bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Machine learning-based identification and cellular validation of *Tropomyosin 1* as a genetic inhibitor of hematopoiesis.

Thom CS[1,2,3]*, Jobaliya CD[4,5], Lorenz K[2,3], Maguire JA[4,5], Gagne A[4,5], Gadue P[4,5], French DL[4,5], Voight BF[2,3,6]*

[1]Division of Neonatology, Children's Hospital of Philadelphia, Philadelphia, PA, USA
[2]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[3]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[4]Center for Cellular and Molecular Therapeutics, The Children's Hospital of Philadelphia, PA, USA
[5]Department of Pathology and Laboratory Medicine, The Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, PA, USA
[6]Institute of Translational Medicine and Therapeutics, University of Pennsylvania, PA, USA
*Corresponding authors: thomc@email.chop.edu, bvoight@pennmedicine.upenn.edu

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## Introductory paragraph

A better understanding of the genetic mechanisms regulating hematopoiesis are necessary, and could augment translational efforts to generate red blood cells (RBCs) and/or platelets *in vitro*. Using available genome-wide association data sets, we applied a machine-learning framework to identify genomic features enriched at established platelet trait associations and score variants genome-wide to identify biologically plausible gene candidates. We found that high-scoring SNPs marked relevant loci and genes, including an expression quantitative trait locus for *Tropomyosin 1* (*TPM1*). CRISPR/Cas9-mediated *TPM1* knockout in human induced pluripotent stem cells (iPSCs) unexpectedly enhanced early hematopoietic progenitor development. Our findings may help explain human genetics associations and identify a novel genetic strategy to enhance *in vitro* hematopoiesis, increasing RBC and MK yield.

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## Main Text

Elucidating genetic mechanisms governing hematopoiesis has broad value in understanding blood production and hematologic diseases.[1] Given interest in generating red blood cells (RBCs) and platelets from *in vitro* culture of induced pluripotent stem cells,[2,3] there is also translational value in harnessing genetic and molecular processes that regulate hematopoiesis. For example, recent advances have increased platelet yield *in vitro*,[2] but generating MKs cost-effectively will require novel strategies based on better knowledge of underlying mechanisms.[2,4,5]

*In vitro* systems might be improved by identifying novel factors from human genetic studies. Genome wide association studies (GWAS) have linked hundreds of single nucleotide polymorphisms (SNPs) with platelet trait variability.[6–9] Because most GWAS SNPs are non-coding, likely influencing transcriptional expression of key genes,[10,11] it has been challenging to derive functional biochemical understanding of the key genes of action,[11–13] and few studies have elucidated biochemical mechanisms for platelet trait variability loci.[14–18] One strategy to narrow focus on candidate genes is to link non-coding variation to expression of nearby genes.[1,19,20] However, platelet trait variation GWAS have thus far implicated >6700 expression quantitative trait loci (eQTL) affecting expression of >1100 genes (**Methods**), highlighting a need to more specifically identify putatively functional sites.

To further narrow studies onto credible candidates for functional follow-up (**Fig. 1a**), we applied a penalized logistic regression model to select a subset of 628 different chromatin features that best distinguished 73 platelet trait GWAS SNPs[6] from matched control SNPs not associated with platelet traits (**Methods, Fig. 1b**, and **Supplementary Table 1**). The resultant predictive model selected 9 epigenomic features and was able to discriminate between positive and negative labeled examples (Area Under the Receiving Operator Curve (AUC) = 0.793, **Fig. 1c**, **Supplementary Fig. 1a**, and **Table 1**). Each selected feature had a positive coefficient, meaning each was more likely to overlap a platelet-associated GWAS SNP than a control.

While some care in interpretation was required, it was encouraging that the model selected biologically plausible features. GATA1 and FLI1 are critical MK transcription factors,[21,22] and most of our features came from hematopoietic cells (primary MK, K562, GM12878; **Table 1**). Furthermore, this set of 9 chromatin features are functionally predicted to identify regulatory elements near and within gene bodies (**Table 1**). This regulatory localization is consistent with previous observations.[14]

We calculated a trait-enrichment score based on SNP overlap with each of the 9 selected features, weighted by our penalized regression model coefficients (**Methods, Table 1**). Resultant scores were significantly higher for training, holdout and validation sets of platelet trait GWAS SNPs, relative to SNPs genome-wide ($p<0.0001$, **Fig. 2a, Supplementary Table 2**). Our regression model performed well compared to other methods[20,23,24] (**Fig. 2b-c, Supplementary Fig. 1b**).

We next assessed biological support for penalized regression scoring beyond the machine learning framework. First, we evaluated the biological specificity of variation prioritized by the model, given practical limitations associated with fine-mapping and cellular validation experiments. The number of high-scoring SNPs from our model fell within the range of other predictors (**Fig. 2d**), and Gene Ontology analysis indicated that

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the nearest genes to penalized regression-prioritized variants were enriched for biologically relevant pathways (**Methods, Fig. 2e** and **Supplementary Tables 3-5**). Second, SNP scores correlated with summary association statistics for platelet trait-GWAS data[6] (**Supplementary Fig. 2a-b**), with variants that were nominally associated but not yet genome-wide significant (and not used during the training or testing phases) having a significantly higher average score compared to SNPs with no clear association ($p<0.0001$, **Methods**, **Supplementary Fig. 2c-d**). This correlation suggested that our scoring algorithm was valid genome-wide and could potentially reveal true biological associations, as had the GWAS itself.[6,14,15,17] Third, FANTOM5 enhancer regions[25] were enriched for high-scoring SNPs, with an average score >0.9 compared with an average score 0.21 genome-wide (**Methods, Supplementary Fig. 3a**), consistent with the hypothesis that functional non-coding SNPs associate with active regulatory regions.[11,26] We further observed that enhancer regions in hematopoietic cell types scored significantly higher than enhancers from irrelevant control cells (**Supplementary Fig. 3a**). This argues for trait specificity in hematopoietic enhancers, consistent with prior studies.[27] Lastly, most high-scoring SNPs from the regression model were in gene bodies or near transcriptional start sites (TSSs, **Supplementary Fig. 3b**), with SNPs near key MK genes scoring significantly higher than SNPs in matched control regions[28] (**Supplementary Fig. 3c**). Collectively, this evidence indicated that our model successfully targeted hematopoietic trait-relevant loci, particularly those near and within gene bodies.

We reasoned that active variants would (i) be in high linkage disequilibrium (LD) with established platelet trait GWAS loci, (ii) score highly relative to other SNPs within that LD block, (iii) regulate target gene(s) as quantitative trait loci, and (iv) overlap GATA binding sites.[29,30] We prioritized GATA binding sites based on the importance of GATA factors in hematopoiesis[21,31] and in our scoring algorithm (**Methods, Fig. 1c**).

This approach led us directly to SNPs known to impact hematopoiesis, MK and/or platelet biology (**Table 2** and **Supplementary Table 6**). For example, rs342293 is a GWAS SNP[6] that regulates *PIK3CG* gene expression[15] (**Fig. 3a-d**). In platelets, *PIK3CG* activity regulates PIK3 signaling[32] and response to collagen.[33] The GATA site is disrupted in the presence of the SNP minor allele (**Fig. 3d**). Individuals harboring this minor allele had increased mean platelet volume (MPV) and decreased platelet reactivity.[15]

This approach also highlighted rs11071720, found within the 3rd intron of the *Tropomyosin 1* (*TPM1*) gene locus. This SNP was in reasonably strong LD with the sentinel GWAS SNP[6] rs3809566 (EUR $r^2$=0.73, **Fig. 3e-h**). The rs11071720 minor allele, which disrupts a near-canonical GATA binding site, is an eQTL associated with decreased *TPM1* expression[34,35] (**Methods**, **Fig. 3h,** and **Supplementary Fig. 4**), higher platelet count, and lower mean platelet volume (MPV).[8] Further, the minor allele for high-scoring rs4075583 (EUR $r^2$=0.71 with rs3809566) was associated with decreased *TPM1* expression in heterologous cells,[36] though not in GTEx tissues.[35] To our knowledge, neither of these SNPs, nor *TPM1*, had been functionally evaluated in the context of human hematopoiesis.

