A Sardinian founder mutation in *GP1BB* that impacts thrombocytopenia.

Authors
Fabio Busonero[1,6*], Maristella Steri[1,6*], Valeria Orrù[1], Gabriella Sole[1], Stefania Olla[1], Michele Marongiu[1], Andrea Maschio[1], Carlo Sidore[1], Sandra Lai[1], Antonella Mulas[1], Magdalena Zoledziewska[1], Matteo Floris[2], Mauro Pala[1], Paola Forabosco[1], Isadora Asunis[1], Maristella Pitzalis[1], Francesca Deidda[1], Marco Masala[1], Cristian Antonio Caria[1], Susanna Barella[3], Goncalo R. Abecasis[4], David Schlessinger[5], Serena Sanna[1], Edoardo Fiorillo[1], Francesco Cucca[1,2*].

Affiliations
[1]Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche (CNR), 09042 Monserrato (Cagliari), Italy. [2]Dipartimento di Scienze Biomediche, Università degli Studi di Sassari, 07100 Sassari, Italy. [3]Ospedale Pediatrico Microcitemico "Antonio Cao" (A.O.Brotzu), 09100 Cagliari, Italy. [4]Center for Statistical Genetics, University of Michigan, Ann Arbor, 48109 Michigan, USA. [5]Laboratory of Genetics and Genomics, National Institute on Aging, US National Institutes of Health, Baltimore, 21224 Maryland, USA. [6]These authors contributed equally to this work.

Correspondence emails
*Correspondence should be addressed to Francesco Cucca (fcucca@uniss.it), Fabio Busonero (fabio.busonero@irgb.cnr.it) or Maristella Steri (maristella.steri@irgb.cnr.it).

Supplemental data table of contents

Supplemental data

Results

Replicated loci

Interrogating 21,832,250 variants, either directly genotyped or successfully imputed, 5 reported genetic associations with platelet count (PLT) at the *ARHGEF3*[16], *HBS1L*[16], *TP53BP1*[17], *GCSAML*[18] and *TPM4*[16] loci were replicated (Suppl. Table 1 and Suppl. Fig. 1), demonstrating the validity of the

experimental design employed. Regional plots and the distribution of platelet count in each of the genotype groups for these loci were shown (Suppl. Figg. 4-8, respectively). Moreover, in addition to fully replicated *ARHGEF3* locus (where the same lead variant was found), conditioning for our leading variants and for consortium's ones[16] at 6q23.3 and 19p13.12 loci, respectively, it was established they were exactly the same signals ($P_{rs9389268cond}$(rs9399137)=0.79 and $P_{rs9399137cond}$ (rs9389268)=0.91 for *HBS1L*, $P_{rs2228367cond}$(rs8109288)=0.66 and $P_{rs8109288cond}$(rs2228367)=0.99 for *TPM4*), as suggested by the high correlation coefficient.

Novel association in the 22q11.21 locus

In addition to the replicated loci, using our population-based GWAS[3] a novel association with platelet count in the 22q11.21 region was found (Suppl Fig. 1). The novel signal encompassed numerous rare variants in linkage disequilibrium (LD) spanning 1 Mb (Suppl. Fig. 9). The variant with the lowest *P* value (22:19940476:A/T, effect=-1.186 s.d., $P=3.428\times10^{-17}$) fell in an intron of the *COMT* gene. Considering the predicted functional impact of 10 variants in the 95% credible set of candidates (Suppl. Table 5), only one coding variant (22:19711445:C/T, MAF=0.0045, r2=0.924 with top variant), mapping in the second exon of the *GP1BB* gene (c.C79T, p.P27S), was considered for downstream functional evaluation. No homozygous and 57 carriers for 22:19711445:C/T rare allele, validated by Sanger sequencing, were found. Noteworthy, this variant was also significantly associated with an increased mean platelet volume (MPV) in a subset of 2,000 individuals ($P=2.13\times10^{-10}$, Suppl. Fig. 10).To verify whether the p.P27S variant in *GP1BB* gene shows different frequencies in the cohort from the Lanusei valley compared to the rest of the island, we further genotyped it in 1,235 whole-genome sequenced individuals from across Sardinia[3]. Four additional heterozygous individuals were found, resulting in a lower MAF (0.00162); according to Hardy-Weinberg expectation, about 4 homozygous BSS cases are predicted in the island. By contrast, p.P27S is completely missing in large sequencing datasets such as UK10K[19], 1,000 Genomes Project (1,000GP)[4], GoNL[5], GnomeAD browser Broad Institute (126,216 exomes and 15,137 genomes)[6], and the Exome Sequencing Project (ESP) data[20] in NHLBI's TOPMed program[7].

