

# Contribution of genetics to hematopoietic stem cell mobilization: a genome-wide association study of 564 healthy donors mobilized with granulocyte colony-stimulating factor


Miriam Stenzinger<sup>1,2</sup> Susanne Beck,<sup>3</sup> Iordanis Ourailidis,<sup>3</sup> Anna-Lena Volckmar,<sup>3</sup> Nagarajan Paramasivam,<sup>4,5</sup> Aysel Ahadova,<sup>6</sup> Aitzkoa Lopez de Lapuente Portilla,<sup>7,8</sup> Yen Phuong La,<sup>7,8</sup> Björn Nilsson,<sup>7,8,9</sup> Justo Lorenzo Bermejo,<sup>10,11</sup> Halvard Bonig<sup>12,13</sup> and Jan Budczies<sup>3</sup>

<sup>1</sup>Institute of Immunology, University Hospital Heidelberg, Heidelberg, Germany; <sup>2</sup>Institute for Clinical Transfusion Medicine and Cell Therapy (IKTZ), Heidelberg, Germany; <sup>3</sup>Institute of Pathology, University Hospital Heidelberg, Heidelberg, Germany; <sup>4</sup>Computational Oncology Group, Molecular Precision Oncology Program, German Cancer Research Center (DKFZ), Heidelberg, Germany; <sup>5</sup>National Center for Tumor Diseases (NCT) Heidelberg, a partnership between DKFZ and Heidelberg University Hospital, Heidelberg, Germany; <sup>6</sup>Department of Applied Tumor Biology, Institute of Pathology, University Hospital Heidelberg, Heidelberg, Germany; <sup>7</sup>Department of Laboratory Medicine, Lund University, Lund, Sweden; <sup>8</sup>Lund Stem Cell Center, Lund University, Lund, Sweden; <sup>9</sup>Broad Institute, Cambridge, MA, USA; <sup>10</sup>Statistical Genetics Research Group, Institute of Medical Biometry, Heidelberg University, Heidelberg, Germany; <sup>11</sup>Laboratory of Biostatistics for Precision Oncology, Institut de Cancérologie Strasbourg Europe, Strasbourg, France; <sup>12</sup>Institute for Transfusion Medicine and Immunohematology, Goethe University, Frankfurt a. M., Germany and <sup>13</sup>German Red Cross Blood Service Baden-Württemberg-Hessen, Institute Frankfurt, Frankfurt a. M., Germany

**Correspondence:** M. Stenzinger  
[miriam.stenzinger@med.uni-heidelberg.de](mailto:miriam.stenzinger@med.uni-heidelberg.de)  
[m.stenzinger@blutspende.de](mailto:m.stenzinger@blutspende.de)

**Received:** February 21, 2025.  
**Accepted:** May 28, 2025.  
**Early view:** June 5, 2025.

<https://doi.org/10.3324/haematol.2025.287637>

©2025 Ferrata Storti Foundation  
Published under a CC BY-NC license 

## **Supplemental Methods:**

### **Genotyping and Quality Control**

A total of 564 of the 612 DNA samples passed the quality control for the genotyping data with quality filters having a sample call rate of at least 95% (42 samples excluded) and passing the heterozygosity filter (4 samples excluded). For this, a coefficient (method-of-moments F coefficient estimates) was calculated and samples for which the value lies outside three times the standard deviation of the mean value were excluded. Another two samples were removed because of too close relationships revealed by kinship estimation with a cutoff of 0.125 (second-degree relations). Variants not covered by the genotyping array were imputed using TOPMed v1.6.6<sup>1-3</sup> resulting in a total of 17,199,566 typed or imputed variants with a strong correlation ( $R^2 \geq 0.8$ ). Variants were filtered by call rate (in at least 95% of the samples), a minor allele frequency (MAF) filter (at least 5%) and exhibiting allele frequencies compatible with the Hardy-Weinberg equilibrium ( $p \geq 1 \times 10^{-6}$ ) further reducing the data set to 6,095,133 variants. Out of these a total of 1,213,011 variants that were in approximate linkage disequilibrium (50 variants window, 10 variants per step,  $R^2 \leq 0.8$ ) were further analyzed for genome-wide association with granulocyte colony stimulating factor (G-CSF) induced stem cell mobilization.