Given that these high-scoring putatively active SNPs impacted *TPM1* expression, we investigated functions for the *TPM1* gene in an *in vitro* human model of primitive hematopoiesis.[37] We anticipated that total gene deletion would show stronger effects than non-coding SNP modification.[38] Using CRISPR/Cas9, we targeted a ~5kb region

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

containing *TPM1* exons 4-8 in iPSCs (**Fig. 4a**), anticipating creation of a null allele.[39] We confirmed deletion by sequencing and western blot (**Fig. 4b** and **Supplementary Fig. 5**). In total, we obtained 3 *TPM1* knockout (KO) clones from 2 separate genetic backgrounds. Karyotype and copy number variation analyses confirmed that engineering these clones did not introduce any *de novo* genomic aberrancies (data not shown).

TPM1 protein was most abundant in iPSCs and downregulated during hematopoietic differentiation (**Fig. 4b**). Early differentiation proceeded normally in KO clones, with normal patterns of primitive streak and mesoderm gene expression (**Fig. 4c**) and pluripotency marker loss (**Supplementary Fig. 6**). The kinetics by which hemogenic endothelium ($KDR^+/CD31^+$) and hematopoietic progenitor cells (HPCs, $CD34^+/CD43^+$) emerged were also normal (**Fig. 4d-e**). In this culture system, hemogenic endothelium yields HPCs.

Remarkably, KO clones showed enhanced formation of hemogenic endothelium (**Fig. 4d**) and KO HPC yield roughly doubled that of WT controls (**Fig. 4e-f** and **Supplementary Fig. 7**) with normal cell surface expression of hematopoietic markers (**Supplementary Fig. 8**). KO HPCs generated normal quantities of mature MKs in liquid expansion culture (**Fig. 4g**). KO MK morphology and activation in response to agonists were normal (**Supplementary Fig. 9-10**). KO HPCs also generated increased numbers of erythroid cells (**Fig. 4h**) and normal quantities of myeloid cells (**Supplementary Fig. 11**).

Microarray gene expression analyses of WT and KO MKs revealed no statistically significant changes in MK genes, though Gene Set Enrichment Analysis (GSEA) showed a trend toward higher MK-related pathway expression in KO MKs (**Supplementary Fig. 12a-e**). Overall, 19 molecular pathways were upregulated in KO MKs (**Supplementary Fig. 12f** and **Supplementary Table 7**).

These data support a model whereby *TPM1* deficiency enhances *in vitro* hematopoiesis and resulting RBC and MK yield, perhaps helping to explain human genetic association data linking SNPs that lower *TPM1* expression[34,36] with increased platelet count (**Fig. 3h and Fig. 4i**).[8] Consistent with an impact on HPCs, *TPM1*-related SNPs have marginal effects on red cell traits in addition to genome-wide significant effects on platelet traits (**Supplementary Fig. 13**). These findings not exclude additional effects on terminal MK or RBC development *in vitro*, nor *in vivo* effects outside the scope of our model.

TPM1 deficiency could leave filamentous actin more 'accessible' to other modulator,[40,41] such as other TPMs. Of these, TPM4 promotes MK development[18] and likely modulates actin dynamics similar to *TPM1*.[42] TPM4 isoforms were upregulated during MK differentiation and significantly increased in *TPM1* KO iPSCs (**Fig. 4b** and **Supplementary Fig. 14**). Increased TPM4 may partially account for our observed enhanced HPC and MK yield.

Enhanced hematopoiesis in *TPM1KO* iPSCs contrasts detrimental effects of *TPM1* deficiency on organism fitness in other contexts.[6,43,44] For example, abrogated *D. rerio* thrombopoiesis with *tpma*-directed morpholinos[6] resembles human *TPM4* deficiency[18] rather than *TPM1* deficiency. This highlights the importance of species-specific genetic validation, particularly given inter-species disparities in hematopoiesis.[45]

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In conclusion, we used penalized regression modeling and cellular validation to define a role for *TPM1* in constraining *in vitro* hematopoiesis. In addition to understanding a genetic modifier of hematopoietic traits,[6,8] application of our results may augment RBC and MK yield *in vitro*. Recent advances increasing per-MK platelet yields[2] have focused a spotlight on increasing cost effectiveness of *in vitro* MK generation. In addition to improved recognition of genes and mechanisms underlying quantitative hematopoietic trait variation, application of the computational approach described herein could also help to specify trait-specific causal genetic variants for virtually any clinically relevant human trait.

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## Methods

***In silico* analysis.** Relevant data sets and coding scripts can be found on GitHub (https://github.com/thomchr/2019.PLT.TPM1.Paper). Human genome version hg19 was used for all analyses, and we utilized the LiftOver script when necessary (https://bioconductor.org/packages/release/workflows/html/liftOver.html). GWAS summary statistics for were obtained courtesy of Nicole Soranzo (for [6]) and are publicly available (http://www.bloodcellgenetics.org/).

**Expression Quantitative Trait Locus analysis.** To estimate the number of eQTLs implicated by prior platelet trait GWAS, SNPs in high LD with established GWAS loci[8] (EUR $r^2$>0.9) were identified using PLINK. From this set of SNPs, eQTLs and affected genes were identified from GTEx V7.[35] Numbers reported in the text reflect unique eQTL SNPs, which often functioned across multiple tissues. The affected gene estimate reflects the number of unique Ensembl gene identifiers (ENSG).

**SNP selection.** Platelet trait GWAS SNPs were identified from Gieger *et al* (see Table 1 in [6]). When two SNPs had been identified in a given region, the SNP with the greater effect size was chosen. The resultant 73 SNPs comprised our training SNP set (**Supplementary Table 2**). The remaining 8 SNPs were designated as a holdout set. From a total of 710 platelet trait (PLT, platelet count; MPV, mean platelet volume; PDW, platelet distribution width; PCT, platelet-crit)-associated GWAS SNPs from a more recent study,[8] 614 had rsIDs that matched our scored genome; these comprised our validation set. All of these SNP sets can be found on GitHub. We used the Genomic Regulatory Elements and GWAS Overlap algoRithm (GREGOR)[46] tool to select control SNPs for our study. GREGOR matched SNPs based on Distance to nearest gene, "LD buddies" (i.e., number of SNPs within a LD block) and Minor allele frequency, and identified controls for each of the 73 training set SNPs.

**Chromatin feature selection.** We collected a subset of available features tracks from ENCODE[47], including data for hematopoietic (K562 and GM12878) as well as other cell types (H1-hESC, HUVEC, HeLa, HepG2). We also collected available feature tracks from primary MKs.[21,48] See **Supplementary Table 1** for a list of these features.

**Penalized regression modeling.** To generate our model, we first analyzed training set GWAS SNPs (73) and matched controls SNPs (780,632) for overlap with 628 chromatin features (data set available on GitHub). Columns representing our 3 baseline parameters (Distance to Nearest Gene, LD Buddies and Minor Allele Frequency) were also included in this data table for each SNP. This chromatin feature overlap data file was then analyzed using the least absolute shrinkage and selection operator (LASSO, L1 regularization, glmnet version 2.0-2)[49,50] with 10-fold cross-validation and forced inclusion of the 3 baseline parameters. Baseline parameters were assigned penalty factors of 0, while other chromatin features were assigned penalty factors of 1. Features and coefficients were taken from the $\lambda_{se}$ (Df 12, %Dev 0.062980, $\lambda$ 6.203e-05). For downstream genome-wide analyses, we scored all SNPs within NCBI dbSNP Build 147.

**Model performance comparison.** We used ROCR[51] to compare prediction model performance. We used public databases to obtain SNP scores for alternative models (CADD v1.3, GWAVA (unmatched score), DeepSea;

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

https://cadd.gs.washington.edu/download, http://www.sanger.ac.uk/resources/software/gwava, http://deepsea.princeton.edu).

