found a very similar enrichment for erythroid traits in megakaryocyte–erythroid progenitor cells (MEPs), which are the natural progenitor cells for erythrocytes (Fig. 1c,d and Extended Data Fig. 1a). The open chromatin regions in MEPs were also more similar to those in K562 cells than other cell types (Extended Data Fig. 1b). These results support the notion that K562 cells share similar chromatin features with primary progenitor cells and could serve as a cellular model for studying the gene-regulatory network associated with erythroid traits.

Among the enriched traits, we selected three traits that are relatively independent, with pairwise genetic correlations ranging from −0.39 to 0.15, for detailed analysis (Extended Data Fig. 1c). We focused primarily on mean corpuscular haemoglobin (MCH), which measures the mean amount of haemoglobin per erythrocyte; but, we also analysed red

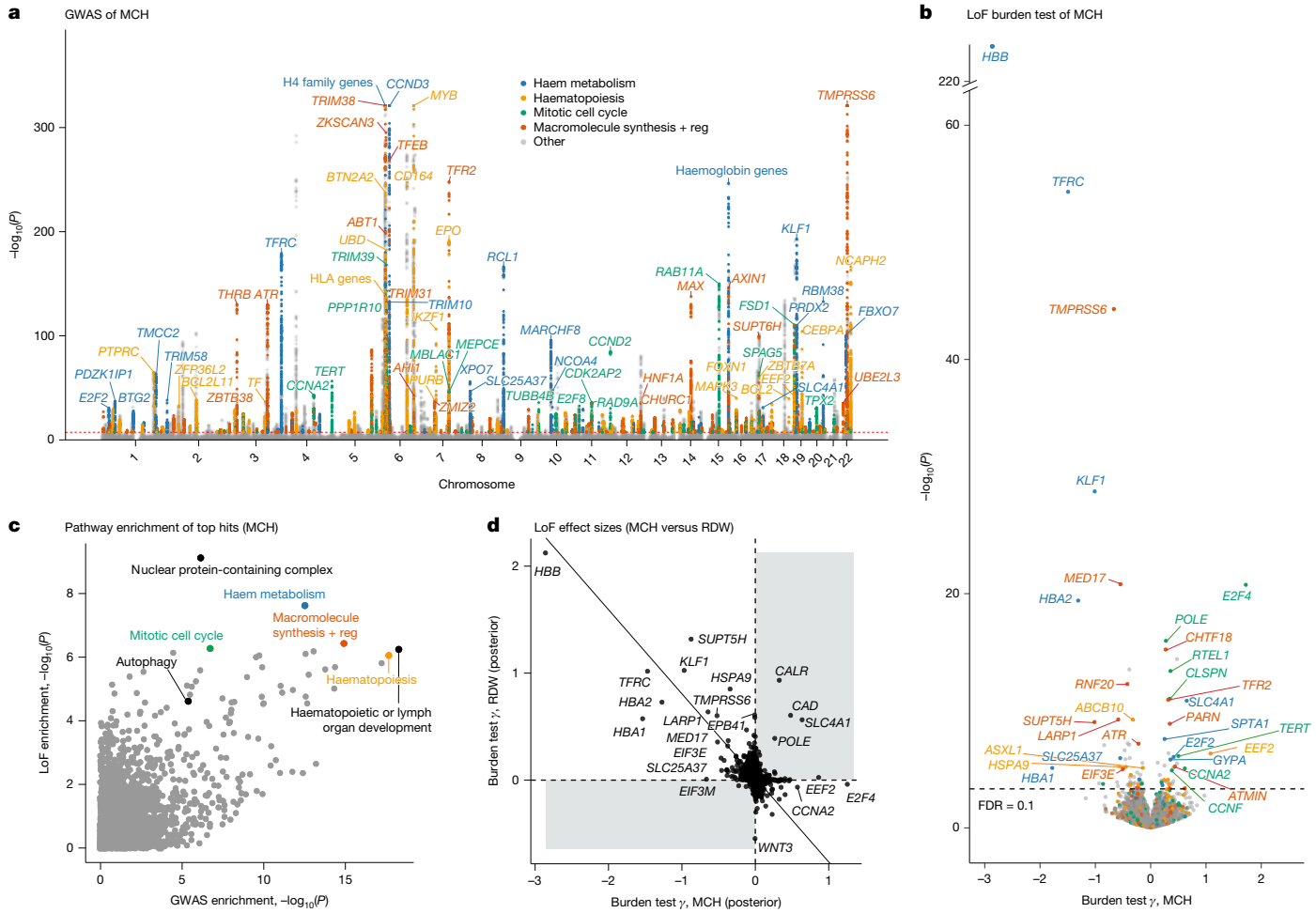



**Fig. 2 | Pathway enrichments for blood trait associations. a**, Genetic associations identified from UKB GWAS for MCH. Variants located within a 100-kb window centred on the transcription start site of the genes in the gene set are coloured. 'Macromolecule synthesis + reg' refers to the positive regulation of the macromolecule biosynthetic process. **b**, Gene associations with MCH from UKB LoF burden tests. The colours indicate the same gene sets as panel **a**. Labelled genes have FDR < 0.01 and belong to the gene sets. **c**, Pathway enrichment of GWAS and LoF burden test top genes. For GWAS, the closest genes from the independent top variants were used. For the LoF burden test, genes were ranked by the absolute posterior effect size from GeneBayes, and the same number of genes as in GWAS was used. *P* values are from one-sided Fisher's exact test. **d**, Comparison of LoF burden test effect sizes after GeneBayes between MCH and RDW. The solid line corresponds to the first principal component.

cell distribution width (RDW)—the standard deviation of the size of erythrocytes per individual—and the immature reticulocyte fraction (IRF). For these traits, a considerable amount of SNP heritability was explained by open chromatin regions in the K562 cell line (53%, 44% and 36% of the total SNP heritability, respectively), further supporting the use of K562 Perturb-seq to interpret their genetic associations (Supplementary Table 2).

## Pathway enrichment for trait associations

Before attempting to build causal models for these traits, we first explored the genetic associations for MCH, RDW and IRF with standard approaches (Fig. 2 and Supplementary Fig. 1). GWAS of MCH in the UKB identified 634 independent genome-wide significant signals. Many of the lead hits fall into a few significantly enriched pathways, including haem metabolism, haematopoiesis and cell cycle (Fig. 2a,c). These enriched pathways are crucially involved in the maturation of erythrocytes. For example, tight control of cell cycle is important at several steps in erythropoiesis[36–39].

In addition to GWAS, UKB has also released whole-exome sequencing data for more than 450,000 participants[40]. Here we focused on the phenotypic effects of LoFs, which are variants such as frameshift and premature stop mutations that are predicted to cause complete LoF of a gene. To estimate the average effect of different LoF variants in the same gene on a phenotype, we compared the phenotypic values for carriers of LoF variants in a given gene versus non-carriers. This approach, known as a burden test, generates a score for each gene that estimates the effect of half loss of gene dosage on the phenotype.

Previously reported burden test statistics for LoF variants[1] identified 90 genes associated with MCH at a false discovery rate (FDR) = 0.1 (Fig. 2b). Although the rankings of top hits differ between GWAS and LoF burden tests (Extended Data Fig. 2a), the lead hits from GWAS and LoF are generally enriched in the same pathways (Fig. 2c). This is consistent with the expectation that common and rare variants associated with a trait act through similar biological pathways, but frequently prioritize different genes[41,42].

As one might expect, LoF variants in the genes that encode components of adult human haemoglobin, *HBB*, *HBA1* and *HBA2*, all show strong negative effects on MCH (Fig. 2b). Clinically, these mutations cause α-thalassaemia or β-thalassaemia, in which a decrease in MCH is characteristic. This highlights a key feature of burden tests: in addition to significance testing, they also provide a quantitative, directional estimate of LoF effects, referred to here as $\gamma$.

The directions of associations in the burden tests also help us to interpret the pleiotropic effects of genes. When looking at genes associated with MCH and RDW, which have a negative genetic correlation in GWAS (Extended Data Fig. 1c), the LoF effects for most genes were associated in opposite directions ($r = -0.53$; Fig. 2d). However, a handful of genes had strong same-direction effects on both traits (Fig. 2d). For instance, *CAD* encodes a multifunctional enzyme of which biallelic mutations cause megaloblastic anaemia[43], whereas heterozygous LoFs increase both MCH and RDW (Fig. 2d). One goal of building a causal mechanistic graph for these traits will be to explain these seemingly discordant associations.

For many genes, the LoF $\gamma$s have large standard errors, due to the low frequency of LoF variants[41]. To improve estimation of the $\gamma$s, we applied an empirical Bayes framework called GeneBayes that we developed previously[44]. Our approach incorporated previous information about gene expression, protein structure and gene constraint to share information across functionally similar genes (Methods). We found that the GeneBayes estimates of $\gamma$ are far more reproducible than naive estimates in the independent All of Us cohort[45] (Extended Data Fig. 2b,c). Furthermore, we observed greater enrichment of genes associated with traits in functional pathways even though we did not directly use that information (Extended Data Fig. 2d,e). These improvements are important for making full use of the beneficial features of LoF burden tests while reducing unwanted noise. Therefore, we used the GeneBayes posterior mean effect sizes in Fig. 2c,d and for the remainder of the paper. For further discussion about the choice of prior information for GeneBayes, see the Supplementary Note.

## Gene regulation shapes genetic signals

Next, we investigated whether Perturb-seq from K562 could allow us to interpret genetic associations in the context of the gene-regulatory network. Perturb-seq estimates the effect of knocking down a gene $x$ on the expression of another gene $y$, which we denote as $\beta_{x\to y}$ (Methods). $\beta_{x\to y}$ represents the total effect of $x$ on $y$, including both direct and indirect pathways through the gene-regulatory network. Previous studies using perturbations to interpret GWAS have identified enrichment of hits in co-regulated gene sets, sometimes referred to as 'programs'[29–33], but have had limited success at identifying GWAS enrichment among program regulators (Supplementary Note).

As an initial proof of concept, we focused on the genes encoding constituents of adult haemoglobin. We focused on the gene *HBA1*, which is the only one abundantly expressed in K562 cells, and which has one of the largest LoF effect sizes for MCH ($\gamma_{HBA1} = -1.5$). We reasoned that if K562 Perturb-seq is relevant for interpreting MCH, then genes that regulate *HBA1* should also be associated with MCH. Moreover, we should be able to predict the direction of effect on MCH from the Perturb-seq data: positive regulators of *HBA1* should, themselves, have promoting effects on MCH, and vice versa for negative regulators. (Note that we refer to genes with negative $\beta$ or negative $\gamma$ from knockdown or LoF, respectively, as promoting and coloured them red; positive $\beta$ and $\gamma$ are considered repressing and coloured blue).

As predicted, we found that across all 9,498 genes that were perturbed and also tested in the LoF burden test, the LoF effect of a gene $x$ on MCH, denoted $\gamma_x$, is significantly positively correlated with the knockdown effects of that gene on *HBA1* expression, $\beta_{x\to HBA1}$ ($\beta$-coefficient = 0.052, $P = 3\times10^{-7}$; Fig. 3a). Of note, among the perturbed genes, of the top ten genes ranked by LoF effects on MCH, seven had nominally significant Perturb-seq effects on *HBA1*, and for all seven, the sign of the Perturb-seq $\beta$ matched what we predicted from $\gamma$.

We also attempted a similar analysis for GWAS hits, testing whether significant GWAS hits were enriched near *HBA1* regulators (Fig. 3b). We observed that GWAS hits were enriched (OR = 2.1 for the top 200 regulators), but to a lesser extent than for significant LoF burden test hits (OR = 6.3 for the top 200 regulators). This cannot be solely explained by inaccurate gene linking, as the same set of GWAS hits showed high enrichment for some of the gene sets (Fig. 2c and Extended Data Fig. 3a). This suggests a benefit of LoF burden tests over GWAS for identifying the trait-relevant regulatory networks.

We were curious whether similar patterns of correlation between LoF effect and Perturb-seq regulatory effects might be found for other genes or other traits. Consistent with the central role of *HBA1* in determining the MCH phenotype, we found that the correlation of $\gamma_x$ with $\beta_{x\to y}$, which we call regulator–burden correlation, was the highest for $y$ = *HBA1* among all genes expressed in K562 cells (Fig. 3c). As a negative control, we also tested for correlations between regulatory effects on *HBA1* with LoF effects on unrelated traits. As expected, we only detected *HBA1* regulator signals for erythroid traits (Extended Data Fig. 3b). These *HBA1* regulator signals for traits were also detected if we used raw burden effect estimates without applying GeneBayes, but with weaker significance (Extended Data Fig. 2f,g), supporting our approach.

Another key question for Perturb-seq studies is whether regulatory relationships learned in one cell type—K562 in this case—are useful for studying traits that are determined by less-related cell types. To examine this, we computed the regulator–burden correlation for all expressed genes, with LoF $\gamma$s for various traits. For each trait, we visualized the distribution of regulator–burden correlations in a two-sided quantile–quantile (QQ)-plot (Fig. 3d).

Starting with our three main erythroid traits, MCH, RDW and IRF, we saw that all three traits show large excesses of both positive and negative correlations compared with the null ($x = y$ line), indicating significant relationships between Perturb-seq and LoF burden tests for many genes. By contrast, there was minimal correlation between regulatory effects and $\gamma$ for other blood traits, including lymphocyte and eosinophil counts (Fig. 3d and Extended Data Fig. 3c). This suggests that cell types that are not differentiated from MEPs cannot be modelled well using K562 cells (Fig. 1c). This observation implies the importance of obtaining Perturb-seq data in trait-relevant cell types.

However, we were surprised to see that some non-erythroid traits, including serum levels of IGF-1 and CRP, as well as body mass index, did show highly significant correlations of regulatory effects with $\gamma$ (Extended Data Fig. 3c). The strongest correlations were seen for insulin-like growth factor 1 (IGF-1), which connects the release of growth hormone to cell growth, acting on many cell types[46]. Further examination revealed that these signals appear to be driven by the regulation of cellular growth markers, including *MKI67*. We hypothesize that essential programs for cellular growth may be broadly shared across cell types that regulate IGF-1 and other traits that share this signal (Extended Data Fig. 3d). Indeed, with further analysis using Perturb-seq in additional cell lines, we confirmed the broad sharing of regulatory effects on the essential programs associated with IGF-1 (Extended Data Fig. 10 and Supplementary Note).

Together, these results confirm the relevance of gene-regulatory relationships learned from Perturb-seq for interpreting complex traits. They highlight the role of both cell-type-specific pathways—for which the cell type used in Perturb-seq must be closely matched to the trait of interest—and broadly active pathways that may be detectable in many cell types.

## Trait-associated program regulations

We next aimed to develop a more comprehensive framework to explain genotype–phenotype associations in terms of the regulatory hierarchy inferred from Perturb-seq data. In principle, one might imagine inferring a complete gene-regulatory network from Perturb-seq that contains all causal gene-to-gene edges. However, the inference of accurate genome-scale causal graphs is extremely challenging, if not infeasible, from current Perturb-seq data.

As a more robust alternative, we followed previous work by clustering genes into co-expressed groups, referred to here as programs[30].