### **Genome-wide association analysis**

A genome-wide association study (GWAS) assessed the CD34<sup>+</sup> cell frequency associations using multivariate linear models in PLINK.<sup>4</sup> CD34<sup>+</sup> cell content was the dependent variable, while variant status, sex, age, body mass index (BMI), and the top 10 principal components were independent variables. Variant status was encoded as trinary variable for homozygous or heterozygous occurrence. Prior to analysis, the dependent and independent variables were standardized (centered to mean = 0 and scaled to standard deviation = 1). A threshold of  $p < 5.0 \times 10^{-8}$  for genomewide statistical significance was used for the data analysis. Candidate variants were selected by p-value threshold ( $p < 0.0001$ ) and annotated with R package SNPlocs.Hsapiens.dbSNP144.GRCh38 (version 0.99.20)<sup>5</sup> and VEP<sup>6</sup>. The variance explained by a variant was calculated as  $((2 \cdot n / (n - 1)) \cdot (MAF \cdot (1 - MAF)) \cdot \beta^2)$ . Afterwards, the selected candidate variants were analyzed in a second cohort of 13,167 unmobilized donors<sup>7</sup> and p-values were corrected for multiple testing. Functional analysis of the list of candidate variants

was performed by the analysis of 16,008 gene ontology categories available from MSigDB v2023.2.Hs.<sup>8, 9</sup> Significance of enrichment was assessed using Fisher's test. P-values for enrichment were corrected for the number of investigated categories using the Bonferroni method. For comparison with published variants, the direction of the association was considered and significance of the association was assessed by the one-sided test.