Sardinian, 1,000GP and HRC reference panels comparison

To assess the effectiveness of our population-specific imputation panel, the association at chr22:19940476 locus was tested using two cosmopolitan reference datasets for imputation: 1,000 Genome Project (1,000GP)[4] and the Haplotype Reference Consortium (HRC)[21]. Interestingly, all of the 32 sequencing rare variants detected using the Sardinian population-specific reference panel (Suppl. Fig. 11A) would have been missed using the 1,000GP reference panel (Suppl. Fig. 11B), as these variants are extremely rare outside Sardinia. In details, only two variants (chr22:19435422 and chr22:19491477) were called with rsqr>0.50, of which chr22:19435422 (MAF=0.0003922 in EUR non-Finnish from GnomAD genomes) turned-out to be the most significantly associated ($P=4.23\times10^{-15}$). Noteworthy, the 1,000GP dataset would not allow us to identify *GP1BB* as the best functional candidate, since the top signal mapped far-away from it (276 Kb). Furthermore, after imputation with HRC reference panel and association testing, the leading variant turned-out to be 22:19747128 (rs72646950-C/T, $P=4.56\times10^{-17}$) (Suppl. Fig. 11C). Due to the presence of eight Sardinian individuals in the HRC panel, this variant mapping in the 5'-UTR region of *TBX1* gene (NM_005992:c.-39C>T) was found in 57 heterozygous carriers (MAF=0.00409, rsqr=0.935). *TBX1* is one of the genes responsible for cardiovascular defects in Velo-Cardio-Facial/DiGeorge Syndrome (DGS, MIM 188,400), a complex developmental disorder, whose patients were hemizygous for a 3 Mb deletion on chromosome 22 comprising several genes, among which *GP1BB*.

Previously reported associations in the GP1BB gene region

Further supporting the relevance of the *GP1BB* gene region in platelet count and volume regulation, this region has been previously associated with platelet-related traits in large GWAS of general population individuals. For example, after imputation of 29.5 million variants in >164,000 Europeans genotyped with Affymetrix arrays, a common variant (rs1059196-C/T) in GP1BB was described to be associated with PLT reduction ($P=1.0\times10^{-12}$, MAF=0.354, effect -0.030±0.0043×109/L) and MPV

increase (P=1.0×10^{-21}, effect +0.040±0.0043 fL)18. Independent from the p.P27S variant (r^2<0.01 in the Sardinian reference panel), rs1059196-C/T had a much smaller effect compared to p.P27S, a typical evidence when comparing the effects of common and rare variants.

Additional results on molecular modeling study of p.P27S mutation

The effects of p.P27S on molecular structure and conformational changes in glycoprotein Ib-β - GPIbβ - were tested by molecular modeling analysis. The reduction of protein stability due to p.P27S, as shown by in-silico Molecular Dynamic simulations, was corroborated with two software tools used to calculate differences in free energy between the WT and mutated protein (ddG=-1,35 and ddG=-0.66, respectively[22,23])(Suppl. Fig. 12A). Furthermore, p.P27S mutation caused a slight increase in solvent accessibility, as calculated with two software tools: 13% for Pro and 18,6% for Ser, 16.3 Å2 for Pro and 20.7 for Ser, respectively[23,24]) as well as changes in the electrostatic potential surface (Suppl. Fig 12B) that likely affects the conformation suitable for normal platelet function, especially influencing the optimal protein-protein interaction for the correct GPIb-IX-V complex assembly.