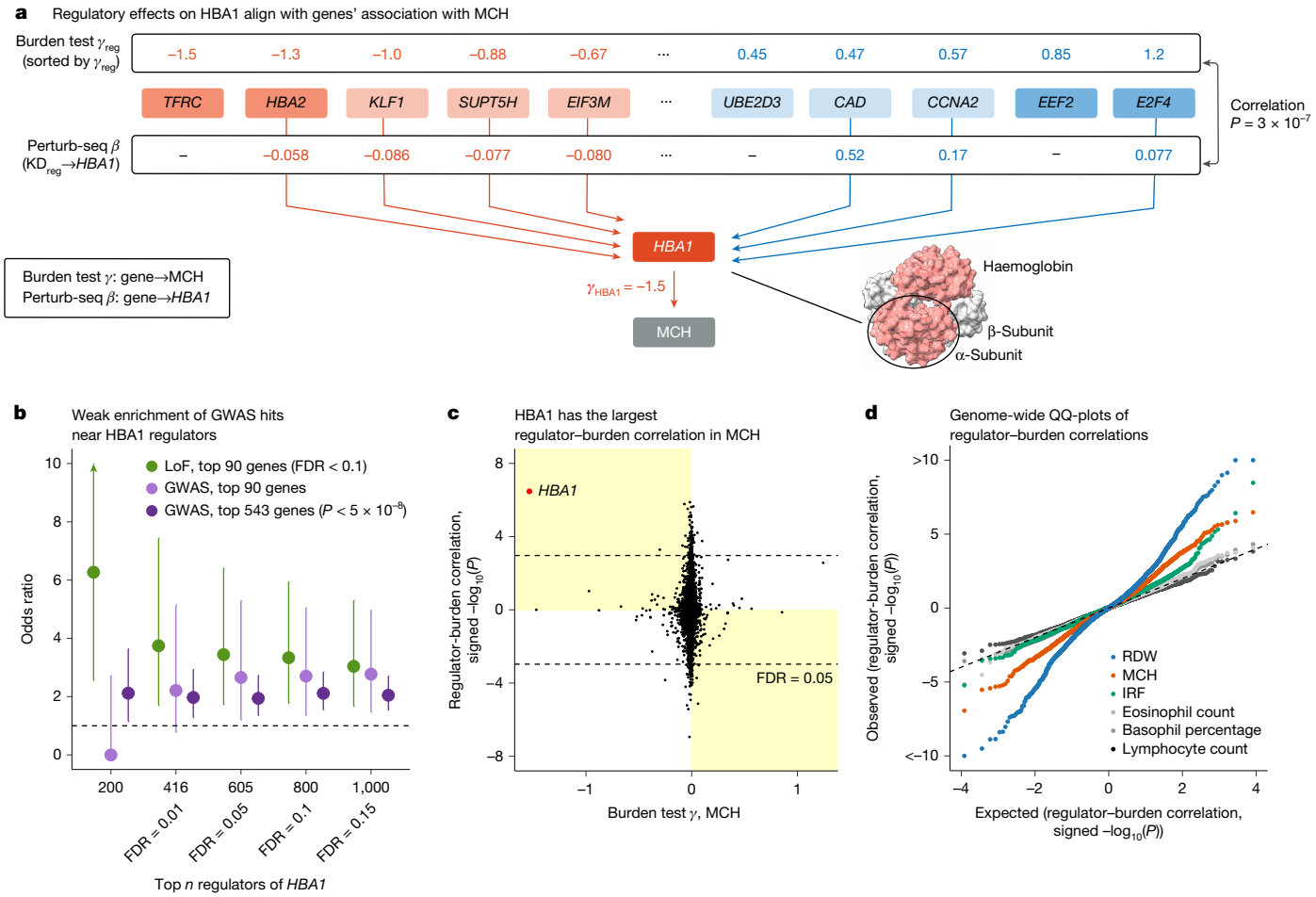



**Fig. 3 | Regulatory effects in Perturb-seq explain genetic association signals. a**, Gene effects on MCH can be predicted by regulatory effects on *HBA1*. Genes perturbed in Perturb-seq experiment are ordered by their effect sizes on MCH from LoF burden test. Perturb-seq $\beta$ refers to log fold change of *HBA1* expression after knockdown of the genes. Significant ($P < 0.05$) regulatory associations in Perturb-seq are connected with arrows. The protein structure of haemoglobin is presented using UCSF ChimeraX[66] based on Protein Data Bank entry 1A3N. The *P* value is from the linear regression and is two-sided. $KD_{reg}$, knockdown of a regulator. **b**, Enrichment analysis testing whether the top *n HBA1* regulators (ranked by *P* values) are enriched at LoF or GWAS top hits. GWAS hits are the closest genes to the independently associated variants (Methods). Points indicate the odds ratio in the exact Fisher's test. Enrichment was calculated with all the perturbed genes in Perturb-seq as a background. The error bars indicate 95% confidence intervals. The *P* values for the enrichment of the top 200 *HBA1* regulators are $9.6 \times 10^{-5}$ for the top 90 LoF hits, 0.65 for the top 90 GWAS hits and 0.01 for the top 543 GWAS hits. **c**, For every expressed gene in K562, regulator–burden correlation is plotted against their $\gamma$ for MCH. The *y* axis shows the $-\log_{10}(P)$ of the regulator–burden correlation, multiplied by the sign of the correlation. The *P* values are from the linear regression. Quadrants with a yellow background correspond to 'concordant' association, in which the sign of regulator–burden correlation aligns with the sign expected from the $\gamma$ of the gene. **d**, Genome-wide QQ-plots for regulator–burden correlations among representative traits. Each dot represents one gene. Traits without significant signals lie along the dotted line. For other traits, see Extended Data Fig. 3c. *P* values are from the linear regression.

To identify programs, we applied consensus non-negative matrix factorization (cNMF)[47] to the gene expression matrix from Perturb-seq (Fig. 4a). This allowed us to quantify the activity of each program in every cell. Similar to ref. 30, we then used the perturbation data to estimate the causal regulatory effects of knockdown of every gene *x* on the activity of each program *P*, denoted as $\beta_{x \to P}$.

On the basis of the preliminary analyses, we chose to model the data using 60 programs (see Methods; Extended Data Fig. 4a). We found that a large fraction of the 60 programs successfully captured biological pathways (Supplementary Table 3). Using external ENCODE data[48], we found evidence for coordinated transcriptional control of many programs: for 49 of the 60 programs at least one transcription factor showed significant binding site enrichment near program genes and knockdown of that transcription factor significantly changed program expression (Extended Data Fig. 4b and Supplementary Table 3).

We next quantified the average effects of programs, and their regulators, on traits (Fig. 4b). To measure program effects, we note that in NMF, the gene loadings on each program are non-negative by definition. Thus, a natural measure of the effect of a program on a trait is simply to compute the average LoF effects ($\gamma$s) of highly loaded genes as a measure of the effect of that program on the trait. We refer to this as the program burden effect. A positive program burden effect is interpreted to mean that the program has a repressing function on the trait; a negative value implies that it is promoting. Significance was determined by permutations (Methods).

To measure the effects of regulators of program *P* on each trait, we needed to account for the fact that distinct regulators can have either positive or negative effects on *P*. Thus, for each program *P*, we computed the correlation across regulators, *x*, of $\beta_{x \to P}$ with $\gamma_x$. We refer to this measure as the regulator–burden correlation; this measure is analogous to the measure of regulatory effects used for single genes above. A positive regulator–burden correlation is interpreted to mean that upregulation of program *P* promotes the trait; a negative value suggests that upregulation of *P* has a repressing effect on the trait.

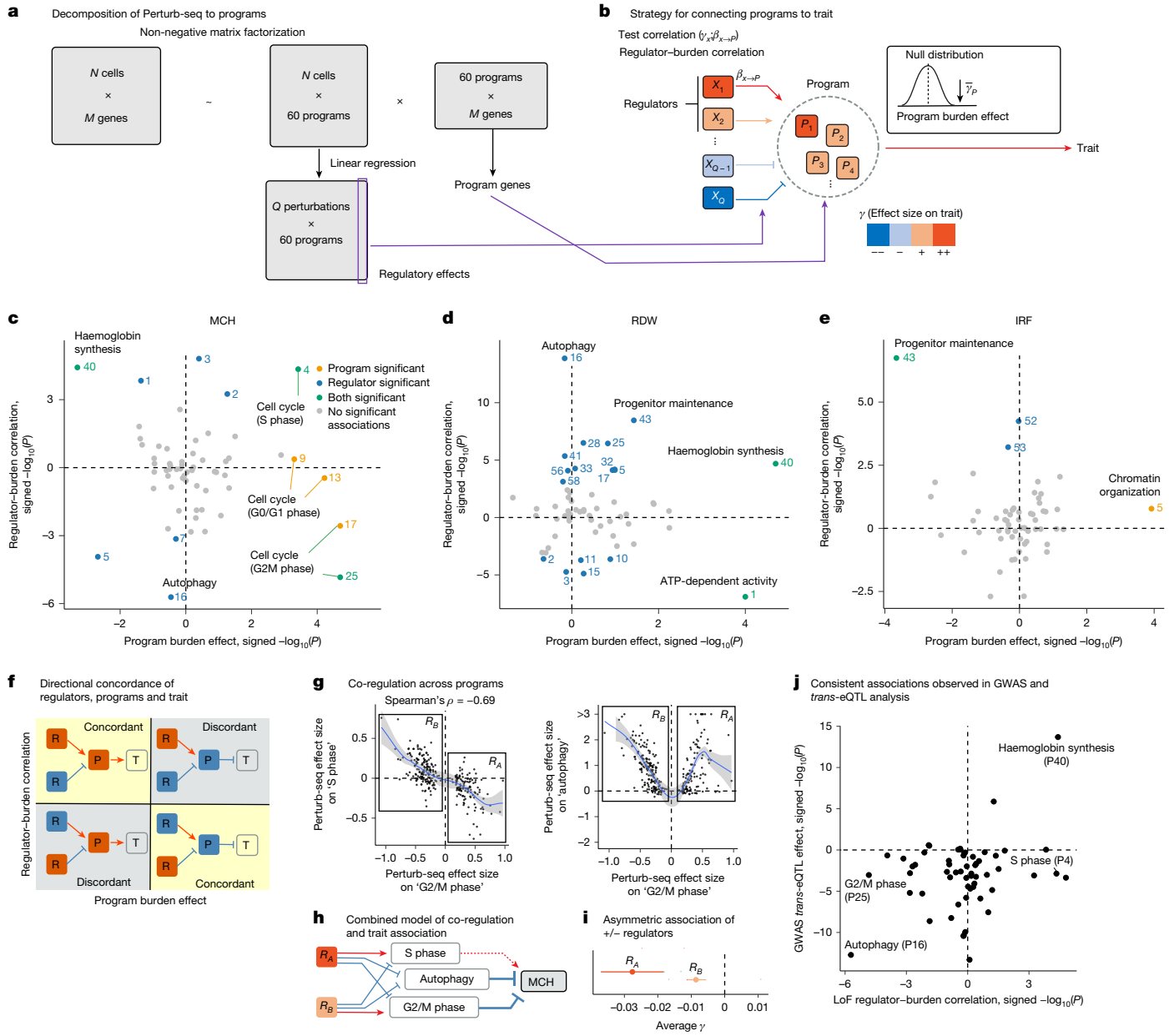



**Fig. 4 | Association of program regulation with blood traits. a**,**b**, Overview of our pipeline for the analysis to find the trait-relevant programs. **c**–**e**, Program burden effects (*x* axis) and regulator–burden correlation (*y* axis) of 60 programs in three blood traits: MCH (**c**), RDW (**d**) and IRF (**e**). Programs with significant associations after Bonferroni correction ($P < 0.05/60$) are coloured. Pathway annotations of representative programs are labelled. For annotations of other programs, see Supplementary Table 3. The *P* values for program burden effects are from the permutation test and are two-sided. The *P* values for regulator–burden correlations are from the linear regression and are two-sided. **f**, Schematic for the concordant and discordant patterns between program burden effects and regulator–burden correlation. P, program; R, regulator; T, trait. **g**, Co-regulation patterns between programs. Each dot represents a gene that has significant regulatory effects on the G2/M phase program. The gene effect size on the program activity was calculated by comparing the program usage of cells between perturbed cells and control cells using linear regression ($\beta_{x\to P}$; Methods). The lines and their 95% confidence intervals are from locally estimated scatterplot smoothing. **h**, A summary of signs of regulatory effects on the programs. **i**, Average $\gamma$ and its standard errors for 115 genes in $R_A$ and 154 genes in $R_B$ MCH. **j**, Program association with MCH in GWAS–*trans*-eQTL analysis (Methods) and LoF–Perturb-seq analysis.

The program effects on each trait are shown in Fig. 4c–e. For MCH, the haemoglobin synthesis program genes and their regulators were both significantly enriched, consistent with our single-gene analysis of *HBA1*. In addition, five programs associated with the cell cycle were all enriched in the program burden effect axis. This mirrors the enrichment of this pathway from the over-representation analysis of GWAS and LoF top hits (Fig. 2c), but here we can confirm the enrichment of both regulators and program genes for these programs (Fig. 4c).

For RDW, the program reflecting ATP-dependent activity was highlighted from both program and regulator axes (Fig. 4d). Iron is incorporated into haem in the mitochondria, and its dysregulation results in high RDW. In extreme cases, mitochondrial dysfunction leads to sideroblastic anaemia, characterized by high RDW[49]. The association of the ATP activity program with RDW is consistent with this biology. For IRF, the program representing the maintenance of the erythroid progenitor population was enriched for both program and regulator axes (Fig. 4e). This program showed the enrichment of binding sites for transcription factors that are important for the maintenance of stem cell and progenitor populations, including TAL1, NFIC, MAX and MNT[50–52] (Extended Data Fig. 4b).

Overall, the Perturb-seq data efficiently captured biological pathways and their regulators, and comparison with gene associations enable us to identify the pathways relevant to each trait.

## Complex interplay of programs

Although the significant programs in Fig. 4c–e provide insight into biological controls of these three traits, they also revealed puzzling inconsistencies. Some programs, including haemoglobin synthesis for MCH, show consistent directional effects for program genes and program regulators, but for other programs, the directions of effects initially appeared to be inconsistent (Fig. 4f). Examination of these programs revealed important principles about the regulatory architecture of programs, and design considerations for building regulatory models of complex traits.

The first principle is revealed by three programs with strong effects on MCH: the S and G2/M phase cell-cycle programs, and the autophagy program. For the G2/M phase, the program and regulator effects have directionally concordant effects on MCH, but for the S phase, the program genes and their regulators imply effects with opposite directions. In addition, for autophagy, only the regulators—but not the program genes—show a signal (Fig. 4c).

One piece of this puzzle is explained by considering patterns of co-regulation across the three programs: (1) regulators of the S phase and G2/M phase programs are shared but affect the programs in opposite directions (Fig. 4g); and (2) most G2/M and S phase regulators also affect autophagy, but the knockdown effect on autophagy is almost always positive (Fig. 4g and Extended Data Fig. 5a). These relationships are intuitive: S phase and G2/M phase are mutually exclusive components of the cell cycle; meanwhile, autophagy is suppressed during mitosis, and cell-cycle regulators are known to have a key role in that suppression[53].

To describe these patterns in a simple way, we defined two sets of regulator genes, denoted $R_A$ and $R_B$, according to their effects on G2/M (Fig. 4g). To determine how the regulators of these three programs affect MCH, we fit their effects jointly in a multiple regression model. This analysis showed that G2/M and autophagy regulators both have independent repressive effects on MCH (Extended Data Fig. 5b). The opposite co-regulation of S and G2/M phase programs explains the opposite correlation of these regulators with $\gamma$ (Fig. 4c). A summary of the joint model of regulator effects is shown in Fig. 4h.

One prediction of this model is that $R_A$ regulators should have stronger (more negative) genetic effects on MCH $\gamma$s than $R_B$ regulators. This is because $R_A$ genes have a repressive effect on both G2/M and autophagy, and both programs have repressive effects on MCH; whereas for $R_B$, the positive regulator effects on G2/M and the negative regulator effects on autophagy partially cancel the effects of each other on MCH. Indeed, consistent with this model, we saw that both $R_A$ and $R_B$ have significantly negative $\gamma$s on average, but $R_A$ is much more strongly negative (Fig. 4i).

These observations emphasize the need for joint modelling of programs and show that the observed effect sizes of regulators on a trait can be modelled as sums of regulatory effects mediated through key pathways. A different form of crosstalk between programs, involving a negative-feedback loop affecting RDW, as well as the distinct relationships of program genes and their regulators with a trait, is discussed in detail in the Supplementary Note.

## Validation with GWAS and *trans*-eQTL

Although the enrichment of GWAS hits to regulators was modest (Fig. 3b), we hypothesized that we might find consistent regulatory effects of GWAS variants on the core pathways if we take the direction of effect into account. We utilized *trans*-eQTL effects in peripheral blood[14] to test the directional regulatory effects of GWAS top hits on the programs identified by Perturb-seq (Fig. 4j, Supplementary Fig. 3 and Supplementary Note). Although the size of each *trans*-eQTL effect is small, MCH GWAS hits had directionally consistent regulatory effects on the haemoglobin synthesis and autophagy programs ($P = 2 \times 10^{-14}$ and $2 \times 10^{-13}$, respectively). The direction of regulation by GWAS top alleles was concordant with what we inferred from our Perturb-seq and LoF burden test model. This indicates that GWAS and the LoF burden test converge on the regulation of shared core pathways.

## Unified graphs from genes to programs to traits

We next aimed to build regulatory maps that link genes, programs and traits into coherent, unified models. Our goals in doing so are twofold: (1) we wanted to understand, in compact form, the main molecular processes that control a set of traits; and (2) we wanted to interpret, and even predict, the directions of effects of important trait-associated genes.

For each trait, we selected the top-ranked programs by program burden effects and, separately, in a joint regression model, the top ranked programs by regulator–burden correlations (Extended Data Fig. 6; Methods). On the basis of our analysis above, we allowed programs and program regulators to have independent effects in the model. After model selection, this procedure resulted in a graph that, for MCH, included five programs and three sets of program–regulators, as well as the inferred direction of effect of each program and regulator set on MCH (Extended Data Fig. 7).

A simplified representation of the MCH graph is depicted in Fig. 5a, showing haemoglobin synthesis, cell cycle and autophagy as critical controls of MCH. The direction of the genetic association of top genes on MCH was generally consistent with this model (43 out of 59 predicted correctly). The overall prediction accuracy was significantly higher than expected under a null model, using both leave-one-out cross-validation ($P = 5 \times 10^{-5}$) and permutation analyses for which we repeated the entire inference procedure ($P < 5 \times 10^{-5}$; Methods; Extended Data Fig. 8a,b). This approach allowed us to connect the gene-level top hits, identified solely from genetic association studies, to their functions in the pathway regulatory map.

Examining the graph, we were intrigued that *SUPT5H*, which is involved in transcriptional elongation[37], has regulatory effects on all three programs. Perturb-seq shows that *SUPT5H* activates haemoglobin synthesis, and inhibits autophagy and the G2/M phase cell cycle, all of which result in increased MCH (Fig. 5a). Thus, our model predicts that *SUPT5H* is a master regulator for MCH, exerting same-direction effects via three different pathways. Indeed, the effect sizes of *SUPT5H* LoFs on MCH are among the largest of all genes (Fig. 2d), and LoFs in this gene can cause a thalassaemia phenotype[54]. Thus, this map can help us to interpret why genes are associated with a trait.

In addition to MCH, we also inferred gene-to-pathway-to-trait maps for RDW and IRF, revealing both shared and independent pathways of regulation across the three traits (Extended Data Fig. 8c–h). There were four programs whose regulators were significantly and independently associated with at least one trait (Fig. 5b): progenitor maintenance, haemoglobin synthesis, autophagy and cell cycle. Previous studies of haematopoiesis confirmed that all four pathways regulate essential aspects of erythrocyte maturation[37,39,55,56] (Fig. 5c).

The multi-trait regulator graph (Fig. 5b) helps us to interpret the concordance and discordance of genetic associations across the traits. Genome-wide, MCH and RDW are negatively correlated in both GWAS data ($r_g = -0.39$) and at significant burden loci (Fig. 2d). We can now interpret these observations as probably driven by opposite direction effects of both autophagy and cell cycle on these two traits. Conversely, RDW and IRF are positively correlated ($r_g = 0.15$; Extended Data Fig. 9a), at least in part because both traits are positively regulated by progenitor maintenance.