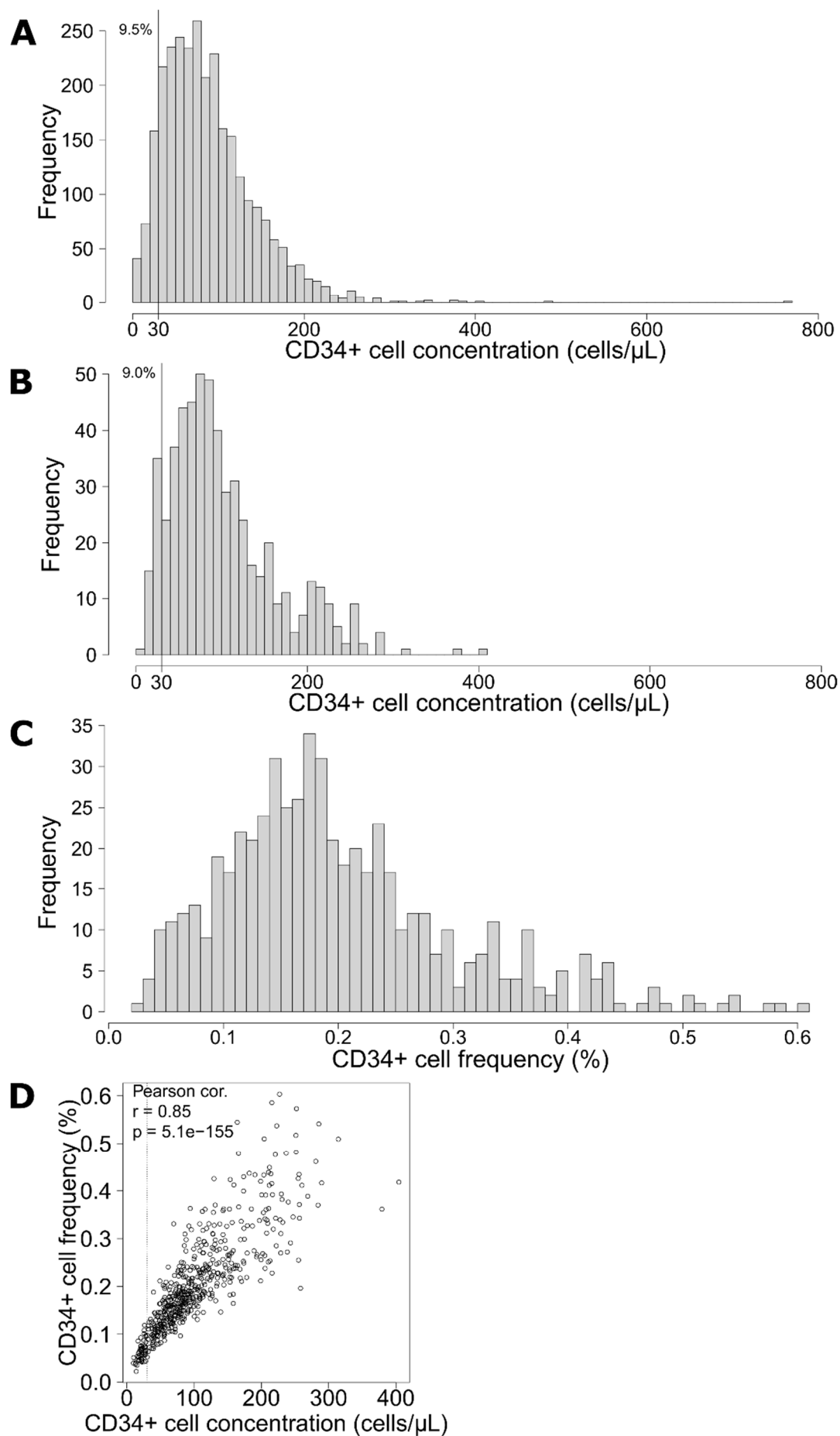
### **Polygenic scores**

Polygenic scores (PS) including multiple variants were calculated from the coefficients in the linear model and the dosages of the variants in the samples. In detail, the PS of sample  $j$  was  $PS_j = \sum_{i=1}^N \beta_i * d_{ij}$  wherein  $\beta_i$  is the coefficient of variant  $i$  and  $d_{ij}$  is the dosage of variant in sample (equal to 1 and 2 for heterozygous and homozygous occurrence, respectively). The PS were trained and validated by 10-fold cross validation.

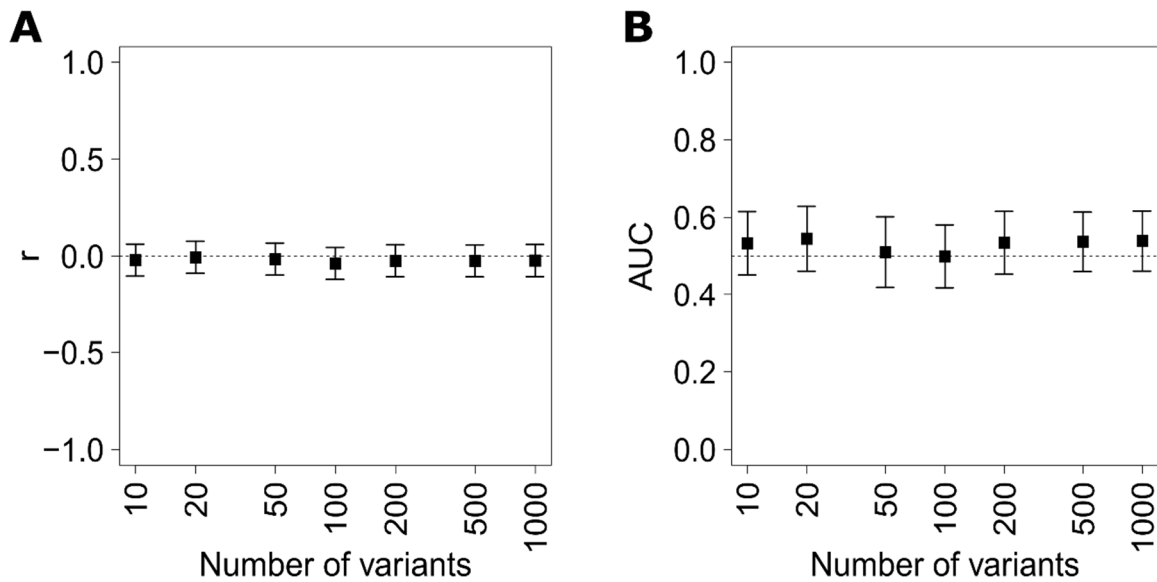
### **Statistics and graphics generation**