Datasets and methods

Cohorts and sample description

The SardiNIA project[3] is a longitudinal cohort study of over 8,000 general population volunteers, recruited in the Lanusei Valley, phenotyped for thousands of quantitative traits, including platelet count, mean platelet volume and other hematological traits, often underlying clinical endpoints for common and rare diseases. All participants gave informed consent to study protocols, which were approved by the Sardinian local research ethic committees: Ethical Committee of ASSL of Lanusei (2009/0016600) and Ethical Committee of ASSL of Sassari (2171/CE).

Blood platelet count and volume measurement

PLT was measured in 6,528 volunteers, while MPV was assessed in a subset of 2,000 of them. Blood samples were collected for each volunteer, and divided into two aliquots; the first used for genomic DNA extraction and the second to characterize several blood phenotypes, including PLT and MPV using the Beckman COULTER LH750 Series Hematology Systems, according to manufacturer's instructions. Over the time, several instruments have been changed, so that PLT and MPV were also measured with Beckman Coulter LH700 2727, and LH750 3435 series. To avoid biases due to different instruments, these were considered as a covariate (1 for the LH700 and 2 for LH750). Given the International System (SI) unit, and data from a recent report[25], the following physiological ranges for number of circulating platelets and volume, were considered: 150-450×10^9/L, and 8-11 femtoliter, respectively.

Genotyping and imputation

Genetic analyses were performed using a genetic map based on 6,602 samples genotyped with four Illumina arrays (OmniExpress, ImmunoChip, Cardio-MetaboChip and ExomeChip) as previously described[26]. Imputation was performed on a genome-wide scale using the Sardinian sequence-based reference panel of 3,514 individuals and the *minimac* software on pre-phased genotypes. After imputation, only markers with imputation quality (RSQR)>0.3 for MAF>=1% or RSQR>0.6 for MAF<1% were retained for association analyses[3], yielding ~22 million variants (20,143,392 SNPs and 1,688,858 indels, for a total of 21,832,250) useful for analyses.

Association analysis

Before performing GWAS, PLT and MPV phenotypes were both inverse-normalized and adjusted by sex, age, age2, smoking status, aspirin consumption and instrument employed for measurements, as covariate. In the same model, the effect on traits of the body mass index was also tested but it was not included in the analyses being not significant. Additive effects were searched using EPACTS (epacts-3.2.6 version)[27], a software that implements a linear mixed model adjusted with a genomic-based kinship matrix calculated using all quality-checked genotyped SNPs with MAF>1%. The advantage of this model is that the kinship matrix encodes a wide range of sample structures, including both cryptic relatedness and population stratification. As a proof of appropriate adjustment

of all confounders, the genomic inflation factor (lambda GC) for PLT GWAS was 0.988. To identify independent signals, a conditional GWAS analysis for each trait was performed by adding the leading SNPs found in the primary GWAS as covariate to the basic model. A SNP reaching the standard genome-wide significance threshold for sequencing-based GWAS ($P<6.9\times10^{-9}$) was considered as significant[3]. To calculate the heritability explained by each variant the following formula was used:

$$Effect^2 \times 2 \times MAF \times (1\text{-}MAF).$$

Gene-prioritization strategy on 22q11.21 associated locus

Statistical fine-mapping of the 22q11.21 region was performed by calculating the 95% credible set, i.e. the minimum set of variants having a 95% summed posterior probability of containing the causal variant (Suppl. Table 5). The credible set was calculated using FINEMAP v1.3 and setting at most one causal variant (--n-causal-snps 1). To predict the functional effects and pathogenicity of the 10 candidate variants in the credible set, several *in silico* methods were considered: Polymorphism Phenotyping v2 (PolyPhen-2), Sorting Intolerant From Tolerant (SIFT), Mutation Taster, and CADD-score annotations included in VEP[28].