We can also use the graph to understand how individual genes affect the different traits. For example, 16 genes in the graph have strong

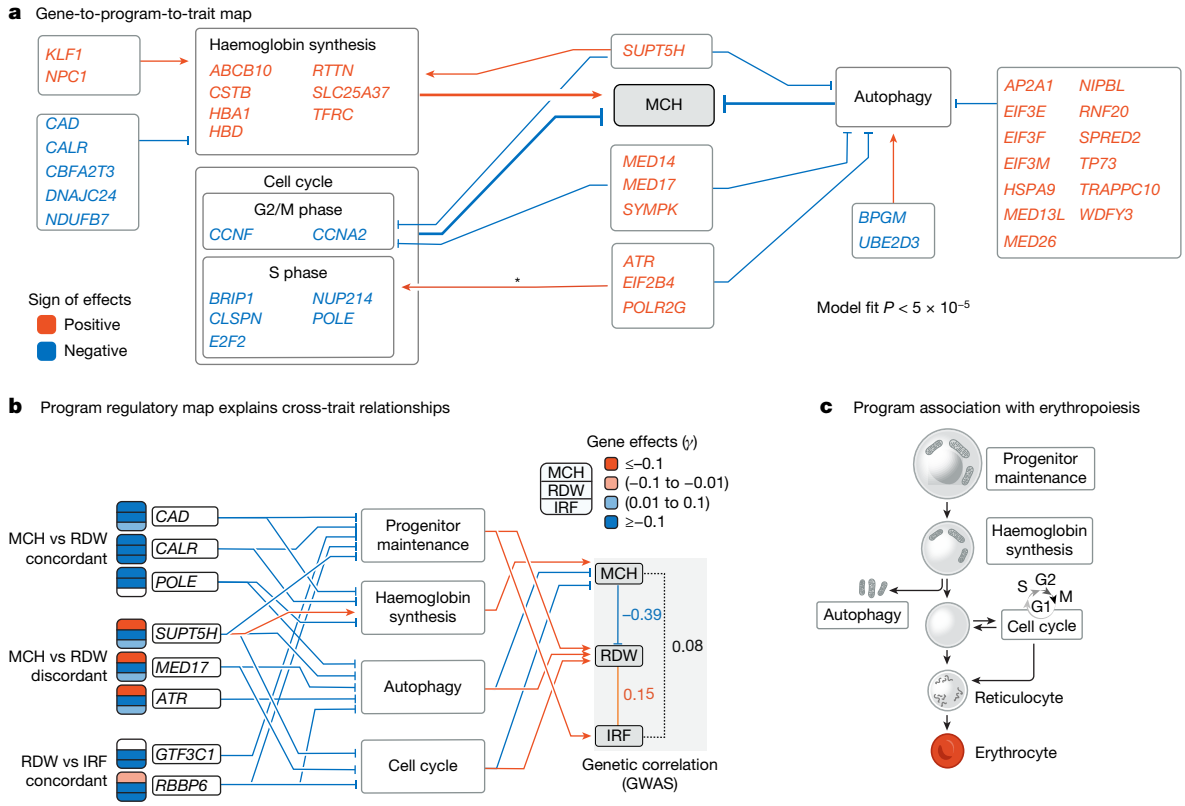



**Fig. 5 | Association map of genes to programs to traits. a**, Regulatory map of MCH. Programs were selected by genome-wide association patterns of regulators or program genes with the trait (Methods). Top hits for MCH ($|\gamma| > 0.1$) whose effect directions were concordant with the model are placed onto the map. The colour of the genes indicate the direction of effects on the trait (sign($\gamma$)); red denotes increase MCH with upregulation of the gene. The arrow with the asterisk was not selected in the initial program selection process. The *P* value is from the permutation test and is one-sided. **b**, Sharing of regulatory networks across traits. Here the arrows from gene to programs indicate the regulatory directions. Programs were selected if their regulators were found to be associated with at least one trait in the gene-to-program-to-trait map. The arrows from programs to traits were determined based on a joint regression model (Extended Data Fig. 9c). Regulatory directions on cell cycle pertain to G2/M phase. *POLE* is also a member of the S phase cell-cycle program. **c**, Programs identified in our model are associated with biological processes that are essential for erythrocyte maturation.

opposite-direction effects on MCH and RDW; our model correctly predicts opposite signs for 14 out of the 16, including *SUPT5H*, *MED17* and *ATR* (Fig. 5b and Extended Data Fig. 9b). For instance, *MED17* inhibits both the G2/M phase cell cycle and autophagy; both effects increase MCH and reduce RDW, with the result that *MED17* increases MCH and reduces RDW.

By contrast, three genes in the graph differ from the genome-wide pattern, showing large same-direction effects on MCH and RDW. Our model correctly predicts two of these, and is suggestive for the third (*POLE*; Supplementary Note). Specifically, *CAD* and *CALR* both have repressive effects on RDW and MCH. Figure 5b suggests why: unlike most genes that affect both RDW and MCH through shared pathways, these genes affect the two traits via independent pathways: progenitor maintenance and haemoglobin synthesis. Both genes inhibit both pathways, but regulation of progenitor maintenance affects RDW and not MCH, whereas haemoglobin synthesis affects MCH but not RDW.

## Extension of the model to other traits

Current availability of genome-wide Perturb-seq data in different cell types and cell conditions is limited. Nonetheless, we assessed the generalizability of our model by analysing Perturb-seq experiments with a limited number of perturbations in multiple additional cell lines—HepG2, Jurkat and RPE1 (refs. 2,57) (Supplementary Table 4)—along with additional complex traits (Extended Data Fig. 10, Supplementary Figs. 4–9 and Supplementary Note). We observed cell-type specificity of regulator–burden correlations, with burden effects for erythroid traits being more enriched in gene-regulatory effects in K562 cells than in other cell lines, whereas burden effects for HDL-cholesterol was more enriched in HepG2 cell-regulatory effects (Extended Data Fig. 10). In addition, we identified trait-specific patterns of regulator association (Supplementary Fig. 5). These findings, along with others (Supplementary Note), indicate that with more diverse and detailed gene regulation information, we can better understand the biology of a broad range of traits.

## Discussion

Genetic associations serve a unique role in studies of human biology, as they can establish causal links from variants or genes to human traits and diseases. Yet, some 20 years after the first GWAS, we still lack genome-scale approaches for inferring interpretable, quantitative models of the biological pathways that connect genes to cellular functions to traits. Here we built on previous work in this area[29–33] to develop the first approach to infer unified graphs linking directional effects of genes on traits via pathways of regulation and cellular functions. Although our work focuses on blood traits that underlie anaemia and related diseases, we anticipate that the principles learned here can be broadly applicable.

One essential feature of this paper is that we built graphs using quantitative gene effects estimated from LoF burden tests instead of unsigned enrichment of GWAS hits. We envisage LoFs and GWAS hits as reflecting the same underlying biological pathways[41,42], but our results are both more significant, and more interpretable, when using LoFs.

Unlike GWAS hits, LoF effect sizes are inherently directional, they are automatically linked to the correct genes, and their magnitudes are comparable across genes. Moreover, compared with common variants with tiny effects, LoFs are probably more functionally similar to CRISPR knockdowns, given the widespread non-linear and even non-monotonic relationships between gene expression and phenotypes[58,59].

Although the model presented here is relatively simple, there will surely be value in future models that add complexity. Future versions could allow for more complex representations of gene-regulatory networks, more explicit modelling of regulatory crosstalk between programs and heterogeneity of gene functions within programs. Many traits are controlled by multiple cell types, and one can envision models in which genetic effects on traits are controlled by a superposition of effects across multiple cell-type-specific networks.

One unexpected result from our model was the finding that the effects of program regulators on a trait may be strongly discordant from the effect of program genes on the same trait. We hypothesize that some programs reflect downstream transcriptional consequences of cell biological processes, and that the genes within a program do not always lie on the causal pathway between the program–regulators and the trait (Supplementary Note). In such cases, the identification of genes in the program can provide useful clues about biological mechanism but the effects of program genes may differ dramatically from the effects of their regulators. Moreover, it is likely that some critical processes may not be detected or may not be interpretable from RNA readouts. Thus, it will be helpful in future analyses to augment Perturb-seq experiments with other types of cell phenotyping such as functional tests, protein measurements or cell painting[60–63].

Finally, one critical challenge for using Perturb-seq to interpret association studies is how closely we need to match the cells used for Perturb-seq to the cells that determine trait variation[30]. Recent work has suggested that gene-regulatory relationships are often shared between closely related cell types, but generally not shared between more distant cell types[57,64]. Consistent with this, our results show that K562 serves as a suitable, although imperfect, model for erythrocyte development, but also that K562 is not suitable for modelling traits related to other blood cell lineages (Fig. 3d). We hypothesize that in general, Perturb-seq data will need to be closely matched to the trait-relevant cell types, but the matching does not need to be perfect.

Although our proof of principle here uses experimental data from K562 cells to model erythrocyte traits, we expect that the next generation of perturbation studies in cells, organoids and tissues[63,65] will provide a critical interpretative framework for human genetics.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-025-09866-3.


1. Backman, J. D. et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
2. Replogle, J. M. et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell* **185**, 2559–2575.e28 (2022).
3. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
4. Cohen, J. C., Boerwinkle, E., Mosley, T. H. & Hobbs, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* **354**, 1264–1272 (2006).
5. Sankaran, V. G. et al. Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. *Science* **322**, 1839–1842 (2008).
6. Minikel, E. V., Painter, J. L., Dong, C. C. & Nelson, M. R. Refining the impact of genetic evidence on clinical success. *Nature* **629**, 624–629 (2024).
7. Smemo, S. et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371–375 (2014).
8. Musunuru, K. et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).
9. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
10. Wang, K. et al. Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn disease. *Am. J. Hum. Genet.* **84**, 399–405 (2009).
11. Sinnott-Armstrong, N., Naqvi, S., Rivas, M. & Pritchard, J. K. GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. *eLife* **10**, e58615 (2021).
12. Suzuki, K. et al. Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. *Nature* **627**, 347–357 (2024).
13. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
14. Võsa, U. et al. Large-scale *cis*- and *trans*-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
15. Sobreira, D. R. et al. Extensive pleiotropism and allelic heterogeneity mediate metabolic effects of IRX3 and IRX5. *Science* **372**, 1085–1091 (2021).
16. Vuckovic, D. et al. The polygenic and monogenic basis of blood traits and diseases. *Cell* **182**, 1214–1231.e11 (2020).
17. Stanford, S. M. & Bottini, N. PTPN22: the archetypal non-HLA autoimmunity gene. *Nat. Rev. Rheumatol.* **10**, 602–611 (2014).
18. Cipolletta, D. et al. PPAR-γ is a major driver of the accumulation and phenotype of adipose tissue Treg cells. *Nature* **486**, 549–553 (2012).
19. Reshef, Y. A. et al. Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. *Nat. Genet.* **50**, 1483–1493 (2018).
20. Liu, X., Li, Y. I. & Pritchard, J. K. Trans effects on gene expression can drive omnigenic inheritance. *Cell* **177**, 1022–34.e6 (2019).
21. Richard, D. et al. Functional genomics of human skeletal development and the patterning of height heritability. *Cell* **188**, 15–32.e24 (2025).
22. Aguet, F. et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
23. Badia-i Mompel, P. et al. Gene regulatory network inference in the era of single-cell multi-omics. *Nat. Rev. Genet.* **24**, 739–754 (2023).
24. Kernfeld, E., Keener, R., Cahan, P. & Battle, A. Transcriptome data are insufficient to control false discoveries in regulatory network inference. *Cell Syst.* **15**, 709–24.e13 (2024).
25. Dixit, A. et al. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–66.e17 (2016).
26. Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
27. Adamson, B. et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882.e21 (2016).
28. Jaitin, D. A. et al. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell* **167**, 1883–1896.e15 (2016).
29. Freimer, J. W. et al. Systematic discovery and perturbation of regulatory genes in human T cells reveals the architecture of immune networks. *Nat. Genet.* **54**, 1133–1144 (2022).
30. Schnitzler, G. R. et al. Convergence of coronary artery disease genes onto endothelial cell programs. *Nature* **626**, 799–807 (2024).
31. Geiger-Schuller K. et al. Systematically characterizing the roles of E3-ligase family members in inflammatory responses with massively parallel Perturb-seq. Preprint at *bioRxiv* https://doi.org/10.1101/2023.01.23.525198 (2023)
32. Yao, D. et al. Scalable genetic screening for regulatory circuits using compressed Perturb-seq. *Nat. Biotechnol.* **42**, 1282–1295 (2023).
33. Weinstock, J. S. et al. Gene regulatory network inference from CRISPR perturbations in primary CD4$^+$ T cells elucidates the genomic basis of immune disease. *Cell Genom.* **4**, 100671 (2024).
34. Neale Lab. *UK Biobank GWAS summary statistics, 2018* (Neale Lab, accessed 1 October 2022); http://www.nealelab.is/uk-biobank/.
35. Andersson, L. C., Jokinen, M. & Gahmberg, C. G. Induction of erythroid differentiation in the human leukaemia cell line K562. *Nature* **278**, 364–365 (1979).
36. Tusi, B. K. et al. Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature* **555**, 54–60 (2018).
37. Martell, D. J. et al. RNA polymerase II pausing temporally coordinates cell cycle progression and erythroid differentiation. *Dev. Cell* **58**, 2112–2127.e4 (2023).
38. Gnanapragasam, M. N. et al. EKLF/KLF1-regulated cell cycle exit is essential for erythroblast enucleation. *Blood* **128**, 1631–1641 (2016).
39. Sankaran, V. G. et al. Cyclin D3 coordinates the cell cycle during differentiation to regulate erythrocyte size and number. *Genes Dev.* **26**, 2075–2087 (2012).
40. Szustakowski, J. D. et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nat. Genet.* **53**, 942–948 (2021).
41. Spence J. P. et al. Specificity, length and luck drive gene rankings in association studies. *Nature* https://doi.org/10.1038/s41586-025-09703-7 (2025).
42. Weiner, D. J. et al. Polygenic architecture of rare coding variation across 394,783 exomes. *Nature* **614**, 492–499 (2023).
43. Koch, J. et al. CAD mutations and uridine-responsive epileptic encephalopathy. *Brain* **140**, 279–286 (2016).
44. Zeng, T., Spence, J. P., Mostafavi, H. & Pritchard, J. K. Bayesian estimation of gene constraint from an evolutionary model with gene features. *Nat. Genet.* **56**, 1632–1643 (2024).
45. Bick, A. G. et al. Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).
46. Hakuno, F. & Takahashi, S. I. 40 years of IGF1: IGF1 receptor signaling pathways. *J. Mol. Endocrinol.* **61**, T69–T86 (2018).
47. Kotliar, D. et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-seq. *eLife* **8**, e43803 (2019).
48. Abascal, F. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

49. Fleming, M. D. Congenital sideroblastic anemias: iron and heme lost in mitochondrial translation. *Hematology* **2011**, 525–531 (2011).
50. Souroullas, G. P., Salmon, J. M., Sablitzky, F., Curtis, D. J. & Goodell, M. A. Adult hematopoietic stem and progenitor cells require either Lyl1 or Scl for survival. *Cell Stem Cell* **4**, 180–186 (2009).
51. Holmfeldt, P., Jennifer, P. & McKinney-Freeman, S. Nfi genes are novel regulators of murine hematopoietic stem-and progenitor cell survival. *Blood* **122**, 2987–2996 (2013).
52. Amati, B. & Land, H. Myc–Max–Mad: a transcription factor network controlling cell cycle progression, differentiation and death. *Curr. Opin. Genet. Dev.* **4**, 102–108 (1994).
53. Odle, R. I. et al. An mTORC1-to-CDK1 switch maintains autophagy suppression during mitosis. *Mol. Cell* **77**, 228–240.e7 (2020).
54. Achour, A. et al. A new gene associated with a β-thalassemia phenotype: the observation of variants in SUPT5H. *Blood* **136**, 1789–1793 (2020).
55. Fader, C. & Colombo, M. I. Multivesicular bodies and autophagy in erythrocyte maturation. *Autophagy* **2**, 122–125 (2005).
56. Hattangadi, S. M., Wong, P., Zhang, L., Flygare, J. & Lodish, H. F. From stem cell to red cell: regulation of erythropoiesis at multiple levels by multiple proteins, RNAs, and chromatin modifications. *Blood* **118**, 6258–6268 (2011).
57. Nadig, A. et al. Transcriptome-wide analysis of differential expression in perturbation atlases. *Nat. Genet.* **57**, 1228–1237 (2025).
58. Milind, N., Smith, C. J., Zhu, H., Spence, J. P. & Pritchard, J. K. Buffering and non-monotonic behavior of gene dosage response curves for human complex traits. Preprint at *medRxiv* https://doi.org/10.1101/2024.11.11.24317065 (2024).
59. Naqvi, S. et al. Precise modulation of transcription factor levels identifies features underlying dosage sensitivity. *Nat. Genet.* **55**, 841–851 (2023).
60. Feldman, D. et al. Optical pooled screens in human cells. *Cell* **179**, 787–799.e17 (2019).
61. Way, G. P. et al. Morphology and gene expression profiling provide complementary information for mapping cell state. *Cell Syst.* **13**, 911–923.e9 (2022).
62. Frangieh, C. J. et al. Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nat. Genet.* **53**, 332–341 (2021).
63. Rood, J. E., Hupalowska, A. & Regev, A. Toward a foundation model of causal cell and tissue biology with a Perturbation Cell and Tissue Atlas. *Cell* **187**, 4520–4545 (2024).
64. Arce, M. M. et al. Central control of dynamic gene circuits governs T cell rest and activation. *Nature* **637**, 930–939 (2024).
65. Engreitz, J. M. et al. Deciphering the impact of genomic variation on function. *Nature* **633**, 47–57 (2024).
66. Meng, E. C. et al. UCSF ChimeraX: tools for structure building and analysis. *Protein Sci.* **32**, e4792 (2023).