Data analysis and visualization was performed using R version 4.1.3. Patient characteristics were investigated by multivariate linear regressions using the R function `lm`. Explained variance per variable in the multivariate model was calculated from standardized beta values as  $sd(x) / sd(y) * \beta^2$ . Receiver operator characteristic (ROC) curves and area under the curve (AUC) were calculated with the R package `ROCR`.<sup>10</sup> Confidence intervals for AUC were calculated with `pROC` R package.<sup>11</sup> Differences between two groups were tested with the Wilcoxon rank sum test. Power and sample size calculations were performed with the R `pwr` package (version 1.3-0) using the function `pwr.f2.test`.<sup>12</sup>

## Supplemental Figures and Tables:



**Supplementary Figure 1: Granulocyte colony stimulating factor (G-CSF) responsiveness of allogeneic stem cell donors.** **(A)** The distribution of CD34+ cell concentration in peripheral blood in a large cohort of 2861 allogeneic stem cell donors, with the 10th percentile corresponding to a value of 30 CD34+ cells/ $\mu$ L. Analysis of peripheral blood **(B)** CD34+ cell concentration and **(C)** CD34+ cell frequency (ratio of CD34+ and CD45+ cells) in the current study cohort of 564 healthy stem cell donors. **(D)** Strong correlation between the CD34+ cell concentration and the CD34+ cell frequency in the current study cohort. Sufficient amounts of stem cells ( $\geq 30$  CD34+ cells/ $\mu$ L) were collected from 513 (91%) healthy donors, while insufficient amounts of stem cells ( $< 30$  CD34+ cells/ $\mu$ L) were collected from 51 (9%) donors (separated by grey dashed line).



**Supplementary Figure 2: Prediction of CD34+ cell frequency using polygenetic scores (PS):** 10, 20, 50, 100, 200, 500, and 1000 variants were included in the analysis. The PS were constructed and validated using ten-fold cross-validation. For each of the ten cross-validation loops, a multivariate linear model including a single variant, age, sex, body mass index, and the ten principal components was fitted to the training dataset. The most significant variants in these models were included into the PS. **(A)** Pearson correlation (r) of CD34+ cell frequency and PS with a 95% confidence interval (CI) could not detect significant correlations. **(B)** For none of the PS were the areas under the curve (AUC) significantly larger than 0.5 with 95% confidence interval (CI) for separating sufficient and insufficient mobilizers.

**Supplementary Table 1: Candidate list of variants that reached the lower significance level of  $p < 0.0001$  (see excel file). (A)** Candidate list of 115 variants that were in linkage disequilibrium and reached the lower significance level of  $p < 0.0001$ . **(B)** None of the 16,008 investigated gene ontology categories were significantly enriched in the genes corresponding to the candidate variants.

**Supplementary Table 2: Comparison of 13 variants investigated in smaller studies of granulocyte colony stimulating factor (G-CSF) mobilized stem cell donors<sup>13, 14</sup> with the current study cohort.** No significant variants could be confirmed after adjustment for multiple testing with a false discovery rate (FDR) of 5% (MAF = minor allele frequency; Beta = regression coefficient; CI = confidence interval; Var<sub>ex</sub> = explained variance).