Platelet staining protocol and Immunophenotyping

Fresh blood samples were collected in Vacutainer tubes containing sodium citrate as anticoagulant. Cell phenotyping was carried-out by flow-cytometry immediately after blood collection to avoid any time-dependent artifacts. Blood samples were then diluted 1:20 with phosphate-buffered saline (PBS, 20 µL of blood and 800 µL of PBS), stained with the antibody mix (Suppl. Table 4) for 20 minutes in refrigerated condition, fixed with 1% formalin, then analyzed by FACS Canto II and Diva software (BD Biosciences). All antibodies were obtained from BD Biosciences. Moreover, to dissect platelet activation status, samples were incubated with adenosine diphosphate (ADP, final concentration 20 µM)[29] for 5 minutes at room temperature, then stained as described for basal conditions.

Flow-cytometry data analyses

Statistical analyses of the cytometric experiments were performed using the R software (R version 3.4.3 (2017-11-30)[30]. The Shapiro-Wilk Normality test was used to verify whether the fluorescence intensities were normally distributed. Markers expression was normalized by CD61 values, to take into account differences in platelet size between p.P27S carriers and controls, while statistical significance was assessed with the two-sided Mann-Whitney test (also known as unpaired two-samples Wilcoxon test). For both CD42a (GPIX) and CD63 (Granulophysin) markers, one volunteer for each of the two experimental groups (C/C and C/T) were discarded from the analyses given their outlier antigen expression values.

Molecular modeling analysis

Molecular dynamics study was performed on the wild type (WT) and mutated (p.P27S) protein using the 3D structure of glycoprotein Ib-β available on Protein Data Bank (PDB code 3RFE)[11]. The structure of p.P27S protein has been carried out using pymol software. NAMD software package[31] was used to perform all-atom molecular dynamics MD and using CHARMM force field[32,33]. The VMD program was used for the step of preparation and analysis of simulations. In details, the two structures were solvated in TIP3P water box of water with 12 Å buffering distance. Assuming normal charge states of ionizable groups corresponding to pH 7, sodium Na^+ and chloride Cl^- ions at physiological concentration of 0.15 mol/L were added to achieve charge neutrality, more closely mimicking a realistic biological environment. Systems were subjected to minimization for 10,000 steps, equilibration for 1 ns and dynamics simulations lasting 100 ns. During minimization and equilibration all C α atoms of the complex were restrained with a one Kcal/mol/$Å^2$. Root-Mean-Square-Fluctuations (*RMSF*) were calculated for WT and p.P27S proteins. Two software tools, Strum[22] and SDM[23], were used to quantify WT and p.P27S proteins stability, while STRIDE[14] and SDM were employed to calculate the solvent accessibility. Electrostatic surface potentials were calculated using APBS method[34]. Individual charges were assigned using pdb2pqr software[35] with the CHARMM force field. Final images were generated with VMD from -10kT to 10kT.

Additional Web Resources

The URLs for data presented herein are as follows

-Protein Data Bank, https://www.rcsb.org/;

-Pymol, www. Pymol.org/;

-RefSeq, https://www.ncbi.nlm.nih.gov/refseq/;

-LocusZoom package, http://locuszoom.org/;