# Methods

## Datasets

**GWAS data.** We downloaded the publicly available GWAS summary statistics and SNP heritability estimates for traits in the UKB from Ben Neale's laboratory (see the URL section below). We focused on traits with SNP heritability estimates exceeding 0.04.

**LoF data.** We used LoF burden test summary statistics from the UKB with 454,787 participants, as previously reported[1]. Specifically, we utilized the gene-level aggregated effect estimates from predicted LoF variants with a minor allele frequency of less than 0.01%. Data were downloaded from the GWAS Catalog[67].

**Perturb-seq data.** We utilized the genome-wide Perturb-seq dataset in K562 reported by Replogle et al.[2]. In this dataset, all expressed genes ($n$ = 9,866) were targeted by a multiplexed CRISPRi sgRNA library in K562 cells engineered to express dCas9–KRAB. Single-cell RNA sequencing was performed to read out the sgRNAs together with the transcriptome. Only cells with a single genetic perturbation were used for the analysis, amounting to a median of 166 cells per gene perturbation and 11,499 unique molecular identifiers per cell. We downloaded the raw count data that the authors uploaded to figshare (see the URLs in the Code availability section).

For additional analyses, we utilized Perturb-seq data for essential genes in K562, RPE1 (ref. 2), HepG2 and Jurkat[57] cell lines. Only cells with a single genetic perturbation were used for the analysis. The number of perturbations and the number of cells per perturbation are summarized in Supplementary Table 4. We downloaded the raw count data uploaded to figshare (see the URLs in the Code availability section) or the Gene Expression Omnibus (GSE264667).

**ChIP–seq data.** We utilized chromatin immunoprecipitation followed by sequencing (ChIP)–seq data in K562 for annotating gene programs. We downloaded 830 transcription factor ChIP–seq narrow peak files from the ENCODE project website[48] (see the URL in the Code availability section). All coordinates were mapped to hg19 with LiftOver[68].

## Linkage disequilibrium score regression

To identify traits whose heritability is enriched in open chromatin regions in K562, we used S-LDSC[9]. All GWAS data were preprocessed with the 'munge_sumstats.py' script provided by the developers (see the URLs in the Code availability section). Variants in the HLA region were excluded from the analysis. The assay for transposase-accessible chromatin using sequencing (ATAC-seq) replicated narrow peak bed file in K562 was downloaded from ENCODE[48] (GSE170378, ENCFF590CPE), and the coordinates were mapped to hg19 using LiftOver[68]. Furthermore, we used narrow ATAC-seq peaks from 18 haematopoietic progenitor, precursor and differentiated cell populations previously reported[69].

For the additional analysis, replicated narrow peak files from ATAC-seq experiments for HepG2 and $CD4^+$ T cells were downloaded from ENCODE[48] (ENCFF439EIO and ENCFF246KRE), and the coordinates were mapped to hg19 using LiftOver[68]. For RPE1 (ref. 70) and Jurkat[71], as narrow peak files for ATAC-seq experiments were not available, we downloaded SRA files from the US National Institutes of Health NCBI Sequence Read Archive (SRR30621812 for RPE1, and SRR12368304 and SRR12368305 for Jurkat) and called the peaks. Specifically, we trimmed the adapter sequence with TrimGalore (v0.5.0)[72], aligned to the hg19 reference with Bowtie2 (v2.3.4.1)[73], filtered duplicates with MACS3 (v3.0.3)[74] and called narrow peaks with the MACS3 (v3.0.3) hmmratac command.

Linkage disequilibrium (LD) scores were calculated for each annotation using the 1000 G Phase 3 European population (ref. 75). The heritability enrichment of each annotation for a given trait was computed by adding the annotation to the baseline LD score model (v1.1) and regressing against trait chi-squared statistics for HapMap3 SNPs. These analyses used v1.0.1 of the LDSC package (see the URL in the Code availability section).

Furthermore, we tested the genetic correlation between specific trait pairs using European LD scores with the LDSC package (v1.0.1).

## Estimation of gene effect sizes with GeneBayes

**Method overview.** LoF burden tests are not well powered, especially for shorter or selectively constrained genes, as the likelihood of having LoF variants in these genes is low. We previously developed GeneBayes[44], an empirical Bayes framework aimed at addressing a similar challenge: the precise estimation of selective constraint on genes, which can be particularly challenging for short genes. Within GeneBayes, we used gene features in a machine learning-based empirical Bayes framework to improve the accuracy of constraint estimates. Diverse gene features, such as gene expression patterns and protein structure embeddings, can enhance the accuracy of these estimates. GeneBayes is a highly adaptable framework, easily extendable to various applications, as outlined in the original article[44]. In this instance, we utilized it to derive more precise effect size estimates for LoF burden tests.

To minimize overfitting when applying GeneBayes to LoF burden test estimates, we first performed feature selection using the BoostRFE function (boost recursive feature elimination) from the shap-hypetune package (see the URL in the Code availability section) to fit XGBoost[76] models on the sign and magnitude of $\hat{\gamma}$, the estimated effect size from LoF burden test summary statistics. We used the predicted sign and magnitude as the features for GeneBayes, which we found to perform better than using the selected features directly; this may be due to differences in training dynamics between XGBoost and the gradient-boosted trees fit using GeneBayes.

Subsequently, we implemented the GeneBayes framework as previously described. Specifically, GeneBayes involves two steps: (1), learning a prior for the effect size of each gene through the utilization of gradient-boosted trees, as implemented in NGBoost[77], and (2), estimating gene-level posterior estimates of the effect sizes using a Bayesian framework. In our application of GeneBayes, we parameterize the prior as follows:

$$\text{sign}(\gamma) \sim \text{Bernoulli}(p)$$

$$\text{magnitude}(\gamma) \sim \text{Gamma}(\alpha, \theta)$$

The parameter $p$ is the probability that $\gamma$ is positive or negative, and $\alpha$, $\theta$ are the shape and scale parameters of the Gamma distribution, respectively. We learned the parameters of the prior using the following likelihood:

$$\hat{\gamma} \mid \gamma \sim \text{Normal}(\gamma, \text{s.e.}(\hat{\gamma}))$$

The summary statistics $\hat{\gamma}$ and s.e.($\hat{\gamma}$) are the estimated effect size and its standard error from the LoF burden tests, respectively.

**Gene features.** We compiled the following types of gene features from several sources: selective constraint of genes ($S_{het}$)[44], gene expression across cell types, protein embeddings and gene embeddings.

$S_{het}$ refers to a reduction in fitness for heterozygous carriers of a LoF variant in any given gene. We utilized the $S_{het}$ estimated in our previous work[44]. Gene expression across 79 single-cell types was downloaded from the Human Protein Atlas[78] (see the URL in the Code availability section). Protein embeddings were adopted from embeddings learned by an autoencoder (ProtT5) trained on protein sequences[79]. Gene embeddings were derived from GeneFormer, a pretrained deep learning model for single-cell transcriptomes[80]. Specifically, we used the Cell×Gene Discover census (see the URL in the Code availability section), and we extracted 1,000 cells per each of the

cell types—'erythroid progenitor cell', 'monocyte', 'erythrocyte', 'fibroblast', 'T cell', 'neutrophil', 'B cell' and 'haematopoietic stem cell'—and computed the average embeddings of each gene for the cellular classifier using the EmbExtractor module (see the URL in the Code availability section).

Finally, we used the posterior mean of the LoF burden test effect size as a point estimate for the following analyses.

**Traits.** As applying GeneBayes to all UKB traits is computationally intensive, we applied this to a subset of traits including all the blood cell-associated traits, blood biomarkers and some of anthropometric traits. A list of traits included in our analyses has been provided in Supplementary Table 5.

### LoF burden test in the All of Us cohort

The All of Us dataset contains whole-genome sequencing together with various laboratory measurements[45]. On 5 February 2025, the values for MCH were reported for 213,787 sequenced individuals after filtering (UKB data-field 30050, AoU ID 3012030). For individuals with data from multiple visits, we took the latest visit, and we excluded outliers (more than 50 or less than 0 pg). Our previous analysis suggested that relatedness and population structure have a minimal effect on burden test results[58]. Therefore, we performed our tests on all individuals that passed our filtering criteria. We included the top 16 genotype principal components, which are provided in the data release. In addition, we generated 20 rare variant principal components using FlashPCA2 (ref. 81) on variants sampled uniformly at random from the rare variant fraction (minor allele count (MAC) > 20, minor allele frequency (MAF) < 1%). We identified high-confidence LoF sites using the Variant Effect Predictor in Ensembl[82] with the LOFTEE plugin[83] and restricted our analyses to variants with MAF < 1%.

We performed burden tests using REGENIE[84], largely following the procedure previously described[58], which is based on ref. 1. We used HapMap SNPs[85] extracted from the ACAF call set (a set of variants provided by All of Us filtered on MAC and MAF) to perform the whole-genome regression in the first step of REGENIE. We included age, sex, age-by-sex, age squared, 16 genotyping principal components and 20 rare variant principal components as covariates in both the first and the second steps. We used the rank-inverse-normal transform on the phenotypes in both steps. The burden mask aggregated all LoF sites with MAC > 5 and MAF < 1% into a single burden genotype for each gene. We added Gaussian noise to summary statistics generated from fewer than 20 individuals to remain in compliance with the All of Us Data User Code of Conduct. The noise was added to the effect size such that all burden tests with fewer than 20 individuals had the same standard error.

### Pathway enrichment analysis of GWAS and LoF top hits

**Clumping of GWAS top variants.** To identify independently associated GWAS variants, we used PLINK (v1.90b5.3)[86] with the –clump flag, a $P$ value threshold of $5 \times 10^{-8}$, a linkage disequilibrium threshold of $r^2 = 0.01$ and a physical distance threshold of 10 Mb. In addition, we merged SNPs located within 100 kb of each other and selected the SNP with the minimum $P$ value across all merged lead SNPs to avoid the false inclusion of genes that have neighbour genes with extremely large effects. This resulted in 634 independent variants associated with MCH. For each independent variant, we annotated the nearest protein-coding gene. To accomplish this, we used the bedtools (v2.30.0)[87] closest module to identify genes that overlap with the variant or have their transcription start site or transcription end site closest to the variant. Furthermore, we excluded genes in the HLA region due to extensive linkage disequilibrium. Finally, we obtained a list of 543 genes possibly associated with MCH GWAS signals.

**Pathway enrichment analysis.** We aimed to compare the pathways enriched in GWAS and LoF top hits for MCH. As pathways, we utilized all ontology terms in Gene Ontology[88] with a minimum of 20 genes and a maximum of 2,000 genes, as well as MsigDB hallmark genesets[89] that include the haem synthesis pathway. We utilized enrichGO and enricher functions in clusterProfiler[90] package in R for the analysis.

Among the enriched pathways, genes in the 'positive regulation of macromolecule biosynthetic process' pathway overlap significantly with those in the 'autophagy' pathway ($P = 2 \times 10^{-8}$), and thus its enrichment may reflect the relevance of autophagy pathway.

**Enrichment analysis of GWAS and LoF top hits to *HBA1* regulators.** For the evaluation of the enrichment of GWAS top hits related to *HBA1* regulators (Fig. 3b), we used the list of 543 closest genes to the independent GWAS hits defined above. We ranked the genes based on the $P$ values of their regulatory effects on *HBA1* expression. For each of the different thresholds for *HBA1* regulators, we evaluated the enrichment using a two-sided Fisher's exact test, using all the genes perturbed in the Perturb-seq as a background. Specifically, the columns of the 2 × 2 table for the test correspond to whether the genes are *HBA1* regulators at each threshold, whereas the rows correspond to whether the genes are GWAS top hits.

In addition, for comparison, we evaluated the enrichment of 90 significant genes in the LoF burden test (FDR < 0.1) and the genes closest to the top 90 independent GWAS hits.

### Estimation of gene-regulatory effects from Perturb-seq

We aimed to estimate gene-to-gene regulatory effects from Perturb-seq. We assessed the total effects of gene knockdown on gene expression by comparing perturbed and non-perturbed cells. After filtering out cells with fewer than 500 genes expressed and genes expressed in fewer than 500 cells, we compared the cells with perturbation of every gene versus the cells with non-targeting control gRNAs. Log-normalized counts of cells were used as input to the limma-trend pipeline[91], while accounting for gel bead-in-emulsion (GEM) group (batch effect), number of genes expressed and the percentage of mitochondrial gene expression as covariates. We utilized the $\log_2$ fold change (logFC) of gene expression in perturbed cells compared with non-targeting cells as a point estimate of the perturbation effect on gene expression ($\beta_{x \to y}$).

### Defining gene programs and the regulatory effects of genes

**Identification of gene programs with cNMF.** From a single-cell gene expression matrix, we identified the co-regulated set of genes. Intuitively, such a set of genes can correspond to genes that determine cellular identity or specific cellular processes, which we call programs. To identify gene programs and their activity in each cell, we applied the cNMF[47] method to the single-cell gene expression matrix from Perturb-seq.

Matrix factorization models the gene expression data matrix as the product of two lower rank matrices, one specifying the proportions in which the programs are combined for each cell, and a second encoding the relative contribution of each gene to each program[47] (Fig. 4a). We refer to the first matrix as a 'usage' matrix[47]. In cNMF, the usage matrix is normalized so that the usage values for each cell sum to 1. We used the normalized matrix as a usage of each program in each cell.

In cNMF, a meta-analysis of multiple iterations of NMF was performed to obtain a 'consensus' result. In cNMF, the number of programs ($K$) is a key model hyperparameter to tune. To determine it, we tested different values of $K$ (30, 60, 90 and 120) and decided to proceed with $K = 60$ based on the error versus stability comparison (Extended Data Fig. 4a), as proposed by the authors. In addition, we used density threshold = 0.5 to filter out the outlier programs.

**Annotation of programs to biological pathways.** From the gene-by-program matrix produced by cNMF, we can obtain the non-negative loadings of each gene to the program. We ranked the genes based on

the loadings and utilized the top-ranked genes for each program to characterize the biological pathways of the program.

Annotating the programs to specific biological processes is a multifaceted task. In this study, for each program, we considered three orthogonal lines of evidence for annotating biological pathways.

**Gene Ontology enrichment of top genes.** We examined the enrichment of the top 200 genes in the Gene Ontology categories and MsigDB hallmark gene sets using the enrichGO and enricher functions in the clusterProfiler[90] package in R. To calculate the enrichment, we utilized genes expressed in K562 cells as a background set to avoid bias.

We tested different thresholds for determining the top genes (100, 200, 300, 400 and 500). The Gene Ontology enrichment results were generally consistent, but we observed a trend: as the number of included genes increased, more categories were enriched in at least one program, with fewer categories specifically enriched for one program (Supplementary Fig. 10).

Capturing a wide range of biological categories, as well as annotating specific categories to the programs, is important for interpretation. Thus, we chose to use the top 200 genes for the Gene Ontology enrichment analysis.

**Enrichment of transcription factor-binding sites.** We can expect that for some programs, the genes within the same program are coordinately regulated by specific transcription factors. Such transcription factors can be used to characterize the programs. To this end, we utilized the ChIP–seq experiments of transcription factors in K562 from the ENCODE project. To convert the information on binding sites to a gene-level regulation score, we calculated the following score for each transcription factor ($i$) for each protein-coding gene ($j$), as adopted from ref. 92:

$$S_{i,j}(d) = \sum_k P_{i,k} \times e^{-x_{i,j,k}/d}$$

where $P_{i,k}$ denotes the strength of peak $k$ for transcription factor $i$ (quantified by $-\log_{10} q$ value for each peak, outputted by MACS2), $x_{i,j,k}$ denotes the distance from peak $k$ to the transcription start site of gene $j$, and $d$ represents the decay distance. The decay distance indicates the effective distance for the transcription factor and can vary depending on the transcription factors. Here we set the value to 1 kb, 5 kb, 10 kb, 50 kb, 100 kb, 500 kb or 1 Mb.

To determine which score was useful for the annotation of programs, we tested the correspondence of the score with differentially expressed genes (DEGs) after knockdown of the same transcription factor. Specifically, for each transcription factor, we listed positive or negative DEGs after knockdown in Perturb-seq (FDR < 0.1) and we compared the ChIP–seq score ($S_{i,j}(d)$) between DEGs and non-DEGs by Mann–Whitney $U$-test.

As a natural consequence, we could annotate each transcription factor as an activator or inhibitor, according to the direction of effects after knockdown. We annotated a transcription factor as an activator if the downregulated DEGs after knockdown had significantly high ChIP scores (FDR < 0.05), and as an inhibitor if the upregulated DEGs after knockdown had significantly high ChIP scores (FDR < 0.05). As a result, ChIP scores for 167 transcription factors showed significant correspondence with their knockdown effects (FDR < 0.05) and were utilized for the annotation of programs. One best decay distance parameter was selected for each transcription factor based on the significance in the overlap with DEGs.

For each program, we compared the top 300 loading genes with other expressed genes in K562 with respect to the 167 ChIP scores using the Mann–Whitney $U$-test. This test evaluates the enrichment of binding sites of the transcription factors to each program genes. Furthermore, we compared the program activity of the transcription factor-knockdown cells with others to see whether the transcription factor had a direct effect on the activity of the program (Extended Data Fig. 4b).