Gene	rsID	Chr	Garciaz et al. (n= 367)		Martin-Antonio et al. (n=112)		study cohort (n=564)			
			MAF (%)	p	MAF (%)	p	MAF (%)	Beta (95% CI)	Var <sub>ex</sub> (%)	p
<i>CXCR4</i>	rs12691874	2	45.4	0.91	NA	NA	47.8	-0.12 (-0.23 - -0.01)	0.71	0.03
<i>CXCL12</i>	rs1413519	12	19	0.15	NA	NA	18	-0.08 (-0.22 - 0.06)	0.19	0.26
<i>ITGA4</i>	rs1449263	2	44.5	0.39	NA	NA	47.3	0.06 (-0.05 - 0.17)	0.17	0.30
<i>CXCL12</i>	rs266087	12	42.3	0.73	NA	NA	37.4	-0.05 (-0.16 - 0.06)	0.12	0.40
<i>VCAM1</i>	rs1041163	1	19.1	0.06	25.4	0.29	14.8	0.06 (-0.1 - 0.21)	0.09	0.46
<i>CXCR4</i>	rs16832740	2	16.5	0.79	NA	NA	20.5	0.05 (-0.09 - 0.18)	0.07	0.50
<i>CXCL12</i>	rs266085	12	42.6	0.67	NA	NA	37.5	-0.03 (-0.14 - 0.08)	0.05	0.59
<i>CXCR4</i>	rs2680880	2	30.9	0.31	37.9	0.3	48	-0.03 (-0.14 - 0.09)	0.03	0.66
<i>CXCL12</i>	rs1801157	12	21.9	0.44	26.8	0.04	20.8	-0.02 (-0.16 - 0.11)	0.02	0.75
<i>CXCL12</i>	rs2297630	12	20	0.28	NA	NA	25.4	-0.02 (-0.15 - 0.11)	0.02	0.75
<i>CSF3R</i>	rs3917924	1	NA	NA	35.3	0.02	39.2	-0.004 (-0.11 - 0.11)	0.001	0.94
<i>CXCR4</i>	rs2228014	2	4.8	0.49	NA	NA	3.4	NA	NA	NA
<i>CD44</i>	rs13347	11	NA	NA	27.5	0.1	NA	NA	NA	NA



**Supplementary Table 3: Comparison of 17 suggestive variants that correlated with the CD34+ cell frequency of untreated individuals in an earlier study<sup>15</sup> with the current study population of granulocyte colony stimulating factor (G-CSF) mobilized healthy donors.** The pre-described variants could not be confirmed in the current study population (MAF = minor allele frequency; Beta = regression coefficient; CI = confidence interval; Var<sub>ex</sub> = explained variance).