**Model specificity analysis.** To compare how restrictive and specific predictive models were for high-scoring SNPs, we first obtained scored NCBI dbSNP Build 147 SNPs for LASSO, GWAVA and CADD models. We quantified the number of SNPs that fell within the top 10% of scores. For GWAVA, this included any SNP scored >0.90 given a maximum score of 1.0. For CADD, this included any SNP scored >32.4 (maximum PHRED score 36). For LASSO, this included any SNP scored >2.32 (maximum score 2.58).

To assess biological specificity, we identified the top 1% highest-scoring SNPs from each model (LASSO, GWAVA, CADD) after excluding platelet trait-associated GWAS loci (81 SNPs from [6] and 710 SNPs from [8]). We then used closestBed (https://bedtools.readthedocs.io/en/latest/content/tools/closest.html) to identify the nearest gene to each of these SNPs. Genes and positioned were defined by BioMart (http://www.biomart.org/). We then used the Gene Ontology resource (http://geneontology.org/) to analyze pathway enrichment. Input analysis settings were Binomial tests and calculated FDR for GO Biological Process complete. Pathways identified with FDR<5% are presented in **Fig. 2e** and **Supplementary Tables 3-5**. Pathways shown in **Fig. 2e** are GO:0045652, GO:1902036, GO:1901532, GO:1903706, GO:0048534, GO:0030097, and GO:0030220.

**Score validation.** Gene Ontology pathways were used to identify key MK genes. A total of 132 “MK genes” were collected from pathways that were returned after a search for the term “megakaryocyte”, including “positive regulation of megakaryocyte differentiation”, “negative regulation of megakaryocyte differentiation”, “regulation of megakaryocyte differentiation”, “megakaryocyte differentiation”, “megakaryocyte development”, “platelet alpha granule”, “platelet formation”, “platelet morphogenesis” and “platelet maturation” (**Supplementary Table 8**). Gene locations for hg19 were obtained from the UCSC Genome Browser Table Browser feature.

The Genomic Regions Enrichment of Annotation Tool[28] (GREAT) was used in combination with the UCSC Genome Browser[52] (Table Browser interface) to analyze SNP locations and proximity to known genes.

Enhancer regulatory regions were defined according to the FANTOM5 data set.[25] Presented FANTOM5 data represent scores for all overlapping SNPs from dbSNP 147.

**Linkage disequilibrium structure assessment.** The SNP Annotation and Proxy Search tool (https://archive.broadinstitute.org/mpg/snap/ldsearch.php), LDlink (https://analysistools.nci.nih.gov/LDlink), and 1000 Genomes Project (phase 3) data were used to measure linkage disequilibrium in the EUR population.

**Transcription factor binding site identification.** Transcription factor binding sites were identified using the Find Individual Motif Sequences (FIMO) and Analysis of Motif Enrichment (AME) tools from MemeSuite (http://meme-suite.org). To identify GATA sites, the genomic sequence contexts for LD blocks containing each GWAS SNP were analyzed for matches ($p<0.001$) by manual curation of canonical or near-canonical GATA binding motif in all orientations (AGATAA, TTATCA, AATAGA, TTATCT).

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**Human iPSC generation.** iPSC models were generated as described from peripheral blood mononuclear cells.[53] The "CHOP10" and "CHOP14" lines were used in this study. CRISPR/Cas9-mediated genome editing was performed as described[54] per protocols from the CHOP Human Pluripotent Stem Cell Core Facility (https://ccmt.research.chop.edu/cores_hpsc.php) with the following guide sequences: 5' (1) ATGACGAAAGGTACCACGTCAGG, 5' (2) TGAGTACTGATGAAACTATCAGG, 3' (1) CCCTTTTCTTGCTGCTGTGTTGG, 3' (2) GGAGAGTGATCAAGAAATGGAGG.

Karyotyping (Cell Line Genetics, Madison, WI) and copy number variation (CHOP Center for Applied Genomics, Philadelphia, PA) analyses were performed per institutional protocols.

**iPSC hematopoietic differentiation and analysis.** iPS cells were differentiated in HPCs and terminal lineages (MKs, erythroid, myeloid) per published protocols.[37,55–57] Validated flow cytometry gating for pluripotency ($SSEA3^+/SSEA4^+$), hemogenic endothelium ($KDR^+/CD31^+$), hematopoietic progenitors ($CD43^+/CD34^+$ and $CD41^+/CD235^+$) and terminal lineages can be found in these references.

**Flow cytometry.** Flow cytometry analysis was performed on a Cytoflex LX and FACS-sorting was performed on a FACS Aria II (BD Biosciences). Flow cytometry data were analyzed using FlowJo 10 (Tree Star, Inc.). The following antibodies were used for flow cytometry: FITC-conjugated anti-CD41 (BioLegend), PE-conjugated anti-CD42b (BD Biosciences), APC-conjugated anti-CD235 (BD Biosciences), PB450-conjugated anti-CD45 (BioLegend), AF488-conjugated anti-SSEA3 (BioLegend, AF647-conjugated anti-SSEA4 (BioLegend), PE-conjugated anti-KDR (R&D Systems), PECy7-conjugated antiCD31 (BioLegend), PECy7-conjugated anti-CD34 (eBioscience) and FITC-conjugated anti-CD43 (BioLegend).

**Gene expression analysis by RT-semiquantitative PCR.** Total RNA was prepared using PureLink RNA micro kits (Invitrogen) in which samples were treated with RNase-free DNase. The reverse transcription of RNA (100 ng-1 µg) into cDNA was performed using random hexamers with Superscript II Reverse Transcriptase (RT) (Life Technologies), according to the manufacturer's instructions. Real-time quantitative polymerase chain reaction (PCR) was performed on QuantStudio 5 Real-Time PCR Instrument (Applied Biosystems). All experiments were done in triplicate with SYBR-GreenER pPCR SuperMix (Life Technologies), according to the manufacturer's instructions. Primers (**Supplementary Table 9**) were prepared by Integrated DNA Technologies or Sigma Aldrich. Dilutions of human genomic DNA standards ranging from 100 ng/µl to 10 pg/µl were used to evaluate PCR efficiency of each gene relative to the housekeeping gene *TATA-Box Binding Protein* (*TBP*).

**Microarray analysis.** For microarray analysis, 50,000 cells were FACS-sorted directly into Trizol. RNA was extracted from using a miRNeasy Mini Protocol (Qiagen). Samples passing quality control were analyzed using the human Clariom D Assay (ThermoFisher Scientific) and analyzed using Transcriptome Analysis Console (ThermoFisher Scientific) Software and Gene Set Enrichment Analysis (http://software.broadinstitute.org/gsea/index.jsp) software.

**Cell analysis and imaging.** For Cytospins, FACS-sorted MKs were spun onto a glass slide and stained with May-Grünwald and Giemsa. Images were obtained on an

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Olympus BX60 microscope with a 40X objective. An Invitrogen EVOS microscope with a 10x objective was used to image cells in culture.

**Western blots.** Cell pellets were resuspended in Laemmli buffer, sonicated for 5 min, and boiled for 5 min at 95 degrees C. Lysates were centrifuged at 10,000 rpm for 5 min at room temperature, and supernatants were used for analysis. Lysate volumes were normalized to cell counts. Samples were run on 4-12% NuPAGE Bis-Tris gels (Invitrogen) and transferred onto nitrocellulose membranes (0.45um pore size, Invitrogen) at 350mA for 90 minutes. Following blocking in 5% milk for 1 h, membranes were incubated with primary antibodies overnight at 4°C. After washing thrice in TBST, membranes were incubated with secondary horseradish peroxidase-conjugate antibodies for 1h at room temperature, washed in TBST thrice, and developed using ECL western blotting substrate (Pierce) and HyBlot CL autoradiography film (Denville Scientific). The following antibodies were used for western blotting: Rabbit anti-TPM1 (D12H4, #3910, Cell Signaling Technologies), Mouse anti-TPM1/TPM2 (15D12.2, MAB2254, Millipore Sigma), Mouse anti-TPM3 (3D5AH3AB4, ab113692, Abcam), (Rabbit anti-TPM4 (AB5449, Millipore Sigma), and Mouse anti-β Actin (A1978, Sigma). Western blot band quantitation was performed using FIJI (https://fiji.sc/).