**Co-expression with marker genes.** In addition, we manually confirmed the co-expression of marker genes for predefined cell types or pathways and the program activity of cells in the uniform manifold approximation and projection (UMAP)[93] space. Markers for red blood cells, myeloid cells and the integrated stress response pathway were adopted from the original Perturb-seq paper[2]. S phase and G2/M phase marker gene sets were adopted from ref. 94. Markers for erythroid progenitors and megakaryocytes were determined from single-cell gene expression data of bone marrow haematopoietic progenitors[95], where we ranked the genes in each corresponding population based on expression specificity ($Z$-score) compared with other populations and selected the top 50 genes. This number of genes was determined to be roughly in the same range as the number of genes in the other gene sets.

After completing these three tests for each program, we defined the curated annotation of each program as follows: initially, when the program corresponded to specific cell types, including cellular marker genes as top-loading genes, it was annotated as the cell type. For the others, we considered them as programs reflecting cellular pathways. We prioritized the most significantly enriched Gene Ontology or MsigDB pathways from the top 10 enriched pathways while avoiding ambiguous pathways for interpretation (such as the 'RNA binding' pathway). In cases in which multiple programs were enriched for the same category, we attempted to distinguish them by their enriched transcription factors or colocalization with marker gene expression. Finally, we curated one annotation per program while considering these factors (Supplementary Table 3).

**Estimation of the regulatory effects of genes on program activity.** From the cell-by-program matrix produced by cNMF, we obtained the usage of each program in each cell. To obtain the effect size of each regulator on the program usage, we standardized the program usage to mean = 0 and s.d. = 1, and we compared perturbed cells with cells with non-targeting control gRNAs with a linear regression model, while accounting for GEM group (batch effect), number of expressed genes and percentage of mitochondrial gene expression as covariates. We utilized the point estimate of the effect size of perturbation on program usage as a regulatory effect of a gene ($\beta_{x\to P}$).

## Comparison of program regulations with genetic associations

**Definition of gene effects on traits and gene regulation.** Unless specified, we utilized the posterior estimate of gene effect size on a trait with GeneBayes as the gene effect on a trait ($\gamma$). For gene-level regulatory effects, we used the logFC of gene expression in perturbed cells compared with non-targeting cells as a point estimate of the perturbation effect on gene expression, as described above ($\beta_{x\to y}$). For program-level regulatory effects, we utilized the effect size of perturbation on program usage as a regulatory effect of a gene, as described above ($\beta_{x\to P}$).

**Correlation of gene regulatory effects with genetic associations.** We started from a simple model in which the effect size of a peripheral gene $x$ was determined by its regulatory effects on a limited set of core genes. In cases in which there was a single or a limited number of core genes $y$, the regulatory effect size of the peripheral gene on the core genes should correlate with the effect size of the peripheral gene on the trait.

We have previously observed a striking correlation between LoF burden test effect sizes and $S_{\text{het}}$ on average across traits[41]. To avoid the confounding effects of selective constraint, we included $S_{\text{het}}$ as a covariate in our linear regression model:

$$\gamma_x \sim \beta_{x\to y} + S_{\text{het},x}$$

where $\beta_{x\to y}$ corresponds to the regulatory effect of gene $x$ on gene $y$. We excluded the effects of gene $y$ itself, that is, $\beta_{y\to y}$, from the comparison because it does not reflect a trans-regulatory effect. For every expressed gene $y$, we evaluated the significance of the coefficient for

the first term. In some of the plots, the significance level was multiplied by the sign of the coefficient.

**Association of program genes with traits.** In the program-level analysis, we quantified the average effects of program genes on traits, which we call program burden effect. Program burden effects are the average $\gamma$ of the genes, which are representative of the program, as determined by the loading for the program in cNMF.

Of note, as a feature of cNMF, the loadings of the genes to the programs are always positive. Thus, the sign of the average $\gamma$ provides interpretable directional information about the program association with the trait.

As selective constraints are positively correlated with $|\gamma|$[41], highly conserved programs, such as those essential for cellular survival, could have larger program burden effects. To avoid confounding, we divided the expressed genes in K562 into ten bins based on $S_{het}$. We then compared the average $\gamma$ of the top loading genes with a 10,000 randomly chosen sets of the same number of genes, while matching for the $S_{het}$ bin. To account for the directional association, we converted the rank of the observed value compared with the random distribution into two-sided $P$ values, while adding the sign of the average $\gamma$ to calculate the signed association $P$ values.

Here the sign of program burden effects corresponds to the average effects of the LoF of program genes on the trait. Thus, positive program burden effects can be interpreted as a repressing association between program $P$ and the trait.

The results were generally not affected by the choice of the number of top genes (100, 200 and 300). However, for some programs including the haemoglobin synthesis program, where the association with MCH was concentrated on a small number of haemoglobin genes, the association was more pronounced with a smaller number of top genes. Therefore, for visualization of program burden effects and regulator–burden correlation (for example, Fig. 4), we chose 100 for defining the top genes.

**Correlation of program regulatory effects with genetic associations.** Next, we aimed to quantify the correlation of regulatory effects of genes on the program with $\gamma$, which we call regulator–burden correlation.

We calculated the correlation of regulatory effects with trait association signals while accounting for $S_{het}$ in the same way as the gene-level analysis:

$$\gamma_x \sim \beta_{x \to P} + S_{\text{het},x}$$

where $\beta_{x \to P}$ corresponds to the regulatory effect of gene $x$ on program $P$.

For every program, we evaluated the significance of the coefficient for the first term. The significance level was multiplied by the sign of the coefficient for visualization.

Here the sign of $\beta_{x \to P}$ corresponds to the effect of the knockdown of gene $x$ on the activity of program $P$. The sign of $\gamma_x$ corresponds to the effect of the LoF of gene $x$ on the phenotype. Thus, a positive regulator–burden correlation can be interpreted as a promoting association between program $P$ and the trait.

**Null distribution of burden effects.** For visualization of the distribution of burden effects of regulators or program genes (Extended Data Fig. 5h), the expected distribution of burden effect sizes was determined by randomly picking up the same number of genes from non-associated genes 10,000 times and taking their average.

### Estimation of causal relationships between programs

While examining the co-regulation patterns across programs, we noticed an asymmetric pattern of co-regulation between programs; that is, the regulators of program A also have effects on program B, but the regulators of program B do not have effects on program A (Extended Data Fig. 5c). Such asymmetry can be explained by a causal directional association from one program to the other. Biologically, this one-way association can be interpreted as positive or negative feedback from one program to the other.

A similar observation—that is, the asymmetric correlation of effects from explanatory variables between two traits—was reported in the GWAS literature[96]. For instance, when LDL cholesterol causally affects the risk of coronary artery disease, but not vice versa, the effect sizes for risk variants of LDL cholesterol show a strong correlation between the two traits, whereas those for risk variants of coronary artery disease do not show such correlation[96].

We adapted the analytic framework for causality from a previous GWAS[96] to our case. Specifically, for a pair of programs, $P_1$ and $P_2$, we identified significant regulators (FDR < 0.05) for each. We then calculate $\rho_{P_1}$, the Spearman's rank correlation of effect sizes for $P_1$ and $P_2$, considering only the regulators of $P_1$. We also calculates $\rho_{P_2}$ for the regulators of $P_2$. Next, we modelled

$$\hat{Z}_{P_1} \sim N\left(Z_{P_1}, \frac{1}{N_{P_1} - 3}\right)$$

where $Z_{P_1} = \text{arctanh}(\rho_{P_1})$ and $N_{P_1}$ corresponds to the number of significant regulators for $P_1$.

Then, we considered four patterns of association, M1: $P_1$ causally associated with $P_2$ ($Z_{P_2} = 0$); M2: $P_2$ causally associated with $P_1$ ($Z_{P_1} = 0$); M3: no relationship between $P_1$ and $P_2$ ($Z_{P_1} = Z_{P_2} = 0$); and M4: correlation does not depend on how the regulators were ascertained ($Z_{P_1} = Z_{P_2}$).

We fit each model by maximizing the corresponding approximate likelihood. We then selected the model with the smaller Akaike information criterion from the two causal models (M1 and M2) and from the two non-causal models (M3 and M4). Finally, we calculated the relative likelihood of the best non-causal model compared with the best causal model.

$$r = \exp\left(\frac{\text{AIC}_{\text{causal}} - \text{AIC}_{\text{non-causal}}}{2}\right)$$

We treated $r < 0.01$ as a threshold for causally associated programs. In the case of programs associated with RDW, the causal association from the haemoglobin synthesis program to the mitochondrial program showed $r = 8.5 \times 10^{-7}$, whereas other pairs of programs had $r > 0.05$ (also refer to Supplementary Note).

### Validation with GWAS and *trans*-eQTL

We downloaded full *trans*-eQTL summary statistics for selected variants in peripheral blood from the eQTLGen[14] website (see the URL in the Code availability section). Here 10,317 trait-associated SNPs were tested for their effects on 19,960 genes that showed expression in blood. Only SNP–gene pairs with a distance greater than 5 Mb were tested. We selected SNPs with significant associations with MCH ($P < 5 \times 10^{-8}$) in the UKB, as well as variants with $P > 0.05$ as control variants.

Using the program genes defined from cNMF in K562 (the top 100 loading genes for each program), we asked whether GWAS hits for MCH have concordant regulatory effects on the program.

Specifically, for each SNP, we derived the MCH-increasing allele based on $\beta$ coefficients from GWAS summary statistics, polarized the *trans*-eQTL Z scores of variants on program genes and calculated the average. We compared the values between GWAS significant variants and control variants using a two-sided Student's $t$-test.

### Validation of multiple program association with the trait

To test whether jointly modelling multiple programs can explain more of the genetic association signals than modelling with a single program, we conducted a cross-validation analysis. We randomly split 80% of the genes into a training set and 20% into a test set, and fitted regression models to explain the gene effects on the trait ($\gamma$) by gene-regulatory

effects on the program (or programs) using the training set. We evaluated the variance of $\gamma$ explained by the model using the test set.

We tested this with the set of multiple programs chosen from the regulator–burden correlations in gene-to-program-to-trait models for MCH and RDW, as well as with the same number of randomly chosen programs, and single program models. The selected multiple program model explained much more variance than any single program model or random combination of programs for MCH and RDW (Extended Data Fig. 6a–c). For IRF, only one program was chosen from the regulator–burden correlation in the gene-to-program-to-trait model, so we did not perform the comparison.

### Construction of the gene-to-program-to-trait model

Prevalent co-regulation across programs, as well as feedback, suggested the need to jointly model multiple programs to identify those whose regulation independently explains the trait association signals. In addition, although program burden effects and regulator–burden correlation sometimes converge on the same program, we have observed cases where either only program content or only regulators are enriched in trait association signals, as well as cases in which both program content and regulators are enriched but through different mechanisms. Therefore, we treated program burden effects and regulator–burden correlation separately to identify trait-associated programs included in the model.

**Step 1: selection of programs based on regulator–burden correlations.** To select programs whose regulators are enriched for the trait association signals, we conducted a stepwise linear regression analysis using the 'regsubsets' function in the 'leaps' package[97] in R. In this analysis, we included gene-regulatory effects on 60 programs ($\beta_{x\to p}$), as well as levels of gene constraint ($S_{het}$; as defined in ref. 44) as potential explanatory variables, with $\gamma_x$ as the dependent variable.

We identified the combination of explanatory variables through exhaustive search to determine the best subsets for predicting $\gamma_x$ in a multiple linear regression model with the given number of explanatory variables. Specifically, for MCH, we changed the number of explanatory variables from 1 to 6, and for each number of explanatory variables, we performed an exhaustive search for the combination of programs that explained the most variance of $\gamma$.

The number of variables to include in the final model was decided by assessing the variance explained in the model upon changing the number of variables (Supplementary Fig. 2a), along with the significance of the model fit in the subsequent permutation test (Supplementary Fig. 2b). For the MCH model, we opted to include three variables together with $S_{het}$: regulators for autophagy, haemoglobin synthesis and G2/M phase cell-cycle programs.

**Step 2: selection of programs based on program burden effects.** For selecting programs with enriched contents for the trait association signals, we followed the following process. First, for each program, we calculated the program burden effects. That is, we ranked the genes based on their loading and selected the top 200 genes and calculated the average of $\gamma$ of these genes. This number was determined by the following test for the model fit. Then, we compared it with randomly selected 10,000 sets of genes expressed in K562 while matching for 10 bins of $S_{het}$ to calculate two-sided enrichment $P$ values. Subsequently, we ranked the programs based on these $P$ values. To determine the number of programs to include in the final model, we varied the number of top programs included and evaluated the model fit in the subsequent permutation test (Supplementary Fig. 2b). Specifically, for the MCH model, five programs were selected: the haemoglobin synthesis program and four programs associated with different phases of cell cycle. These five programs largely corresponded to those that had significant program burden effects after Bonferroni correction in the previous test (Fig. 4c).

**Step 3: predicting the signs of associations for the regulators and program genes in the model.** After selecting programs from both regulator and program content associations with the trait, we assigned the predicted signs of effects to each gene in the model. Specifically, for regulators, we considered genes that exhibited significant regulatory effects on the selected programs (FDR < 0.05). In cases in which a regulator had regulatory effects on multiple programs, we calculated the total effects of a gene on the model by summing the product of the effect sizes of the selected programs on the trait in the multiple linear regression model ($w_p$) and the gene effects on the program ($\beta_{x\to p}$; Extended Data Fig. 6d). The sign of this product was utilized as the regulatory direction of the gene to the trait predicted from the model.

For program contents, we assigned the sign of the association of the program (that is, the sign of the average $\gamma$ of the top loading genes) to the top 200 loading genes. If a gene belongs to both program and regulator genes, although a such case was relatively rare, we assigned the sign from the program enrichment test because of the potentially larger effect sizes of program function on the trait (Supplementary Note).

**Step 4: assessing the directional concordance of the associations of top hits with the model.** To assess how well the predicted model can explain the directional genetic associations, we evaluated it in two ways: leave-one-out cross-validation and permutation testing.

For leave-one-out cross-validation, we left out one gene at a time, selected the programs based on program burden effects and regulator–burden correlation using the other genes, and predicted the sign of the left-out gene as described above. We then assessed the enrichment of correctly predicted genes among the top hits (genes with $|\gamma| > 0.1$), compared with genes with minimal associations (genes with $|\gamma < 0.01$), using Fisher's exact test. In this test, the enrichment is influenced by both (1) the enrichment of the top genes among the genes selected in the model (significant regulators or program genes in the model), and (2) the accuracy of the predicted signs among the genes in the model. Our result for the MCH model showed that the top genes were enriched in both (1) selected genes in the model (OR = 1.8), and (2) sign concordance (OR = 1.9), with an overall enrichment of $P = 5 \times 10^{-5}$ and OR = 2.2. This result supported the use of Perturb-seq for predicting the directed gene associations.

For the permutation test, we created 20,000 sets of permuted $\gamma$ by permuting gene labels. We then followed the same program selection and sign assignments processes, while fixing the number of selected programs from both the program burden effects and the regulator–burden correlation. In each permutation, we counted the number of top genes whose sign of association was correctly predicted by the model and evaluated the enrichment over other genes using Fisher's exact test. Finally, we compared the Fisher's test $P$ value of the observed data to those of the permuted sets and calculated the permutation $P$ value (Extended Data Fig. 8b,d,f). Similar to leave-one-out cross-validation, we observed that the observed genetic association data had many more concordant genes, along with a higher ratio of concordant to discordant predicted signs than the permuted data (Extended Data Fig. 8a,c,e). The permutation test can evaluate the fit of our model to the genetic association signals.

For the permuted dataset, we slightly modified the way for program selection. Here, instead of matching for $S_{het}$, we compared the distribution of $\gamma_x$ between the top loading genes and randomly selected genes expressed in K562 using the Mann–Whitney $U$-test to calculate enrichment $P$ values. Subsequently, we ranked the programs based on these $P$ values and selected the same number of top programs. This helps to greatly speed up the process, although the resulting permutation $P$ value for the model is potentially conservative.

We ran the permutation tests while differing the parameters for the modelling. The model fit to the data was robust to the choice of the number for defining program genes (100, 200 or 300) and to different thresholds for defining high-effect genes ($|\gamma|$; Supplementary Fig. 2c).

Although the enrichment was not very sensitive to the number of top genes, 200 genes resulted in slightly more stable enrichment across a range of $\gamma$ thresholds. On the basis of these results, we chose to use the top 200 genes for creating the gene-to-program-to-trait map. In addition, we chose the threshold for $|\gamma|$ to be 0.1 based on the fit of the model.

**Step 5: drawing the gene-to-program-to-trait map.** Finally, we aimed to draw a map to interpret the functions of the trait-associated genes. Here we included all the top hits with $|\gamma| > 0.1$ whose direction of association was concordant with that predicted from the model into the map (Fig. 5a). When regulators have concordant regulatory effects on multiple programs, we included all paths in the map.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

Data generated by or processed for this article are available on Zenodo[98] (https://doi.org/10.5281/zenodo.14751877). Public data used in this study are accessible via the URLs cited in the Code availability section.

## Code availability

The codes used for this article are available on Zenodo[98] (https://doi.org/10.5281/zenodo.14751877). The following URLs can be accessed: Neale laboratory UKB data (http://www.nealelab.is/uk-biobank); Replogle et al. Perturb-seq data (https://plus.figshare.com/articles/dataset/_Mapping_information-rich_genotype-phenotype_landscapes_with_genome-scale_Perturb-seq_Replogle_et_al_2022_processed_Perturb-seq_datasets/20029387; linkage disequilibrium score regression software (https://github.com/bulik/ldsc); the ENCODE database (https://www.encodeproject.org/); the shap-hypertune package (https://github.com/cerlymarco/shap-hypetune); gene expression in single-cell types (https://www.proteinatlas.org/humanproteome/single+cell+type); the CellxGene Discover census (https://chanzuckerberg.github.io/cellxgene-census/); the GeneFormer embedding extractor module (https://geneformer.readthedocs.io/en/latest/geneformer.emb_extractor.html); and the eQTLGen *trans*-eQTL data (https://www.eqtlgen.org/trans-eqtls.html).