Additional supplementary references

16. Soranzo N., Spector T.D., Mangino M., Kühnel B., Rendon A., Teumer A., Willenborg C., Wright B., Chen L., Li M., et al. (2009). A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nat Genet 41, 1182-1190.
17. Auer P.L., Johnsen J.M., Johnson A.D., Logsdon B.A., Lange L.A., Nalls M.A., Zhang G., Franceschini N., Fox K., Lange E.M. et al. (2012). Imputation of exome sequence variants into population-based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. Am J Hum Genet 91, 794–808.
18. 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).
19. Kaye J., Hurles M., Griffin H., Grewal J., Bobrow M., Timpson N., Smee C., Bolton P., Durbin R., Dyke S., et al. (2014). Managing clinically significant findings in research: the UK10K example. Eur. J. Hum. Genet. 22, 1100-4.
20. Guidelines for Large-Scale Sequence-Based Complex Trait Association Studies: Lessons Learned from the NHLBI Exome Sequencing Project. P.L. Auer, A.P. Reiner, G. Wang, H.M. Kang, G.R. Abecasis, D. Altshuler, M.J. Bamshad, D.A. Nickerson, R.P. Tracy, S.S. Rich, NHLBI GO Exome Sequencing Project, and S.M. Leal (2016). A.J.H.G. 99(4), 791-801.
21. the Haplotype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics 48, 1279-1283.
22. Quan L., Lv Q., and Zhang Y. (2016). STRUM: structure-based prediction of protein stability changes upon single-point mutation. Bioinformatics 32, 2936-2946.
23. Pandurangan AP, Ochoa-Montaño B, Ascher DB, Blundell TL (2017). SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Res. 45(W1):W229-W235.
24. Heinig M. and Frishman D. (2004). STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Res. 32, W500-W502.
25. P. Patel, A. Shah, K. Mishra and K. Ghosh (2019). Prevalence of Macrothrombocytopenia in Healthy College Students in Western India. Indian J Hematol Blood Transfus 35(1):144-148.
26. Pistis G., Porcu E., Vrieze S.I., Sidore C., Steri M., Danjou F., Busonero F., Mulas A., Zoledziewska M., Maschio A., et al. (2015). Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. Eur J Hum Genet. 23, 975-83.
27. Kang H.M., Sul J.H., Service S.K., Zaitlen N.A., Kong S.Y., Freimer N.B., Sabatti C., Eskin E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42, 348-354.
28. McLaren W., Gil L., Hunt S.E., Riat H.S., Ritchie G.R., Thormann A., Flicek P., Cunningham F. (2016). The Ensembl Variant Effect Predictor. Genome Biology 17(1), 122.
29. Michelson, A.D. and Furman, M.I. (1999). Laboratory markers of platelet activation and their clinical significance. Curr. Opin. Hematol. 6, 342-348.
30. R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

31. Phillips J.C., Braun R., Wang W., Gumbart J., Tajkhorshid E., Villa E., Chipot C., Skeel R.D., Kalé L., Schulten K. (2005). Scalable molecular dynamics with NAMD. J. Comput. Chem. 26, 1781-1802.
32. MacKerell AD et al., (1998). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. J Phys Chem B. 102(18), 3586-3616.
33. MacKerell AD Jr., Banavali N, Foloppe N (2000). Development and current status of the CHARMM force field for nucleic acids. Biopolymers 56: 257-265.
34. Baker N.A., Sept D., Joseph S., Holst M.J., McCammon J.A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. Proc. Natl. Acad. Sci. USA. 98, 10037-41.
35. Dolinsky TJ et al., (2004). PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res. 32, W665-7.
36. Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K. & Tolstoy, I. (2014). RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids Res 42, D553–D559.
37. Pruim, R. J. et al., (2010). LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26, 2336–2337.

Supplementary figure titles and legends

Suppl. Figure 1 *Manhattan plot of the g-w association findings.*

Shown are the association results for mean platelet levels, obtained from imputation using the Sardinian sequencing data, at QCed genotyped and imputed autosomal markers. *P* value results from the GWAS were plotted on a $-\log_{10}$ scale (*y* axis), according to its genomic coordinates (*x* axis), for all SNPs. The blue horizontal dotted line marked the Bonferroni threshold for declaring a locus genome-wide to be significant, and SNPs in loci exceeding this threshold were highlighted in green. To declare new genome-wide signals, a threshold of $P=6.9\times10^{-9}$, which was calculated empirically estimating the number of independent tests in a Sardinian genome[3], was used.