**MK activation assay.** MKs were pelleted and resuspended in Tyrode's Salts (Sigma) with 0.1% bovine serum albumin (BSA) containing FITC-conjugated PAC-1 (BD Biosciences), PacBlue-conjugated CD42a (eBioscience) and APC-conjugated CD42b (eBioscience) at a concentration of roughly 100,000 cells per 50μl. Following addition of Convulxin (Enzo Biochem) or Thrombin (Sigma), cells were incubated at room temperature in the dark for 10 min. Cells were then incubated on ice for 10 min. An additional 100μl Tyrode's Salts containing 0.1% BSA were added and cells were immediately analyzed by flow cytometry.

**Data presentation.** Genome-wide SNP Scores were loaded as custom tracks into the UCSC Genome Browser.[52] Images depicting genomic loci were generated using this tool, as well as Gviz.[58] Other data were created and presented using R, Adobe Illustrator CS6 or GraphPad Prism 6.

**Statistics.** Statistical analyses were conducted using R or GraphPad Prism 6.

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**Acknowledgements**

We are grateful for thoughtful suggestions from Drs. Mortimer Poncz, Michele Lambert, and members of the Voight laboratory, as well as technical support from Tapan Ganguly and Hetty Rodriguez (University of Pennsylvania Microarray Core Facility), and the Penn Medicine Academic Computing Services. We thank Dr. Nicole Soranzo for sharing summary level GWAS data[6] and Osheiza Abdulmalik for generous use of his microscope for Cytospin imaging.

This work was supported through R01DK101478 (BFV), a Linda Pechenik Montague Investigator Award (BFV), R01HL130698 (DLF, PG), T32HD043021 (CST), a Children's Hospital of Philadelphia Neonatal and Perinatal Medicine Fellow's Research Award (CST), an American Academy of Pediatrics Marshall Klaus Neonatal-Perinatal Research Award (CST) and a Children's Hospital of Philadelphia Foerderer Award (CST).

**Author Contributions**

CST and BFV conceived of this study. CST, CDJ, KL, JAM, DLF, and BFV conducted and/or analyzed experiments. CST and BFV wrote the manuscript. BFV oversaw the work.

**Competing Interests**

The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## Figure Legends

**Fig. 1. A machine learning approach to identify genes and loci that regulate megakaryocyte and platelet development.**
**a**, Schematic outline of our approach. We identified a machine learning-based predictive scoring algorithm based on platelet trait GWAS loci and applied the resultant scoring algorithm genome-wide to predict causal variants and genes. We validated this model computationally and by establishing biochemical activity of novel loci. **b,** Platelet trait GWAS SNPs[6] and matched control SNPs[46] were analyzed for overlap with 628 chromatin features (e.g., histone marks and TF binding sites) from 7 different cell types (K562, GM12878, HeLa, HUVEC, H1-hESC, HepG2, primary MKs).[21,59] LASSO[49] analysis identified chromatin features that best discriminated GWAS SNPs. **c**, The best LASSO model identified 9 chromatin features from the indicated cell types, in addition to background characteristics. Bar heights are LASSO coefficients, indicating the relative importance of each feature.

**Figure 2. A LASSO-based model most specifically identifies SNPs relevant to megakaryocyte and platelet biology.**
**a**, SNP scores for the training holdout,[6] and validation[8] sets of platelet trait-associated GWAS SNPs were significantly higher than genome-wide SNP scores. Bars represent mean+-SEM, ****p<0.0001 by ANOVA. **b,c**, Performance comparison of our LASSO-based model to DeepSEA,[23] GWAVA,[20] and CADD[24] for (**b**) training or (**c**) holdout SNP sets. AUC values are shown in the legend with each plot. **d,** Bar heights represent the percent of all SNPs scored in the top numerical decile for the indicated models. The number of SNPs that fell within this top score decile is indicated. **e,** Hematopoietic pathways[60] identified by the highest-scoring 1% of SNPs identified by LASSO, GWAVA, or CADD, excluding established platelet trait loci[6,8] (FDR, False Discovery Rate).

**Figure 3. Fine-mapping GWAS-implicated regions using LASSO-based SNP scores and putative GATA binding sites identifies active loci and eQTLs**.
**a,b,c**, Part (**a**) shows SNP scores genes near platelet trait GWAS SNP rs342293.[6] This GWAS SNP is shown in black and linked SNPs (EUR $r^2$>0.7) in cyan. Bar heights depict SNP scores. The boxed region is expanded in (**b**) to show genomic context, including key chromatin features (Fig. 1c). The delineated area (red dashed line) is expanded in (**c**) to show the local DNA sequence, including rs342293 (underlined) and canonical GATA binding site[29] (red). **d**, The major and minor rs342293 alleles and associated phenotypes.[15] **e,f,g**, SNP scores and genes (**e**) near platelet trait GWAS SNP rs3809566.[6] This GWAS SNP is shown in black and linked SNPs (EUR $r^2$>0.7) in cyan. Bar heights depict SNP scores. The boxed region is expanded in (**f**) to show local genomic context, including *TPM1* exons and key chromatin features (Fig. 1c). The delineated area (red hashed line) is expanded in (**g**) to show the local DNA sequence, including rs11071720 (underlined) and a putative GATA binding site[29] (red). **h**, The major and minor rs11071720 alleles and associated phenotypes.[8,34] Allele percentages based on UCSC genome browser and dbSNP.

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**Fig 4. *TPM1* deficiency enhances iPSC-derived HPC and MK generation.**
**a**, A 5 kb region (*TPM1* exons 4-8, red box) was targeted for CRISPR/Cas9-mediated deletion to create KO iPSCs. **b**, Western blots showing TPM1-4 expression in wild type (WT) and KO iPSCs, hematopoietic progenitor cells (HPC, differentiation d8) and FACS-sorted MKs ($CD41^+/CD42b^+$, expansion d3). TPM1 antibodies targeted exon 4 (top) or exon 9d ($2^{nd}$ panel). **c,** Expression of primitive streak and mesoderm genes are normally expressed in differentiating KO iPSCs. **d,e**, KO cells yield (**d**) hemogenic endothelium ($KDR^+/CD31^+$) and (**e**) hematopoietic progenitor cells ($CD43^+/CD34^+$) with normal kinetics, but in enhanced abundance. Percent (%) cells within boxed regions are shown from a representative experiment. **f**, Quantification of WT and KO non-adherent HPCs on differentiation day 8. Bars represent fold change in HPCs (mean±SD) vs WT for ≥4 experiments. (Top) Culture images on differentiation d8, with HPCs (light color) floating above an adherent monolayer. Scale bar, 20mm. **g**, WT and KO HPCs put into MK expansion culture generate equivalent numbers of MKs. Points represent $CD41^+/CD42b^+$ MKs percentage multiplied by total cell count, normalized to cell count on day 0. **h**, KO HPCs put into erythroid (Ery) expansion culture generate more erythroid cells than WT HPCs. Points represent $CD235^+$ percentage multiplied by total cell count, normalized to cell count on day 0. **i**, Model in which KO iPSCs yield more HPCs than WT, generating more total MKs and RBCs. ns, not significant, *p<0.05, **p<0.01 by ANOVA.

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## Table Legends

**Table 1. Nine chromatin features discriminate platelet trait GWAS SNPs from controls**. In addition to background characteristics, penalized regression modeling identified 9 independently enriched chromatin features. The cell type for each feature data set and function for each feature[61–66] are displayed. Coefficients for each feature, indicating relative importance, are shown at right. Ery, erythroid. FLI1, Friend Leukemia Virus Integration 1. RbBP5, Retinoblastoma Binding Protein 5.