67. Sollis, E. et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2022).
68. Hinrichs, A. S. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
69. Ulirsch, J. C. et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* **51**, 683–693 (2019).
70. Yuan, Y. et al. Single-cell analysis of the epigenome and 3D chromatin architecture in the human retina. Preprint at *bioRxiv* https://doi.org/10.1101/2024.12.28.630634 (2024).
71. Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
72. Krueger, F. et al. TrimGalore. *Zenodo* https://zenodo.org/record/7598955 (2023).
73. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
74. Zhang, Y. et al. Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9**, R137 (2008).
75. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
76. Chen, T. & Guestrin, C. Xgboost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (eds Krishnapuram, B. et al.) 785–794 (Association for Computing Machinery, 2016).
77. Duan, T. et al. Ngboost: natural gradient boosting for probabilistic prediction. In *Proc. 37th International Conference on Machine Learning* (eds Daumé, H. III & Singh, A.) 2690–2700 (PMLR, 2020).
78. Karlsson, M. et al. A single-cell type transcriptomics map of human tissues. *Sci. Adv.* **7**, eabh2169 (2021).
79. Elnaggar, A. et al. Prottrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2021).
80. Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).
81. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of biobank-scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).
82. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
83. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
84. Mbatchou, J. et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
85. Consortium, I. H. et al. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
86. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
87. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
88. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
89. Liberzon, A. et al. The molecular signatures database hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
90. Xu, S. et al. Using clusterProfiler to characterize multiomics data. *Nat. Protoc.* **17**, 3292–3320 (2024).
91. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Pprecision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
92. Ouyang, Z., Zhou, Q. & Wong, W. H. ChIP-seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc. Natl Acad. Sci. USA* **106**, 21521–21526 (2009).
93. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018).
94. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
95. Pellin, D. et al. A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nat. Commun.* **10**, 2395 (2019).
96. Pickrell, J. K. et al. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).
97. Thomas, L. based on Fortran code by Miller, A. leaps: regression subset selection. R package version 3.2. *CRAN* https://CRAN.R-project.org/package=leaps (2024).
98. Ota, M. Code repository for "Causal modeling of gene effects from regulators to programs to traits" (version v3). *Zenodo* https://doi.org/10.5281/zenodo.14751877 (2025).
99. Gilbert, L. A. et al. Genome-scale CRISPR-mediated control of gene repression and activation. *Cell* **159**, 647–661 (2014).



**Acknowledgements** This research has been conducted using the UKB resource under application number 52374. We utilized the All of Us resource under workspace ID aou-rw-c30ba93b. We thank V. G. Sankaran for valuable feedback on an earlier draft of the manuscript; J. Engreitz, K. Cromer, H. Mostafavi, R. Zhu, R. Lopez, N. Milind, C. J. Smith and the members of the Pritchard laboratory for helpful conversations; C. Theodoris for help with accessing GeneFormer gene embeddings; T. Tolpa for help with the figure design; and the reviewers for their constructive feedback. This work was funded by grants R01HG008140, R01HG011432, U01HG012069 and 1R01HG014005. A.M. received funding from the Simons Foundation, the Lloyd. J. Old STAR Award (Cancer Research Institute), the Parker Institute for Cancer Immunotherapy, the Innovative Genomics Institute, the Larry L. Hillblom Foundation (grant 2020-D-002-NET), the Northern California JDRF Center of Excellence, the Byers family, K. Jordan and the CRISPR Cures for Cancer Initiative. M.O. is supported by the Astellas Foundation for Research on Metabolic Disorder and the Chugai Foundation for Innovative Drug Discovery Science. E.D. is supported by an EMBO Post-doctoral Fellowship.


**Author contributions** M.O., A.M. and J.K.P. conceived and designed the study. M.O. performed all the data analyses. J.P.S. and E.D. contributed to the design and interpretation of the statistical analyses, and provided intellectual contributions to all aspects of the study. T.Z. contributed to the data analysis with GeneBayes. N.M. contributed to the data analysis with the All Of Us cohort. M.O. and J.K.P. wrote the manuscript, with critical input from all authors. J.K.P. and A.M. supervised the study and acquired funding.

**Competing interests** A.M. is a cofounder of Site Tx, Arsenal Biosciences and Survey Genomics; serves on the boards of directors at Site Tx and Survey Genomics; is a member of the scientific advisory boards of network.bio, Site Tx, Arsenal Biosciences, Cellanome, Survey Genomics, NewLimit, Amgen and Tenaya; owns stock in network.bio, Arsenal Biosciences, Site Tx, Cellanome, NewLimit, Survey Genomics, Tenaya and Lightcast; has received fees from network.bio, Site Tx, Arsenal Biosciences, Cellanome, Spotlight Therapeutics, NewLimit, AbbVie, Gilead, Pfizer, 23andMe, PACT Pharma, Juno Therapeutics, Tenaya, Lightcast, Trizell, Vertex, Merck, Amgen, Genentech, GLG, ClearView Healthcare, AlphaSights, Rupert Case Management, Bernstein and ALDA; is an investor in and informal advisor to Offline Ventures; and is a client of EPIQ. The Marson laboratory has received research support from the Parker Institute for Cancer Immunotherapy, the Emerson Collective, Arc Institute, Juno Therapeutics, Epinomics, Sanofi, GlaxoSmithKline, Gilead and Anthem and reagents from Genscript, Illumina and Cellanome. The remaining authors declare no competing interests.

**Additional information**

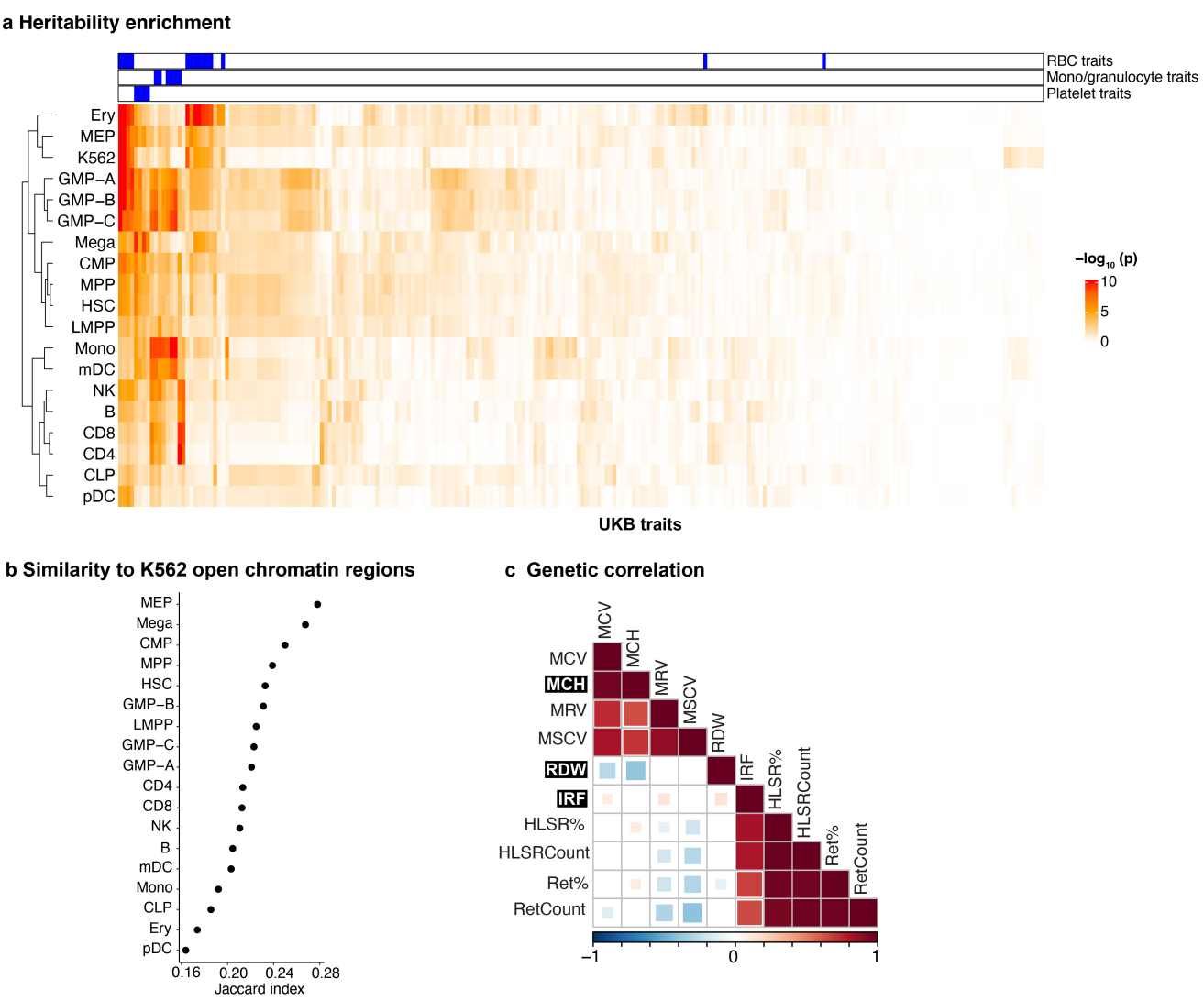

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41586-025-09866-3.



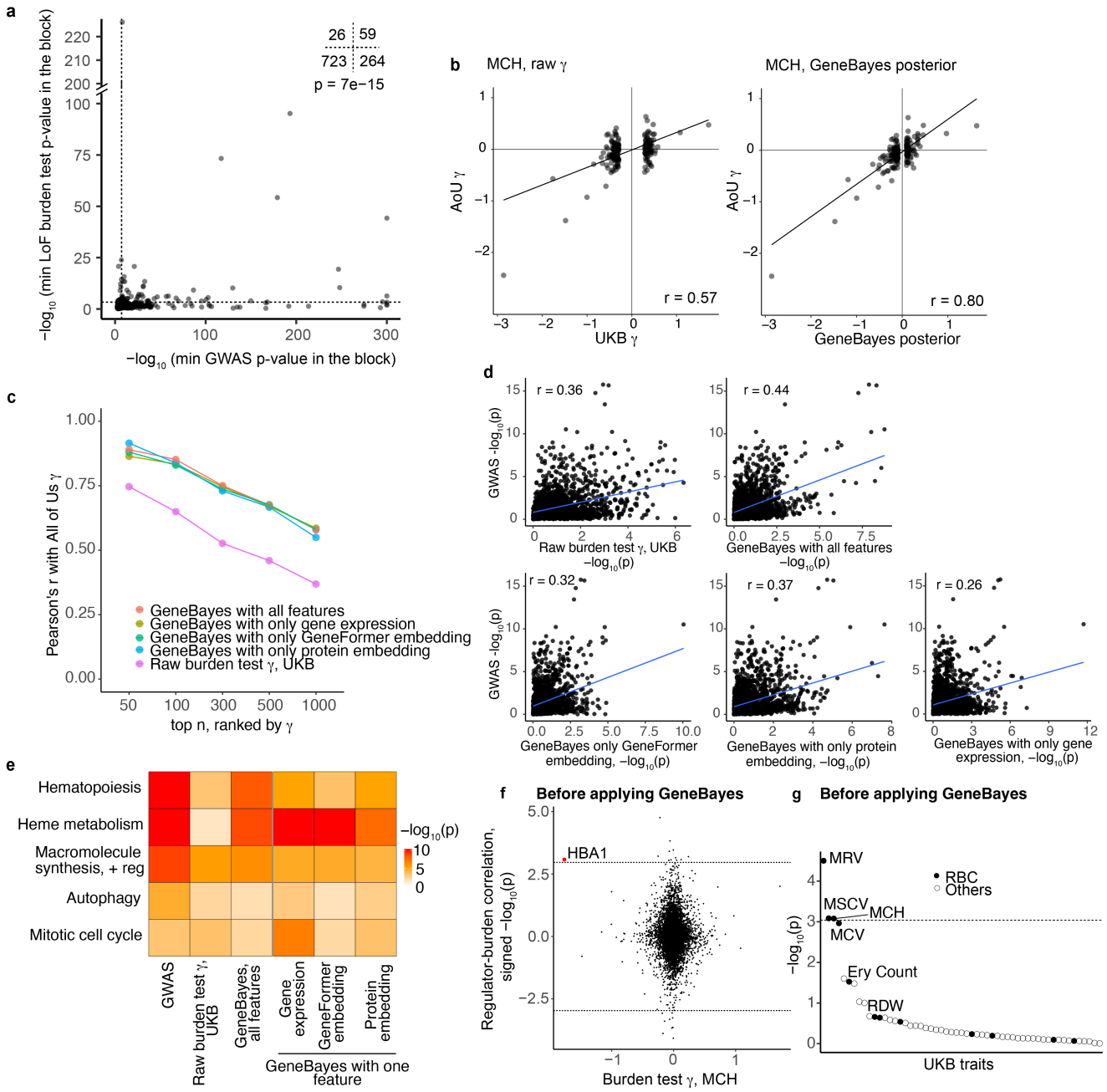
**Correspondence and requests for materials** should be addressed to Mineto Ota, Alexander Marson or Jonathan K. Pritchard.



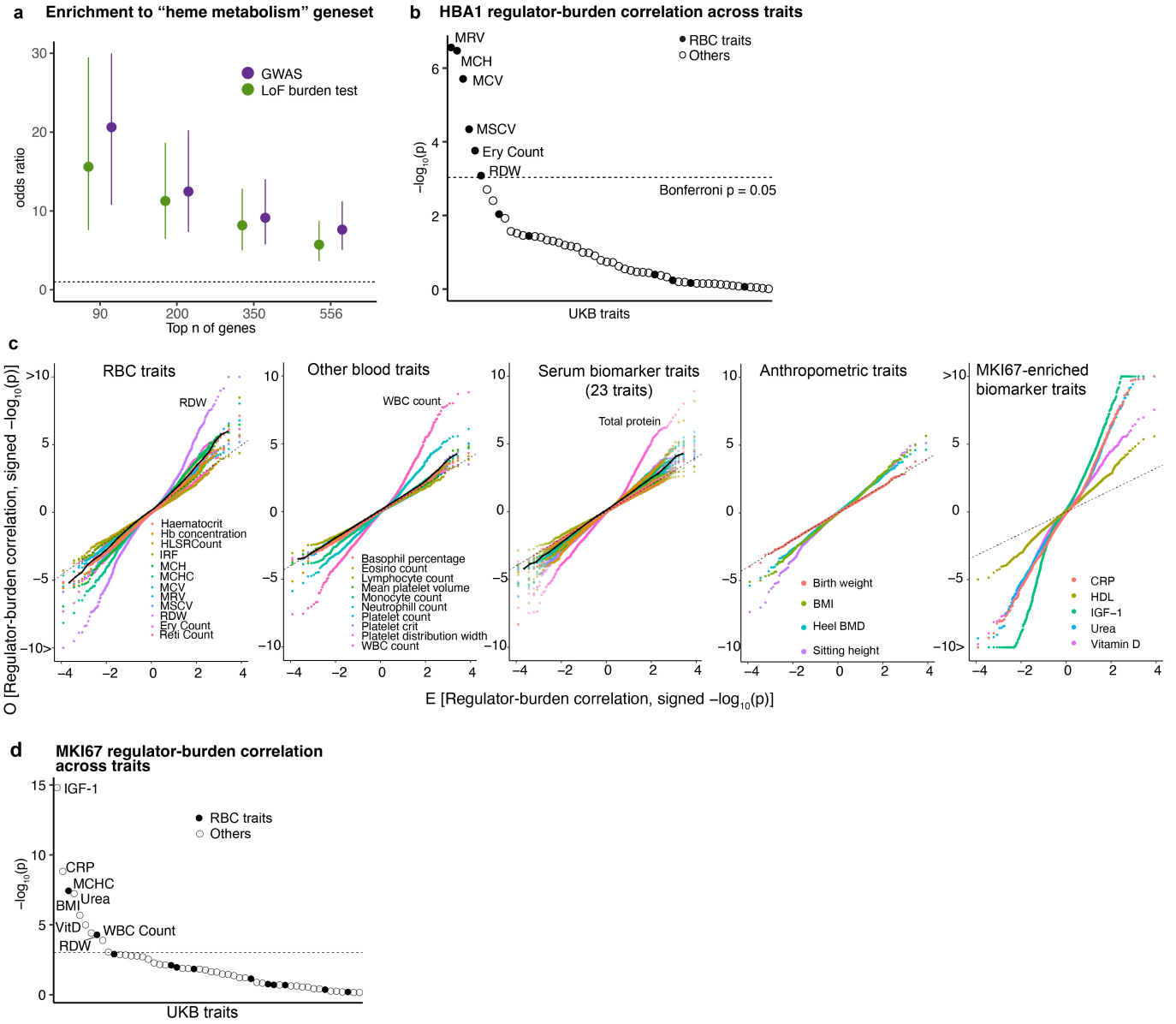
**Peer review information** *Nature* thanks Hailiang Huang, Gosia Trynka who co-reviewed with Olivier Bakker, and George Vassiliou for their contribution to the peer review of this work. Peer reviewer reports are available.