Suppl. Figure 2 *Gating strategy identifying platelets.*

Platelets were distinguished from other cells based on dot-plot patterns (***A***) of side light scatter (SSC-A) versus forward light scatter (FSC-A) characteristics on a log/log scale (first gate), and by positive staining with monoclonal antibodies, such as CD41a/BV510 (***B***) and CD61/PerCP (***C***), directed against the αIIbβ3 integrin (second gate). In all samples, CD61 was analyzed to confirm the identity of platelets population and correct gating, which was positive in all patients. ****P*<0.001.

Suppl. Figure 3 *Molecular modelling analyses*

Shown is the impact of p.P27S mutation on GPIbβ glycoprotein mobility. RMSF analysis indicates the fluctuations (y axis) for the WT (***red***) and the mutated protein (***green***) for each amino acid present in the 3D structure (x axis). Considering the PDB model, position 1 corresponded to the amino acid 26. Major changes in fluctuations are identified in loop 2 adjacent to the mutation p.P27S.

Suppl. Figure 4 *Association plot for PLT at the chr3:56849749 locus.*

Shown is the regional association plot (***A***) for platelet levels-associated variants at the *ARHGEF3* locus using the Sardinian sequencing-based reference panel. The association strength (expressed as $-\log_{10}$ *P* values; *y* axis) is plotted versus the genomic position (on the hg19/GRCh37 genomic build; *x* axis) around the most significant SNP chr3:56849749, which is indicated with a dot purple. Other variants in the region are color-coded to reflect their extent of linkage disequilibrium with the top variant (taken from pairwise r2 values calculated on Sardinians haplotypes), or gray if LD was<0.01. Symbols reflect genomic functional annotation, as indicated in the inner box. The right y-axis corresponds to recombination rate (cM/Mb), plotted as a solid blue line overlying the plot. The box at the bottom shows specification on genes, exons, as well as the direction of transcription, according to RefSeq[35]. This plot is drawn using the standalone version LocusZoom package[36].

Shown are box plots (***B***) indicating the distribution of the platelet levels within each genotype of chr3:56849749:T/C when considering the raw traits (***left***) and the normalized traits adjusted for the covariates (***right***) as in the association model. Box plots refer to the 6,528 samples used for the GWAS analysis.

Suppl. Figure 5 *Association plot for PLT at the chr6:135419631 locus.*
Shown are the association results for platelet levels at the *HBS1L* locus. Plots and specific features are described in Suppl. Figure 4.

Suppl. Figure 6 *Association plot for PLT at the chr15:43753426 locus.*
Shown are the association results for platelet levels at the locus *TP53BP1*. Plots and specific features are described in Suppl. Figure 4.

Suppl. Figure 7 *Association plot for PLT at the chr1:247719769 locus.*
Shown are the association results for platelet levels at the locus *GCSAML*. Plots and specific features are described in Suppl. Figure 4.

Suppl. Figure 8 *Association plot for PLT at the chr19:16197320 locus.*
Shown are the association results for platelet levels at the locus *TPM4*. Plots and specific features are described in Suppl. Figure 4.

Suppl. Figure 9 *Association plot for PLT at the chr22:19711445 locus.*
Shown are the association results for platelet count at the locus *GP1BB*. Plots and specific features are described in Suppl. Figure 4.

Suppl. Figure 10 *Association plot for MPV at the chr22:19711445 locus.*
Shown are the association results for mean platelet volume at the locus *GP1BB*. Plots and specific features are described in Suppl. Figure 4.

Suppl. Figure 11 *Regional association plots for PLT at the 22q11.21 locus with imputation performed using: (**A**) Sardinian, (**B**) 1000GP or (**C**) HRC reference panels.*
Shown are association results imputed using the Sardinian (***left***), 1,000 Genomes Project (***center***) or Haplotype Reference Consortium (***right***) reference panels. The association strength ($-\log_{10}$(*P* value); *y* axis) is plotted versus the genomic positions (on the hg19/GRCh37 genomic build; *x* axis) around the most significant SNP, which is colored in purple. Other variants in the region are color-coded to reflect their extent of linkage disequilibrium with the top variants (chr22:19940476 for SardiNIA, chr22:19435422 for 1,000GP, and 22:19747128 for HRC)(taken from pairwise r^2 values calculated on Sardinians, 1,000 Genomes Project, or HRC haplotypes), or gray if LD was <0.01. Symbols reflect genomic functional annotation, as indicated in the inner box in ***Fig. 1***. The right y-axis corresponds to recombination rate (cM/Mb), plotted as a solid blue line overlying the plot. The box at the bottom shows specification on genes, exons, as well as the direction of transcription, according to RefSeq[35]. This plot is drawn using the standalone version LocusZoom package[36].