**Table 2. LASSO-based fine-mapping identifies eQTLs in established platelet trait GWAS loci that overlie GATA binding sites.** Listed are SNPs within GWAS LD blocks (EUR $r^2$>0.7) scoring in the top 5% (LASSO) that overlap canonical or near-canonical GATA binding sites and are eQTLs for at least 1 gene (GTEx V7).[35] Associated GWAVA scores and percentiles are shown. Genes in **bold** have evidence supporting function in human hematopoiesis, megakaryocytes, and/or platelets, except where indicated.

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## References


1. Ulirsch, J. C. *et al.* Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* **51**, 683–693 (2019).
2. Ito, Y. *et al.* Turbulence Activates Platelet Biogenesis to Enable Clinical Scale Ex Vivo Production. *Cell* **174**, 636-648.e18 (2018).
3. Giani, F. C. *et al.* Targeted Application of Human Genetic Variation Can Improve Red Blood Cell Production from Stem Cells. *Cell Stem Cell* **18**, 73–78 (2016).
4. Noh, J.-Y. *et al.* Inducible Gata1 suppression expands megakaryocyte- erythroid progenitors from embryonic stem cells. *J. Clin. Invest.* **125**, 2369–2374 (2015).
5. Sim, X., Poncz, M., Gadue, P. & French, D. L. Understanding platelet generation from megakaryocytes: Implications for in vitro-derived platelets. *Blood* **127**, 1227–1233 (2016).
6. Gieger, C. *et al.* New gene functions in megakaryopoiesis and platelet formation. *Nature* **480**, 201–208 (2011).
7. Soranzo, N. *et al.* A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat. Genet.* **41**, 1182–1190 (2009).
8. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415-1429.e19 (2016).
9. Meisinger, C. *et al.* A Genome-wide Association Study Identifies Three Loci Associated with Mean Platelet Volume. *Am. J. Hum. Genet.* **84**, 66–71 (2009).
10. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362–9367 (2009).
11. Tak, Y. G. & Farnham, P. J. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics Chromatin* **8**, 57 (2015).
12. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.* **93**, 779–97 (2013).
13. Xu, Z. & Taylor, J. a. SNPinfo: Integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Res.* **37**, 600–605 (2009).
14. Simon, L. M. *et al.* Integrative Multi-omic Analysis of Human Platelet eQTLs Reveals Alternative Start Site in Mitofusin 2. *Am. J. Hum. Genet.* **98**, 883–897 (2016).
15. Soranzo, N. *et al.* A novel variant on chromosome 7q22.3 associated with mean platelet volume, counts, and function. *Blood* **113**, 3831–3837 (2009).
16. Polfus, L. M. *et al.* Whole-Exome Sequencing Identifies Loci Associated with Blood Cell Traits and Reveals a Role for Alternative GFI1B Splice Variants in Human Hematopoiesis. *Am. J. Hum. Genet.* **99**, 481–488 (2016).
17. Nurnberg, S. T. *et al.* A GWAS sequence variant for platelet volume marks an alternative DNM3 promoter in megakaryocytes near a MEIS1 binding site. *Blood* **120**, 4859–4868 (2012).
18. Pleines, I. *et al.* Mutations in tropomyosin 4 underlie a rare form of human macrothrombocytopenia. *J. Clin. Invest.* **127**, 814–829 (2017).
19. Shihab, H. A. *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31**, 1536–43 (2015).
20. Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat. Methods* **11**, 294–296 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

21. Tijssen, M. R. R. *et al.* Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. *Dev. Cell* **20**, 597–609 (2011).
22. Pimkin, M. *et al.* Divergent functions of hematopoietic transcription factors in lineage priming and differentiation during erythro-megakaryopoiesis. *Genome Res.* **24**, 1932–1944 (2014).
23. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
24. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
25. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–61 (2014).
26. Farh, K. K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2014).
27. Petersen, R. *et al.* Platelet function is modified by common sequence variation in megakaryocyte super enhancers. *Nat. Commun.* **8**, 16058 (2017).
28. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
29. Mathelier, A. *et al.* JASPAR 2016: A major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **44**, D110–D115 (2016).
30. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–8 (2011).
31. Freson, K. *et al.* Platelet characteristics in patients with X-linked macrothrombocytopenia because of a novel GATA1 mutation Platelet characteristics in patients with X-linked macrothrombocytopenia because of a novel GATA1 mutation. **98**, 85–92 (2011).
32. Hawkins, P. T., Stephens, L. R., Lim, W., Mayer, B. & Pawson, T. PI3Kg Is a Key Regulator of Inflammatory Responses and Cardiovascular Homeostasis. *Science (80-. ).* **318**, 52–58, 324–328 (2014).
33. Pasquet, J. M. *et al.* Thrombopoietin potentiates collagen receptor signaling in platelets through a phosphatidylinositol 3-kinase-dependent pathway. *Blood* **95**, 3429–34 (2000).
34. Fehrmann, R. S. N. *et al.* Trans-eqtls reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet.* **7**, (2011).
35. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-. ).* **348**, 648–660 (2015).
36. Savill, S. A., Leitch, H. F., Daly, A. K., Harvey, J. N. & Thomas, T. H. Polymorphisms in the tropomyosin TPM1 short isoform promoter alter gene expression and are associated with increased risk of metabolic syndrome. *Am. J. Hypertens.* **23**, 399–404 (2010).
37. Sim, X. *et al.* Identifying and enriching platelet-producing human Stem cell-derived megakaryocytes using factor V uptake. *Blood* **130**, 192–204 (2017).
38. Sankaran, V. G. & Orkin, S. H. Genome-wide association studies of hematologic phenotypes: a window into human hematopoiesis. *Curr Opin Genet Dev* **23**, 339–344 (2013).
39. Schevzov, G., Whittaker, S. P., Fath, T., Lin, J. J.-C. C. & Gunning, P. W. Tropomyosin isoforms and reagents. *Bioarchitecture* **1**, 135–164 (2011).
40. Gunning, P. W., Hardeman, E. C., Lappalainen, P. & Mulvihill, D. P. Tropomyosin - master regulator of actin filament function in the cytoskeleton. *J. Cell Sci.* **128**,