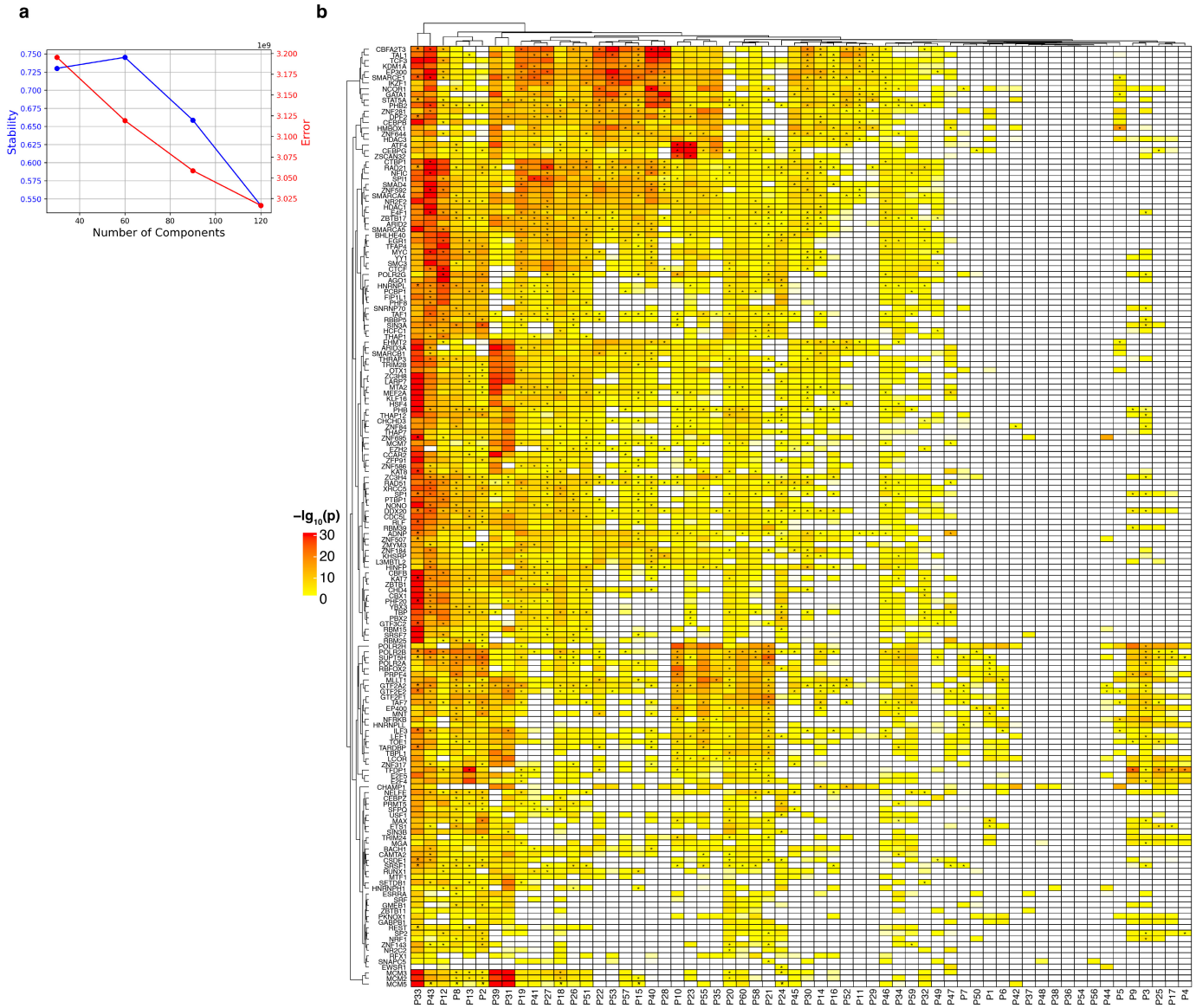

**Extended Data Fig. 1 | Analysis of the heritability of multiple traits from GWAS, related to Fig. 1. a)** Heritability enrichment of UKB traits to 18 primary hematopoietic cell types[69] and K562. Heritability enrichment was estimated with S-LDSC by adding each annotation to the baseline model. Traits associated with the morphology or quantity of RBC, monocyte/granulocyte or platelet are labeled on top. Both cell types (rows) and traits (columns) are hierarchically clustered based on their patterns of enrichment. K562 showed the closest similarity to MEP. **b)** Similarity of open chromatin regions of primary cell types to K562. Plotted are Jaccard index, which captures the proportion of open chromatin regions that are shared with K562. **c)** Genetic correlation across traits which were enriched to K562 in S-LDSC analysis.



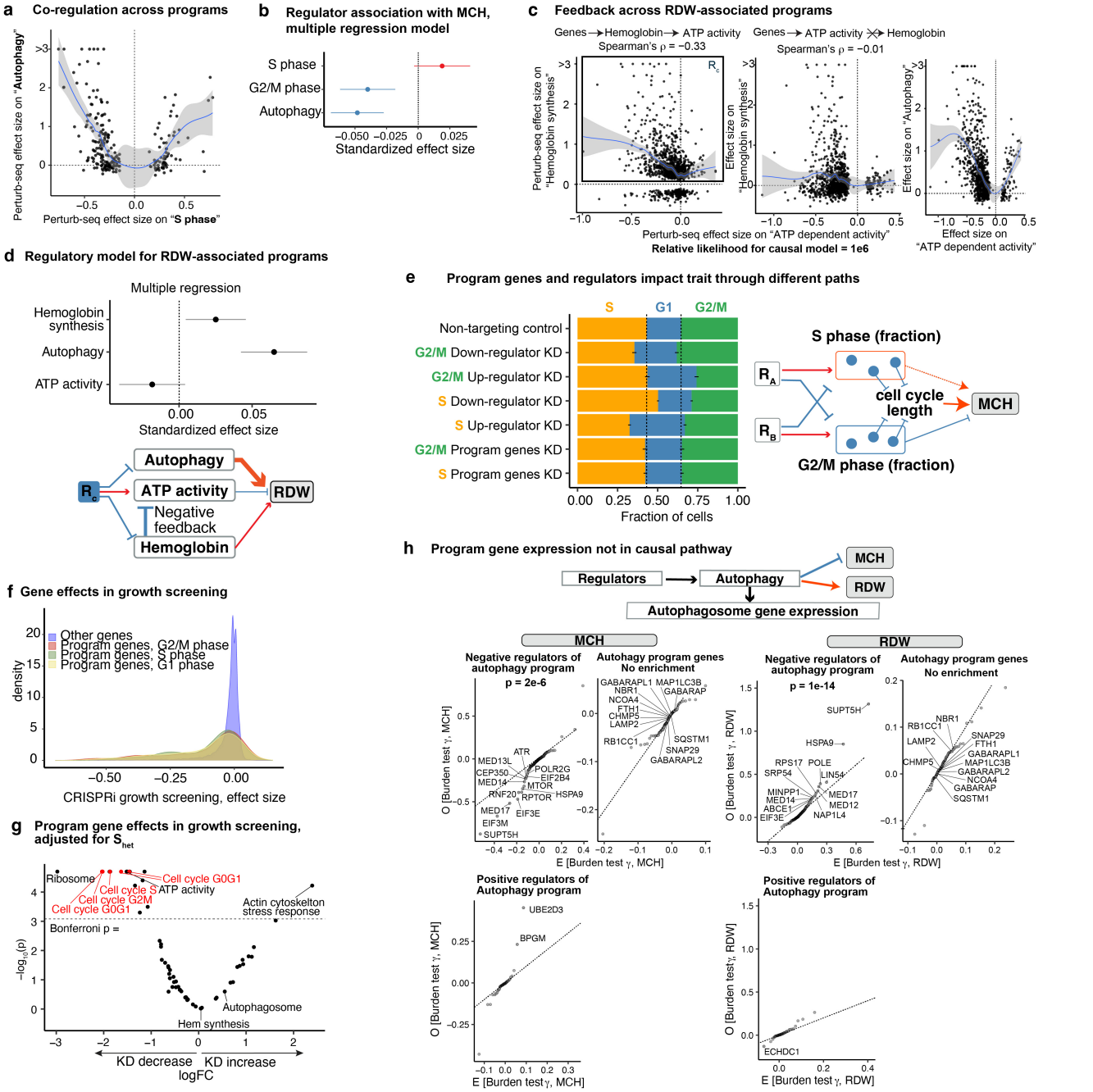

**Extended Data Fig. 2 | Evaluation of GeneBayes, related to Figs. 2 and 3.** **a)** Comparison of GWAS and LoF burden test associations for MCH. We took the minimum GWAS p-value within an LD block, and the minimum LoF burden test p-value for any gene that overlaps the LD block. Dotted lines indicate $p = 5 \times 10^{-8}$ for GWAS and $p = 5 \times 10^{-4}$, which corresponds to an FDR of 0.1 for the LoF burden test. Each dot corresponds to the LD block. Numbers of blocks in each quadrant are depicted on the top right corner. P-value is from a two-sided Fisher's exact test. **b)** Correlation of burden test $\gamma$ with All of Us. Plots are for the top 200 genes ranked by absolute values of raw $\gamma$ (left) or GeneBayes posterior (right). **c)** Correlation of burden test $\gamma$ with All of Us with different prior information. The result is for MCH. We ranked the genes based on absolute burden test effect size in UKB, either with or without applying GeneBayes with various patterns of prior information. **d)** Enrichment of GO and MsigDb hallmark pathways to top hits for MCH. The enrichment of the top 200 genes from the LoF burden test and GWAS is compared. For GWAS, the closest genes to the lead hits were ordered by p-values. For the LoF burden test, whether or not GeneBayes was applied, genes were ordered by absolute effect sizes. The GeneBayes posterior from various patterns of priors is also compared. **e)** Enrichment of top 200 genes from GWAS or LoF burden test with or without applying GeneBayes to representative pathways. The result is for MCH. **f)** Regulator-burden correlation for MCH is compared with their $\gamma$ for MCH. Same comparison with Fig. 3c, but this time using $\gamma$ before applying GeneBayes. Dotted lines indicate the same threshold with Fig. 3c. **g)** Correlation significance of *HBA1* regulatory effects with gene effects across a variety of traits. Same comparison with Extended Data Fig. 3b, but this time using $\gamma$ before applying GeneBayes. Dotted line indicates the same threshold with Extended Data Fig. 3b.



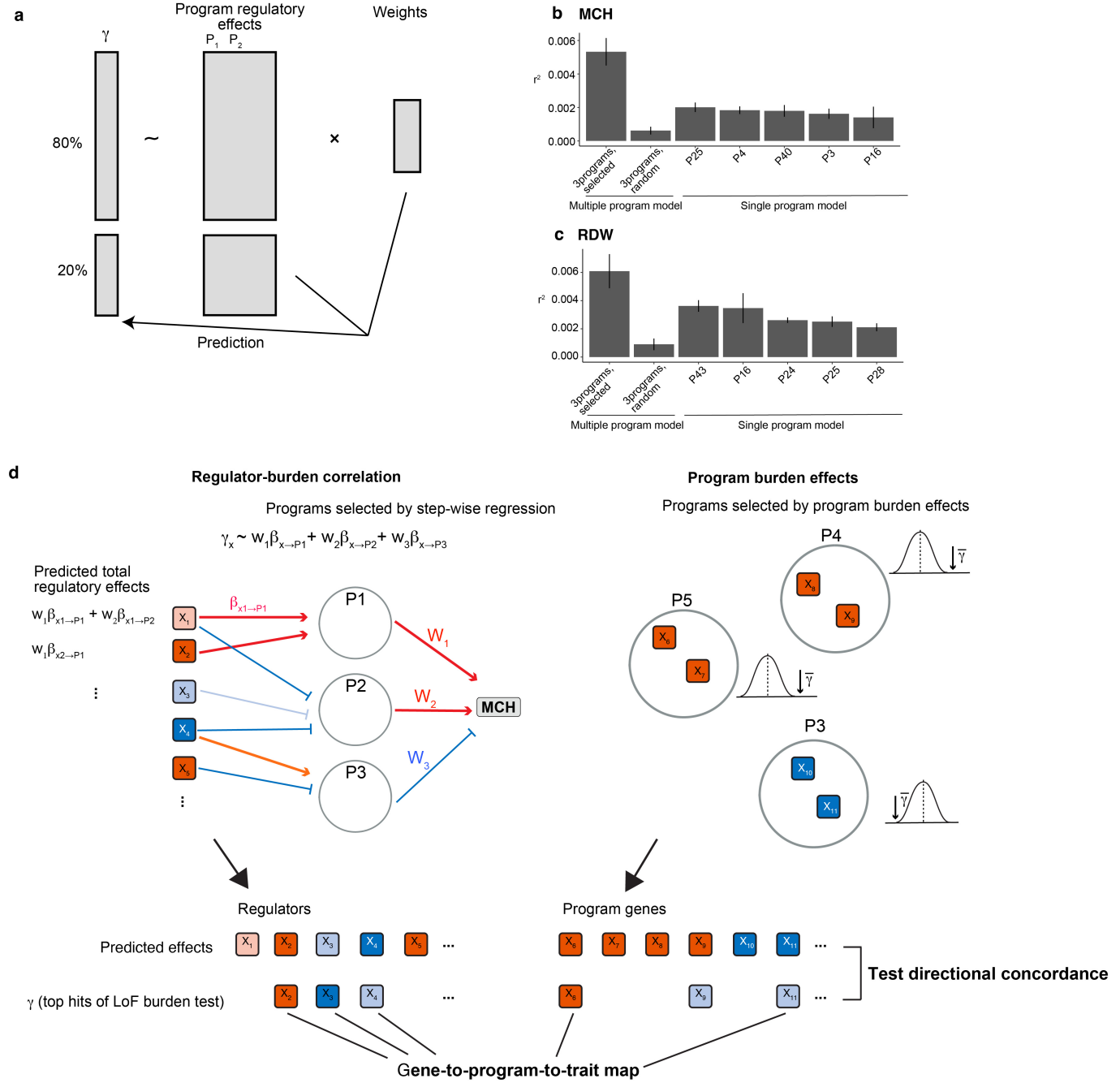

**Extended Data Fig. 3 | Relevance of gene regulatory effects on trait associations, related to Fig. 3. a)** Enrichment of hemoglobin metabolism gene set for GWAS and LoF lead hits. For both GWAS closest genes and the LoF burden test, genes were ranked by association p-values, and top gene enrichment for the gene set was assessed using Fisher's exact test. Error bars indicate 95% confidence intervals. **b)** Correlation significance of *HBA1* regulatory effects with gene effects ($\gamma$) across a variety of traits. **c)** Genome-wide QQ-plots for burden-regulator correlations for a wide variety of traits. Each dot indicates one gene. Black solid line indicates the median across each category of traits. For serum biomarker traits, 5 traits which showed extensive association with *MKI67* regulatory effects are plotted separately. **d)** Correlation significance of *MKI67* regulatory effects with gene effects across a variety of traits. Dotted line indicates the threshold for Bonferroni significance.



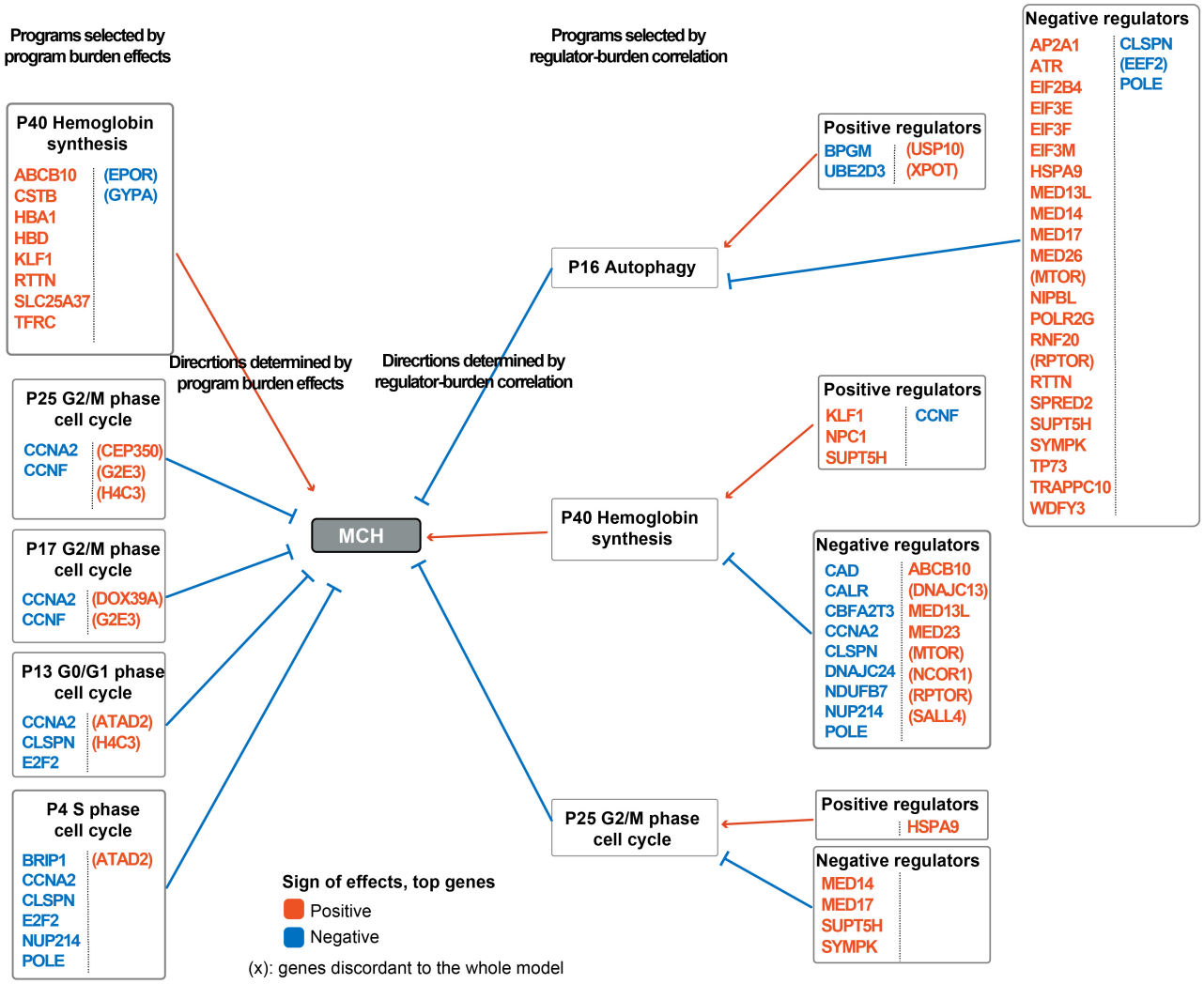

**Extended Data Fig. 4 | Annotation of programs by transcription factor binding sites, related to Fig. 4. a)** Number of cNMF components against solution stability measured by the Euclidean distance silhouette score of the clustering, and Frobenius error of the consensus solution, outputted by cNMF. **b)** Enrichment of transcription factor binding sites to program genes. Narrow peaks from ChIP-seq of transcription factors (TF) in K562 cells were used to calculate the enrichment (Methods). For significantly enriched TF-program pairs (FDR < 0.05), we tested the effect of knockdown of the TF on program activity and marked an asterisk if the KD also had an effect in the expected direction; that is, if the KD of an activator transcription factor decreased the program activity ($p < 0.05$), we marked it, and vice versa for repressor.