Suppl. Figure 12 *Electrostatic surface potential calculation.*
Shown are the electrostatic surface potentials (***A***) of the first frame of the WT (***left***) and p.P27S protein (***right***). Negative and positive charges are indicated in red and blue, respectively (-10kT and 10kT). In B, the first (***left***) and last frames (***right***) of the mutated protein (***B***).

Supplementary table titles and legends

Suppl. Table 1 *List of associated regions.*
Shown are the independently associated variants for each locus, along with the association parameters. From left to right: chromosomal position (chr:pos) on hg19/GRCh37 genomic build of the lead variant and the corresponding SNP identification number (rsID), if available; the closest genes mapping in the region; the effect of substitution at nucleotide or protein level: the reference sequences being used were *ARHGEF3*, NM_001128615.1; *C1orf150*, NM_145278.4:exon2:c.89+1G>A; *TUBGCP4*, NM_001286414.1; and *TPM4*, NM_001145160.1. The major and the minor alleles (A1 and A2), and the minor allele frequency (MAF); the *P* value; the effect size or Beta (standard error) in standard deviation units per each copy of A2 allele.

G=genotyped and I=imputed; RSQR is the imputation accuracy as reported by minimac. Known SNP are previous associated variants, as listed in the GWAS catalog; r2 is the correlation between the leading SNP, as listed in the GWAS catalog, and the SardiNIA top SNP.

Suppl. Table 2 *Associated loci from GWAS Catalog.*

Shown are variants associated with PLT and reported in GWAS Catalog, along with their association parameters as in GWAS catalog. From left to right: the locus number, the sample ancestry, the cytogenetic region, the chromosome number and the physical position on hg19/GRCh37 genomic build; the reported gene(s), the strongest SNP with its tested allele, the tested allele frequency, the *P* value, beta and 95% CI of the tested SNP, the PMID for the related paper.

Suppl. Table 3 *Heritability explained in the SardiNIA study by variants reported in GWAS Catalog.*

The variants associated with PLT reported in GWAS Catalog are tabulated, along with the association parameters in the SardiNIA study. Indicated are, from left to right: the SNP identification number (rs ID) when available, the chromosome number and physical position on hg19/GRCh37 genomic build; the major and the minor alleles (A1 and A2); the genotype counts for A1 homozygous, heterozygous and A2 homozygous individuals, respectively; and the minor allele frequency (MAF); the *P* value; the effect size and the standard error in standard deviation units per each copy of allele A2; the percentage of heritability and the percentage of phenotypic variance explained by the variant; the PMID for the association result.

Suppl. Table 4 *7-color antibody panel assessed by flow cytometry.*

Shown are antibody, fluorochrome, titre (in microL/test), and manufacturer (catalog number) for each Glycoprotein (GP) being investigated.

Suppl. Table 5 *Functional relevance of variants in the 95% credible set.*

The list of variants included in the 95% credible set were reported, along with their association parameters and their functional relevance as found in VEP. Indicated are, from left to right: the chromosome number and physical position on hg19/GRCh37 genomic build for each variant; the major and the minor alleles (A1 and A2); the minor allele frequency (MAF); the effect size and the standard error in standard deviation units per each copy of allele A2; the association *P* value; the Posterior probability (PP) and the cumulative PP from the bayesian test; the PolyPhen, SIFT and MutationTaster scores for changes to protein sequence; the CADD score; the functional consequence for each variant on the protein sequence and the gene affected by the variant.