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

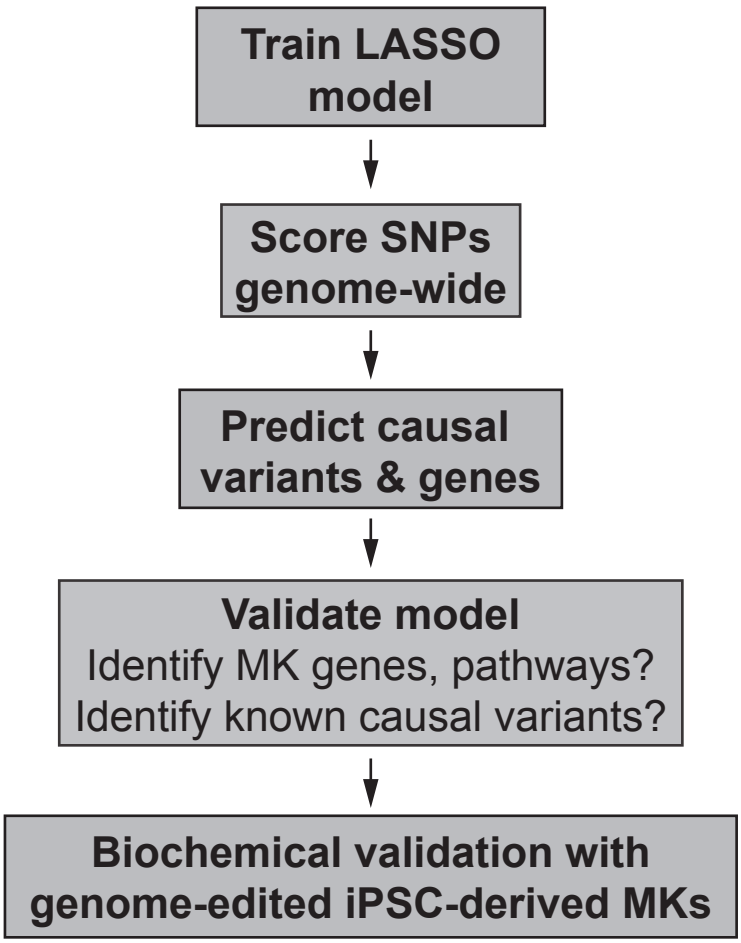
2965–2974 (2015).
41. Gateva, G. *et al.* Tropomyosin Isoforms Specify Functionally Distinct Actin Filament Populations In Vitro. *Curr. Biol.* **27**, 705–713 (2017).
42. Gunning, P. W. & Hardeman, E. C. Tropomyosins. *Current Biology* **27**, R8–R13 (2017).
43. Rethinasamy, P. *et al.* Molecular and Physiological Effects of α-Tropomyosin Ablation in the Mouse. *Circ. Res.* **82**, 116–123 (1998).
44. Anyanful, A., Sakube, Y., Takuwa, K. & Kagawa, H. The third and fourth tropomyosin isoforms of Caenorhabditis elegans are expressed in the pharynx and intestines and are essential for development and morphology. *J. Mol. Biol.* **313**, 525–537 (2001).
45. Pishesha, N. *et al.* Transcriptional divergence and conservation of human and mouse erythropoiesis. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 4103–8 (2014).
46. Schmidt, E. M. *et al.* GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics* **31**, 2601–6 (2015).
47. Feingold, E. *et al.* The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (80-. ).* **306**, 636–40 (2004).
48. Paul, D. S. *et al.* Maps of open chromatin highlight cell type-restricted patterns of regulatory sequence variation at hematological trait loci. *Genome Res.* **23**, 1130–1141 (2013).
49. Tibshirani, R. Regression Selection and Shrinkage via the Lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
50. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320 (2005).
51. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940–1 (2005).
52. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
53. Maguire, J. A. *et al.* Generation of human control iPS cell line CHOPWT10 from healthy adult peripheral blood mononuclear cells. *Stem Cell Res.* **16**, 338–341 (2016).
54. Maguire, J. A., Cardenas-Diaz, F. L., Gadue, P. & French, D. L. Highly Efficient CRISPR-Cas9-Mediated Genome Editing in Human Pluripotent Stem Cells. *Curr. Protoc. Stem Cell Biol.* **48**, e64 (2019).
55. Paluru, P. *et al.* The negative impact of Wnt signaling on megakaryocyte and primitive erythroid progenitors derived from human embryonic stem cells. *Stem Cell Res.* **12**, 441–451 (2013).
56. Mills, J. A., Paluru, P., Weiss, M. J., Gadue, P. & French, D. L. Hematopoietic Differentiation of Pluripotent Stem Cells in Culture. in *Hematopoietic Stem Cell Protocols* (eds. Til, N. P. Van & Wagemaker, G.) **1185**, 311–319 (Humana Press, 2014).
57. Mills, J. A. *et al.* Clonal genetic and hematopoietic heterogeneity among human-induced pluripotent stem cell lines. *Blood* **122**, 2047–51 (2013).
58. Hahne, F. & Ivanek, R. Visualizing Genomic Data Using Gviz and Bioconductor. in *Statistical Genomics: Methods and Protocols.* (eds. Mathé, E. & Davis, S.) 335–351 (Springer, 2016). doi:10.1007/978-1-4939-3578-9_16
59. Consortium, E. P. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
60. Ashburner, M. *et al.* Gene Ontology: Tool for The Unification of Biology. *Nat. Genet.* **25**, 25–29 (2000).

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

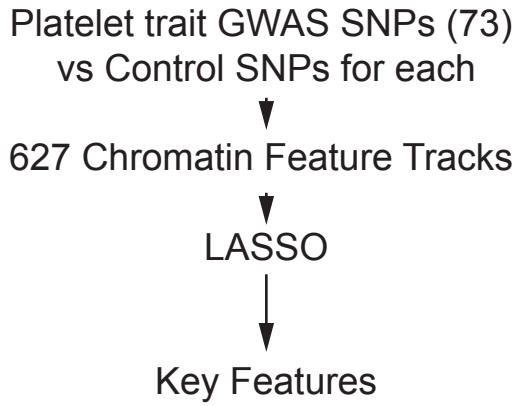
61. Kimura, H. Histone modifications for human epigenome analysis. *J. Hum. Genet.* **58**, 439–445 (2013).
62. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
63. Ernst, P. & Vakoc, C. R. WRAD: Enabler of the SET1-family of H3K4 methyltransferases. *Brief. Funct. Genomics* **11**, 217–226 (2012).
64. Calo, E. & Wysocka, J. Modification of Enhancer Chromatin: What, How, and Why? *Mol. Cell* **49**, 825–837 (2013).
65. Wang, Y., Li, X. & Hu, H. H3K4me2 reliably defines transcription factor binding regions in different cells. *Genomics* **103**, 222–228 (2014).
66. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21931–6 (2010).
67. Cunha, M. L. R. *et al.* Whole exome sequencing in thrombophilic pedigrees to identify genetic risk factors for venous thromboembolism. *PLoS One* **12**, e0187699 (2017).
68. Ouseph, M. M. *et al.* Autophagy is induced upon platelet activation and is essential for hemostasis and thrombosis. *Blood* **126**, 1224–33 (2015).
69. Kumari, S., Chaurasia, S. N., Nayak, M. K., Mallick, R. L. & Dash, D. Sirtuin Inhibition Induces Apoptosis-like Changes in Platelets and Thrombocytopenia. *J. Biol. Chem.* **290**, 12290–9 (2015).
70. Lv, Y. *et al.* Aberrant expression of NLRP3, NLRC4 and NLRP6 inflammasomes in patients with primary immune thrombocytopenia. *Thromb. Res.* **176**, 101–103 (2019).
71. Buitrago, L., Tsygankov, A., Sanjay, A. & Kunapuli, S. P. Cbl proteins in platelet activation. *Platelets* **24**, 419–27 (2013).
72. Karniguian, A., Zahraoui, A. & Tavitian, A. Identification of small GTP-binding rab proteins in human platelets: thrombin-induced phosphorylation of rab3B, rab6, and rab8 proteins. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 7647–51 (1993).
73. De Luca, G., Barberi, I., Ruggeri, P. & Di Giorgio, R. M. Cathechol-O-methyl transferase activity in the individual human platelet populations. *Ital. J. Biochem.* **25**, 213–8

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a




b




c

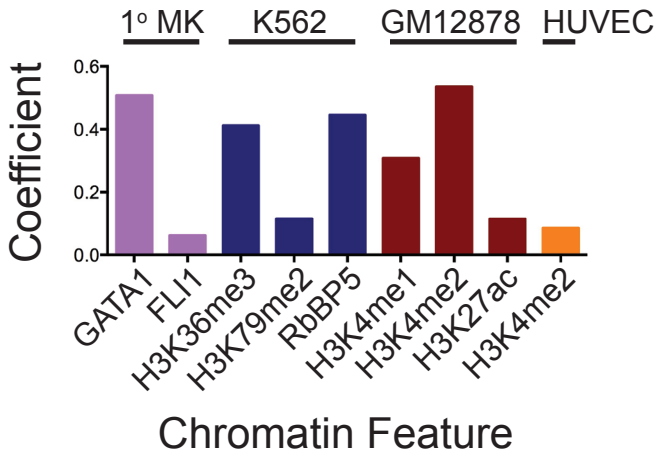
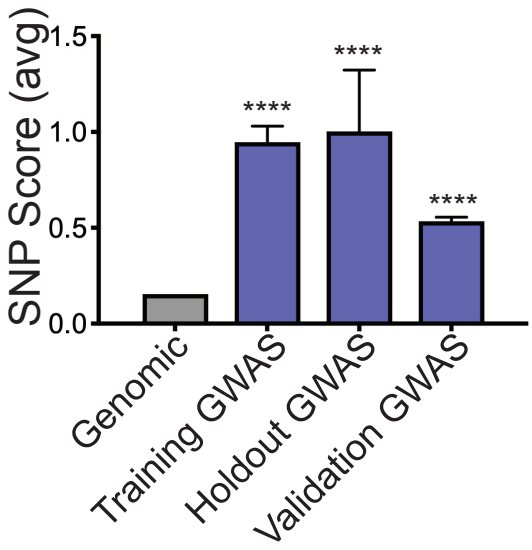