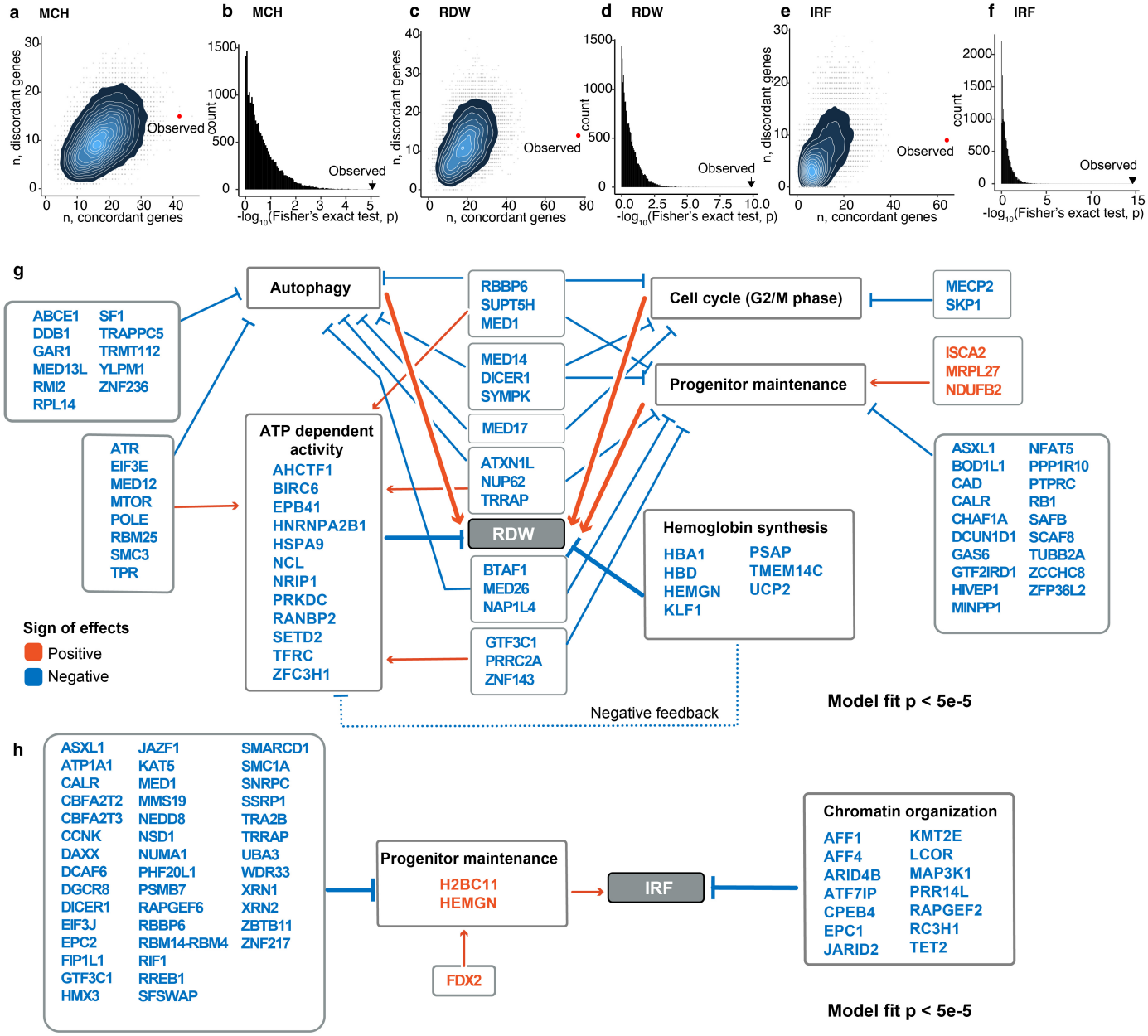

**Extended Data Fig. 5 | Association of programs and regulators with traits, related to Fig. 4. a)** Co-regulation pattern between S phase and autophagy programs. Each dot is a gene that has significant regulatory effects on S phase program. **b)** Correlation of regulatory effects on three programs with MCH $\gamma$ in the multiple regression model. Error bars indicate 95% CI. **c)** Co-regulation pattern between ATP dependent activity, hemoglobin synthesis and autophagy programs. Genes with regulatory effects on hemoglobin program activity also had effects on ATP activity, but the opposite was not true. **d)** Correlation of regulatory effects on three programs with RDW $\gamma$ in the multiple regression model. Error bars indicate the 95% CI. Bottom: model that combines the co-regulation pattern and trait association of the programs. **e)** The fraction of cells in different cell cycles in the groups of cells with perturbations (left) and the model for explaining the cell cycle program association with MCH (right). Error bar indicates standard error estimated from Jackknife resampling. **f)** Effects of cell cycle program genes KD on cellular growth. Growth screening data were obtained from an independent experiment using K562 (ref. 99). The effect size is a normalized measure of the impact of KD on cellular growth compared to wild type, denoted as gamma in the original manuscript. **g)** Effects of program genes KD on cellular growth[99]. Here, for each program, we created 100,000 sets of control genes matched for $S_{het}$ and compared the mean effects on cellular growth. **h)** Distribution of burden test effect sizes for MCH (left) and RDW (right). The plots show significant regulators of the autophagy program (FDR < 0.05, divided into positive and negative regulators) and the top 100 genes for the autophagy program by loading weights. P-values are from the regulator-burden correlation test.



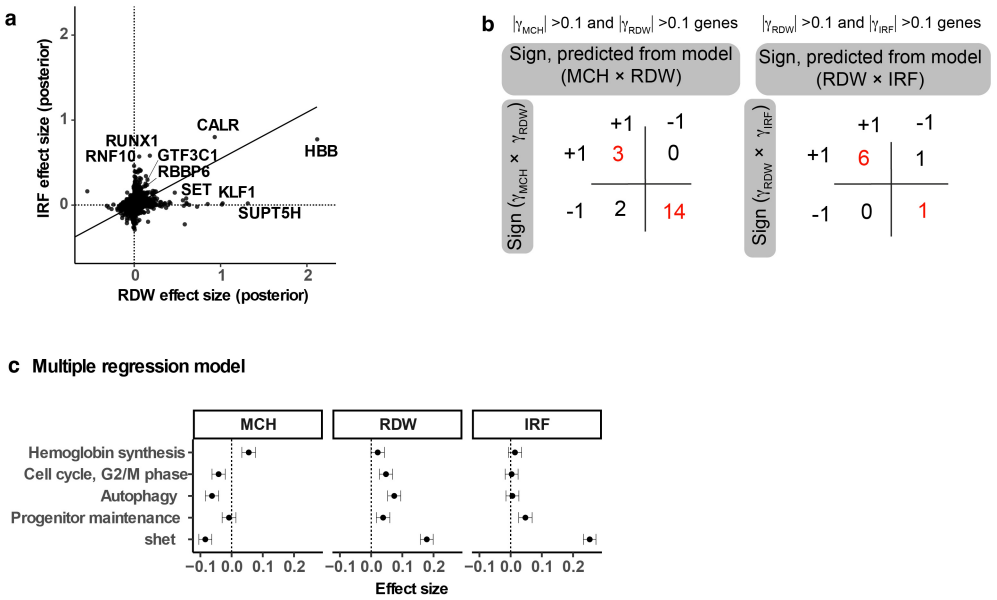

**Extended Data Fig. 6 | Multiple program association model. a)** We split the genes into a training set and a test set, and fitted multiple or single regression models to test the association between the gene regulatory effects on the program(s) and the gene effects on the trait ($\gamma$). We evaluated the variance explained by the model using the test gene set. **b-c)** Variance explained by the regression models for MCH (b) and RDW (c). "Programs, selected" refers to the programs selected from the regulator-burden correlations in the gene-to-program-to-trait map. "Programs, random" refers to the randomly selected sets of multiple programs. Single programs shown are the top 5 programs as to the variance explained. Error bars indicate 1.96 × standard errors. **d)** Schematics for making multiple program association model.



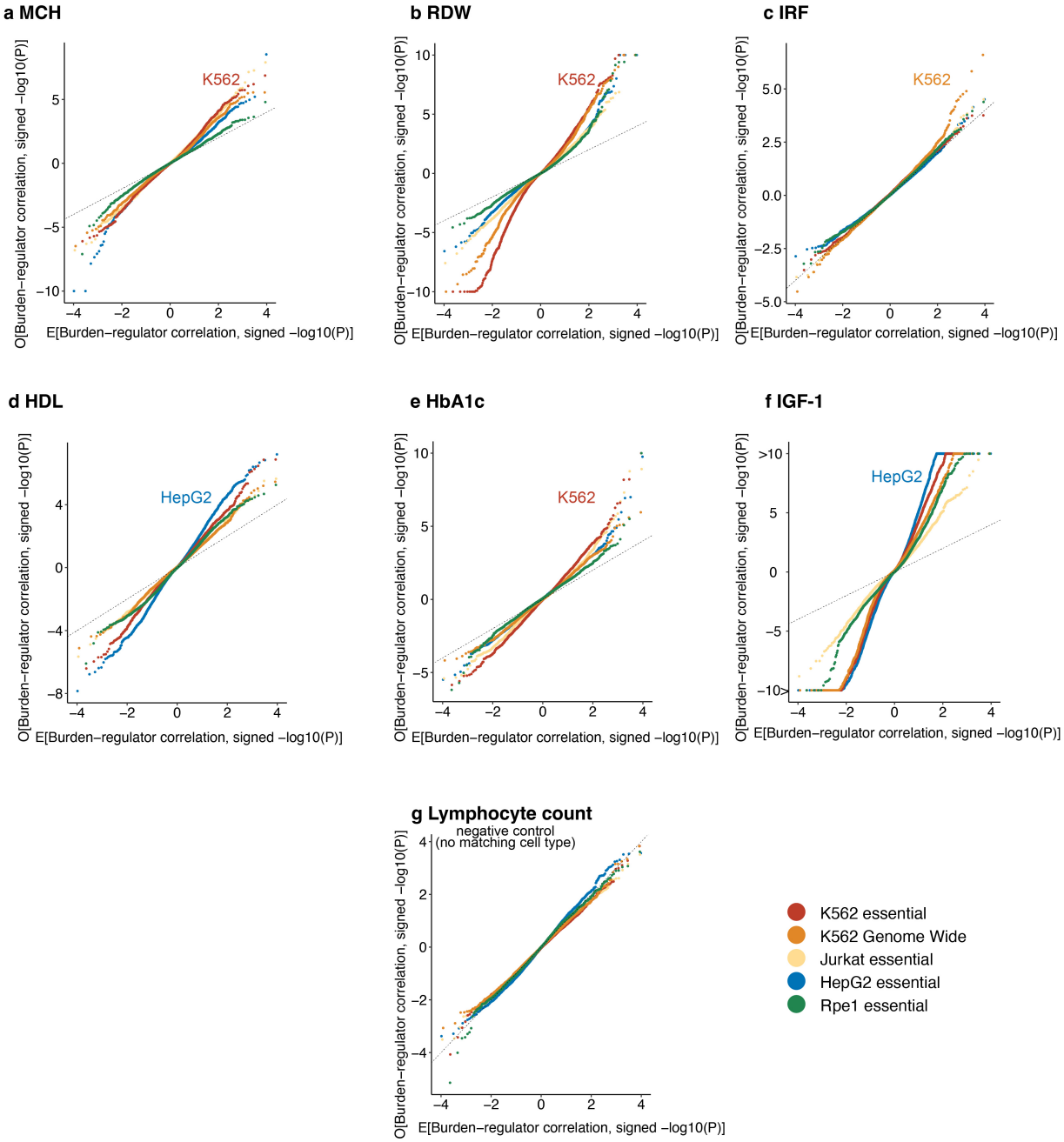

**Extended Data Fig. 7 | Programs selected for modeling MCH associations.** Programs were selected based on program burden effects (left) or regulator-burden correlations (right). For each program, top hits ($|\gamma| > 0.1$) for MCH that overlap with the top 200 loading genes (for program genes) or regulators (FDR < 0.05, for regulator genes) are listed. The color of genes correspond to the sign of $\gamma$. Genes in parentheses are discordant from the predicted directions from the overall model. Some of the genes are associated with multiple programs or regulators.




**Extended Data Fig. 8 | Gene to program to trait maps, related to Fig. 5.** **a)** Number of top hits ($|\gamma| > 0.1$) for MCH whose direction of associations were concordant or discordant with that predicted from the model. Grey points and their density plot are the results from 20,000 permutations. Red point shows the observed data. **b)** Distribution of top hits concordance p-values in permutation tests for MCH. In each permutation, we counted the number of top hits concordant with the model and evaluated its enrichment (Methods). The observed result showed the highest concordance compared to permuted sets. **c-f)** Same plots as (a) and (b), for RDW (c,d) and IRF (e, f). **g-h)** Gene to program to trait map for RDW (g) and IRF (h).




**Extended Data Fig. 9 | Cross-trait comparisons of gene effects, related to Fig. 5. a)** Comparison of LoF burden test effect sizes after GeneBayes between IRF and RDW. The solid line corresponds to the first principal component. **b)** Cross-trait directional relationships of gene effects in the predicted gene-to-program-to-trait model and raw data from the LoF burden test. The left table shows the comparison between MCH and RDW, while the right table shows the comparison between RDW and IRF. For each table, only genes that have strong effects in both traits ($|\gamma| > 0.1$) and selected in the predicted model for both traits are considered. For instance, +1 means that the gene has strong effects for both traits in the same direction. MCH and IRF share few genes with strong effects and could not be compared. **c)** Correlation of regulatory effects on four programs or $S_{het}$ with $\gamma$. For each trait, correlation coefficients were estimated with the multiple regression model. Error bars indicate 95% CI.




**Extended Data Fig. 10 | Regulator-burden correlation of genes in different cell lines.** Each dot represents one of the genome-wide expressed genes. For essential gene Perturb-seq, the correlation between regulatory effects and burden effects across essential genes is plotted. For the genome-wide Perturb-seq, the correlation across all perturbed genes is plotted.

nature portfolio

Corresponding author(s): Mineto Ota
Alexander Marson
Jonathan K. Pritchard

Last updated by author(s): Sep 18, 2025

# Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☐ | ☒ | A description of all covariates tested |
| ☐ | ☒ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☐ | ☒ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | No software was used for data collection |
| Data analysis | Scripts for analyzing and plotting the data are provided on Zenodo (https://doi.org/10.5281/zenodo.14751877). All statistical analyses were performed using R v. 4.2.0 with packages clusterProfiler v4.4.4, leaps v3.1, limma v3.52.4; and python v. 3.9.0 with packages shap-hypertune v0.2.6, xgboost v1.7.6, ngboost v0.3.13. We performed additional analyses using S-LDSC v. 1.0.1, PLINK v1.90b5.3,bedtools v2.30.0, cNMF v1.4, TrimGalore v0.5.0, Bowtie2 v2.3.4.1, MACS3 v3.0.3, and REGENIE v3.5. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data used for the analyses are deposited on Zenodo with https://doi.org/10.5281/zenodo.14751877. Public data used in this study are accessible via URLs cited at appropriate locations in the Methods.

## Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

| | |
|---|---|
| Reporting on sex and gender | N/A |
| Reporting on race, ethnicity, or other socially relevant groupings | N/A |
| Population characteristics | N/A |
| Recruitment | N/A |
| Ethics oversight | N/A |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | All analyses were performed on previously collected data from the UK Biobank or publicly available data. In all analyses we used all data that passed filtering criteria as described in the methods. |
| Data exclusions | We excluded traits with SNP heritability < 0.04 from further analysis because we cannot expect the reported genetic associations to explain much of the trait biology. |
| Replication | The regulator-burden correlations and program burden effects for MCH identified with GeneBayes posterior using the UK Biobank were replicated with AoU burden test statistics as well as UK Biobank burden test statistics without applying GeneBayes. The regulator-burden correlation for MCH was also validated using GWAS and trans-eQTL data. |
| Randomization | This study utilizes summary statistics from publicly available GWAS data and LoF burden test data. We were not involved in data collection for these GWAS and burden test analyses, making randomization irrelevant to this study. |
| Blinding | This study utilizes summary statistics from publicly available GWAS data and LoF burden test data. We were not involved in data collection for these GWAS and burden test analyses, making blinding irrelevant to this study. |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |
| ☒ | ☐ Plants |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

# Plants

Seed stocks

*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*

Novel plant genotypes

*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*

Authentication

*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*