Gene	rsID	Chr	CD34+ cell frequency, steady state (Cohen et al., n=1,595)				CD34+ cell frequency, G-CSF mobilized (study cohort, n=564)			
			MAF (%)	Beta	Var <sub>ex</sub> (%)	p	MAF (%)	Beta (95% CI)	Var <sub>ex</sub> (%)	p
<i>DGKB</i>	rs976760	7	26	0.25	0.03	2.86 x 10 <sup>-6</sup>	25.2	0.023 (-0.102 - 0.147)	0.02	0.36
<i>DSC3</i>	rs1943535	18	25	-0.25	0.03	4.61 x 10 <sup>-6</sup>	11.2	0.151 (-0.028 - 0.331)	0.46	1
<i>DSC3</i>	rs2591120	18	25	-0.25	0.03	4.62 x 10 <sup>-6</sup>	11.2	0.151 (-0.028 - 0.331)	0.46	1
<i>DSC3</i>	rs2591119	18	25	-0.25	0.03	4.65 x 10 <sup>-6</sup>	11.2	0.151 (-0.028 - 0.331)	0.46	1
<i>DSC3</i>	rs2591118	18	25	-0.25	0.03	4.65 x 10 <sup>-6</sup>	11.2	0.151 (-0.028 - 0.331)	0.46	1
<i>DSC3</i>	rs2733151	18	23	-0.23	0.02	4.89 x 10 <sup>-6</sup>	12.4	0.106 (-0.065 - 0.277)	0.25	1
<i>DSC3</i>	rs2591117	18	23	-0.23	0.02	4.93 x 10 <sup>-6</sup>	12.4	0.106 (-0.065 - 0.277)	0.25	1
<i>DSC3</i>	rs2591116	18	23	-0.23	0.02	4.94 x 10 <sup>-6</sup>	12.4	0.106 (-0.065 - 0.277)	0.25	1
<i>DSC3</i>	rs2733152	18	23	-0.23	0.02	4.99 x 10 <sup>-6</sup>	12.4	0.106 (-0.065 - 0.277)	0.25	1
<i>DSC3</i>	rs2733150	18	23	-0.23	0.02	4.90 x 10 <sup>-6</sup>	12.3	0.102 (-0.07 - 0.274)	0.23	1
<i>DSC3</i>	rs2105759	18	23	-0.23	0.02	4.99 x 10 <sup>-6</sup>	12.3	0.102 (-0.07 - 0.274)	0.23	1
<i>ENO1, RERE</i>	rs11121245	1	49	-0.17	0.03	3.10 x 10 <sup>-6</sup>	47.3	0.043 (-0.067 - 0.154)	0.09	1
<i>ENO1, RERE</i>	rs11590606	1	49	-0.17	0.03	3.17 x 10 <sup>-6</sup>	47.5	0.041 (-0.07 - 0.152)	0.08	1
<i>ENO1, RERE</i>	rs10864368	1	49	-0.17	0.03	3.18 x 10 <sup>-6</sup>	47.6	0.04 (-0.071 - 0.151)	0.08	1
<i>ENO1, RERE</i>	rs6577536	1	47	-0.18	0.03	9.78 x 10 <sup>-7</sup>	47.3	0.028 (-0.084 - 0.14)	0.04	1
<i>ENO1, RERE</i>	rs11121242	1	48	-0.19	0.03	8.81 x 10 <sup>-7</sup>	47.2	0.028 (-0.084 - 0.139)	0.04	1
<i>OR4C12</i>	rs2183383	11	20	-0.22	0.02	6.67 x 10 <sup>-7</sup>				

### Supplemental References:

1. Das S, Forer L, Schonherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet.* 2016;48(10):1284-1287.
2. Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. *Bioinformatics.* 2015;31(5):782-4.
3. Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature.* 2021;590(7845):290-299.
4. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:7.
5. Pages H. SNPlocs.Hsapiens.dbSNP144.GRCh38: SNP locations for Homo sapiens (dbSNP Build 144). 0.99.20.  
<https://bioconductor.org/packages/release/data/annotation/html/SNPlocs.Hsapiens.dbSNP144.GRCh38.html>. Accessed April 1, 2021.
6. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol.* 2016;17(1):122.
7. Lopez de Lapuente Portilla A, Ekdahl L, Cafaro C, et al. Genome-wide association study on 13 167 individuals identifies regulators of blood CD34+cell levels. *Blood.* 2022;139(11):1659-1669.
8. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics.* 2011;27(12):1739-40.
9. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* 2005;102(43):15545-50.
10. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. *Bioinformatics.* 2005;21(20):3940-1.
11. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics.* 2011;12:77.

12. Champely S. pwr: Basic Functions for Power Analysis. 1.3-0. <https://CRAN.R-project.org/package=pwr>. Accessed April 6, 2021.
13. Garciaz S, Sfumato P, Granata A, et al. Analysis of a large single institution cohort of related donors fails to detect a relation between SDF1/CXCR4 or VCAM/VLA4 genetic polymorphisms and the level of hematopoietic progenitor cell mobilization in response to G-CSF. PLoS One. 2020;15(3):e0228878.
14. Martin-Antonio B, Carmona M, Falantes J, et al. Impact of constitutional polymorphisms in VCAM1 and CD44 on CD34+ cell collection yield after administration of granulocyte colony-stimulating factor to healthy donors. Haematologica. 2011;96(1):102-9.
15. Cohen KS, Cheng S, Larson MG, et al. Circulating CD34(+) progenitor cell frequency is associated with clinical and genetic factors. Blood. 2013;121(8):e50-6.