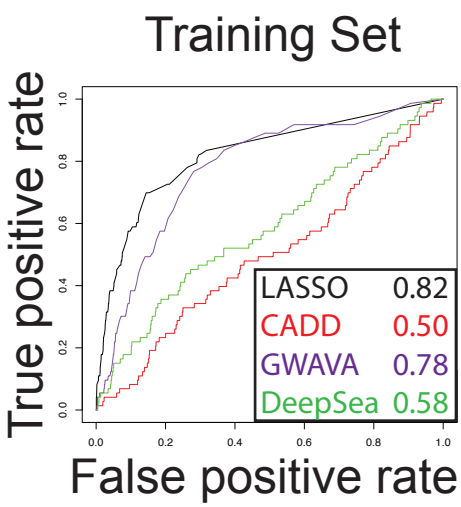

Figure 1. Thom et al

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

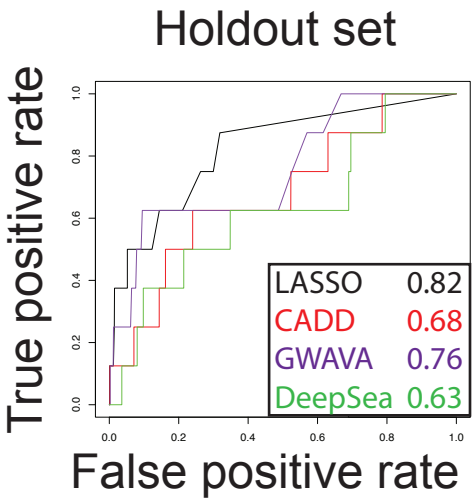
a



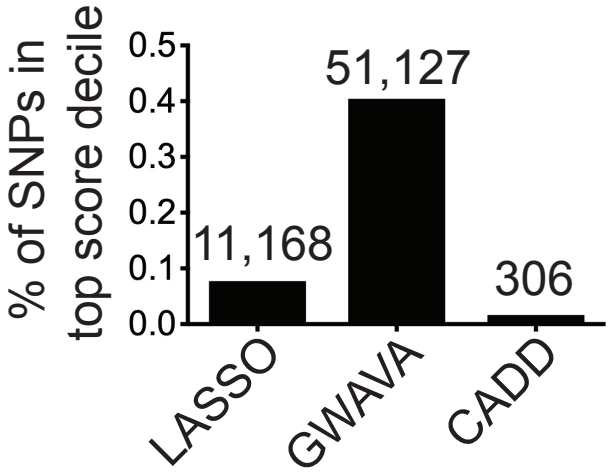

b




c




d



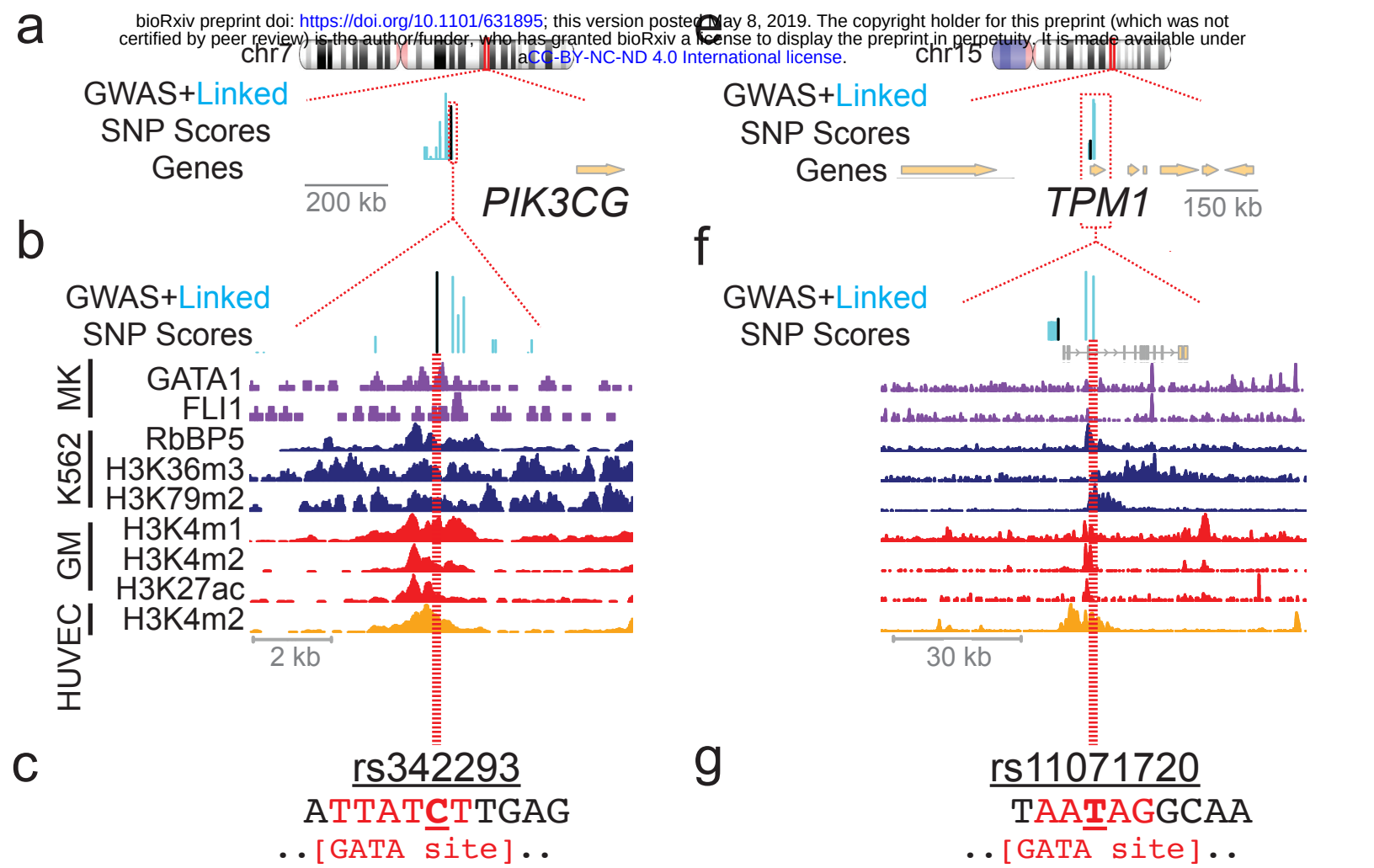

e

## Hematopoietic pathways identified by highest scoring SNPs

| GO Biological Process | Fold Enrichment | FDR |
|---|---|---|
| **LASSO** | | |
| Regulation of MK differentiation | 3.98 | 2.60E-5 |
| Regulation of hemopoiesis | 1.86 | 2.17E-5 |
| Hematopoietic or lymphoid organ development | 1.65 | 1.78E-4 |
| Hemopoiesis | 1.64 | 4.65E-4 |
| Hemostasis | 1.86 | 1.23E-3 |
| Platelet activation | 2.29 | 1.71E-3 |
| | | |
| **GWAVA** (none) | | |
| | | |
| **CADD** (none) | | |

Figure 2. Thom et al

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.




c

rs342293

ATTATCTTGAG

..[GATA site]..

g

rs11071720

TAATAGGCAA

..[GATA site]..

d

Major allele (64%)

..TTATCT..
..AATAGA..

→ Normal *PIK3CG*

Minor allele (36%)

..TTATGT..
..AATACA..

→ ↑ *PIK3CG* level, ↑ MPV, ↓ Platelet function

h

Major allele (51%)

..AATAG..
..TTATC..

→ Normal *TPM1*

Minor allele (49%)

..AACAG..
..TTGTC..

→ ↓ *TPM1* level, ↓ MPV, ↑ PLT#

Figure 3. Thom et al

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

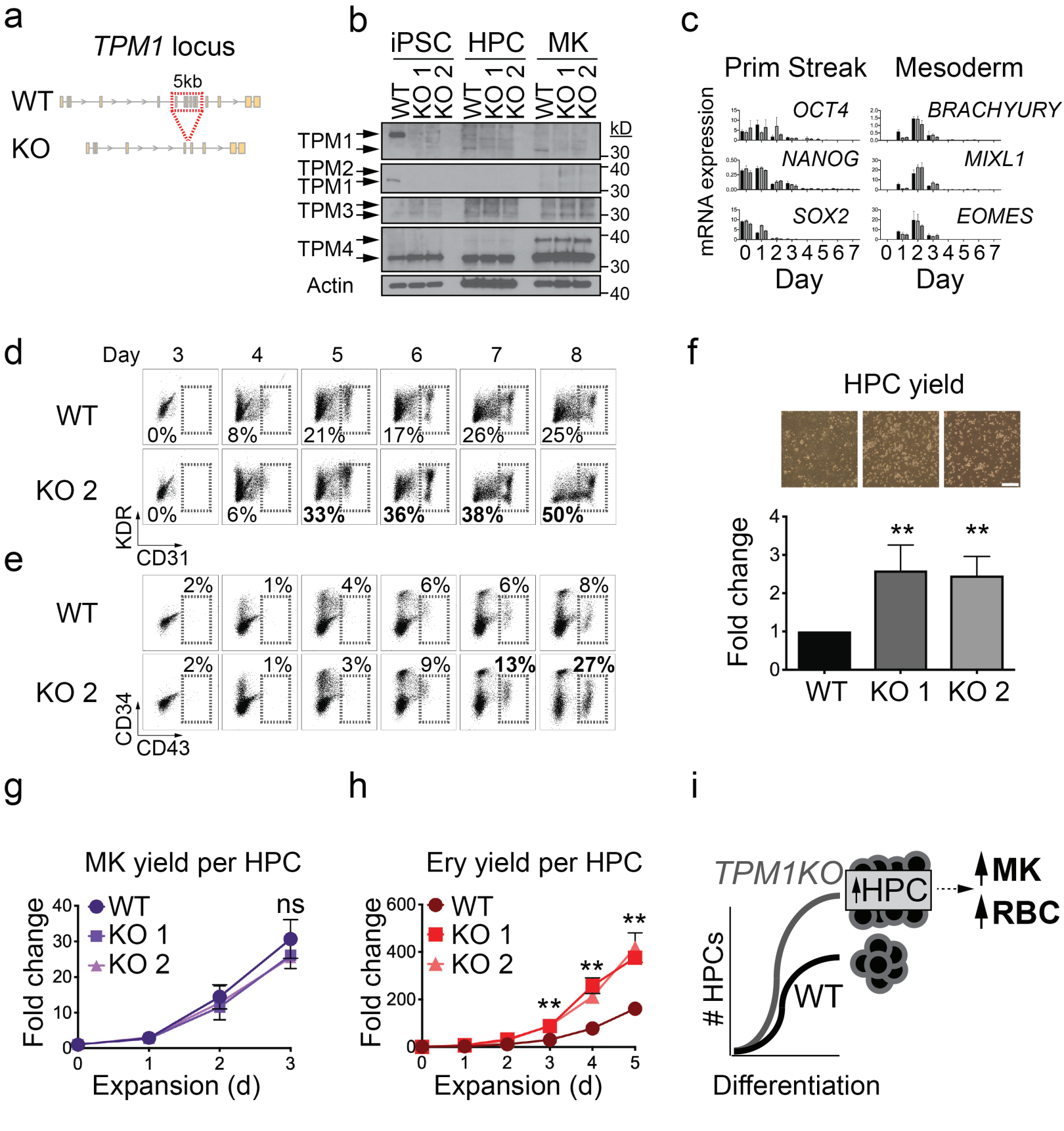



Figure 4. Thom et al

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Table 1. Thom et al

| Cell Type | Mark | Function | Coefficient |
|---|---|---|---|
| 1° MK | GATA1 | Ery/MK TF | 5.08e-01 |
| 1° MK | FLI1 | Ery/MK TF | 6.17e-02 |
| K562 | H3K36me3 | Active gene bodies | 4.11e-01 |
| K562 | H3K79me2 | Enhancers | 1.15e-01 |
| K562 | RbBP5 | SET1 methylation complex | 4.45e-01 |
| Gm12878 | H3K4me1 | Enhancers, 5' active genes | 3.08e-01 |
| Gm12878 | H3K4me2 | Enhancers, promoters, TF binding sites | 5.36e-01 |
| Gm12878 | H3K27ac | Active genes, enhancers, TF binding sites | 1.14e-01 |
| HUVEC | H3K4me2 | Enhancers & promoters | 8.59e-02 |

bioRxiv preprint doi: https://doi.org/10.1101/631895; this version posted May 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Table 2. Thom et al

| rsID[#] | Chr | Pos (Mb) | LASSO (%ile) | GWAVA (%ile) | Nearest Gene | eQTL Gene | Support |
|---|---|---|---|---|---|---|---|
| rs625132 | 2 | 31.5 | 1.29 (96$^{th}$) | 0.19 (73$^{rd}$) | *EHD3* | *EHD3* | |
| rs72879290 | 2 | 43.6 | 1.37 (96$^{th}$) | 0.27 (81$^{st}$) | *THADA* | ***PLEKHH2**, RN7SL531P* | 67 |
| rs342293 | 7 | 106.4 | 2.00 (99$^{th}$) | 0.94 (99$^{th}$) | *CC71L* | ***PIK3CG**** | 15 |
| rs7088799 | 10 | 65.0 | 2.01 (99$^{th}$) | 0.27 (81$^{st}$) | *JHD2C* | ***NRBF2**, MRPL35P2* | 68 |
| rs7899657 | 10 | 65.3 | 1.03 (95$^{th}$) | 0.17 (69$^{th}$) | *REEP3* | *REEP3* | |
| rs17655663 | 11 | 0.27 | 1.58 (98$^{th}$) | 0.72 (97$^{th}$) | *NLRP6* | ***SIRT3**, **NLRP6**, BET1L, RIC8A, PSMD13, SCGB1C1, AC136475.3* | 69,70 |
| rs72882962 | 11 | 0.27 | 1.68 (98$^{th}$) | 0.48 (91$^{st}$) | *NLRP6* | ***SIRT3**, **NLRP6**, BET1L, RIC8A, PSMD13, SCGB1C1* | 69,70 |
| rs6589734 | 11 | 119.2 | 1.90 (99$^{th}$) | 0.95 (99$^{th}$) | *MUC18* | ***CBL**, NLRX1, HINFP, HMBS* | 71 |
| rs941207 | 12 | 57.0 | 2.08 (99$^{th}$) | 0.68 (96$^{th}$) | *BAZ2A* | *BAZ2A, RBMS2* | |
| rs11071720 | 15 | 63.3 | 1.59 (98$^{th}$) | 0.58 (94$^{th}$) | *TPM1* | ***TPM1**, **RAB8B**, LACTB, APH1B* | 6**, 72 |
| rs4819526 | 22 | 20.0 | 1.90 (99$^{th}$) | 0.89 (99$^{th}$) | *ARVC* | ***COMT**, ARVCF, TANGO2* | 73 |

[#]SNP names and locations refer to hg19 genome.
*eQTL in human platelets,[18] though not in GTEx tissues.[38]
**Function suggested by *D. rerio* morpholino experiments.